OCT4 regulated transcription networks in human embryonic stem cells and human embryonal carcinoma cells

Dissertation zur Erlangung des akademischen Grades des
Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie
der Freien Universität Berlin

vorgelegt von

Dipl.
Humanbiologe
Marc Jung
aus Bad Harzburg

September 2009

1. Gutachter: Prof. Dr. Hans Lehrach,
Max-Planck-Institut für Molekulare Genetik
2. Gutachter: Prof. Dr. Thomas Schmülling
Freie Universität Berlin

Disputation am 09.12.2009

"Ich fürchte, daß nichts von allem wirklich ist. Das nennt man existentielle Angst und ist in der Regel nur ein Übergang zu einer neuen Erkenntnis."

(Jostein Gaarder, "Sophies Welt")

Acknowledgment

Ich danke Herrn Professor Lehrach für die Betreuung, das Interesse und die Unterstützung beim Durchführen dieser Arbeit sowie für die Überlassung des Themas und die Möglichkeit die Dissertation am Max-Planck-Institut für Molekulare Genetik anzufertigen. Des Weiteren danke ich Herrn Professor Schmülling für das Interesse und die Unterstützung beim Durchführen dieser Arbeit. Ein besonderer Dank gebührt Herrn Dr. James Adjaye für die Korrektur, Unterstützung und Anleitung bei der Anfertigung meiner Dissertation sowie für die angenehme Zusammenarbeit während der letzten Jahre. Den Mitgliedern meiner Arbeitsgruppe sowie allen anderen Mitarbeitern und Doktoranden des Max-Planck-Instituts für Molekulare Genetik danke ich für die gute Zusammenarbeit und das gute Arbeitsklima. Meiner Familie danke ich herzlich für vielerlei Unterstützung, Aufmunterung und Anerkennung während der zurückliegenden Zeit. Ich danke herzlich meinen Freunden für ihre Unterstützung und Aufmunterung während der zurückliegenden Zeit. Danke Raed und Harald für die tatkräftige Unterstützung bei der Korrektur und viele anregende Gespräche. Besonderen Dank an Hedi, für die vielen inspirierenden Gespräche und die besonders gute Zusammenarbeit.

Contents

Abbreviations

Abbreviations

ATP	adenosine triphosphate
BSA	bovine serum albumin
CC	cellular compartment
cDNA	complementary deoxyribonucleic acid
ChIP	Chromatin Immunoprecipitation
ChIP-chip	Chromatin Immunoprecipitation followed by microarray hybridisation
Ct	threshold cycle
Cy	cytochrome
DMSO	Dimethyl Sulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
EB	embryoid body
ECC	Embryonal carcinoma cell
eGFP	enhanced green fluorescent protein
ESC	embryonic stem cell
FDR	false discovery rate
GO	gene ontology
hESC	human embryonic stem cells
HMM	hidden Markov model
HMG	High Mobility Group
ICM	inner cell mass
IP	immuno precipitation
LFDR	local false discovery rate
LIF	leukaemia inhibitory factor
MORE	More Palindromic-Oct-factor-Recognition-Element
mRNA	messenger ribonucleic acid
PBS	phosphate buffered saline
PCR	Polymerase chain reaction
PGC	primordial germ cells

PORE	Palindromic-Oct-factor-Recognition-Element
POU	Pituary Octamer Unc family protein
PWM	Poly weight matrix
RNA	ribonucleic acid
RNAi	RNA interference
RT	room temperature
SDS	sodium dodecyl sulfate
TSS	transcription start site

Semantics

°C	degrees Celsius
µg	micrograms
µl	microlitres
µM	micromol
g	grams / gravity
hr/hrs	hours
M	mol
mA	milliampere
mg	milligrams
min	minutes
ml	millilitres
mm	millimetres
mM	millimol
ng	nanograms
nm	nanometres
pg	picograms
pmol	picomol
s	seconds
U	units
V	volt

Abstract

Understanding the network of transcription factors, controlling pluripotency in human embryonic stem cells (ESCs) and human embryonal cancer cells (ECs), is essential for possible future therapies in medicine. Connecting the expression levels after ablation of OCT4 with potential binding sites allows a higher predictability of motif specific driven expression modules important for self-renewal and differentiation.

In this study several peak analysis programs have been used to access a refined list of OCT4 targets in human EC cells and this data was connected to ES cell specific OCT4 binding and expression. A highly enriched POU-motif could be verified, discovered by a *de novo* approach, thus enabling connections to the distribution of OCT4 connected motifs like for the dimerisation factor SOX2. Selected targets have been validated, containing an OCT4-SOX2 binding site in their proximal promoter, and targets not connected to the classical HMG motif. Of those USP44 was further examined, containing a highly conserved POU-motif and GADD45G, having an impact on cell cycle regulation. The overexpression of GADD45G in EC cells resulted in an enrichment for upregulated genes, connected to differentiation pathways. Additionally preferred distances for the HMG and the POU motif could be observed, giving cause for additional binding modes than the classical HMG-POU consensus sequence.

New OCT4 connected targets were discovered, and their importance in ESC differentiation and pluripotency was highlighted. Through a highly connected database, everyone can test now simple hypotheses based on their target genes. The use of NCCIT cells as a model to test pluripotency associated pathways in terms of potential functional binding sites has been demonstrated.

Furthermore array based comparisons of gene expression levels between ES and EC cells have been conducted and new links have been established for further functional characterisation of these cells.

Finally a ChIP-seq study revealed an unbiased genome wide view on putative OCT4 bound regions and suggested a genome wide binding pattern for OCT4 which is not centered for five prime proximal promoters.

Abstract (german)

Für die Entwicklung möglicher, zukünftiger Therapien in der Medizin ist das Verständnis jener Transkriptionsfaktoren, die die Pluripotenz kontrollieren und die Differenzierung blockieren, von entscheidender Bedeutung.

Nach einer RNAi vermittelten Herunterregulierung von OCT4 wurde eine höhere Vorhersagbarkeit von Motiv spezifischen Expressionsmodulen, die wichtig für den Selbsterhalt und die Differenzierung der Zellen sind durch die Verknüpfung von differentiell regulierten Genen mit potentiellen Bindungsstellen ermöglicht.

In dieser Arbeit wurden mehrere Programme für die Berechnung von Bindungsstellen verwendet und kombiniert, um eine Algorithmen unabhängigere Liste von potentiellen OCT4-Bindungsstellen in humanen embryonalen Karzinomzellen zu erhalten. Die daraus resultierenden Daten wurden mit spezifischen OCT4-Bindungsstellen und Expressionsmustern in embryonalen Stammzellen verknüpft. Durch den Einsatz von *De Novo* Motiverkennungsprogrammen konnte ein hoch angereichertes POU-Motiv verifiziert werden. Dadurch wurde wiederum die Analyse der Verteilungsmuster von OCT4 korrelierten Motiven ermöglicht, wie das HMG Motiv von SOX2, einem Heterodimerisierungspartners von OCT4.

Es wurden Kandidaten validiert, die ein OCT4-SOX2 Bindungsmotiv im proximalen Promoterbereich enthielten, als auch solche, die nicht mit dem klassischen HMG Motiv verknüpft waren. Von diesen wurde USP44 weitergehend untersucht. Dieses Gen enthält eine hoch konservierte OCT4 Bindungsstelle. Des Weiteren wurde das Gen GADD45G untersucht, das einen Einfluss auf die Regulierung des Zellzyklus ausübt. Die Überexprimierung von GADD45G in EC Zellen führte zu einer Anreicherung von hochregulierten Genen, die mit Signalwegen der Differenzierung verknüpft sind.

Zudem wurde ein Hinweis auf bevorzugte Entfernungsbeziehungen zwischen dem POU und dem HMG Motiv gefunden, die Grund zur Annahme geben, dass es neben dem klassischen HMG-POU Motiv zusätzliche Bindungsvarianten von OCT4 gibt. Auf der Grundlage Expressionsarray basierter Vergleiche zwischen zwei humanen Stammzellen- und zwei humanen embryonalen Karzinomzelllinien

wurden neue Verknüpfungen zu annotierten funktionalen Signalwegen etabliert, um eine weitere Charakterisierung dieser Zellen zu ermöglichen.

Mit Hilfe der ChIP-seq Technik wurde die genomweite Bindungsverteilung von OCT4 in humanen EC Zellen analysiert um einen weniger verzehrten Blick zu ermöglichen. Diese weist darauf hin, dass OCT4-Bindungsstellen nicht abhängig von der Entfernung zum Transkriptionsstartpunkt sind.

Es konnten neue Kandidatengene identifiziert werden, die mit OCT4 verknüpft sind, und ihre Bedeutung für Differenzierungsprozesse und Pluripotenz hervorgehoben. Durch die Einrichtung einer hochvernetzten Datenbank, die alle relevanten OCT4 verwandten Daten verknüpft, ist es nun möglich, einfache Hypothesen basierend auf spezifischen Genlisten zu testen. Der Einsatz der EC Zelllinie NCCIT als Modellsystem für die Untersuchung Pluripotenz-assoziierter Signalwege im Hinblick auf potentielle funktionelle Bindungsstellen wurde erfolgreich demonstriert.

1 Introduction

1.1 OCT4 regulated networks in pluripotent cells

1.1.1 Early Development – how a single cell develops to a complex organism

It is still an unsolved question how the fertilized egg can give rise to a multicellular organism. Only very slowly the knowledge about the processes by which a fertilized egg divides (cleavage), forms a ball of cells (morula), develops a cavity (blastocyst stage), forms the three primary germ layers of cells that will ultimately give rise to all the cell types of the body (gastrula stage), and ultimately generates all the specialized tissues and organs of a mature organism is being deciphered. However, there is still little knowledge about the specific genes that regulate these early events or how interactions among cells or how cellular interactions with other molecules in the environment of the early embryo affect the early development stages. The process by which an egg develops into an embryo is called embryogenesis and includes coordinated cell division, cell migration, programmed cell death and cell specialization. In that process a so called totipotent cell, will lead to more committed pluripotent cells and finally to unipotent, differentiated cells.

Pluripotency can be generally defined as having more than one potential outcome. In biology, cells which show this trait, have the potential to differentiate into any of the three germ layers including endoderm (interior stomach lining, gastrointestinal tract, the lungs), mesoderm (muscle, bone, blood, urogenital), or ectoderm (epidermal tissues and nervous system). Pluripotent stem cells can give rise to any fetal or adult cell type. However, alone they cannot give rise to all cell types because they lack the potential to contribute to extraembryonic tissue, such as the placenta. Before one can understand the network of transcription factors operating in these early steps one needs to understand the biology of the earliest steps after the fertilization of the egg [1,2].

After the fertilization events, the zygote will locate to the uterus, a process which takes three to four days in mice and five to seven days in humans. During this time, the zygote will divide. After the first cleavage two identical cells are produced and then this process repeats to produce four cells. Separating these cells at this stage would result in genetically identical embryos which is the basis of identical twinning. If the cells remain together they divide asynchronously to produce 8 cells, 16 cells, and so on [3]. The early rounds of cell division take approximately 36 hours [4]. During the eight-cell stage, the embryo compacts, meaning that the cells come together in a tight array that is spaced by gap junctions. These trans-membrane proteins consist of an array of six protein molecules called connexins, which form a pore that allows the exchange of ions and small molecules between cells [5].

When the cells reach a compacted 16-cell stage, the embryo is termed a morula (see figure 1.1). During this stage, Gilbert and colleagues could show in mice that cells have become specialized. This process occurs when the outer cells of the 16-cell morula divide to produce an outer rim of cells—the trophectoderm—and an inner core of cells, the inner cell mass [3]. The signals within the 16-cell morula that regulate the differentiation of the trophectoderm are largely unknown. It has been shown that the outer cells of the morula are polarized, meaning one side of the cell distinguish from the other side. Thus, in the first differentiation event of embryogenesis, the outer, polar cells give rise to trophectoderm and the inner, apolar cells will become the inner cell mass (ICM), meaning that a part of the cells of the morula show a specific intrinsic polarity [5]. Thus the morula will develop into a cavity like structure called the blastocyst by embryonic day 3 (E3.0) in the mouse and days 5 to 6 in human development [6], whereas the cavity is called the blastocoel which is filled by a watery fluid, where the outer cells of the cavity will form the trophectoderm and an attached mass of small round cells will form the ICM in the cavity [3,7].

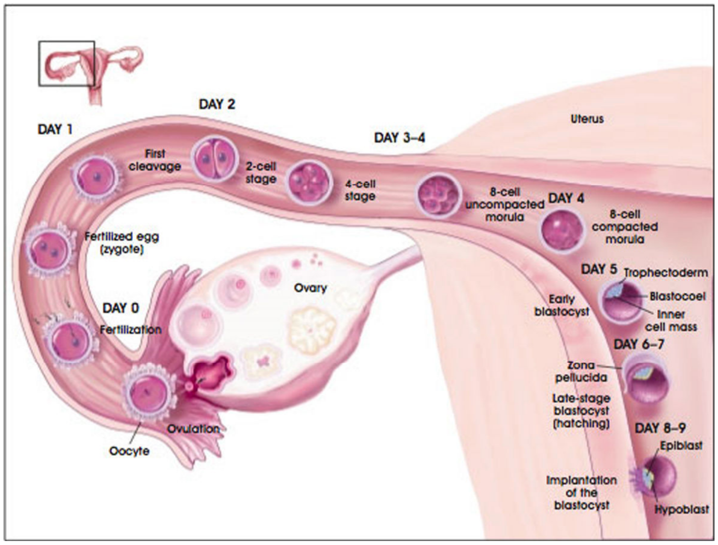


Figure 1.1 Human preimplantation development. After fertilization, the one-cell embryo undergoes a series of cleavage divisions and forms a blastocyst at about six days of development. The blastocyst is composed of an inner cell mass and a trophectoderm. (© 2001 Terese Winslow)

The trophectoderm will generate the trophoblast cells of the chorion, the embryo's contribution to the extraembryonic tissue which is known as the placenta [3,8].

During these stages, the cells of the ICM and trophectoderm will continue to divide. Previous studies of mouse embryos have shown that the two tissues need to interact; the ICM helps maintain the ability of trophectoderm cells to divide, and the trophectoderm supports the ongoing development of the ICM [9]. Secreted paracrine factors, which are molecular signals that affect other cell types, including fibroblast growth factor-4 (FGF-4), which is released from inner cell mass cells [10], help direct embryogenesis at this stage. FGF-4 signalling also helps regulate the division and differentiation of trophectoderm cells [11].

By day 4 in mice, and between 5 to 7 days post fertilization in humans, the blastocyst reaches the uterus. It has not yet implanted into the uterine wall and is therefore still a pre-implantation embryo. When this structure arrives in the uterus, the blastocyst moves out of the zona pellucida, the structure in which the oocyte was originally contained and that also prevented the implantation of the blastocyst into the wall of the oviduct [3]. As in humans, the access to the uterus is limited, it proved to be difficult to study human embryogenesis. Therefore other models were needed in order to study this process. In this regard it turned out that by studying germ cell tumors, some insights into this process could be made.

Thus, in order to study early development processes, several *in vitro* cell culture models have been developed in the last 30 years, starting with murine embryonic carcinoma cells.

Murine and human embryonic carcinoma cells

Germ cell tumours originate typically as benign tumours in the ovary, but also in the testis, where they are always malignant. These tumours are called teratoma if benign, or teratocarcinoma if malignant. Both tumour types are histological complex and contain a wide range of different tissues [12,13].

In contrast to the teratomas, the teratocarcinomas contain stem cells, the embryonic carcinoma (EC) cells, which are capable of differentiating into the various cell types found in these tumours and initiate malignancy [14].

EC cell lines were originally derived from a testicular germ cell tumour, which had been developed from premalignant and non-invasive intratubular germ cell neoplasia [15,16] (for comparison with normal development, see figure 1.2). The differentiation potential of these cells has been already demonstrated in the mid seventies in mice, by injecting embryonal carcinoma cells derived from the central portions of mouse embryoid bodies into blastocysts. Mintz and colleagues demonstrated that EC cells can contribute to the development of most of the tissues and cell lineages in the newly formed mosaics [17,18].

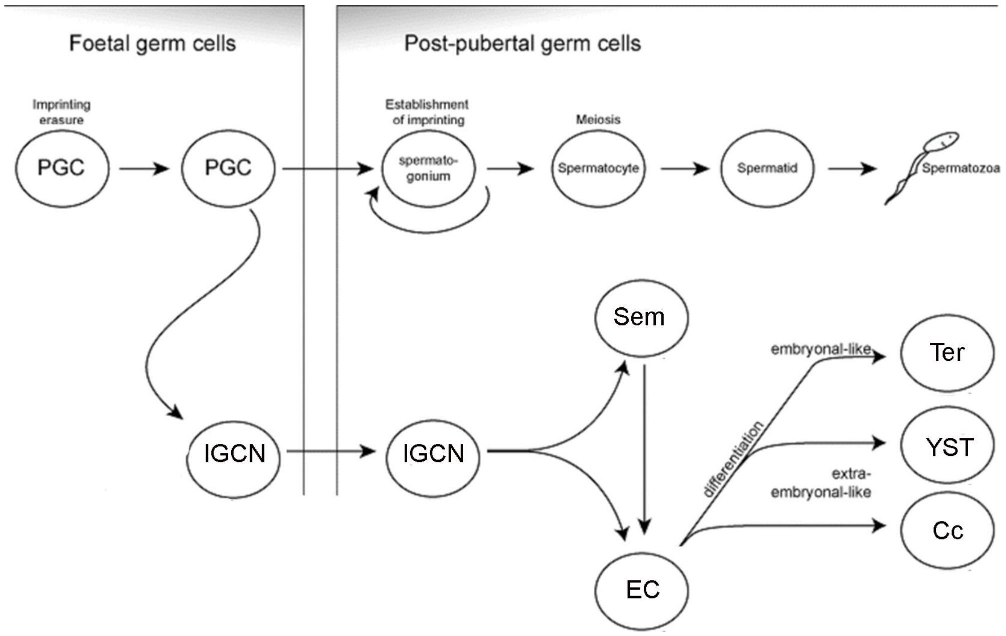


Fig. 1.2: The histological development and gene-specific DNA promoter hypermethylation of testicular germ cell tumours (TGCTs). The precursor stage intratubular germ cell neoplasia (IGCN) is believed to be initiated already during fetal life from a primordial germ cell (PGC) and does not develop into invasive TGCT until after puberty. IGCN can develop into seminoma (Sem) or embryonal carcinoma (EC) cells. The latter cell type has pluripotent capabilities and may differentiate along an embryonal-like lineage into highly differentiated teratoma (Ter) or along extra-embryonal-like lineages into yolk sac tumour (YST) or choriocarcinoma (Cc). Adopted from Guro et al. 2007.

Human embryonic development is not only limited because of the accessibility of the developing embryo, but also because of significant ethical issues. On the other hand, there exist significant differences between human embryos compared to murine embryos in various respects. A full understanding of the human embryonic development can not be achieved from the murine system alone. Human EC (hEC) cells, derived from human teratocarcinomas can narrow the gap and provide an useful model for learning about human embryogenesis [19]. Human EC cells were derived from teratocarcinomas, which are predominantly found as testicular cancers in young men. Many human teratocarcinoma cell lines were established in culture, but most appeared to be near nullipotent. One reason might be the typically high aneuploidy, commonly with about 60 chromosomes, including many rearrangements [20]. Given the high number of chromosomal aberrations, this long period of teratocarcinomas might have shifted selection for cell variants that had lost the ability to differentiate [19,21,22]. However a few human EC cell lines retained the potential to differentiate.

Human undifferentiated and pluripotent embryonal carcinoma cells have the capacity to differentiate into various lineages, where cell lines like NTERA2 and NCCIT cells differentiate extensively in culture in response to morphogens like retinoic acid (RA) [23,24]. In contrast, other established human embryonal carcinoma cell lines, such as the 2102Ep, are relatively nullipotent, meaning they remain in an undifferentiated state even after RA treatment. Furthermore, human EC cells typically express the glycolipid antigens SSEA3 (stage-specific embryonic antigen-3) and SSEA4, but not SSEA1, which are both present in human ES cells, the high molecular mass proteoglycan antigens TRA-1-60, TRA-1-81 and GCTM2 and the protein antigens Thy1 and MHC class 1. In contrast, murine EC and ES cells express SSEA1 but not the other markers [21]. Finally, strengthening the relationships between these cell types, concerning surface markers, human ICM cells from blastocysts also express similar patterns of surface antigen expression to both human EC and ES cells, emphasizing that the differences from the corresponding mouse cells most probably represent species differences in embryogenesis [25,26].

Most human EC cell lines do not require special culture conditions such as feeder layer support or the addition of extrinsic factors. This, however, may be due to their aneuploid nature resulting in part from an adaptation to the culture environment. In fact, hES cells tend to acquire similar chromosomal abnormalities as hEC cells do [27-29]. However, closer to the ICM are embryonic stem cells, which are directly derived from the ICM.

Murine and human embryonic stem cells

The techniques obtained while studying EC cells made it possible that in 1981 Evans and Kaufmann and independently Martin derived murine ES cell lines from the ICM of a blastocyst onto feeder cell layers [30,31]. The first characterisations of these cells unravelled the similarities with EC cells in terms of morphology, high level expression of cell surface antigens, including SSEA1 and their differentiation potential when removed from the feeder cells or giving rise to teratomas, when injected into mice. More importantly, concerning the biological function, when implanted into blastocyst and allowed to develop to term, they gave rise to chimeras [32].

However, ES cells resembled more closely the pluripotent cells of the blastocyst than EC cells and the contribution to the whole embryo is more efficient, while frequently contributing to the germ line.

Experiments for deriving ES cell lines were essential for designing “knock out” animal models, meaning experiments which made it possible to delete a functional gene and ultimately obtain a mouse line lacking this gene, due to the potential of ES cells to contribute to the germ line. Using this approach, the gene of interest is rendered non-functional by homologous recombination in ES cells. After selection of these cells, they are propagated and finally injected into a blastocyst to yield chimeras after transferring into the uterus of a pseudo pregnant mouse.

It took 14 years, due to logistical and ethical problems, until Thomson and colleagues could derive ES cells from rhesus monkeys in 1996 [33]. Afterwards, they derived the first human ES cells in 1998 [34]. Human ES cells (hES cells) were expected to reflect human embryogenesis more closer than the human EC cells and still offer an opportunity for regenerative medical therapies, moreover hES cells can form teratomas and give rise to all three germ layers [33-35].

Similar to EC cells and to murine ES cells, hES cells can be induced *in vitro* to differentiate. Concerning regenerative approaches it was very encouraging that these cells could be differentiated into a variety of lineages including neural, endothelial, pancreatic and haematopoietic [36]. New and more defined protocols for a directed differentiation are and will be established.

Similar to cancer cells, the telomerase positive hES cells can be maintained indefinitely *in vitro,* meaning they do not undergo senescence like normal human telomerase negative somatic cells [34,37].

In summary ES as well as EC cells are excellent *in vitro* systems to study self-renewal and early differentiation events.

Indispensably linked to this phenomenon is the transcription factor OCT4, as the expression of OCT4 is tightly linked to pluripotent cells as ES and EC cells and its ablation will always ultimately lead to differentiation events.

1.1.2 The POU family transcription factor OCT4

Early development of the mammalian embryo is controlled by regulatory genes, some of which regulate the transcription of other genes. These regulators encode so called transcription factors which activate or repress genes that mediate phenotypic changes during stem cell differentiation as well as embryonal carcinoma differentiation [38-40].

The expression level of the transcription factor OCT4 is essential for defining the transition of totipotent OCT4 positive cells in the morula stage towards OCT4 negative trophectoderm cells and the still OCT4 positive inner cell mass cells (see figure 1.2). Targeted disruption of OCT4 in mice has produced embryos devoid of a pluripotent ICM [9] (see Figure 1.3). Additional support comes from a study by Adjaye et al., where differential microarray analysis established OCT4 as a marker gene for the ICM [41]. Furthermore, in the ES cell model it seems that OCT4 acts dose dependent to maintain pluripotency by fulfilling gene regulatory functions. Overexpression of OCT4 drives ES cells towards the extra-embryonic mesoderm or endoderm lineages, while knockouts or knockdowns of OCT4 differentiate into trophectodermal lineage expressing CDX2. ES cells with a normal level of OCT4 remain pluripotent [42,43]. Ryo Matoba et al. showed that that at least 418 genes, 30 of which are primary targets, are regulated in a peculiar manner: the same gene is activated or repressed depending on the amount of OCT4. The presence of these “bell-shaped” and “inverse bell shaped” gene expression regulation relationships indicate that the maintenance of appropriate OCT4 levels is built into the gene regulatory network in mouse ES cells.

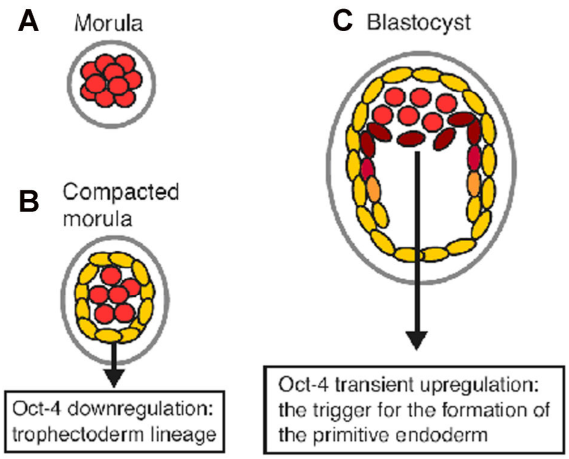

A
Morula
B
Compacted morula
C
Blastocyst
Oct-4 downregulation: trophectoderm lineage
Oct-4 transient upregulation: the trigger for the formation of the primitive endoderm

Figure 1.3 Events in the formation of the mouse blastocyst with respect to Oct-4 expression levels. A) Prior to compaction, all the cells of the morula express similar amounts of Oct-4 protein (orange colour). B) The formation of the trophoblast tissue is accompanied by down regulation of Oct-4 in the outer cells (yellow). C) Differentiation of primitive endoderm cells is preceded by transient up regulation of Oct-4 and subsequent shutdown (red). Adopted from Pesce et al. [44]

In contrast to the cell models, at maturity, long after the first differentiation events of the inner cell mass, OCT4 expression becomes confined exclusively to the developing germ cells [39,40]. In cell culture models, using hES cells, OCT4 knockdowns lead to the induction of differentiation, supposedly regulated by ACTIVIN, BMP, fibroblast growth factor, and WNT signaling pathways [45].

Concerning a pathological function of OCT4 expression, there is a correlation with some cancer types. The first cancer types, OCT4 was found expressed were germ cell cancers [46-52]. Based on the fact that the origin of germ cell development are primordial germ cells, where OCT4 is a marker gene [53], it became apparent that these cells might be the cells of origin in certain cancers of the gonads. The importance of OCT4 in this cancer progression was underlined first by an aberrant expression of OCT4, contributing to PGCs' malignant transformation [54] and second by over expressing OCT4 in a mouse teratoma model causing a more malignant histological phenotype while forced down regulating prevented tumour growth [49]. Concerning a possible role in cancer formation of OCT4 in human cancer, OCT4 expression has been shown for different breast cancer cell lines [55] and a colon cancer cell line [56] in contrast to untransformed cell lines. Furthermore, the identification of the reoccurrence of OCT4 in diverse cancer types has led to the assumption that a re-expression of OCT4 might be linked to the progeny of certain cancer stem cell types [57,58].

Finally there seems to be a correlation of OCT4 expressing human cancer types and an embryonic stem cell–like gene expression signature in poorly differentiated tumours, giving more support to the notion that OCT4 re-expression might be an important event for certain cancer stem cell types [57].

In summary OCT4 expression is critical for forming of the ICM, for development of the later embryo and for the development of the gonads. Finally, there is a correlation to the occurrence of certain cancer types. However many of these functions can not be operated by OCT4 alone.

OCT4 is known to interact with other transcription factors to activate and repress gene expression in mouse embryonic stem (ES) cells [44]. For example, OCT4, which is a member of the POU (PIT/OCT/UNC) class of homeodomain proteins, can heterodimerize with the high mobility group (HMG) box transcription factor, SOX2, to affect the expression of several genes in mouse ES cells as well as human ES cells [59,60]. The cooperative interaction of the POU homeodomain factor and the HMG factors is thought to be a fundamental mechanism for the developmental control of gene expression [61]. Like OCT4, if SOX2 expression is reduced in hES cells it results in the loss of the undifferentiated stem cell state. Again, similar to an OCT4 reduction, this can be indicated by a change in cell morphology, altered stem cell marker expression, and increased expression of trophectoderm markers [62]. These results were consistent with a dominant-negative form of mouse SOX2, which could induce trophectoderm differentiation and progressive polyploidy in mouse ES cells [63].

The importance of the OCT4-SOX2 complex can be demonstrated by reported biofeedback loops, which have been reported, meaning that the heterodimer will control the gene expression of its own transcription factors. In this regard Chew et al. showed by chromatin immunoprecipitation assay that both OCT4 and SOX2 bind directly to the promoters of *POU5F1* and *SOX2* in mouse and human ESCs, uncovering a positive regulatory loop for maintaining OCT4 and SOX2 expression [64].

These feedback loops extend to other important factors, required for maintaining the self renewal state of ES cells like the homeodomain-containing transcription factor NANOG. Rodda et al. could show that both OCT4 and SOX2 are required for the expression of NANOG in F9 embryonal carcinoma cells, embryonic germ cells and mouse ES cells [65].

More recently both OCT4 and SOX2 binding to the target promoter have been shown to be required for the expression of a micro-RNA called miR-302a, which represses the translation of cyclin D1, an important G1 regulator [66]. This is consistent with the observation that the length of the G1 phase of the cell cycle in ES cells is significant shorter, compared to differentiated, somatic cells [67-69].

In summary, both the ability to form heterodimers and act by regulatory feedback loops are two reasons which lead to a complex pattern where this transcription factor exerts its function.

1.1.3 The diversity of binding site recognition motifs of OCT4

Each transcription factor can be characterised by its ability to bind to a certain cis-element, meaning a specific DNA sequence. For OCT4 several cis-elements could be characterised.

One of these motifs is called the octamer motif and consists of 8 basepairs (the term “motif” refers in this study henceforth to a model of a transcription factor’s DNA binding specifity). It was first identified in the promoters of the histone H2B and the light and heavy chain immunoglobulin genes. The apparent paradox that the same element is required for both ubiquitous and B-cell-specific gene expression was resolved when two different proteins interacting with this sequence were characterised and cloned. OCT1 was present in all cell types tested, while OCT2 was detected only in B lymphocytes. Subsequently other POU class proteins have been discovered, including OCT4 [40].

Besides the recognition of octamer motifs, POU-class proteins have the ability to bind to different sequence elements. This is possible due to their inherent versatility in how they regulate transcription. According to Alexey Tomilin this is due to four often interdependent, factors: The first one is flexible amino acid–base interaction. A second reason are variable orientation, spacing, and positioning of DNA-tethered POU subdomains relative to each other [70]. Next, there are posttranslational modifications, and finally possible interactions with heterologous proteins [71].

POU domain proteins are able to bind to DNA cooperatively, thus conferring additional functional variability. The homo- and heterodimerization of OCT1 and OCT2 on immunoglobulin (Ig) heavy chain promoters (VH) provided evidence of cooperativity, with a dimer arrangement [72-74]. The cis-elements are considered to consist of low-affinity heptamer and high-affinity octamer sites separated by two nucleotides.

Another mechanism outlining cooperative DNA binding by POU proteins was determined during the course of an OCT4 target gene characterization [59]. The Palindromic-Oct-factor-Recognition-Element (PORE) with the recognition sequence ATTTGAAATGCAAAT (15 bp), of the Osteopontin (OPN) enhancer interacts with an Oct-4 dimer, thereby mediating strong transcriptional activation in preimplantation mouse embryos. Homo- and hetero-dimerization of other Oct factors like Oct-1 and Oct-6 on the PORE has also been demonstrated.

Yet there exists another palindromic DNA motif called MORE (More PORE) with the recognition sequence ATGCATATGCAT. This motif can assemble homodimers of OCT4 and heterodimers of OCT4/OCT6 and OCT4/OCT1, thus vastly expanding the diversity of OCT4 mediated direct DNA binding [70] (See Figure 1.4).

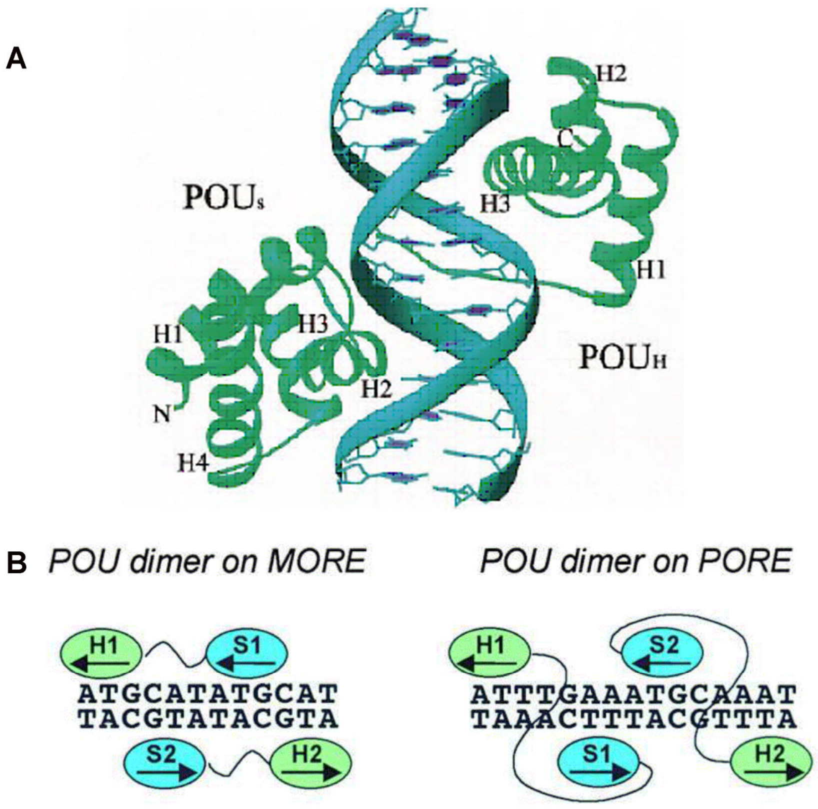


Figure 1.4 Different OCT4 binding modes
A 3 dimensional structure of the OCT4 DNA complex, the POU$_H$ domain fits to the big groove and the POU$_S$ domain fits into the small groove of the DNA helix.
B 2 dimensional model of how OCT4 dimers bind to the PORE and MORE sequences

Another known heterodimerization partner, mentioned above, is SOX2. The OCT4-SOX2 heterodimer will bind to a sox-oct element, in which the SOX2 motif and the OCT4 motif need to be positioned in close proximity, as has been suggested by using experimental high-resolution structure determination [75].

Given that OCT4 proteins could bind to several recognition sites, an emerging question became apparent if certain OCT4 cis-elements have a distinct functional role. In this regard, Jonathan et al. suggested that OCT4 binding to the PORE sequence could be a major mechanism of transcriptional control of stem cell self-renewal pathways using P19 mouse EC cells [76]. More specifically it has been demonstrated that OCT4 has the capacity along with OCT1 to respond to stress signals by selectively altering the affinity for complex binding sites *in vitro*, using mouse ES cells [77].

Taken together, this illustrates the special characteristics of this transcription factor to form a multitude of different direct DNA binding complexes. Some of these complexes have been demonstrated to have a specific functional role.

1.1.4 Transcriptional network for pluripotency

Important for reconstructing transcriptional networks are studies in a broad range of eukaryotes, which have shown that transcriptional regulators have key roles in cellular processes and frequently regulate other regulators associated with that process [78,79].

Using ES cells as a model and based on their unique expression patterns and the fact, that their presence is essential for the early development, the transcription factors OCT4, SOX2, and NANOG are thought to be central to the transcriptional regulatory hierarchy that specifies ES cell identity. Furthermore, they are the earliest-expressed set of genes known to maintain pluripotency [9,40,80-82]. Boyer et al. mapped OCT4, SOX2, and NANOG to their binding sites within known promoters. One conclusion was the revelation that these regulators collaborate to form in hES cells regulatory circuitry consisting of specialized autoregulatory and feedforward loops, explaining in part their profound role in developmental processes [83,84].

A ChIP-PET (Paired-End diTagging procedure enables to obtain sequence information from both termini of any contiguous DNA fragment) study using antibodies against OCT4 and NANOG identified 32 genes that were bound by OCT4 and NANOG in both mouse and human ES cells. The small overlap was partly explained by the unbiased approach by detecting binding sites in the whole mouse genome through PET sequencing and different technology platforms and reagents. However, among this list 18 encode transcription regulators including key pluripotency markers like NANOG, SOX2 and RIF1.

More recently, the lab of Stuart Orkin showed that in addition to the above mentioned transcription factors, other factors such as TCF3, DAX1, NAC1, KLF4, ZFP281, REX1 and MYC seem to be involved in a cooperative manner, meaning that at least 6 of these transcription factors are bound at the same promoter, especially in the putative regulation of genes related to developmental processes [85]. Thus at least for mouse ES cells, an expanded network of transcription factors needs to be assembled at their target promoters for preventing the differentiation of these cells.

Some of these targets are involved in signalling pathways, which are then again regulating the core circuitry factors for maintaining pluripotency. For example, it was shown that both TGFbeta and FGF signals synergize to inhibit BMP signalling; sustain expression of pluripotency-associated genes such as *NANOG, POU5F1,* and *SOX2*, and promote long-term undifferentiated proliferation of human ESCs. Furthermore it was shown that both TGFbeta- and BMP-responsive SMADs could bind to the *NANOG* proximal promoter. The conclusion was that *NANOG* promoter activity is enhanced by TGFbeta/Activin and FGF signalling and is decreased by BMP signalling [86,87] (see Figure 1.4). In mouse ES cells, Suzuki and colleagues showed that NANOG physically interacts with SMAD1 in mouse ESCs, thus interfering with the recruitment of coactivators to active Smad transcriptional complexes, and repressing the expression of BMP-responsive genes [88].

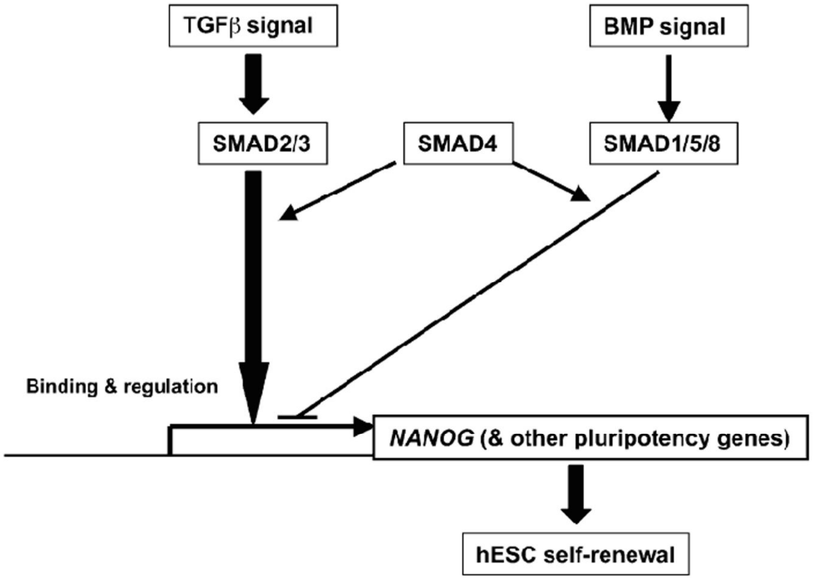


Figure 1.4 Model of SMAD Regulation of NANOG Transcription in Human ESCs. Arrows represent induction, and hammer-ended lines represent inhibition (Adapted from Xu et al., 2008)

It should be noted at this point, that there are striking differences between human and mouse ES cells in the way signalling pathways support their self renewal. An example is a member of the interleukin-6 related family of cytokines, so called leukaemia inhibitory factor (LIF), which is essential for mouse self-renewal and dispensable for human ES cell culture [34,89]. Furthermore, most of what is known of OCT4 related protein interactions and DNA binding motifs has been discovered in the mouse. The mouse and human OCT4 orthologs have a highly conserved nucleotide sequence and genomic organization [90,91].

In summary, there is a tight link between the key factors for pluripotency, OCT4, SOX2 and NANOG and signalling pathways which need to be activated or suppressed for the pluripotent capability of the cell.

1.2 Analysis of transcription factor binding sites

For the aim of studying transcription factor binding sites on a genomic scale two techniques need to be combined. One is called Chromatin-Immunoprecipitation followed by hybridization on arrays or mass sequencing, the other is referred as RNA-interference induced knockdown of target genes.

1.2.1 RNA interference

The expression status of the genes in the vicinity of the binding sites can give information on the overall association between the investigated factor and transcription. However, in case of transcription factors the effect on each target gene may be different. The method of choice is to compare the expression status of cells depleted of the transcription factor (TF) with that of normal cells. Thereby information can be gained on all genes influenced by the presence of the TF. In combination with ChIP-chip, direct targets can be separated from downstream pathways and the influence on each target gene can be determined. Common methods to achieve such a depletion are: knockdown by RNA interference (RNAi) or antisense methods (e.g. phosphorothioate-linked DNA [92], morpholinos [93] or genetic knockouts [94]. RNA interference (RNAi) is an intrinsic cellular mechanism which is conserved in most eukaryotic species. It plays a role in the regulation of gene expression, differentiation and defense against viral infections. RNA interference plays an important role in determining cell fate and survival. The relevance of the field has recently been acknowledged by the Nobel prize in Physiology or Medicine 2006 to Andrew Z. Fire [95] and Craig C. Mellow [96] for the first description of the phenomenon [97,98]. The natural mechanism has been utilized to artificially silence particular genes and thereby to gain insight into their functions [99].

RNA-Interference is a mechanism of eukaryotic cells, in which the expression of a target gene is suppressed by small interfering RNA (siRNA) [97,100-102]. Evolutionarily this mechanism was thought to develop as a protection strategy of the immune system against viruses or transposons. If a long double stranded

RNA (dsRNA) passes into a eukaryotic cell, this dsRNA will be degraded to small RNA-duplexes between 21 and 23 bp long, by an enzyme called RNase III Dicer [103-105]. Afterwards the Dicer-siRNA complex will be recruited by the protein transactivating response RNA-binding protein (TRBP) and transported to the RNA-induced silencing complex (RISC) [106,107]. This will be followed by the incorporation of the 5 prime end of that siRNA strand with the lower binding energy into the RISC complex, which contains a helicase activity. This is the antisense strand of the siRNA, which will be recognized by the RISC complex and will lead to the recruitment of a complementary target messenger-RNA (mRNA). The target mRNA will be cut by a catalytic subunit of RISC, called Argonaut 2 (AGO2). For this reaction the piwi domain of AGO2 is essential, containing a RNase H like endonuclease activity, which will cut the single phosphodiester bonds in the backbone of the target mRNA [108-110] (A scheme of the whole process can be seen in Figure 1.5).

This mechanism can be exploited as binding of a transcription factor (TF) to a given promoter, discovered by an ChIP-chip approach is insufficient in showing a putative regulatory effect e.g. activation or repression *in vivo*. However, such a function can be analyzed by coupling RNA interference with the analysis of the transcriptome.

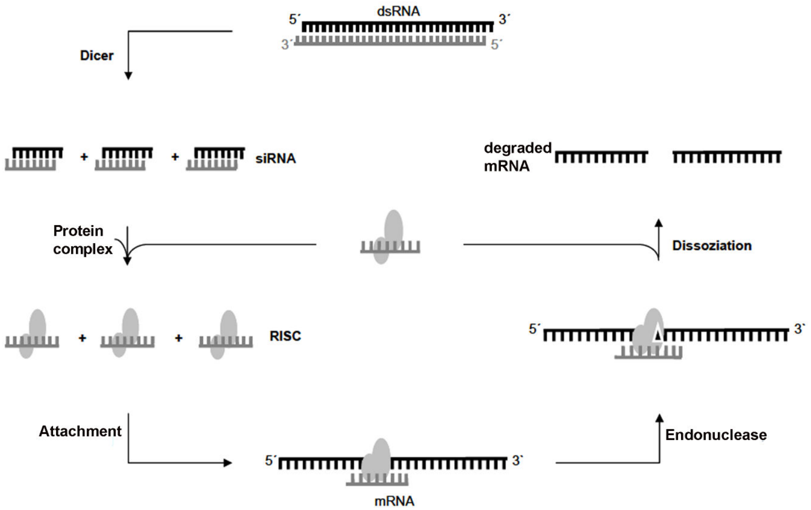


Figure 1.5: Mechanism of gene silencing in eukaryotic cells.

1.2.2 Chromatin Immunoprecipitation

Traditional methods for analyzing protein-DNA interactions include in vivo footprinting and chromatin immunoprecipitation (ChIP).

One method is called bandshift assay, also called gel shift or electrophoretic mobility shift. A band shift is observed when a protein forms a complex with a DNA fragment, because complexes of protein and DNA migrate through a non-denaturing polyacrylamide gel more slowly than free DNA fragments or double-stranded oligonucleotides.

Another method is called DNase footprinting [66], which allows one to compare the cleavage pattern of isolated DNA against that of the DNA in the presence of proteins. If the protein binds the DNA, the corresponding stretch is protected against DNase I cleavage and therefore fewer cleavage sites are found. In combination with gel-shift assays, the protected sites can be separated from the cleaved sites. This method allows the determination of the precise location of the protein binding sites.

In comparison, ChIP has the advantage of detecting the binding site on DNA in the natural genomic state and as such combined by Polymerase Chain Reaction could be used for the detection of binding sites in the whole genome.

The basic principle of a ChIP is that the proteins are cross linked to the DNA double helix by using crosslinking agents like formaldehyde. Formaldehyde is a tight (2 Å) crosslinking agent that efficiently produces both protein–nucleic acid and protein–protein crosslinks in vivo. Formaldehyde is a very reactive dipolar compound in which the carbon atom acts as a nucleophilic centre. Amino and imino groups of amino acids (lysines, arginines and histidines) and of DNA (primarily adenines and cytosines) readily react with formaldehyde leading to the formation of a Schiff´sche base. This intermediate can further react with a second amino group and condense to give the final crosslink. These reactions take place *in vivo* within minutes after addition of formaldehyde to living cells or embryos [111]. Although other crosslinking reagents have been employed [112], formaldehyde remains the most widely used as the reaction can be reversed by heat. This is achieved primarily by protonation of imino-groups at low pH in aqueous solution. After cross-linking the chromatin, the cells are either directly lysed or the nuclei are extracted. The chromatin is sheared into fragments of the desired size by sonication or through micrococcal nuclease digest to a size of usually 0.2-1.0 kb. For ChIP-chip applications, a smaller size is essential if a higher resolution of the subsequent analysis is desired. The fragments bound to the protein of interest are usually enriched by immunoprecipitation with an antibody against the respective protein. Protein-specific antibodies require optimizing immunoprecipitation (IP) conditions of each individual antibody necessary and often these may show unwanted cross-reactivity. Additionally, using different polyclonal antibodies for the same protein may show a different preference for epitopes, resulting in a possible different selection of cross-linked loci. Monoclonal antibodies would be preferable due to their specific epitope selection, but it is more difficult to obtain functional ChIP grade antibodies by this approach. As a control one sample is processed with the pre-immune serum from the host organism of the specific antibody used for the IP. This control identifies unspecific fragments enriched e.g. by adhesion to the samples tubes. The formaldehyde cross-links are then reversed and the precipitated DNA fragments are purified. Yields from ChIP are usually low but sufficient for subsequent PCR or qPCR analysis.

1.2.3 Chromatin Immunoprecipitation followed by microarray hybridization (ChIP-chip)

As the traditional methods had failed to create high-resolution, genome-wide maps of the interaction between a DNA-binding protein and DNA, the combination of chromatin immunoprecipitation (ChIP) and whole-genome DNA microarrays (ChIP-chip) circumvented these limitations by creating high-resolution genome-wide maps of the in vivo interactions between DNA-associated proteins and DNA.

The ChIP-chip technique was first used to identify binding sites for individual transcription factors in *Saccharomyces cerevisiae* [78,111,113]. More recently a c-Myc epitope protein tagging system was used to map the genome-wide positions of 106 transcription factors in yeast [114].

For microarray-based detection of immunoprecipitated DNA, amplification of the DNA is generally necessary, as the DNA yield, obtained after the pulldown is not sufficient for hybridization. Ideally the ChIP reactions are scaled up and amplifications are avoided. Three amplification methods have so far been widely used: randomly primed [115], ligation-mediated PCR [116] as well as amplification on the basis of T7 DNA polymerase [117]. Before adding the antibodies for the pulldown reaction, a part of the fragmented chromatin will be retained as total genomic reference DNA. Although these samples usually give enough material for microarray hybridization they should also be amplified to avoid any amplification bias. The enriched and the reference DNA are then fluorescently labeled. Although one color platforms, where both samples have the same label e.g. Cy3 are hybridized on separate arrays the use of two color platforms is often preferred, as this minimizes the influence of microarray batch effects on the experimental results. In this case the ChIP DNA is labeled with different fluorescent dyes and the samples are combined and hybridized to a single DNA microarray. The relative intensities of the two dyes allow the detection of the fragments that are enriched in the immunoprecipitation, thereby enabling the identification of protein-DNA interaction sites (see figure 1.6). For a comprehensive analysis, microarrays used in ChIP-chip applications represent ideally the entire genome of the organism in form of overlapping fragments. In this case the limitation will be the obligatory selection of preferred probe sequences for optimal hybridization, which in turn defines the maximal resolution of the tiling array. Furthermore, for larger genomes such as for higher eukaryotes these are not available or only at very high monetary cost. Therefore arrays are often custom designed for specific applications. The resolution of the identified binding sites depends on the size of the sheared DNA and the size and spacing of the probes on the arrays. For example, typical yeast experiments achieve a resolution of about 1 kb, which is sufficient to assign binding to the regulation of a single gene. Once the bound regulatory region is identified, the exact binding site can often be inferred by computational methods.

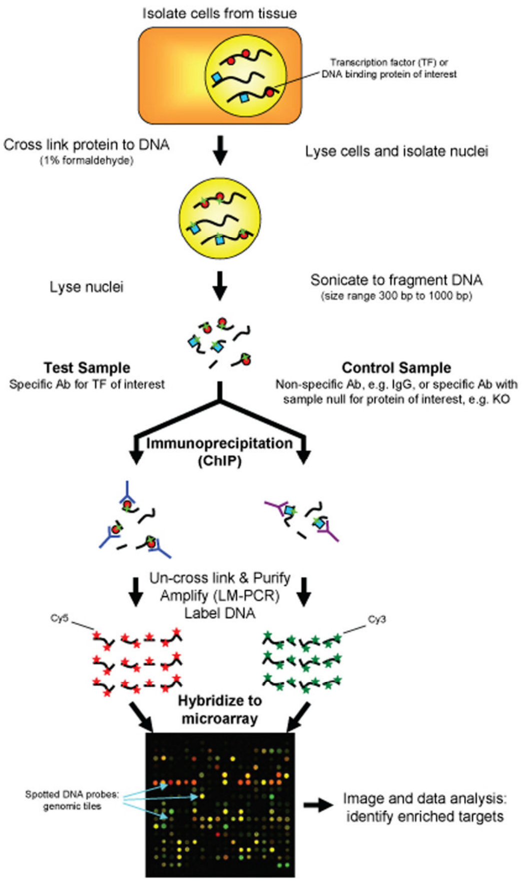

Isolate cells from tissue
Transcription factor (TF) or DNA binding protein of interest
Cross link protein to DNA
(1% formaldehyde)
Lyse cells and isolate nuclei
Lyse nuclei
Sonicate to fragment DNA
(size range 300 bp to 1000 bp)
Test Sample
Specific Ab for TF of interest
Control Sample
Non-specific Ab, e.g. IgG, or specific Ab with sample null for protein of interest, e.g. KO
Immunoprecipitation (ChIP)
Un-cross link & Purify
Amplify (LM-PCR)
Label DNA
Cy5
Cy3
Hybridize to microarray
Spotted DNA probes: genomic tiles
Image and data analysis: identify enriched targets

Figure 1.6 Principle of a ChIP-on-Chip experiment (Adapted from Peter White, PhD)

1.2.4 ChIP followed by sequencing (ChIP-seq)

One obvious disadvantage of the ChIP- chip application is the unavoidable bias, obtained when using arrays designed for selected promoter regions. The variation of binding sites is huge; the size of the region where cis-regulatory elements are found can vary by nearly three orders of magnitude from a few hundred bp to more than 100 kb. Regulatory regions have also been found downstream, in introns and even in exons of genes. The actual transcriptional regulation is achieved through a complex, combinatorial set of interactions between transcription factors at their binding sites [118].

ChIP-seq, in comparison to ChIP-chip offers a genome wide view of potential binding sites for a given transcription factor. Robertson et al. showed that the Solexa sequencing technology provides short read length sequences of ~30 base pairs that are optimized for characterizing ChIP-derived fragments [119]. Resulting sequences were mapped back to the reference genome, whereby the most frequently sequenced fragments formed peaks at specific genomic regions. ChIP-seq offers important advantages over ChIP-chip, including lower cost, minimal hands-on processing and a requirement for fewer replicate experiments as well as less input material. Furthermore, ChIP-seq offers a rapid analysis pipeline, as long as a high-quality genome sequence is available for read mapping. ChIP-seq also provides the potential to detect mutations in binding-site sequences, which may directly support any observed changes in protein binding and gene regulation.

By using the Solexa system, single molecules are covalently attached to a planar surface and amplified *in situ*. Sequencing by synthesis is carried out by adding a mixture of four fluorescently labeled reversible chain terminators and DNA polymerase to the template. This results in addition of a single reversible terminator to each template. The fluorescent signal is detected for each template, and the fluorophore and the reversible block are removed. The terminator–enzyme mix is then added to start the next cycle, and the process is reiterated until the end of the run. Given that all four nucleotides are present in the reaction, the risk of mis-incorporation is minimized, increasing sequencing accuracy. Accuracy is also independent of sequence context, and a discrete signal is generated for every base [120].

As mentioned above, the Solexa sequencing technology [120] provided short read length sequences of approx 30 base pairs that were ideal for characterizing ChIP-derived fragments. Robertson et al. mapped the resulting sequences back to the reference genome, whereby the most frequently sequenced fragments formed peaks at specific genomic regions. They then analyzed sequences under these peaks by comparison known STAT1 binding site sequences and locations, and for their proximity to genes. They also compared the results to previous STAT1 ChIP-chip data. Their comparison of STAT1 binding locations in human HeLa S3 cells stimulated by interferon gamma versus unstimulated cells showed that stimulated cells provided evidence of 4-fold more STAT1-bound sites, and that specific sites bound in these interferon gamma–stimulated cells correlated well with expectations from previous studies [121,122]. Peaks for both unstimulated and stimulated cells showed the highest density at approximately 100 base pairs upstream of the transcriptional start sites of nearby genes. a similar study by Johnson and colleagues [123] that defined genome-wide binding sites of the neuron-restrictive silencer factor (NRSF), provide initial evidence that next-generation sequencing platforms like Solexa are being coupled with previously applied techniques to generate genome-wide views of protein-binding phenomena.

1.2.5 Identification of enriched sequences in ChIP-chip experiments as a way to access potential binding sites

With the advent of detecting enrichment signals in a global scale it became imminent to develop algorithms, called peak finding algorithms, which would assist in the screening of genomic loci, which were correlated to a putative transcription factor binding site. In recent years, several methods have been developed to detect peak regions [85,124-126]. Cawley et al. [125] and Keles et al. [127] applied the Wilcoxon rank sum test and t-test, respectively, to generate test-statistics for sliding windows. Cawley et al. used a fixed p-value cutoff to select peak regions. Whereas Keles et al. employed the Benjamini and Hochberg step-up procedure [128] to control false discovery rate (FDR). In addition to the requirement for experimental replicates, Gottardo et al. [126] identified the absence of powerful multiple testing adjustment methods as a limitation of these methods. Li et al. [129] proposed a hidden Markov model (HMM) approach to identify peak regions, assuming model parameters could be estimated from previous experiments. Ji et al. [130] used a modified t-statistic with a more robust estimate of variance to measure probe-level binding signal, then used either moving window averaging or HMM to estimate window-level binding signal, and finally estimated local false discovery rate (LFDR) of each peak region [131]. Estimation of LFDR requires dissection of the mixture distribution of ChIP-chip signals, which includes the distribution of ChIP enriched signals (or peak signals) and the background (null) distribution. Ji et al. [130] estimated the mixture distribution by unbalanced mixture subtraction, which requires additional information to construct the unbalanced mixtures. Instead of concentrating exclusively on the strengths of binding signals, Zheng et al. [132] identified peaks using both signal strength and signal pattern. Specifically, they modeled the DNA fragmentation process with a Poisson point process and concluded that if the binding signal is transformed to log scale, isolated "peaks" should exhibit a triangular shape allowing development of a double regression method, Mpeak, to identify triangular patterns from ChIP-chip data [133] (See Fig 1.7 for some peak examples).

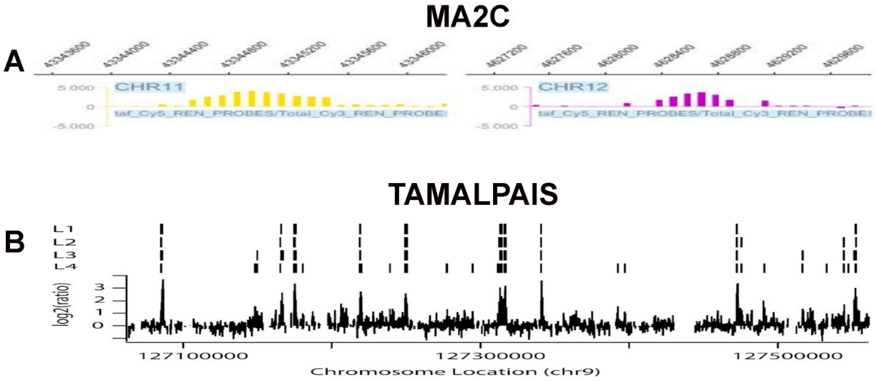


Figure 1.7 Peak profiles of two different algorithms
A Two MA2C detected peaks
B TAMALPAIS detected peaks for different threshold levels (L1 –L4) for a specific region of chromosome 9.

The computational challenge of applying such algorithms is to normalize the data properly and to detect confident enriched regions by filtering out false peaks. One of these programs which is capable of doing that step, developed in 2007 is MA2C [134]. The normalization method of MA2C was more effectively than median scaling and removes much of the GC-effect, as two-color arrays exhibited a sequence bias, particularly dependent upon the GC content of probes. More precisely, probes with high GC counts tended to have a high intensity. Incorporating these effects improved the detection of true positive peaks. Another algorithm, which was developed to screen peak regions was first used in a study of Peggy J. Farnham and colleagues in 2006, which has become available as a web service meanwhile [135]. They sought an approach to peak detection that made minimal assumptions about the shape and amplitude of peaks representing true binding sites. The binding sites should appear in the data as runs of consecutive points (each point representing a 50-mer) with enhanced amplitude. Compared to previously used algorithms, [85], this left open the question of setting an appropriate combination of threshold and width for each array. Clearly, a threshold requirement for an array that shows strong signals should be very different than for an array that shows weaker signals. Therefore, for a threshold they used a percentile for each array (95th and 98th percentile) of log2 oligomer ratios. Use of this percentile “normalizes” the threshold values for each array to reflect both the amplitudes and distribution of signal in the arrays and, furthermore, presented a consistent, nonarbitrary way to set thresholds for different arrays.

1.2.6 Motif analysis and modules

One of the major challenges in molecular biology is the unravelling of the complex system that regulates the expression of genes. Essential for this aim is the ability to identify regulatory elements, specifically the binding sites of transcription factors.

Gene expression is regulated by transcription factors binding to specific transcription factor binding sites in regulatory regions associated with genes or gene clusters. Identification of regulatory regions and binding sites (called motifs) is a prerequisite for understanding gene regulation, and as experimental identification and verification of such elements is challenging, much effort has been put into the development of computational approaches. However, the recently tremendous increase of the diversity of motif discovery programs, each of them having its own advantages and disadvantages, makes it especially difficult to find the best solution for a given task, e.g. the discovery of motifs linked to enriched sequences of a ChIP-chip or ChIP-seq experiment [136].

Modules are often referred as clusters of binding sites for cooperating TFs. One part of describing a module can be achieved by defining the distance between single motifs and the occurrence a set of motifs will appear in a given context, e.g. correlating with peak sequences. This theory has been applied successfully to human EC cells in finding a new cooperation partner of OCT4. The principle was to screen OCT4 associated peak sequences for the enrichment of other transcription factor binding sites, resulting in the identification of an OCT4 and SRY regulatory module [137]. New software tools, freely accessible, such as Cisgenome are recently becoming more available, allowing experimental biologists to ask questions how peak sequences are related to a potential pattern of motif inherent to a specific Chip on chip or ChIP-seq study [138].

1.3 Aim of this work

The aim of this work was to gain more detailed insights into the understanding of the OCT4 dependent transcriptional network in pluripotent cell lines by detecting new putative targets of OCT4, which function as hubs. Hubs are on top of a functional or regulatory hierarchy and were to be identified by combining all detected binding sites in OCT4 knockdown experiments. The second aim was to draw differences between human EC and ES cells and to investigate in more detail the role of key factors that are needed for maintaining the pluripotent state using knockdown or overexpression, coupled with subsequent microarray analysis. Preferred were factors with functions in apoptosis and cell cycle pathways as recent studies have discussed links to the core factors, regulating pluripotency [67,68,139]. Furthermore, the differentially regulated genes were to be functionally characterized and screened for links to self-renewal pathways. Finally, the distribution and prevalence of the OCT4 binding patterns in terms of different binding modes should be analyzed.

The rationale to use specifically NCCIT cells came from a study, which demonstrated the differentiation potential by using esiRNA against OCT4, SOX2 and NANOG [38]. This study had supported the idea that hECC and hESC share a number of characteristics in the context of maintaining their cellular identity.

A technical focus of this work was to choose and define algorithms or programs that could predict with high sensitivity and specificity the occurrence of potential binding sites in a more unbiased way, concerning the applied algorithms. Some of these binding sites were to be confirmed by using *in vitro* methods such as band shift assays.

Furthermore, the data obtained from these large scale experiments were to be connected to published datasets and possible differences between the cell lines were to be characterized.

2 Material and Methods

2.1 Molecular biology

2.1.1 Polymerase Chain Reaction

PCR-Reactions for the specific amplification of DNA-fragments were done in the following order:

add 25 µl bi-dest water
2,5 µl 10X-buffer (see below)
0-2,5 µl 25mM $MgCl_2$ [1]
0,2 µl dNTP-Mix (dATP, dCTP, dGTP, dTTP at each 25mM)
every 0,25µl 100µM primer
1,25 µl DMSO [2]
0,5 µl 10U/µl Taq/Pfu-Polymerase-Mixture [3]
0,5-5 µl DNA-Templat [4]

[1] 0 µl only if the chosen PCR buffer already contained $MgCl_2$
[2] optional
[3] *Pfu*-part varies between 0 and 100%
[4] genomic DNA (≈50ng), cDNA (≈50ng-RNA-equivalent) or diluted PCR-Product (5-10ng)

The choice of suitable primer sequences were done in the way that the hybridization temperature calculated by the program NetPrimer (http://www.premierbiosoft.com/netprimer/) for oligo lengths between 22-28 bp was between 60 °C and 66 °C and the primers were cross and self dimerized only in an acceptable way for ΔG: >-5 kcal/mol for 3' end self-dimer and ΔG >-6 kcal/mol for internal dimmers. Using genomic DNA as a template, repetitive parts of the sequence of aim were excluded by using the program RepeatMasker (http://www.repeatmasker.org/). The mainly used 10X buffer contained:

500 mM Tris-Cl pH 8,8
200 mM $(NH_4)_2SO_4$
15 mM $MgCl_2$
0,1% (v/v) Tween 20

PCR reactions were performed as touchdowns PCR, meaning the annealing temperature was decreased step wise to a given value. Reactions were started

"hot", meaning the starting temperature was already 95℃ when the samples were inserted.
Here is a depicted a typical program:

94℃ 2'30''
94℃ 30''
68℃ 45'' 12 repititions, T_{ann}
* -1℃/cycle
72℃ 1'/kb product length
94℃ 30''
56℃ 45'' 15-19 repititions
72℃ 1'/kb
72℃ 5'
7℃ ∞ *_annealing_-temperature

2.1.2 Isolation of plasmid DNA

3ml LB media supplemented with appropriate selection antibiotics were inoculated with a single colony and grown overnight at 37℃ on a shaker. Cells were centrifuged at 3,000 rpm for 10 min. Plasmid DNA was isolated using the QIAGEN Plasmid Mini Kit, which is based on a modified alkaline lysis procedure, followed by binding of plasmid DNA to an anion-exchange resin under appropriate low salt and pH conditions, according to the manufacturer's protocol. The DNA pellet was washed two times with 70% ethanol and dissolved in 30µl 1x TE buffer. For Maxipreps, Plasmid DNA was isolated using the NucleoBond Xtra Maxi Plus EF Kit, according to the manufacturer's protocol. For this purpose, 3ml LB media supplemented with appropriate selection antibiotics were inoculated with a single colony and grown for 8 h at 37℃ on a shaker. The volume was then transferred to 300ml LB media and grown overnight at 37℃ on a shaker. Finally, the precipitated Plasmid DNA in the filter was eluted with 1ml TE buffer, passing 2 times through the filter. For mammalian cell transfections, plasmids with an OD 260/280 > 1.9 were used.

2.1.3 Gel extraction and PCR purification

The gel extraction kit from Qiagen was used according to the manufacturer's instructions for PCR product purification and to extract DNA fragments from agarose gels.

2.1.4 Cloning and Sequencing of PCR-Products

PCR products were cloned by using primers containing given restriction sites for the target construct. For the sequencing of PCR products, they were isolated by running on an agarose gel, cutting the desired region out and cleaning by the use of the MinElute PCR Purification Kit (QIAGEN). The sequencing reaction and analysis was done by providing the vector construct to the service group of the MPI for Molecular Genetic.

2.1.5 Ligation

The final reaction volume for ligation was 20µl. 100ng of vector was used with the molar ratio of vector to insert being set at 1: 3 to 1:5.

Vector 100ng
3 x insert 1µl
10 x buffers 1µl
T4-DNA-ligase 1µl
H2O to 20µl

The reaction mixture was incubated at 24 °C for 1h. 10µl of ligated mixture was used for transformation into competent bacteria.

2.2 RNA analyses

2.2.1 Total RNA isolation using RNeasy® Mini Kit

Using the RNeasy® Mini Kit

To adherent cell lines like the NCCIT cells 350µl of RLT buffer (Qiagen) with 1% ß-mercapto-ethanol were added. The cell lysate was homogenised by passing through a 19G gauge needle 5 times. RNA isolation from the homogenates was performed using the RNeasy® Mini Kit (Qiagen) including DNase I on column treatment to get rid of trace amounts of genomic DNA following the manufacturer’s protocol.

2.2.2 RNA and cDNA quantification

The quantity of RNA and DNA was determined using the NanoDrop (NanoDrop Technologies, Wilmington, DE, USA). 1–2µl of sample was applied to the

NanoDrop and measured. If the concentrations exceeded measurable values the samples were either concentrated by speed vac centrifugation or diluted by adding dH_2O, respectively.

2.2.3 Agarose gel electrophoresis

Agarose gel electrophoresis and ethidium bromide staining enabled the visualizing of RNA and DNA for quality control. By mixing 0.5-1.5g of agarose (Life Technologies, Paisley, Scotland) and 50ml of 1x TAE, gels of 1-3% were obtained. 1µl of ethidium bromide (10mg/ml; Invitrogen) was added directly to the gel and mixed before solidifying. To assign the length of the amplicons the GeneRuler^{TM} 1kb DNA ladder (Fermentas, St. Leon-Rot, Germany) was used. Prior to loading of samples, a third of the volume of 6x loading buffer (Fermentas) was added to the samples. Gels were run in an electrophoresis chamber with 50V for 30 to 60min. Nucleotides were visualized with UV light using the AlphaImager^{TM} (Alpha Innotech, San Leandro, CA, USA).

2.2.4 Reverse transcription

Using Superscript II

For reverse transcription using Superscript II (Invitrogen), 1.0µl (1µg/µl) RNA was added to 1.0µl of50 µM Oligo-dT primer plus 8.0µl of dH_2O. The mixture was spun briefly, heated to 70°C for 5min and cooled on ice. 10.0µl of master mix were added including the following components per reaction: 4.0µl of 5x RT buffer, 2.0µl of 0.1M DTT, 2.0µl of (10mM) dNTP, 1.0µl (200U/µl) Superscript II and 1.0µl of dH_2O. After pulse spinning, incubation was carried out at 42°C for 1.5hrs.

Using M-MLV reverse transcriptase

For reverse transcription using M-MLV reverse transcriptase (Promega, Madison, WI, USA), 2.0µl (1µg/µl) RNA was added to 0.5µl of Oligo-dT primer

(1µg/µl; Invitek, Berlin, Germany) plus 7.0µl of dH_2O. The mixture was spun briefly, heated to 70℃ for 3 min and cooled on ice. 15.0µl of master mix were added including the following components per reaction: 5.0µl of 5x reaction buffer (Promega), 0.5µl of (25 mM) dNTP, 0.1µl of M-MLV reverse transcriptase (200U/µl; Promega) and 9.4µl of dH_2O. After pulse spinning, incubation was carried out at 42℃ for 1.0hrs and then stopped at 65℃ for 10min.

2.2.5 Real-time polymerase chain reaction (Real-Time PCR)

Real-Time PCR was performed in 96-Well Optical Reaction Plates (Applied Biosystems, Foster City, CA, United States). The PCR mix in each well included 10µl of SYBR®Green PCR Master Mix (Applied Biosystems), 5µl dH_2O, 1.5µl each of the forward and reverse primers (5ng/µl; Invitek) and 2 µl of single strand cDNA (2.5ng/µl) in a final reaction volume of 20µl. Triplicate amplifications were carried out per gene with three wells as negative controls without template. GAPDH and ACTB were amplified along with the target genes as endogenous controls for normalization. The PCR reaction was carried out on the ABI PRISM 7900HT Sequence Detection System (Applied Biosystems) using the following program, stage 1: 50℃ for 2min, stage 2: 95℃ for 10min, stage 3: 95℃ for 15s and 60℃ for 1min, for 40 cycles and, stage 4: 95℃ for 15s, 60℃ for 15s and 95℃ for 15s. The last heating step in stage 4 was performed with a ramp rate of 2% in order to enable the generation of a dissociation curve of the product.

The output data generated by the Sequence Detection System 2 software were transferred to Excel (Microsoft, Redmond, WA, USA) for analysis. The differential mRNA expression of each gene was calculated with the comparative Ct (threshold cycle) method recommended by the manufacturer.

2.2.6 Illumina bead chip hybridisation

Biotin-labeled cRNA was produced by means of a linear amplification kit (Ambion, Austin, TX, USA) using 300ng of quality-checked total RNA as input. Chip hybridisations, washing, Cy3-streptavidin staining, and scanning were performed on an Illumina BeadStation 500 platform (Illumina, San Diego, CA, USA) using reagents and following protocols supplied by the manufacturer.

cRNA samples were hybridised on Illumina human-8 BeadChips. We hybridised the samples in biological triplicates, and in biological duplicates.

2.3 Protein analyses

2.3.1 Protein isolation

NCCIT Cells were homogenized in 500µl lysis buffer (25% glycerol, 0.42M NaCl, 1.5mM $MgCl_2$, 0.2mM EDTA, 20mM HEPES) and with addition of 5µl protease inhibitor

2.3.2 Protein quantification (Bradford)

Protein samples were quantified using the Bradford method. 10x bovine serum albumin (BSA, 1µg/µl; Sigma-Aldrich, Munich, Germany) was used as a standard and the samples were diluted 1:5 in 1x PBS before use in the assay. Standards and samples were brought to 50µl by adding dH_2O. For the samples 1µl of sample was mixed with 49µl of dH_2O. The standards were mixed in 7 different dilutions to enable a standard curve. The volume of 10x BSA was 0, 2, 4, 6, 8, 10 and 12µl mixed with the needed volume of dH_2O to get the final volume of 50µl. Bradford solution (Bio-Rad Protein Assay; Bio-Rad, Hercules, CA, USA) was diluted 1:5 with 1x PBS and 950µl of the diluted Bradford solution were added to each standard and sample. The mixtures were incubated on the bench for 5min at RT. Afterwards they were transferred to 1ml cuvettes (Sarstedt, Nümbrecht, Germany) and measured using the Ultrospec 3100 pro (GE Healthcare, Munich, Germany) and the provided Bradford programme of the photometer.

2.3.3 SDS-PAGE gel electrophoresis

Protein gels were poured in Bio-Rad protein chambers. To get good separation for the target protein a 10% gel was used. A 10% resolving gel was prepared by sequentially adding 2.45ml of dH_2O, 1.25ml of resolving buffer (see appendix I), 50µl of 10% SDS, 1.25ml of 40% acrylamid (Rotiphorese® Gel 40; Carl Roth, Karlsruhe, Germany), 25µl APS (Ammoniumperoxodisulfate; Carl Roth) and

2.5µl TEMED (Carl Roth) followed by well mixing and transfer to the chamber. For the time of solidifying the gel was covered with isopropanol to get an even edge. After setting the isopropanol was discarded and a 5% stacking gel was prepared by sequentially adding 1.5ml of dH_2O, 0.6ml of stacking buffer (see appendix I), 25µl of 10% SDS, 0.3ml of 40% acrylamid, 25µl APS and 5µl TEMED followed by well mixing, transfer to the chamber and applying a comb.
21µg of protein were loaded by mixing 7µl of protein (~3µg/µl) with 3.5µl of 3x loading buffer (see appendix I). Prior use genomic DNA in the protein samples was disrupted by pipetting up and down 5 times with a BD Microlance^{TM} 3 injection needle (Becton Dickinson, Madrid, Spain). The samples and 10µl prestained protein marker (New England Biolabs, Beverly, MA, USA) were heated to 95℃ for 5min and afterwards cooled on ice for 1min before loading. The gel was run in 1x running buffer (see appendix I) with 110V until the loading buffer front did pass the whole gel. Gels were then used for western blotting.

2.3.4 Western blotting

Proteins were transferred from the gel to an Amersham Hybond^{TM} ECL^{TM} nitrocellulose membrane (GE Healthcare) by building up a blot in the following order; filter paper, membrane, gel, filter paper. The blot was then covered with cellular material from both sides and placed in the transfer chamber (Bio-Rad). The blot was run in ice cooled 1x transfer buffer (see appendix II) with a constant 350mA for 1 h. After blotting the protein quality was checked by Ponceau Red staining of the membrane using Ponceau S Solution (Sigma-Aldrich). The membrane was shortly washed with dH_2O and then blocked with blocking solution (see appendix II) by shaking for 5 min at RT and then over night at 4℃.
After short washing with 1xTBST primary antibody was applied to the membrane by shaking 1hrs at RT in 0.5g BSA dissolved in 10ml 1x TBST plus 2µl primary antibody. Afterwards the membrane was again shortly washed with 1x TBST and then extensively washed by shaking 4 times for 5min in 1x TBST. The secondary antibody was then applied by shaking 1h at RT in 10ml blocking solution plus 2µl secondary antibody. Afterwards the membrane was again shortly washed with 1x TBST and then extensively washed by shaking 4 times for 5 min in 1x TBST.

250µl of detection reagent 1 and 250µl of detection reagent 2 (GE Healthcare) were mixed in a tube and kept in the dark until use. The membrane was placed on foil and the mixture was dispensed on the membrane. The membrane was directly covered with foil by avoiding air bubbles and incubated for 1 min. The liquid was then disposed from the membrane and the membrane was placed in a Hypercassette^{TM} (Amersham). In a dark room BioMAx XAR film (Kodak, Stuttgart, Germany) was exposed for 20-60s to the membrane and directly developed using the Curix 60 develop machine (Agfa, Cologne, Germany).

2.3.5 Chromatin Immunoprecipitation (ChIP)

Human NCCIT cells were grown to a final count of 5×10^7–1×10^8 cells for each location analysis reaction. Cells were chemically crosslinked by the addition of one-tenth volume of fresh 11% formaldehyde solution for 10 min at room temperature. Cells were rinsed twice with 1 × PBS and harvested using a silicon scraper and flash frozen in liquid nitrogen and stored at -80°C prior to use. Cells were resuspended, lysed in lysis buffers, and sonicated to solubilize and shear crosslinked DNA. Sonication conditions vary depending on cells, culture conditions, crosslinking, and equipment. We used a BRANSON 250 and sonicated at power 3 for 11:00 min with 30% Duty Cycle at 4°C while samples were immersed in an ice bath. The resulting whole cell extract (WCE) was incubated overnight at 4°C with 100µl of Dynal Protein G magnetic beads that had been preincubated with 10µg of OCT4 antibody (insert). Beads were washed five times with RIPA buffer and once with TE containing 50mM NaCl. Bound complexes were eluted from the beads by heating at 65°C with occasional vortexing, and crosslinking was reversed by overnight incubation at 65°C. Whole-cell extract DNA (reserved from the sonication step) was also treated for crosslink reversal. Immunoprecipitated DNA and whole-cell extract DNA were then purified by treatment with RNaseA, proteinase K, multiple phenol:chloroform:isoamyl alcohol extractions and precipitation with ethanol. Purified DNA was amplified using a one-stage random PCR protocol. For ChIP-on-chip assay three biological replicate ChIP experiments were performed. Labelling and hybridisation of ChIP-DNA was done by NimbleGen. Using the NimbleGen human promoter tiling arrays (HG18) we screened 6517 putative promoter regions more, with a median probe spacing of 100bp, compared to the

OCT4 ChIP-on-Chip done by Boyer et al. Though the chip was covering only 4250bp, these probes were within the most abundant TF binding sites, using TRANSFAC [3].

2.3.6 Amplification of ChIP and Input DNA

Linear amplification of ChIPed DNA and input control was carried out on the basis of random primer amplification developed by Bohlander et al. [115] and which was subsequently modified for ChIP applications [114] except only one round of amplification with 20 cycles was performed. Amplified samples were purified using Wizard SV PCR purification kits according to the manufacturer's instructions. DNA quality was confirmed by real time analysis of 4 downstream targets of OCT4 and 5 OCT4 independent targets. Samples were labeled and hybridized according to NimbleGen standard procedure.

2.3.7 Band shift assays (EMSA)

For the Bandshift, nuclear extracts were prepared from NCCIT cells, using the method of Dignam et al., with the modifications of Rodda et al. dsDNA oligonucleotides (INVITEK) labeled with cy5 at the 5´termini of both strands. Sense strand sequences are provided in supplementary:

For DNA binding reactions 4µl (40µg) of nuclear extract was added to a 40µl reaction (final) containing 50nM cy5 oligonucleotide and 5µg poly-dGdC (Amersham). The final binding buffer composition was 60% with 1 µg/µl BSA. Where specified 1µM unlabeled ds competitor was also included prior to the addition of nuclear extracts. Where specified 2µl anti-Oct4 (sc-9081x, Santa Cruz) antibody were added. Binding reactions were resolved on pre-run 6% native PAGE gels in 0.5X TBE for overnight at 50V. Gels were imaged directly using a Fuji film FLA-5100-R radioluminographic scanner.

2.3.8 Pulldown assays using biotinylated DNA

50µl streptavidin conjugated Dynabeads (Dynal) were washed with PBS-BSA (PBS, pH 7.4, 0.1% BSA) for each sample. Biotinylated USP44 promoter fragment DNA (100 pmol) was incubated with the streptavidin beads for 4h at 4 °C with rotation. Dynabead-DNA complexes were extensively washed with PBS-BSA to remove unbound DNA. Beads were added to 1000µg Nuclear

Extract of NCCIT cells (in Buffer D: 20mM Hepes, pH 7.9, 20% glycerol, 100mM KCl, 0.83mM EDTA, 1.66 mM Dithiothreitol, 1% protease inhibitor mixture, 50µl polyGdC and 300X scrambled oligo). Samples were incubated for 8h at 4℃ with rotation. Dynabead-DNA-protein complexes were magnet separated and washed three times with ice cold Buffer D, adding 300X scrambled oligos each time. Samples were transferred to fresh microfuge tubes prior to final wash to avoid eluting plastic bound proteins. Dynabead-DNA-protein complexes were eluted in SDS-reducing sample buffer by heating at 95℃. Duplicate samples were pooled and equal volumes loaded onto 10% polyacrylamide gels for SDS-PAGE. Samples were transferred to nitrocellulose membranes and subjected to Western blot analysis. Western blotting was performed according to standard procedures and using chemiluminescence detection (ECL – Amersham). Antibodies used were Santa Cruz sc-8629 (OCT4) and PARP1 (sc-7150).

2.3.9 Chromatin Immunoprecipitation followed by sequencing (ChIP-seq)

For ChIP-seq analysis, the three biological replicates used for ChIP-chip analysis were pooled for each the OCT4 enriched and for the control DNA, and prepared for the sequencing reactions according to the manufacturing manual (Illumina).

2.4 Cell culture

2.4.1 Embryonal carcinoma cells, NCCIT cells

hEC cells were grown in high-glucose DMEM supplemented with 10% FCS (Biochrom, Berlin/Germany), 2 mM glutamine, and penicillin/streptomycin on conventional tissue culture plastic surfaces. Different lines were compared with regards to their growth properties.

2.4.2 Transient Transfections

For transient transfection, ExGen 500 was used, according to the manufacturer’s instructions. ExGen 500 is a sterile solution of linear polyethylenimine molecules in water. For transfecting NCCIT cells, 2µg of vector DNA were mixed with 6,6µl of ExGen 500 in 100µl OptiMem and incubated for 10 minutes at room temperature. 50µl were applied to 20-30% subconfluent cells in a 12-well plate. Transfection efficiency was checked by use of a reporter plasmid, expressing GFP.

2.5 Data analysis

2.5.1 OCT4 ChIP array analysis

Peak finding algorithms

Brute force

Based on the quantile normalized data, for each oligonucleotide a fold-enrichment was calculated by dividing the signal intensity from the immunoprecipitated sample by the signal intensity of the whole-genome sample. For each array, the total ChIP/IP ratio distribution was examined in order to obtain array specific threshold values for the upper 0.01 and for the upper 0.05 quantile. A potential binding event is defined with respect to the estimated average fragment size of the sonicated DNA (550bp) in relation to the distance of oligonucleotides relative to the promoter regions of the examined TSSs (distances between oligonucleotides is 100bp). Therefore, a potential binding event is defined as at least three oligonucleotides that fulfil the following criteria: a centre oligonucleotide has a ChIP/IP ratio within the upper 0.01 quantile of the total ratio distribution and one upstream and one downstream neighbour each within a distance of max. 1000bp have a ChIP/IP ratio within the upper 0.05 quantile. All identified peaks are connected to the closest transcription start site (TSS), if one exists within a distance of 8kb. Genomic positions of transcription start sites are based on Ensembl4 and were downloaded via biomart5.

MA2C

MA2C was used with standard settings for first normalizing our PairData files for each of the five experiments and thereafter searched for peaks [134]. Promoter1 3 replicates had 269, 504, 460 peaks. Promoter2 two replicates had 1366 and 915 peaks. When providing MA2C with all three replicates simultaneously and using replicate function the program identified 830 peaks for promoter1 and 1208 for promoter2. When all three programs identify a peak close to a gene then the peaks found by MA2C tend to have the highest motif scores for OCT4 associated to it.

TAMALPAIS

We used a web version of the Tamalpais program for analysing already normalized files provided by NimbleGen [135]. Tamalpais searches for peaks in each array separately and lists as an output all peaks and their occurrences in different replicates. We chose the lowest stringency set of L4 for further analysis. We had for promoter1 1036 peaks in total, 54 that were found in all three replicates (max gap allowed between peaks is 50bp), 93 that were found in two and 889 identified in only one replicate. For promoter two we had only two biological replicates and for these we found 505 peaks, 32 of which were found in both and 419 that were found only in one replicate.

De Novo Motif Search

The new OCT4 seqlogo was made by mapping the motifs that had levenshtein distance which measures the changes that have to be made (insertions, deletions, substitutions) to make two sequences equal to the ACGTAAAT OCT4 consensus sequences, allowing a maximum of 2 mismatches. We mapped all those motifs back to all the peak regions, took the longest matches allowing at most 1 bp gap between two motifs from the input set. We then aligned these motifs and produced a pwm and a sequence logo.

2.5.2 Microarray expression analysis

All basic expression data analysis was carried out using the manufacturer's software BeadStudio 3.1.3.0 (Illumina). Raw data were background-subtracted and normalised using the "rank invariant" algorithm, by which negative intensity values may arise. Normalized data were then filtered for significant expression on the basis of negative control beads. Selection for differentially expressed genes was performed on the basis of arbitrary thresholds for fold changes plus statistical significance according to an Illumina custom model [140].

Differentially expressed genes were further filtered according to Gene Ontology terms or mapped to KEGG pathways using DAVID 2006 (http://david.abcc.ncifcrf.gov). For analysis, we used GenBank accession numbers represented by the corresponding chip oligonucleotides as input.

2.5.3 ChIP-seq *in silico* methods

Mapping of reads

In total, 38512647 Mio Solexa 36mer single-end reads from three lanes of the OCT4 ChIP samples and 25263427 Mio Solexa 36mer single-end reads from two lanes of the Input samples were obtained. 18734743 Mio ChIP reads could be mapped to the human genome (hg18 [141] downloaded from UCSC [142]) using MAQ [143] with default parameter settings. Analogous, 12559090 Mio Input reads were mapped. These reads were further filtered by selecting only those that have a MAQ single-end quality score ≥10 ending up with 11739324 Mio ChIP and 7957923 Mio Input reads. Moreover, all reads that were aligned at exactly the same position and represented such positions with only one read were removed. Finally, 11194815 Mio high-quality reads from the OCT4 ChIP samples and 7435342 high-quality reads from the Input samples were received.

Data quality control

Reads were extended to a length of 500bp (250bp bandwidth). The human genome was divided into windows of 50bp length and for each window the number of overlapping reads was counted. Chromosomes were concatenated and the resulting vectors were compared via scatter plots and Pearson correlation coefficients using the R environment.

Saturation analysis

The total set of high quality mapped ChIP reads was divided into two distinct random sets. From both sets, 100000 reads were randomly selected, extended (250bp bandwidth), distributed over genome wide 50bp windows, chromosomes were concatenated and the resulting vectors were compared via Pearson correlation. This process was iteratively repeated by adding 100000 random selected additional reads from the according distinct sets at each step. Input reads were processed analogously.

The available set of high quality mapped ChIP reads was doubled and afterwards divided into to distinct random sets. Based on these artificially increased set, the saturation analysis was performed as previously described. Because reads are considered twice, a higher correlation of the compared randomly selected subsets is expected. In order to test the scale of this effect, we randomly selected 2.8 Mio unique reads (and 5.6 Mio unique reads, respectively) from the original set of ChIP reads. The selected subset was artificially doubled and the estimated saturation analysis was performed as previously described.

Peak identification

The total set of high-quality mapped ChIP and Input reads were used as input for CisGenome [138], an integrated software system for analyzing ChIP-Seq data. The provided hg18 genome database was selected as reference genome. First, the exploration step was performed with default parameter settings. Peak detection was performed with the *Sequencing Two Sample Analysis* module based on the results of the exploration step. Parameters were set as follows: dP0_hat=0.599875, W=100, M=5, S=100, Max Gap=0, Min Peak Length=0, FDR≤0.9.

Motif mapping

Position-specific count matrices were retrieved from TRANSFAC [144] for the octamer (M00795) and the SOX-OCT joint (M01124 and M01125) motifs. Matrices were transformed into the pseudo-count format required for CisGenome [138]. Motifs were mapped to the peak regions using CisGenome's *Known Motif Mapping Single Matrix->cod* module with default parameter settings.

Conservation analysis

Conservation analysis was performed by the *Get Conservation Summary* module of CisGenome with default parameter settings. The total set of received peaks was analyzed and the mean conservation score for conserved non-repeat positions was used for the following analyses.

Peak Annotation

In order to connect each peak to its closest TSS, the gene file (*refGene.txt*) provided by CisGenome was accessed. For each peak, the distance to the closest (downstream or upstream) TSS and the according gene name were stored.

Enrichment analysis

Enrichment analysis was conducted with the DAVID platform [145]. Official gene symbols were used as input, the *Homo sapiens* species was selected as background and DAVID was executed with default parameter settings.

3 Results

3.1 ChIP-Chip data analysis

The first aim of the study was to establish a ChIP-Chip with NCCIT cells, to discover OCT4 related binding sites. Therefore in a first step anti-OCT4 antibodies had to be tested for efficient chromatin immunoprecipitation.

3.1.1 Specifity of the OCT4 antibody and the impact amplification bias has on site specific enrichment

Antibody specificity is essential for chromatin immunoprecipitation experiments. Although the antibodies used in this study were previously described to be specific in the same application [83] the quality may be lot-dependent. Therefore, Western blot analysis using two polyclonal antibodies against OCT4, one termed N19, the other H-134, was carried out. Figure 3.1 shows that each antibody only gives one band corresponding to the expected size. However antibody N19 shows a better signal and was thus chosen for subsequent immunoprecipitation reactions.

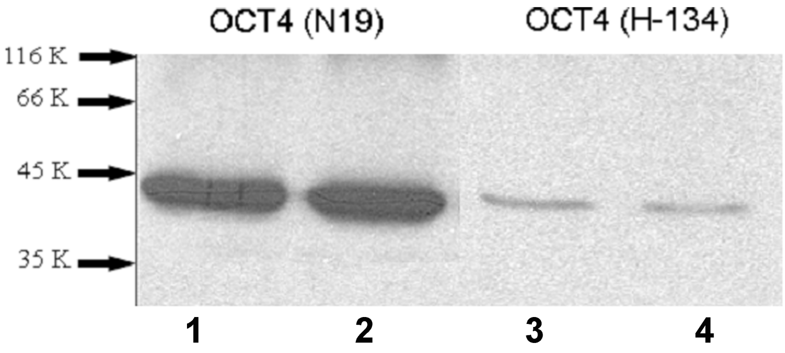


Figure 3.1 Control of antibody specificity by Western blot analysis of OCT4. Lane 1 and 2: polyclonal antibody against OCT4, type N19 (Santa Cruz), lane 3 and 4: polyclonal antibody against OCT4, type H134).

For efficient hybridization onto a promoter array, approximately 5 µg DNA is required. Therefore amplification of the enriched DNA samples is unavoidable. This was accomplished by a random PCR approach. Comparing the distributions of non-amplified with amplified DNA an enrichment of fragments in the size range as seen in Figure 3.2 was noted.

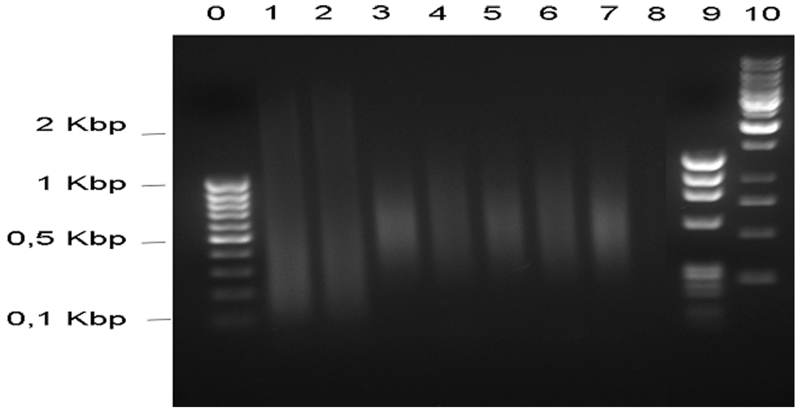


Figure 3.2 DNA-fragment distributions before and after amplification employing Random-PCR.

Loaded with 300 ng per lane.

Lane 1 and 2 are DNA from whole cell extracts (WCE).

Lane 3: INPUT – Antibody control

Lane 4: ENRICHMENT –Antibody control

Lane 5: INPUT Oct4-N19

Lane 6: ENRICHMENT Oct4-N19

Lane 7: INPUT Oct4-H134

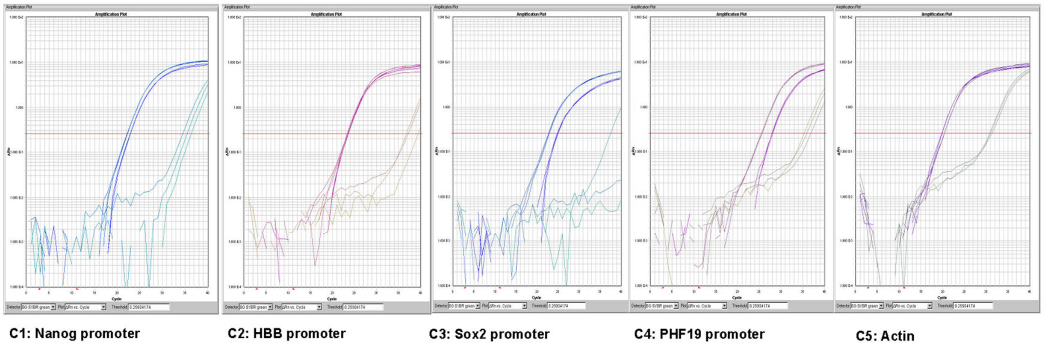
Lane 8: ENRICHMENT Oct4-H134

Lane 9: Fermentas 1 Kbp Marker

Lane 0: Fermentas 100 bp Marker

The Amplification step does not introduce a significant bias

To test for a bias, which might occur after the randomized amplification we compared non-amplified with amplified DNA samples by qPCR for a selection of previously reported OCT4 binding sites and negative control loci. The Ct-difference was in all cases below 2,5 (see Figure 3.3). An enrichment of selected regions could not be found, meaning none of the sequences was preferred during the PCR-amplification. As the random PCR is an exponential based amplification, a delta Ct of 2,5 can still be considered valid for further succeeding applications i.e. hybridization on arrays.



C1: Nanog promoter C2: HBB promoter C3: Sox2 promoter C4: PHF19 promoter C5: Actin

Figure 3.3 Comparison of 5 chosen regions with nonamplified and amplified DNA fragments with the Random-PCR method. Delta Ct was ranging between 0,1 and 2,4. The amplified DNA had in all cases a higher Ct-value *NANOG* and *SOX2* promoters contained a binding motif for OCT4. *PHF* and *HBB* primers were located in the 5´prime promoter region.

3.1.2 Real time validation of known and putative OCT4 targets

To test if previously reported potential OCT4 binding sites in human ES cells could be also verified in NCCIT cells, potential binding sites related to the genes *NANOG, SOX2, LEFTY2, FGF2, HISTH2AM* and *GAP43* were chosen. As a control a number of regions in Exons and proximal promoters not reported to be correlated with OCT4 binding were added for real time validation (see Figure 3.4).

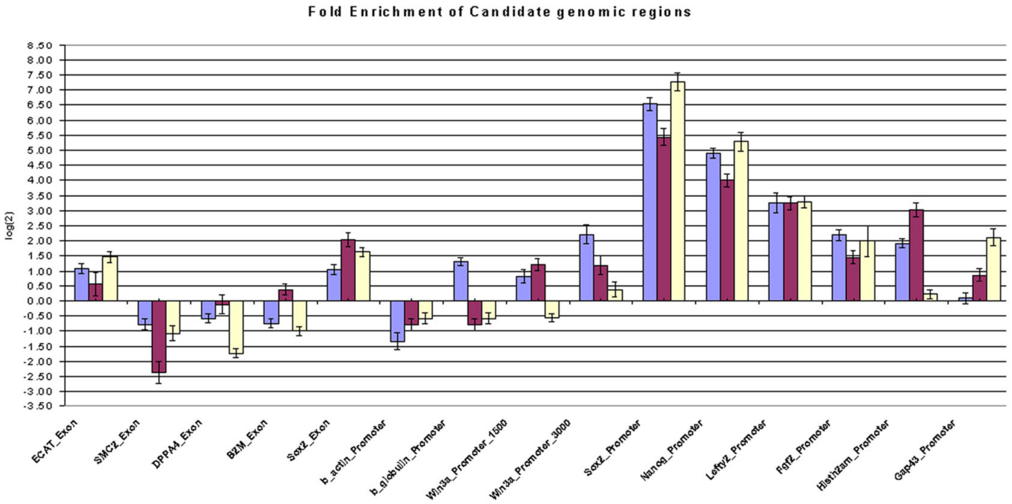


Figure 3.4 qPCR confirmation of known target sites of OCT4 with 3 biological replicates, containing a SOXOCT motif, compared to other genomic regions. Note the strong enrichment for *NANOG*.

In none of the controls a significant enrichment in all three biological replicates could be observed. In comparison, apart from *GAP43*, at least 4 fold enrichment could be detected in at least two biological replicates for 5´proximal promoter regions of the genes *NANOG, SOX2, LEFTY2, FGF2* and *HISTH2AM*, which were confirmed to be OCT4 target genes in human ES cells in a study by Boyer et al. [83]. Thus a global scale analysis of binding sites, using these samples was justified.

3.1.3 ChIP-chip raw data normalization and quality control

For array hybridization, a commercial 2-array set from NimbleGen was chosen. The set consisted of 59357 represented transcripts (using the HG18 version) with an average region size of 50 bp every 100 bp for an average region size of 4700 bp relative to the transcription start site (TSS) and for most sites starting at 4000 bp upstream of the TSS. As three biological replicates were taken, this resulted in a total of 12 arrays for both Cy3 and Cy5 labeled samples (see table 3.1).

CHIP_ID	DYE	DESIGN_NAME	SAMPLE_DESCRIPTION
87866	Cy3	HG18_promoter_1of2	total III
87866	Cy5	HG18_promoter_1of2	experimental III
89715	Cy3	HG18_promoter_2of2	total II
89715	Cy5	HG18_promoter_2of2	experimental II
95313	Cy3	HG18_promoter_1of2	total I
95313	Cy5	HG18_promoter_1of2	experimental I
95758	Cy3	HG18_promoter_2of2	total I
95758	Cy5	HG18_promoter_2of2	experimental I
95760	Cy3	HG18_promoter_1of2	total II
95760	Cy5	HG18_promoter_1of2	experimental II
95935	Cy3	HG18_promoter_2of2	total III
95935	Cy5	HG18_promoter_2of2	experimental III

Table 3.1 Number and name of arrays used for hybridisation. total: randomised DNA, taken before Immunoprecipitation for each replicate (I – III). experimental: enriched DNA fraction after immunoprecipitation for each replicate (I – III).

After hybridization and scanning of the arrays, the first aim was to evaluate the quality of the raw data and chose a suitable normalization method. For this reason Pearson correlations of the scatter plots for the enriched fraction (Cy5 labeled) against the control (Cy3 labeled, see Figure 3.5) as well as for the different biological replicates (see Figure 3.6) were calculated.

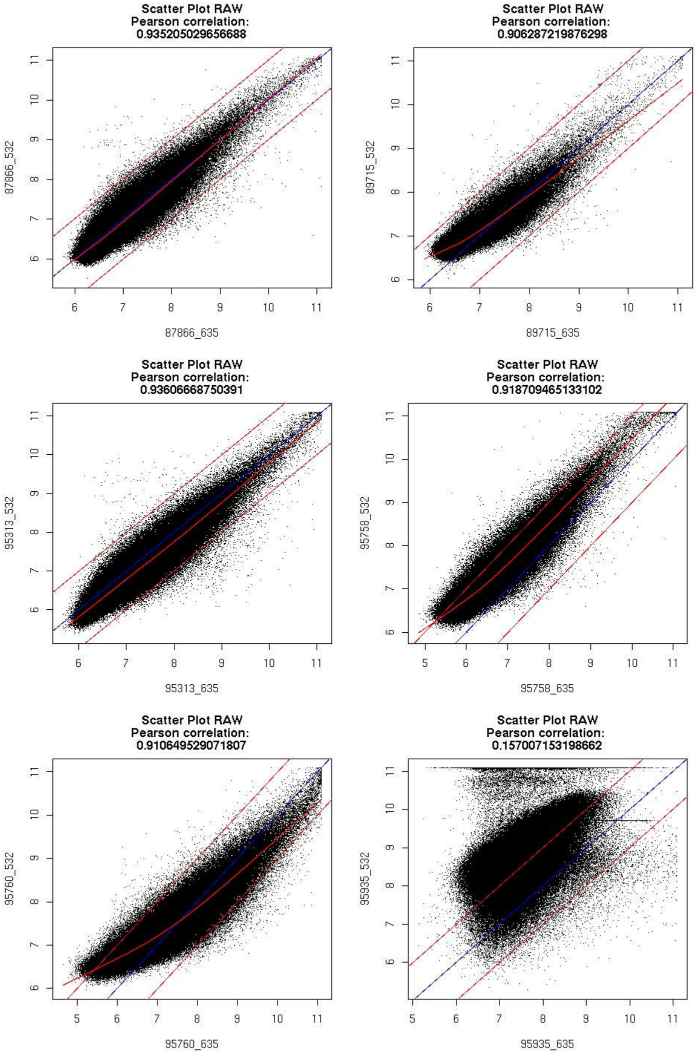


Figure 3.5 Scatter plots of all array experiments, showing the Pearson correlations. Blue line represents theoretical mean, middle red line is measured mean. The flanking red lines represents the 2-fold changes between Cy3 (532) and Cy5 (635), represented for each array (87866 – 95935).

This resulted in scores between 0.90 and 0.94. One array showed a Pearson score of 0.15 and a reconstruction of the TIF image showed a gradient of the intensities from the upper right to the lower left corner, which argued for errors during the hybridization process. Conclusively this array was excluded for further array analysis.

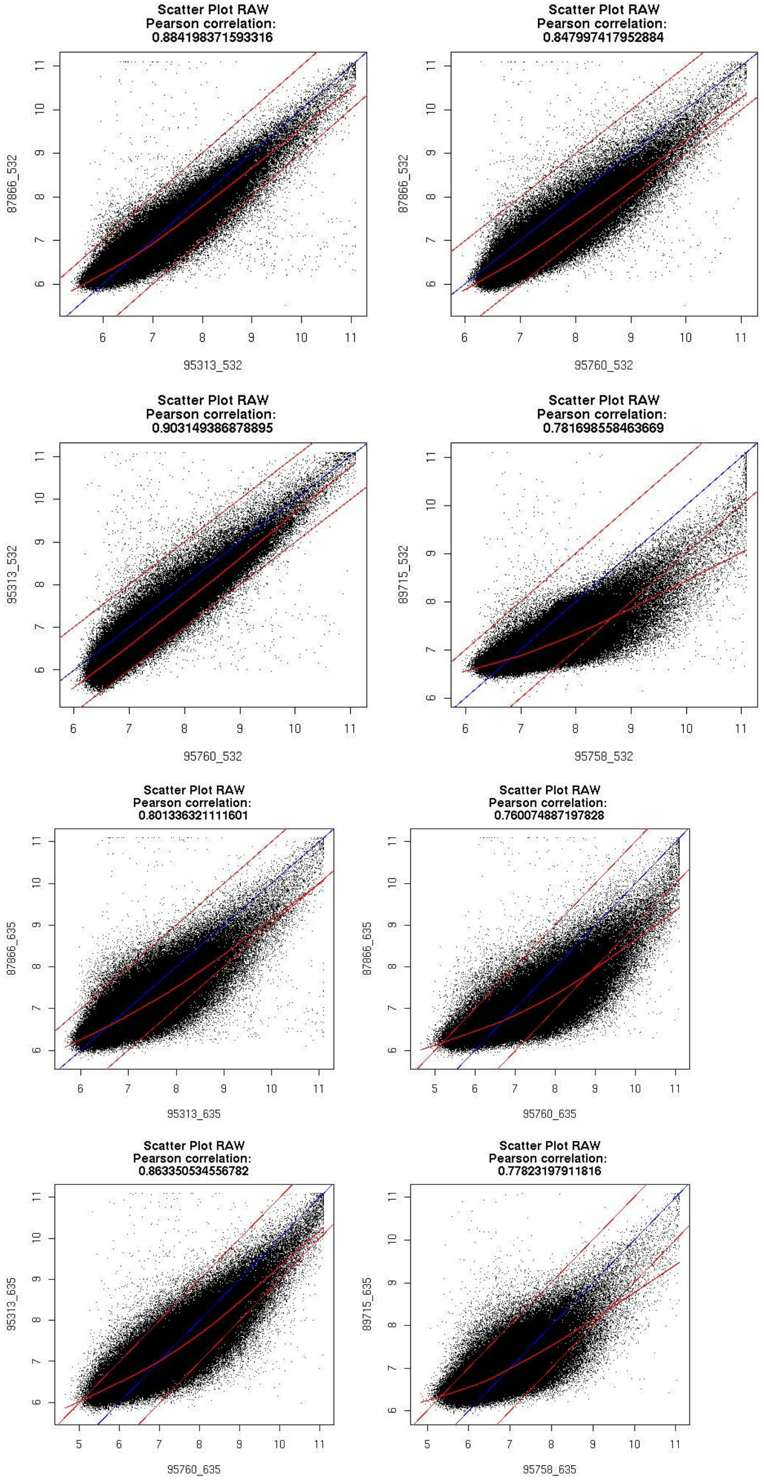


Figure 3.6 Scatter Plots showing the variance between the single biological replicates

Furthermore quantile based normalisation showed the best performance when compared to three other, already established methods, Median, VA and Lowess normalisation, and was conclusively chosen for further peak analysis (see

Figure 3.7). In summary, five of the six arrays used showed raw correlation coefficients (Cy3 vs. Cy5) in the range of 0.91-0.94 with correlation coefficients always slightly higher after applying quantile normalization.

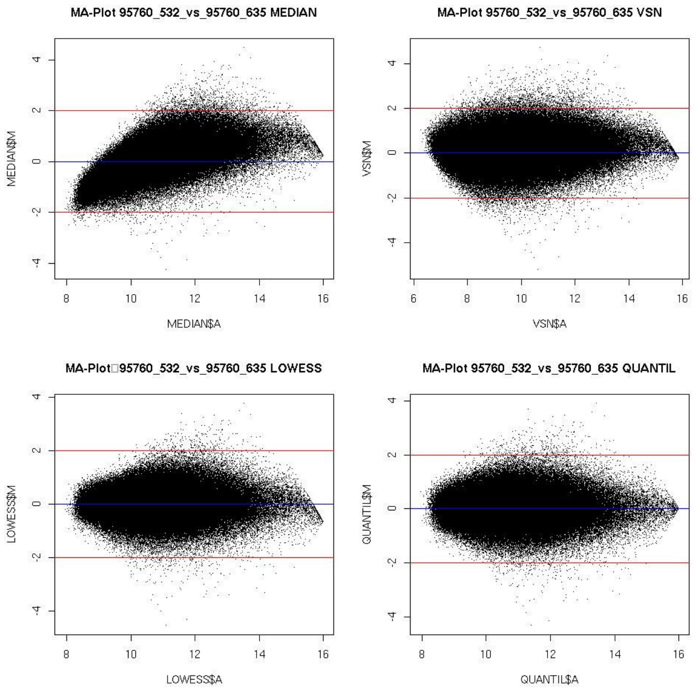


Figure 3.7 Example of MA-Plots of different normalization methods (MEDIAN, VSN, LOWESS and QUANTIL) used.

In order to get an overview of the overall probe intensity score distribution and the extend of background noise effects, random probes were compared with the total probe set in a scatter plot, showing the Cy5 labeled probes on the Y-axis and the Cy3 labeled ones on the X-axis (see Figure 3.8).

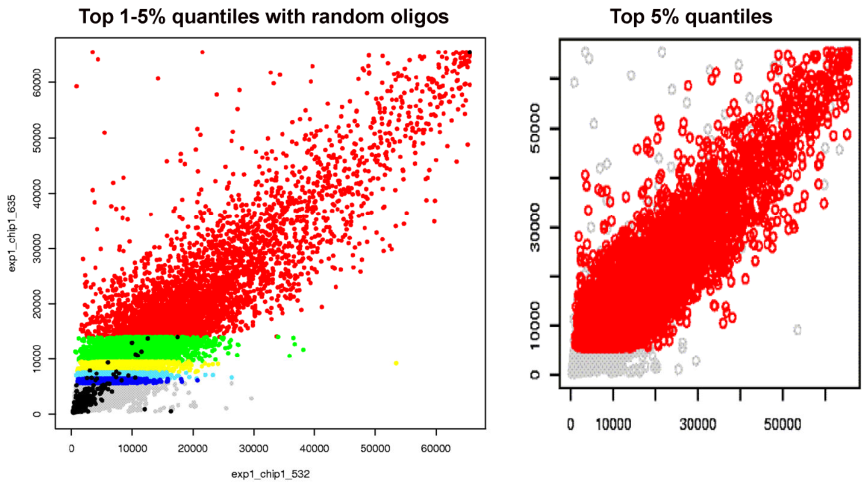


Figure 3.8 The left scatter plot shows the distribution of the enriched probes (Y-axis) against the INPUT probes (X-axis) for Chip ID 95313. The red to dark blue dots represent the top 1% to top 5% quantiles based on the absolute intensities of the array. The black dots represent random probes, used as an internal control of the array. The right scatter plot shows the overlap of the three different biological replicates.

Most of the random oligos were scattered between an intensity score of 0 and 5000. The overlap in the first quantile was high; indicating that enrichment correlated with these probes was consistent.

3.1.4 Impact of different peak detection strategies on peal quality

Pure ratio based algorithms are insufficient for reliable peak prediction

One of the simplest algorithms for detecting peak scores is to compare the ratio of the intensities for each probe for a given sequence length and obtain a true value for peak identification once a number of n probe ratios are above a certain threshold. For the ChIP-chip analysis the first applied algorithm had the following condition for a potential binding event: For a window of 350 bp at least 3 probes had to have a threshold of at least 1.5 for at least 2 biological replicates. Such an approach identified known binding sites such as the 5´proximal promoter region of the *NANOG* gene, which had high intensity scores for the enriched probes. In general, peaks, which contained probes in the top 1% quantile (see Figure 3.8), showed high ratios between the enriched and the control samples (see *NANOG, YAF2* and *PHC1* in Figure 3.9). However this approach had the flaw of peak detections for regions, which could not be distinguished from the background noise, thus resulting in a high number of false positives (see *ESX1, PAK1* and *SLCO1A2* in Figure 3.9). Furthermore for low threshold levels, overlapping peak windows could occur easily, resulting in wrong, artificially long peak length (see *ESX1* in Figure 3.9).

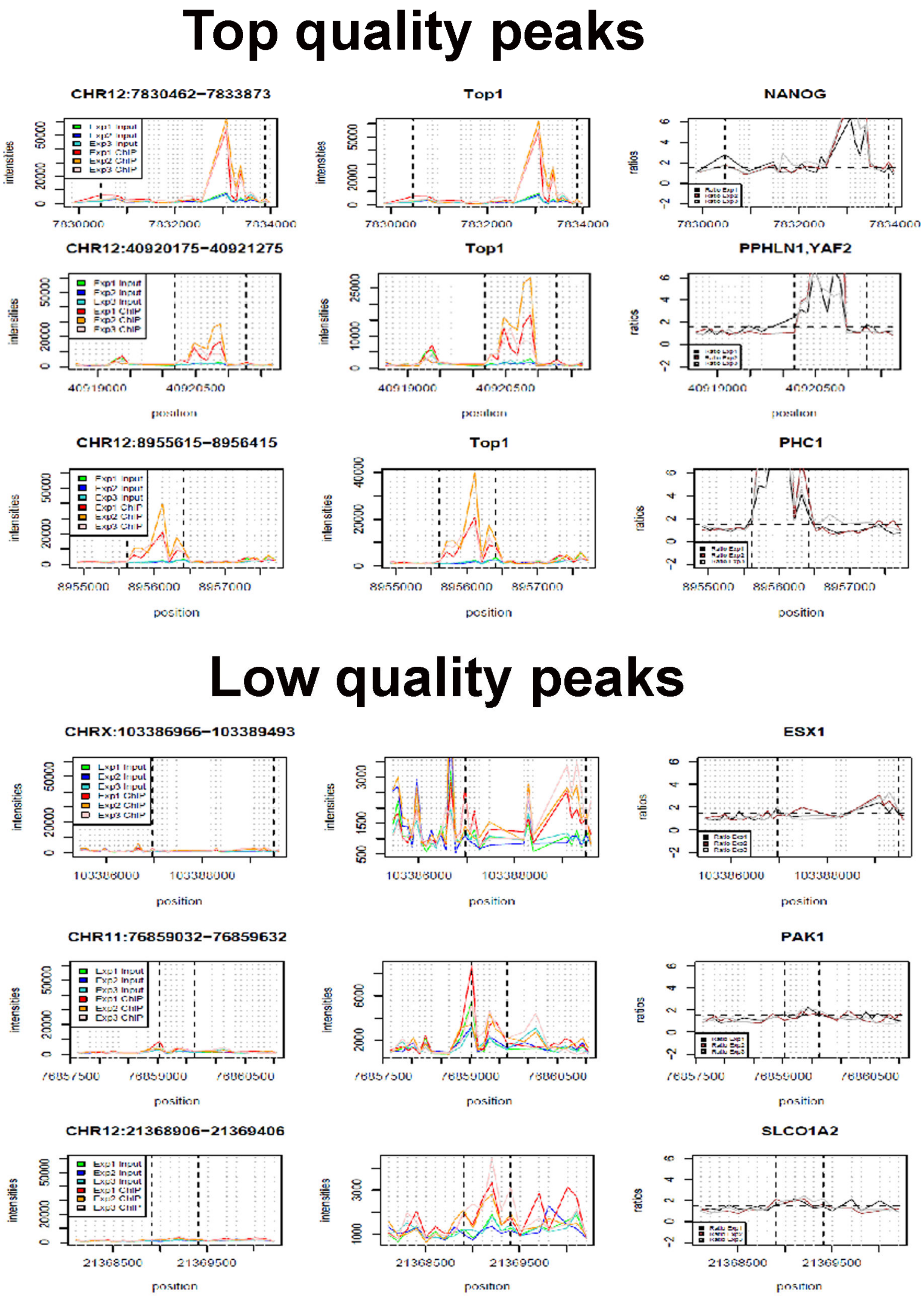


Figure 3.9 Tiling maps of identified peak regions, showing the intensities of the control sample (input) and the enriched DNA fraction (ChIP). The upper 3 examples included probe intensities in the upper 1% quantile. The lower 3 examples illustrate the disadvantage of the algorithm used as no real shoulder-head-shoulder formations could be detected and the intensities of the enriched fraction is not significant higher (for example at least two standard derivations) than the average signal of the random probes as seen in Figure 3.8.

In conclusion, for further downstream analysis, this algorithm was prone to generate false positive detected regions and thus needed to be replaced by more enhanced algorithms.

Algorithms including quantile percentage of the probe scores

As for a more reliable peak detection the upper quantile percentages clearly need to be included, a more advanced algorithm was formulated. For this case a peak is defined as a triple of oligonucleotides, where the ratio of the centre oligonucleotide must be in the upper 0.01 quantile of all ratios and the two flanking oligonucleotides must each have a ratio within the upper 0.05 quantile of all ratios, thus including shoulder-head-shoulder patterns and probe intensities significantly beyond the background. Centre and flanking oligos had to be within a window of 550 bp, reflecting the fragment distribution after the amplification (see Figure 3.2). If a peak was detected within –8000 bp to +2000 bp relating to the transcription start site of an annotated gene, following the NCBI36 build, this gene was associated with the peak. A final target was defined, when the peak was assigned at least to two replicates, respectively one replicate for the promoter array 2, due to the one bad hybridized array. For this approach 473 genes could be detected. Compared to the previously mentioned algorithm (529 targets) the overlap was only 160 genes.

3.1.5 Comparison of different peak algorithms and rank based peak detection

In order to evaluate the performance of the algorithm mentioned above, which was an in-house developed peak analysis tool for ratio distribution dependent interval analysis, referred to as brute-force, it was compared with two independent peak analyses programs, including MA2C [134] and TAMALPAIS [135] which are publicly available. Comparing the three different peak analysis programs, a significant number of targets in NCCIT cells were identified exclusively by one program, for peaks detected in up to 3 biological replicates

(Fig. 3.9 A-C). This was in accordance to a previous study performed by Johnson et al. showing that variation in performance between labs, protocols, and algorithms within the same array platform was greater than the variation in performance between array platforms [146]. In conclusion, in order to evade a potential bias due to specialized peak finding algorithms a peak finding program was devised, which was including the information of each detected peak, whereas a given algorithm was not preferred. Each program was considered equally for the purpose of peak finding and reasoned that a peak identified by three separate programs in each replicate was equivalent to a peak identified by one program in three biological replicates.

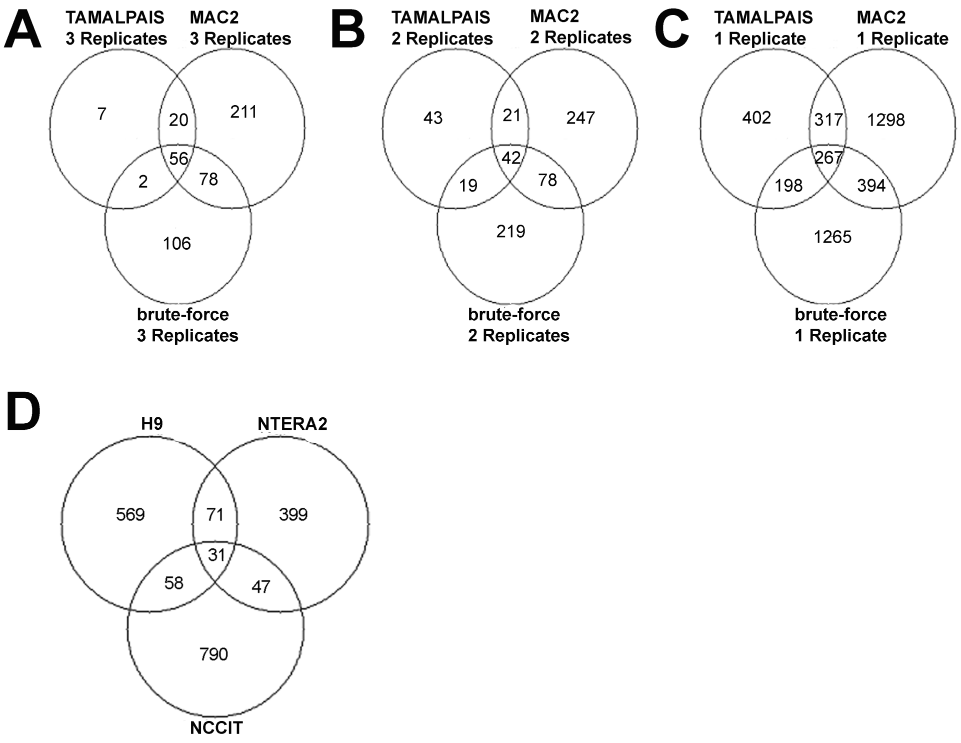


Figure 3.9 Venn diagrams, illustrating the overlaps between different peak analysis programs according to refseq IDs.
A-C: sorted replicate-wise
D: showing the overlap between different cell lines- NCCIT, this study, H9 and NTERA2.

The key pluripotency factors OCT4, NANOG and SOX2 contained an identified binding site with a score higher than 0,5. For a peak score of 0.33 and above, 15 OCT4 target genes of our CHIP-chip with NCCIT cells out of 16 genes, which were reported before to have a critical role downstream of OCT4 were the same, compared with mouse and human ES cells [84].

SYMBOL	87866		95313		95760		89715		95758		Score
POU5F1	3	LMT	3	LMT	3	LMT					1.00
NANOG							3	LMT	3	LMT	1.00
EOMES	3	LMT	3	LMT	3	LMT					1.00
GBX2	3	LMT	3	LMT	3	LMT					1.00
REST	2	MT	2	MT	2	MT					1.00
CDYL	3	LMT	2	LM	3	LMT					0.88
RIF1	2	LMT	3	LMT	2	LMT					0.78
SOX2	3	LMT	1	M	2	LT					0.66
FOXC1			3	LMT	3	LMT					0.66
RFX4							1	M	2	LM	0.50
HOXB1									2	LM	0.33
LHX5									2	LM	0.33
ISL1	3	LMT									0.33
JARID2			2	LM	1	L					0.33
REST	1	L	1	L	1	L					0.33
MYF5					3	LMT			1	L	0.16

Table 3.2 Chosen examples of potential binding sites and the number of detected peaks in biological replicates and peak analysis programs. The 5 numbers of the header represent the corresponding arrays. 1 – 3 indicates how many programs identified a peak. L: brute-force M: MA2C T : TAMALPAIS.

3.1.6 Functional analysis and comparison with literature

Using a threshold of 0.33 resulted in 927 Refseq DNA IDs, which is close to the number of OCT4 targets found in the H9 cell line (729) and almost twice as much as detected with the NTERA2 cell line (548). Comparing the Refseq DNA identifiers from the OCT4 ChIP-on-chip targets with another EC cell line NTERA2 [137] and with a human ES cell line H9 [83] uncovered a set of 31 targets amongst which are both positively regulated (including OCT4, SOX2 and NANOG) and negatively regulated genes (see Figure 3.9 D and Table 3.3). Notably, this list contains a significant enrichment of developmental factors (4,4E-8 for multicellular organism development, using DAVID, see supplemental material).

HGNC symbol	RefSeq DNA ID	Description
GCNT2	NM_001491	N-acetyllactosaminide beta-1,6-N-acetylglucosaminyl-transferase (EC 2.4.1.150) (N-acetylglucosaminyltransferase) (I-branching enzyme) (IGNT).
CTGF	NM_001901	Connective tissue growth factor precursor (Hypertrophic chondrocyte-specific protein 24).
DUSP6	NM_001946	Dual specificity protein phosphatase 6 (EC 3.1.3.48) (EC 3.1.3.16) (Mitogen-activated protein kinase phosphatase 3) (MAP kinase phosphatase 3)
GAP43	NM_002045	Neuromodulin (Axonal membrane protein GAP-43) (Growth-associated protein 43) (PP46) (Neural phosphoprotein B-50).
TNC	NM_002160	Tenascin precursor (TN) (Tenascin-C) (TN-C) (Hexabrachion)
NKX2-2	NM_002509	Homeobox protein Nkx-2.2 (Homeobox protein NK-2 homolog B).
POU5F1	NM_002701	POU domain, class 5, transcription factor 1 (Octamer-binding transcription factor 3) (Oct-3) (Oct-4).
PPP2R3A	NM_002718	Serine/threonine-protein phosphatase 2A regulatory subunit B'' subunit alpha (Serine/threonine-protein phosphatase 2A 72/130 kDa regulatory subunit B)
CCL2	NM_002982	C-C motif chemokine 2 precursor (Small-inducible cytokine A2)
SOX2	NM_003106	Transcription factor SOX-2.
TAL1	NM_003189	T-cell acute lymphocytic leukemia protein 1 (TAL-1) (Stem cell protein) (T-cell leukemia/lymphoma protein 5).
HESX1	NM_003865	Homeobox expressed in ES cells 1 (Homeobox protein ANF) (hAnf).
EPHA1	NM_005232	Ephrin type-A receptor 1 precursor (EC 2.7.10.1) (Tyrosine-protein kinase receptor EPH).
EOMES	NM_005442	Eomesodermin homolog.
SIX1	NM_005982	Homeobox protein SIX1 (Sine oculis homeobox homolog 1).
MAN2C1	NM_006715	Alpha-mannosidase 2C1 (EC 3.2.1.24) (Alpha-D-mannoside mannohydrolase) (Mannosidase alpha class 2C member 1) (Alpha mannosidase 6A8B).
PIP5K1C	NM_012398	Phosphatidylinositol-4-phosphate 5-kinase type-1 gamma (EC 2.7.1.68) (Phosphatidylinositol-4-phosphate 5-kinase type I gamma)
ATAD2	NM_014109	ATPase family AAA domain-containing protein 2.
IL1RAPL1	NM_014271	X-linked interleukin-1 receptor accessory protein-like 1 precursor (IL1RAPL-1) (Oligophrenin-4) (Three immunoglobulin domain-containing IL-1 receptor-related 2)
CXorf26	NM_016500	UPF0368 protein Cxorf26.
SHC3	NM_016848	SHC-transforming protein 3 (SH2 domain protein C3) (Src homology 2 domain-containing-transforming protein C3) (Neuronal Shc) (N-Shc) (Protein Rai).
GALNT8	NM_017417	Probable polypeptide N-acetylgalactosaminyltransferase 8 (EC 2.4.1.41)
LHX5	NM_022363	LIM/homeobox protein Lhx5 (LIM homeobox protein 5).
HOXB4	NM_024015	Homeobox protein Hox-B4 (Hox-2F) (Hox-2.6).
NANOG	NM_024865	Homeobox protein NANOG (Homeobox transcription factor Nanog) (hNanog).
ZIC4	NM_032153	Zinc finger protein ZIC 4 (Zinc finger protein of the cerebellum 4).
SPRED1	NM_152594	Sprouty-related, EVH1 domain-containing protein 1 (Spred-1) (hSpred1).
OLIG3	NM_175747	Oligodendrocyte transcription factor 3 (Oligo3) (Class B basic helix- loop-helix protein 7) (bHLHB7).
LRRTM3	NM_178011	Leucine-rich repeat transmembrane neuronal protein 3 precursor.
CSMD3	NM_198123	CUB and sushi domain-containing protein 3 precursor (CUB and sushi multiple domains protein 3).
SP8	NM_198956	Transcription factor Sp8 (Specificity protein 8).

Table 3.3 list of core genes common between OCT4 ChIP-chip experiments performed in this study with NCCIT cells compared with H9 human ES cells and Ntera2 human EC cells.

A functional annotation of genes for which peaks have been identified only in H9 as well as in NCCIT cells using g:profiler [147], identified genes contributing to neural crest cell development, developmental processes, with an enrichment of genes involved in DNA dependent regulation of transcription. The functional term “DNA dependent regulation of transcription“ was also enriched for OCT4 putative target genes for only H9 hES cells. Neural crest cell development might reflect the propensity of EC cells to differentiate to the neuronal lineage upon stimulation with retinoid acid [148].

Additionally, performing a functional annotation with genes correlating with identified peaks in their promoter region for NCCIT cells, the most stringent annotations (p-value < 0.01) were homeobox, transcriptional repressors and activators, neuronal differentiation and segmentation. For homeobox-containing proteins (see Table 3.4), 17 out of the 31 specific targets identified in NCCIT cells, were detected as OCT4 targets in the human ES cell line- H9 as well.

HGNC symbol	Description	Occupied by OCT4 in H9
TPRX1	Tetra-peptide repeat homeobox protein 1	
HOXB4	Homeobox protein Hox-B4	+
HOXC10	Homeobox protein Hox-C10	
TGIF2LX	Homeobox protein TGIF2LX (TGFB-induced factor 2-like protein, X-linked) (TGF(beta)induced transcription factor 2-like protein) (TGIF-like on the X)	
ADNP	Activity-dependent neuroprotector homeobox protein (Activity-dependent neuroprotective protein)	
SIX1	Homeobox protein SIX1 (Sine oculis homeobox homolog 1)	+
OTX2	orthodenticle homeobox 2	
MEIS2	Homeobox protein Meis2 (Meis1-related protein 1)	
MEIS1	Homeobox protein Meis1	+
ISL1	Insulin gene enhancer protein ISL-1 (Islet-1)	+
LHX5	LIM/homeobox protein Lhx5 (LIM homeobox protein 5)	+
PITX3	Pituitary homeobox 3 (Homeobox protein PITX3)	
HOXB6	Homeobox protein Hox-B6 (Hox-2B) (Hox-2.2) (HU-2)	+
HOXB1	Homeobox protein Hox-B1 (Hox-2I)	+
PHOX2A	Paired mesoderm homeobox protein 2A (Paired-like homeobox 2A) (Aristaless homeobox protein homolog) (ARIX1 homeodomain protein)	
PITX2	Pituitary homeobox 2 (RIEG bicoid-related homeobox transcription factor) (Solurshin) (ALL1-responsive protein ARP1)	
HESX1	Homeobox expressed in ES cells 1 (Homeobox protein ANF) (hAnf)	+
GSC	Homeobox protein goosecoid	+
HOXA3	Homeobox protein Hox-A3 (Hox-1E)	+
POU5F1	POU domain, class 5, transcription factor 1 (Octamer-binding transcription factor 3) (Oct-3) (Oct-4)	+
ZHX3	Zinc fingers and homeoboxes protein 3 (Zinc finger and homeodomain protein 3) (Triple homeobox protein 1)	
MEOX2	Homeobox protein MOX-2 (Mesenchyme homeobox 2) (Growth arrest-specific homeobox)	
TGIF2	Homeobox protein TGIF2 (5'-TG-3'-interacting factor 2) (TGF(beta)-induced transcription factor 2) (TGFB-induced factor 2)	+
NANOG	Homeobox protein NANOG (Homeobox transcription factor Nanog) (hNanog)	+
TSHZ1	Teashirt homolog 1 (Serologically defined colon cancer antigen 33) (Antigen NY-CO-33)	
NKX2-2	Homeobox protein Nkx-2.2 (Homeobox protein NK-2 homolog B)	+
BARX2	Homeobox protein BarH-like 2	
HOXD13	Homeobox protein Hox-D13 (Hox-4I)	
HOXD11	Homeobox protein Hox-D11 (Hox-4F)	+
HOXD8	Homeobox protein Hox-D8 (Hox-4E) (Hox-5.4)	
HIPK1	Homeodomain-interacting protein kinase 1 (EC 2.7.11.1)	
GBX2	Homeobox protein GBX-2 (Gastrulation and brain-specific homeobox protein 2)	+
PROX1	Prospero homeobox protein 1 (Homeobox prospero-like protein PROX1) (PROX-1)	+

Table 3.4: Examples of Homeobox-domain containing genes bound by OCT4 in NCCIT and H9 cells.

To determine if these genes potentially exist as an OCT4-gene regulatory network, this list of genes was submitted to the STRINGS network analysis tool [149]. The resulting network (Fig. 3.10) consisted of a distinct self-renewal cluster composed of NANOG, SOX2, FOXD3, OCT4 (OTF3C) and differentiation-inducing network clusters regulated by transcription factors such as NKX2-2, OLIG3, LHX5, HOXB4 and GATA1, which are themselves negatively regulated by OCT4 [8].

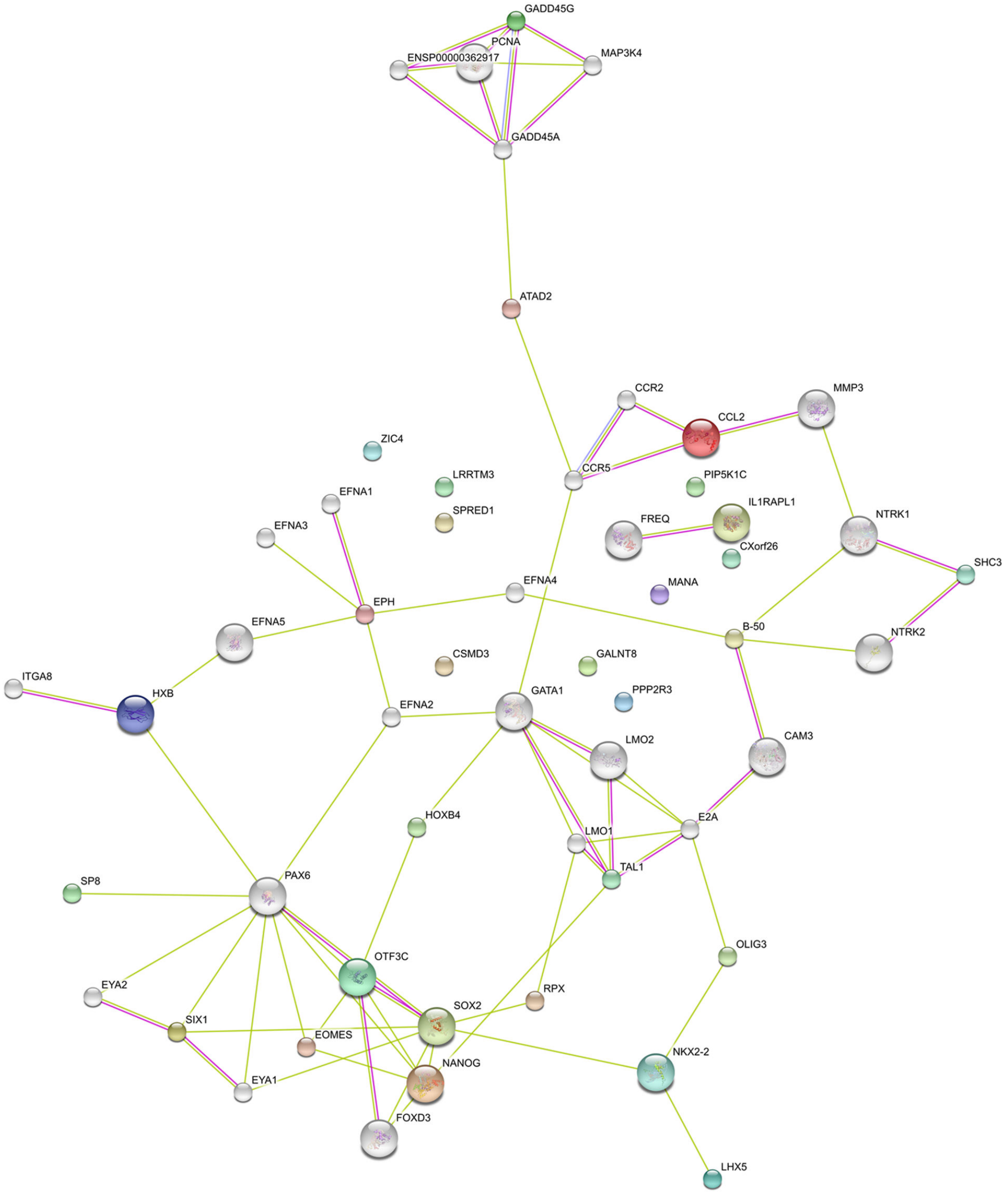


Figure 3.10 A gene regulatory network based on the 31 genes (coloured) common in OCT4 ChIP-Chip targets derived from NCCIT, NTERA2 and H9 cells. GADD45G was also included in this analysis. The network was generated using the web-based program STRINGS.

Pink lines: connectivity based on experimental evidence.

Green lines: connectivity based on text mining.

3.1.7 Six distinct OCT4 binding modules

To investigate if most of the targets obtained in this study contain an octamer motif, all the peak regions of 497 target genes for OCT4 motifs were screened, using a peak-score of at least 0.5 to ensure that peaks were detected at least for two biological replicates and ranked them based on a significance score.

Early studies in mouse showed that a strong enhancer element for OCT4 binding is the octamer motif [9]. Thus, based on the algorithm applied for peak detection (see 3.1.6), the correlation of octamer motifs and the peak score value were investigated. For this reason, indeed, the octamer motif could be reconstructed, using the sequences of the detected potential binding sites for peak-scores of 0,5 and furthermore for peaks falling in the first 5% quantile and using the *de novo* motif discovery program MEME, the OCT4 motif for the OCT4 and SOX2 heterodimer (referred to as oct-sox motif) was detected and showed a good correlation with the published oct-sox motif discovered by a ChIP-PET study, performed in mouse [150] (see Figure 3.11).

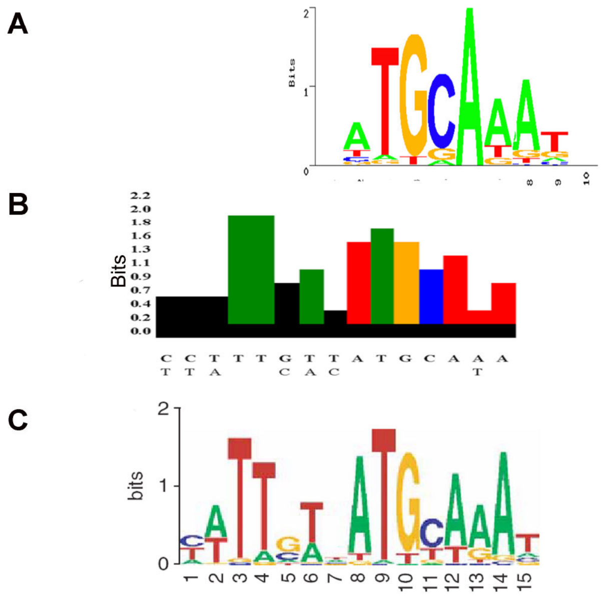


Figure 3.11 Detected OCT4 motifs in NCCIT cells, compared with mouse ES cells

A: reconstructed octamer motif in NCCIT cells

B: detected oct-sox motif in NCCIT cells

C: detected oct-sox motif in mouse ES cells [150]

As seen in Figure 3.12, 50% of all potential octamer motifs fall within peak scores starting at 0.5. The median for the motif scores was 7.3 and was used as a threshold for subsequent motif analysis.

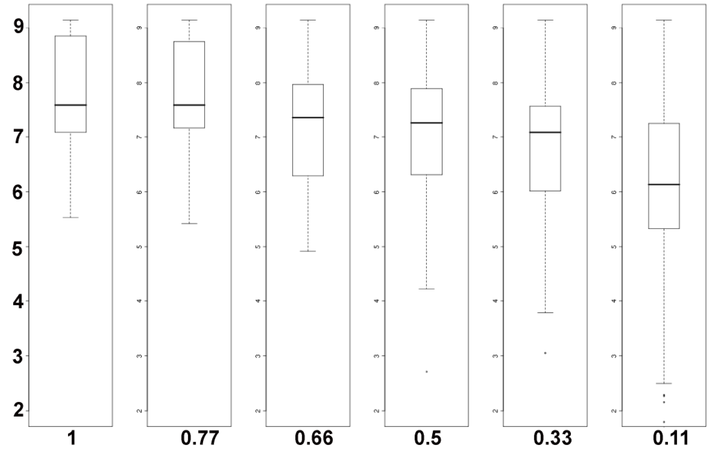


Figure 3.12 Box plot, showing the distribution of the quality of octamer motifs in relation to our defined peak score. For a peak-score of 0.5, half of the motifs will have a motif-

score of 7.3 and above. The average motif score will decrease slightly for a peak score of 0.33 and a significant drop in the motif score can be perceived for a peak score of 0.11.

Thus, genes with motif scores of 7.3 and above were defined as potential direct targets of OCT4. We then sorted all targets with an OCT4 and a SOX2 motif above the threshold level, resulting in a list of 372 genes. The comparison of this list with the target list from Boyer et al [3] that had a SOX2 and an OCT4 peak region (332 targets), resulted in an overlap of 293 targets.

Additionally in order to investigate if those target genes containing a motif score below 7.3 could be regulated by another transcription factor, these regions were scanned with a *de novo* motif discovery program [23-25], which resulted in the discovery of several known transcription factors which could potentially bind to the *de novo* discovered motifs (see below for module 4). In addition to OCT4 motifs, peak regions were screened for the presence of SOX2 motifs, as it is known to form a heterodimer with OCT4. This analysis led to the identification of 6 distinct putative modules of OCT4-binding and transcriptional regulation (Fig. 3.13).

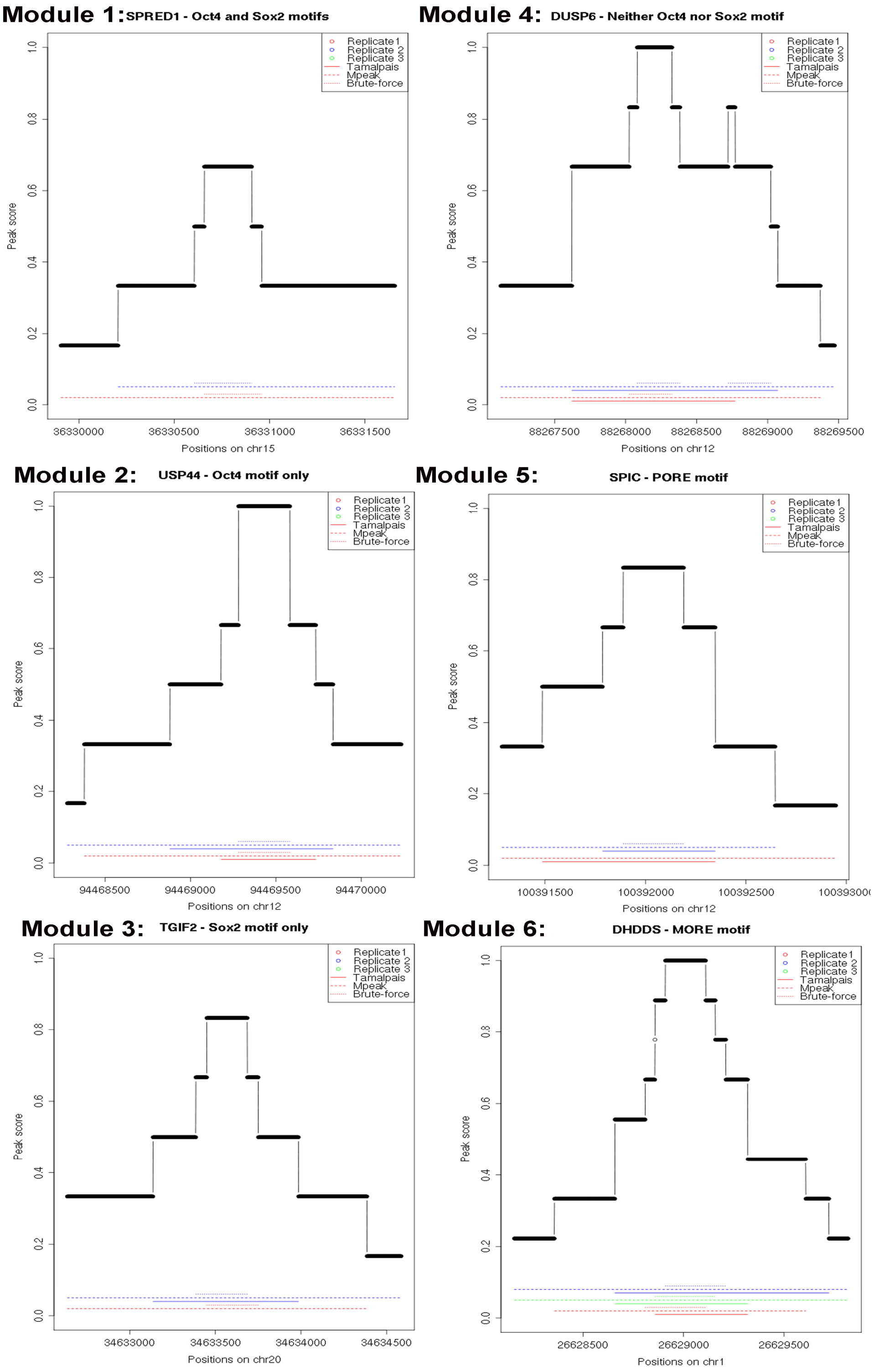


Figure 3.13 Six distinct OCT4 binding modules. Shown are the peak scores, relative to the overlap between MAC2, TAMALPAIS, the in-house developed algorithm - brute-force

and the biological replicates. Peak profiles could be screened for the octamer and SOX2 motifs.

In all of these modules the corresponding genes with their significant 2-fold up- or down-regulated expression upon OCT4 knockdown in NCCIT cells were correlated [5] (supplemental material).

Module 1: OCT4-SOX2 binding motif

This group consisted of 39 genes in total. Within this module, *CTGF* and *TXNRD1* were up-regulated whilst *TPST2, PAK1* and *NANOG* were down-regulated in OCT4 depleted NCCIT cells [5]. The binding of OCT4 to the OCT4-SOX2 motif within the proximal promoter of the *NANOG* gene were confirmed in NCCIT cells using a bandshift assay (see 3.1.10 for details).

Module 2: OCT4 binding motif but lacking a SOX2 binding motif

This module consisted of 122 genes in total, of these *FOXC1, RUNX1, LGALS3, NR2F2, CRABP1, CAMK2D, GFOD1* and *HN1* were up-regulated whilst *GAGE7, GAGE8, ZNF398, USP44* and *DPPA4* were down-regulated in OCT4 depleted NCCIT cells. Another gene harbouring this module in its promoter was *GADD45G* (see 3.1.11 for details).

Module 3: SOX2 binding motif but lacking an OCT4 motif.

This set consisted of 65 genes in total, of these *EMP1, RIN2, TNC, KLHL5, FOXB1, PKD1L2, GPC6 and CBR3* were up-regulated whilst *GSPT2, HESX1, RHCE, RHD, SFRP2* and *GDF3* were down-regulated in OCT4 depleted NCCIT cells.

Module 4: SOX2 and OCT4 binding motif not present

This is a very interesting module suggesting that within 3,5 kb upstream and 750 bp downstream of the TSS of the 271 genes identified, OCT4 might be part of a protein complex with yet unknown transcription factor(s) physically contacting the promoter regions of these target genes. Of these genes, *IL1, COL4A1, PLAU, TPM1, SYTL2, CDC42EP1, KDELR3, KLNK10, H2AFY, SLC7A7, LGI1, BAG3, PACS1, MAP3K8, TOM1L2, LBR, KCTD10, ZFP90,* *EPHB3*, and *WDR1* were up-regulated whilst, *SCGB2A2, GABRA5, FRAT2, RAB25, CSPG5, MAD2L2, SPTBN2, C20orf12, PHC1, MYCN, TUB, GPR3* and *TIMP4* were down-regulated in OCT4 depleted NCCIT cells. For these regulated genes, it was investigated if within the respective promoter regions where OCT4 binding sites could be confirmed, an enrichment of known transcription factor binding sites could also been detected by adopting a de novo motif discovery approach. The hypothesis was that some of these sites might recruit OCT4 into a complex, which is not dependent on direct OCT4-DNA interaction for activating or repressing downstream target genes. Using this strategy, four significant motifs were predicted to be the binding sites for transcription factors such as *REST, TCF3, NR2F1, TP53, NFKB1, LF-A1, RUNX1* and *PAX5* were identified (Fig. 3.14).

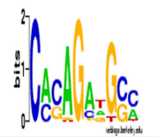
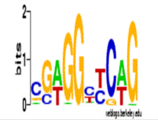
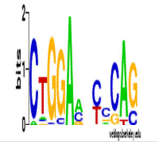
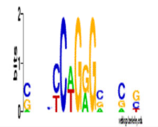
ID	Motif	Complexity	Bits	Similar to known motifs in ***TRANSFAC*** **and Jaspar**	OCT4 regulated genes harbouring these motifs
1		1.8483	11.4	*Sry-beta, Tal-1beta-ITF-2, Sn, NF-muE1, Tal-1beta-E47*, HMG-IY, TAL1-TCF3, Pax5, sna, NHLH1	IL11, KDELR3, IL2RG, TOM1L2, KCTD10 TIMP4, GABRA5
2		1.8483	11.01	*XPF-1, core-binding, LF-A1, TEIL, Dde*, REST, RUNX1, Fos, NR2F1, RXRA-VDR	CDC42EP1, KDELR3, KCTD10
3		1.7977	12.32	*c-Rel, NF-kappaB, STAT5B, Pax*, TP53, Roaz, SRF, REST, Spz1	SLC7A7, MAP3K8, KCTD10 MAD2L2
4		1.1090	10.60	*EBF, p53, Eve, XPF-1, LUN-1*, HMG-IY, DI 2, RELA, NFKB1, NF-kappaB	IL11, CDC42EP1, KDELR3, KCNK10, BAG3,A1BG, PACS1, C20ORF132, LBR, DUSP6, EPHB3 MAD2L2, TUB, GPR3

Figure 3.14 *De novo* motif discovery for genes identified as OCT4 indirect targets and differentially regulated (2-fold and above) in NCCIT cells but lacking the OCT4 and SOX2 motif within the promoter region analysed. The 4 most significant motifs identified and the related potential transcription factor binding sites related to these motifs are displayed. In addition, putative regulated genes harbouring these motifs in their promoter regions shown. Red depicts up-regulated and green down-regulated in response to the ablation of OCT4 activity in ES and EC cells.

Module 5: PORE motif

The PORE sequence (Palindromic Oct factor Recognition Element ATTTGAAATGCAAAT) shown to co-operatively bind two OCT4 molecules was first identified within the first intron of the Osteopontin gene [35]. In this study 4 PORE target genes were identified: *ATXN3, CIR, FLJ16611* and *SPIC*. However, none of these genes were significantly regulated upon knockdown of OCT4 in NCCIT cells.

Module 6: MORE motif

This motif (More PORE- ATGCATATGCAT) was discovered after the PORE sequence was identified. Like the PORE motif, two OCT4 molecules co-operatively bind to the MORE sequence [36]. OCT4 targets bearing this motif include *ATPBD4, C14orf94, CLLU1, DHDDS, SNX20, ORFA17, REM2, SERPINB7, UBE2C* and *GSPT2*. Interestingly *GSPT2*, which encodes a GTP-binding protein that plays an essential role at the G1 to S-phase transition in human cells is also regulated by OCT4 under module 3 (conserved SOX2 binding motif but lacking OCT4). Furthermore, knockdown of OCT4 in NCCIT resulted in a down-regulated expression of GSPT2 and UBE2C.

To further analyse these modules *in silico* the sequences under the respective OCT4 binding peak regions of selected genes within each module were aligned (Fig. 3.15).

Module 1: OCT4 and SOX2 motif present:

```
CTGF      TCGGCTTTAGAAACAATGCATTATAGATTTATCA..[62]…..GGCTTTAGAAACAATGCATTATAGATT
SPRED1    CTTTGTCTTCTAACAATGCATTCGTGACAGCTGT..[190]..TTGTCTTCTAACAATGCATTCGTGACT
```

Module 2: OCT4 motif present, no SOX2 motif present:

```
FOXC1     TTATAAAACACAAATGCAAATATCTTTTCCTGGCG
USP44     GTAGCCTGCTTCTATTTGCATCGTAACCTCTTGGG
```

Module 3: no OCT4 motif present, SOX2 motif present:

```
TGIF2     GAGGAGGGGCTCCAGGTCTTTGTTATGCAGCTCTCAGGTGGAGGAGGAGGCAGCACTCCCCC
HESX1     CAATACTACTTGCTAGGCATTGTTCTAGGTGTGAGGGATACAGTGGTGAAAACAAAGACAATTGTTTA
```

Module 6: PORE motif present:

```
MAN2C1    ACTGACCAAGTTTACAACATTTCAAATGCAGATCACACGGA
```

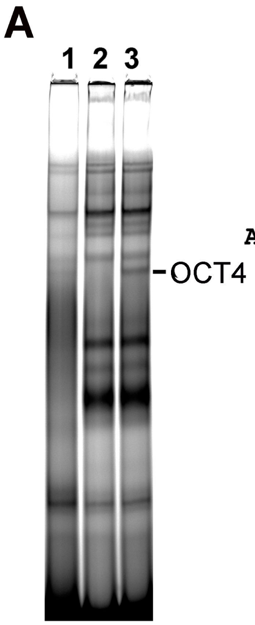
Figure 3.15 Sequence alignments of selected OCT4-regulated genes under the distinct modules. The OCT4 motif is represented in red and the SOX2 motif in green.

A summary of genes contained in the 6 modules and at least 2 fold differentially expressed upon OCT4 knockdown can be found in the supplemental material.

3.1.8 Validation of selected conserved OCT4 binding sites

3.1.8.1 Validation of selected conserved OCT4 binding sites in the NANOG 5 prime proximal promoter

To test if indeed, OCT4 would bind to the oct-sox motif of the *NANOG* promoter in NCCIT cells as has been demonstrated before with mouse [65], a bandshift was performed using the human homologue sequence for the region, spanning the classical HMG-POU consensus sequence (oct-sox motif, see Figure 3.16).



B

OCT4-SOX2

```
    Human  TCTGCAGCTAC--TTTTGCATTACAATGGCCTTGGTGAG
   Rhesus  TCCGCAGCTAC--TTTTGCATTACAATGGCCTTGGTGAG
    Mouse  CTTACAGCTTC--TTTTGCATTACAATGTCCATGGTGGA
      Dog  CCGGCAGCGAC-CTTTTCCATTACAATGGCCTCGGTGAG
    Horse  TCGGCAGCTAC-CTTTTGCATTACAATGGCCTTGGTGAG
Armadillo  TGGGCTGCTCTCCGTTTGCATTACAATGGCCTTGGTGAG
  Opossum  TCG-----TTT--ATTTTCATGATAATGGCCTTGGGGAG
```

Figure 3.16 The *NANOG* promoter harbours an evolutionary conserved binding sites for OCT4 (red) and a SOX2 (bold).

A: Bandshift showing a supershift with OCT4 antibody, using NCCIT-derived nuclear extracts and a Cy5 labelled probe in the 5´region of the *NANOG* promoter bearing the oct-sox motif.

Binding specificity was tested using oligonucleotide competitors.

1) 20-fold excess of unlabelled competitor

2) Supershift with OCT4 (sc9081) antibody

3) Nuclear extract with Cy5-labelled probe

B: Alignment of the OCT4-SOX2 binding sequence in multiple species

OCT4 binding to the *NANOG* promoter was specific in NCCIT cell extracts.

3.1.8.2 in the USP44 5 prime proximal promoter

The binding site for OCT4 and SOX2 of the *NANOG* promoter was conserved. To test if there were octamer motifs with higher conservation scores, peak regions with a peak score of at least 0.5 were screened for their phyloP44wayPrimate C-scores. Using this approach, 43 octamer motifs were detected with higher conservation scores (See Table 3.5).

Symbol	Position of octamer motif	C-score
LRRTM3	chr10:68355702-68355709	0.86498425
RAD54B	chr8:95518105-95518112	0.86250875
SOX6	chr11:16381290-16381297	0.8602815
CIR_HUMAN	chr2:174912215-174912222	0.860212375
CDH10	chr5:24680835-24680842	0.859212625
SLC12A8	chr3:126482634-126482641	0.856962625
SOX6	chr11:16381288-16381295	0.8555375
TRAF3IP2	chr6:112033803-112033810	0.849153375
RASSF8	chr12:26001587-26001594	0.84653525
TXNRD1	chr12:103221204-103221211	0.84404925
SPIC	chr12:100392157-100392164	0.825083875
IL1RAPL1	chrX:28516103-28516110	0.82115575
HSA-LET-7A-2	chr11:121491534-121491541	0.799249125
LMO4	chr1:87568720-87568727	0.798337625
EBF3	chr10:131654436-131654443	0.796267625
DHDDS	chr1:26629016-26629023	0.794768625
BARX2	chr11:128660123-128660130	0.79452175
BARX2	chr11:128660123-128660130	0.79452175
PIPOX	chr17:24394074-24394081	0.7826555
ST8SIA4	chr5:100266434-100266441	0.7811125
AEBP2	chr12:19482320-19482327	0.777708875
C2ORF61	chr2:47259102-47259109	0.7687855
SLC10A2	chr13:102517242-102517249	0.760951
ID3	chr1:23759265-23759272	0.755868
NR2F2	chr15:94669982-94669989	0.753581625
BLMH	chr17:25644530-25644537	0.751254
HN1	chr17:70663100-70663107	0.745100125
Q5T4H8_HUMAN	chr10:21857392-21857399	0.744831625
SFRS18	chr6:99981065-99981072	0.741701625
TXNRD1	chr12:103221206-103221213	0.739501
PAK1	chr11:76859464-76859471	0.7369025
CLEC4A	chr12:8178050-8178057	0.723058875
USP44	chr12:94469489-94469496	0.719515625
ID2	chr2:8738877-8738884	0.701818875
FOXC1	chr6:1552720-1552727	0.697346625
FIGN	chr2:164300383-164300390	0.694212875
C11ORF63	chr11:122255608-122255615	0.69246575
DYM	chr18:45161281-45161288	0.687757125
CCDC102A	chr16:56129004-56129011	0.6869625
RBM32B	chrX:72137577-72137584	0.684763875
OR52E4	chr11:5877570-5877577	0.677344625
PPP2R3A	chr3:137165924-137165931	0.675052
NP_775824.1	chr4:189266437-189266444	0.674706875
NANOG	chr12:7833144-7833151	0.67239675

Table 3.5 OCT4 octamer correlated peak regions and conservation using phyloP44wayPrimate C-scores

The binding of OCT4 to such another evolutionary conserved OCT4 motif was found in the proximal promoter region of the *USP44* gene, an ubiquitin specific protease. Interestingly, the octamer motif was contained here in a 54 bp long conserved contig (See Figure 3.17).

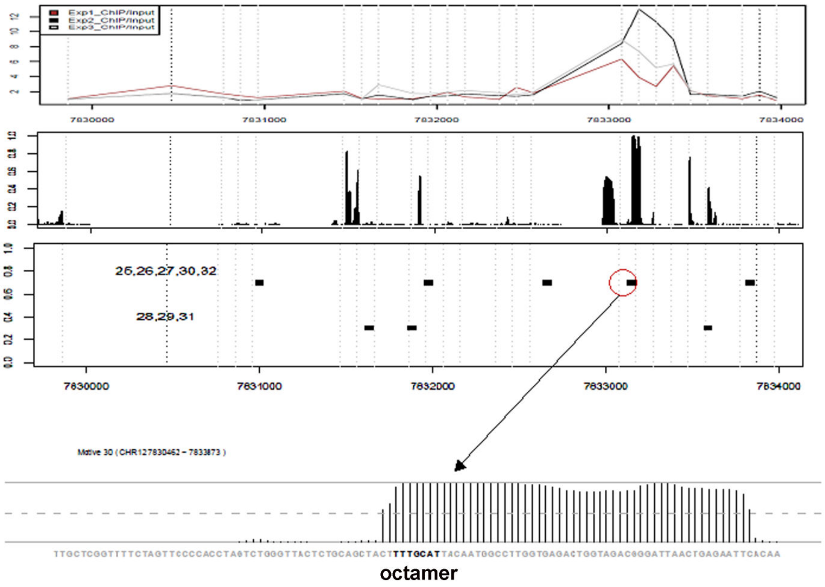


Figure 3.17 *In silico* mapping of the conserved octamer in the 5´promoter region of Usp44. Shown are the ratios of enrichment for all three experiments. Below is a graph, showing the C-scores. The black boxes below indicate detected octamer motifs. The sequence of the conserved contig of the most conserved octamer motif is shown below.

A recent publication by Stegmeier reinforced the role of USP44 as an essential enzyme involved in the control of the anaphase promoting complex [151]. Furthermore, the transcriptional level of USP44 decreases significantly upon OCT4 knockdown in hECs (NCCIT) and hES cells [45]. Additional evidence for a functional role of USP44 in pluripotent cells was revealed by comparing published microarrays, comparing EC cells, undifferentiated human ES cells, embryoid body (EB) formation of these cells and somatic tissues like testis. The gene seems to be specific for several human ES cells like HUES8, hHES-3, H1 and H9. H1 and H9 EB differentiation leads to a significant down regulation of USP44 (Fig. 3.17). Concerning somatic tissues, transcription seems to be limited to oocytes and testis, transcription in other somatic cells could not been detected. In summary there seems to be a correlation between OCT4 expressing cells and the transcriptional level of USP44.

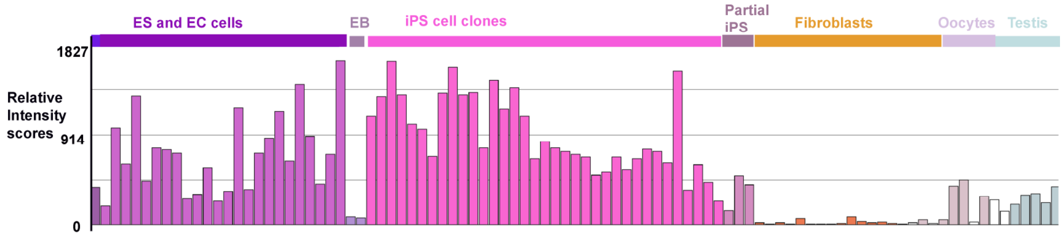


Figure 3.17 Comparison of different microarray experiments, related to the relative transcriptional level of USP44. The gene is transcribed in human ES and EC cells and furthermore in oocytes and testis. After EB differentiation, the transcriptional level goes down significantly (Picture adapted from Amazonia database, http://amazonia.transcriptome.eu/).

Therefore, it was aimed at investigating a possible correlation between OCT4, USP44 and cell cycle control with respect to maintaining self-renewal in these cells. Using the conserved fragment as bait, the enrichment of OCT4 in a pull-down assay was demonstrated (Fig. 3.18 D). Furthermore, the signal obtained by ChIP-real-time-PCR could also been confirmed (Fig. 3.18 B).

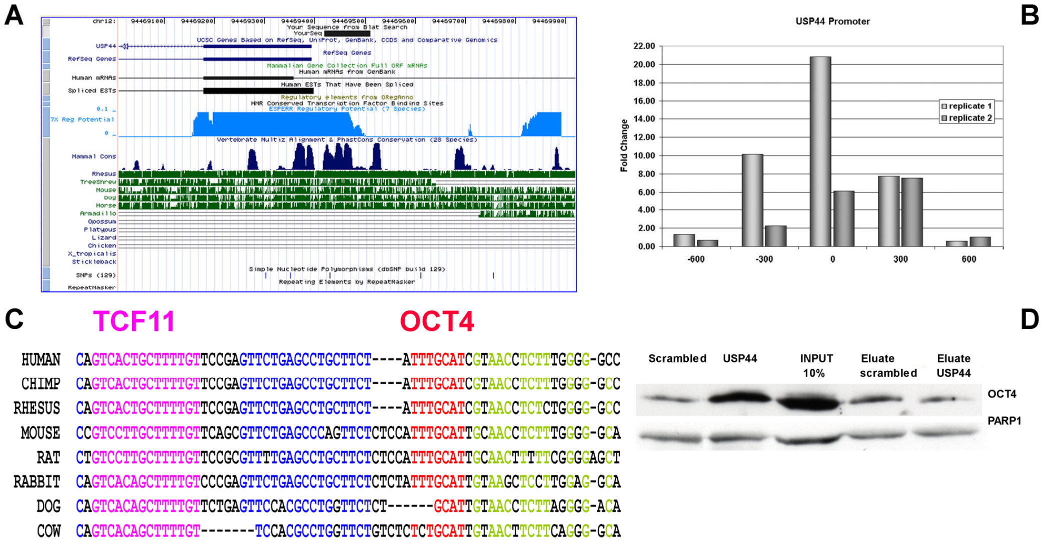

A
B
USP44 Promoter
22.00
20.00
18.00
16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
Fold Change
replicate 1
replicate 2
-600
-300
0
300
600
C
TCF11
OCT4
HUMAN CAGTCACTGCTTTTGTTCCGAGTTCTGAGCCTGCTTCT----ATTTGCATCGTAACCTCTTTGGGG-GCC
CHIMP CAGTCACTGCTTTTGTTCCGAGTTCTGAGCCTGCTTCT----ATTTGCATCGTAACCTCTTTGGGG-GCC
RHESUS CAGTCACTGCTTTTGTTCCGAGTTCTGAGCCTGCTTCT----ATTTGCATCGTAACCTCTCTGGGG-GCC
MOUSE CCGTCCTTGCTTTTGTTCAGCGTTCTGAGCCCAGTTCTCTCCATTTGCATTGCAACCTCTTTGGGG-GCA
RAT CTGTCCTTGCTTTTGTTCCGCGTTTTGAGCCTGCTTCTCTCCATTTGCATTGCAACTTTTTCGGGGAGCT
RABBIT CAGTCACAGCTTTTGTCCCGAGTTCTGAGCCTGCTTCTCTCTATTTGCATTGTAAGCTCCTTGGAG-GCA
DOG CAGTCACAGCTTTTGTTCTGAGTTCCACGCCTGGTTCTCT------GCATTGTAACCTCTTAGGGG-ACA
COW CAGTCACAGCTTTTGT-------TCCACGCCTGGTTCTGTCTCTCTGCATTGTAACTTCTTCAGGG-GCA
D
Scrambled
USP44
INPUT 10%
Eluate scrambled
Eluate USP44
OCT4
PARP1

Figure 3.18 The *USP44* promoter harbours the evolutionary conserved OCT4 binding site but lacks the SOX2 motif

A: Sequence containing the conserved POU site as displayed by the UCSC genome browser

B: Real time PCR confirmation of the presence of the OCT4 binding site.

Position 0 indicates the conserved region seen in (A)

C: Western blot analysis of proteins bound to biotinylated oligos representing the promoter fragment shown in (A). The OCT4 antibody shows higher binding intensity to the USP44- specific probe compared to the corresponding scrambled oligo.

D: Multiple alignments showing evolutionary conservation of the OCT4-bound region

The sequences depicted in blue and green to date uncharacterised with respect to transcription factor recognition and binding.

Notably within the same conserved region a potential binding site for TCF11 could be detected, which has been implicated in the regulation of the antioxidant response [26] and its function is vital during embryonic development [27]. Concerning a potential functional role of USP44 in NCCIT a knockdown, using esiRNA, which reduced the level of USP44 mRNA to 40% did not change the levels of OCT4, NANOG and SOX2 (See Figure 3.19).

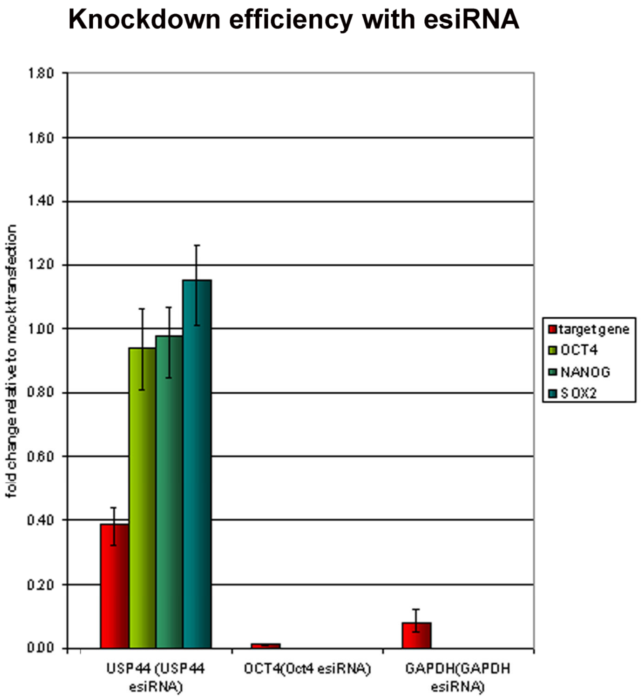


Figure 3.19 Real time analysis after knockdown of USP44. A knockdown of 40% did not show any significant alterations in the expression of OCT4, NANOG and SOX2. As proof of principles knockdown efficiencies are shown for GAPDH and OCT4.

3.1.9 Validation of an octamer site in the 5 prime proximal promoter of GADD45G

Another gene harbouring an octamer motif in its 5 prime proximal promoter is *GADD45G*, a regulator of the cell cycle at the G2/M transition [28] and also recently identified as a putative OCT4/PORE target gene [29]. However this octamer motif was not conserved (Fig. 3.20B). *GADD45G* was not discovered as a potential OCT4 target gene in NCCIT cells but was detected in the Boyer dataset [3]. Furthermore, it has been shown to be one of the earliest OCT4-responsive target genes [30] and was significantly up regulated in OCT4 knockdown experiments [38]. To confirm *GADD45G* as a bona fide direct target of OCT4, a ChIP-real-time-PCR reaction was performed, and the fold enrichment immediately flanking the OCT4 motif compared to neighbouring sites was confirmed. Fold changes of above 2 for 2 replicates with a peak approximately 1 kb upstream of the OCT4 motif (Fig. 3.20) were obtained. For additional confirmation of binding, a bandshift assay using 2 oligos flanking the core OCT4 motif was performed (Fig. 3.20A).

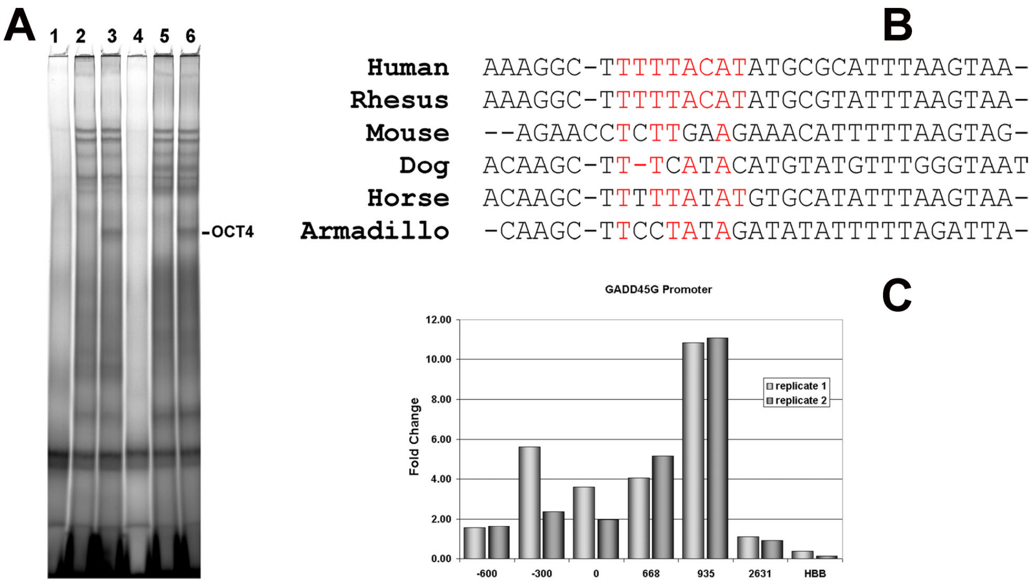


Figure 3.20 The *GADD45G* promoter harbours the evolutionary conserved OCT4 binding sites.

A: Bandshifts showing supershifts with Oct4 antibody using NCCIT cells derived- nuclear extracts using two probes in the 5´region of the *GADD45G* promoter containing an OCT4 motif at positions 9 -15 (lane 1–3) and 17-23 (lane 4-6) of 31 nucleotides.

Lane 3,6: Nuclear extract plus labelled probe

Lane 2,5: same as lanes 3 and 6 but with the addition of OCT4 antibody (sc-9081). Lane 1,4: same as lanes 3 and 6 but with the addition of a 20-fold increase in unlabelled competitor oligo.

B: Multi-species alignment of the selected region chosen for the bandshift assay, the conserved OCT4 binding site is highlighted in red.

C: Real time PCR confirmation of the presence of the OCT4 binding site.

Position 0 indicates the relative position of the octamer motif.

A supershift with OCT4 antibody for both sets of oligos was obtained, thus supporting specific binding of OCT4 to this locus.

3.2 Induction of GADD45G expression in NCCIT cells by Genistein

As shown before, *GADD45G* is a potential target gene of OCT4. To get an idea of the functional implication of GADD45G in NCCIT cells, the ability of Genistein to upregulate *GADD45* gene expression was used in a first approach. Genistein is an isoflavonoid present in soybeans that exhibits anti-carcinogenic properties. GADD45G and GADD45A are regulators of the cell-cycle at the G2/M transition [152] and act as tumour suppressors [153]. The direct effect of Genistein on GADD45A and GADD45G gene expression has been shown before [153]. A direct effect for NCCIT cells should be demonstrated. Thus, human embryonic carcinoma (NCCIT) cells were treated with 50 µM, 100 µM Genistein and DMSO as control. Further growth was carried out for 48 h and as morphological changes were already visible, RNA isolated and the expression of *GADD45A* and *GADD45G* analysed by Real-Time PCR. As shown in Figure 3.22A, Genistein induces transcription of these genes as well as down regulation of *NANOG*. Furthermore a flattened morphological phenotype of the NCCIT cells treated with Genistein could be observed (Fig. 3.21).

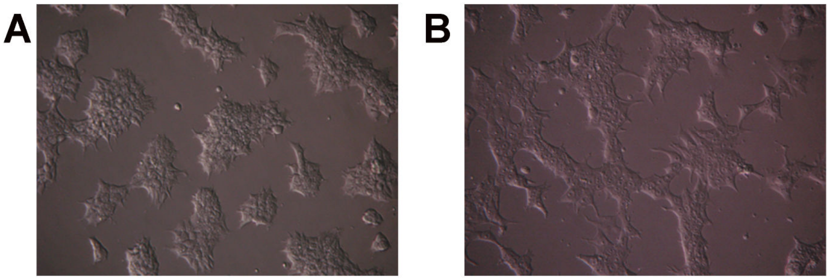


Figure 3.21 NCCIT cells, mock treated 48h with DMSO (A) and with 50µM Genistein

To test, if Genistein treatment also alters the protein-levels of known markers of pluripotency, western-blot analysis of OCT4, SOX2 and NANOG in treated and untreated NCCIT cells were performed. As shown in Figure 3.22B, decreased protein levels of NANOG correlate with the results from RT-PCR analysis, however OCT4 protein level was decreased significantly compared to the mock control in contrast to an only slight decrease in the real time (Figure 3.22A).

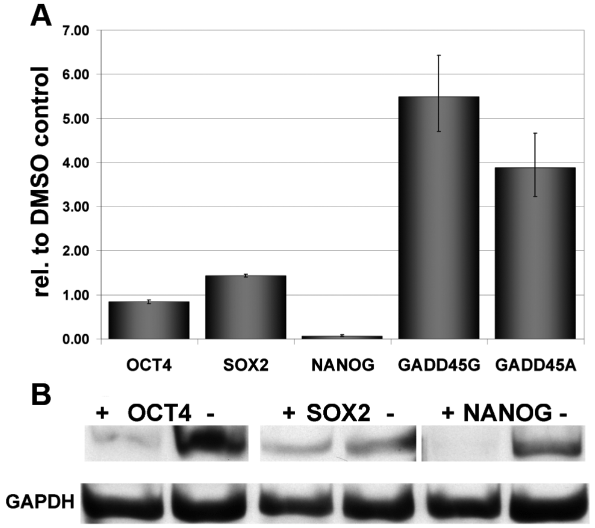


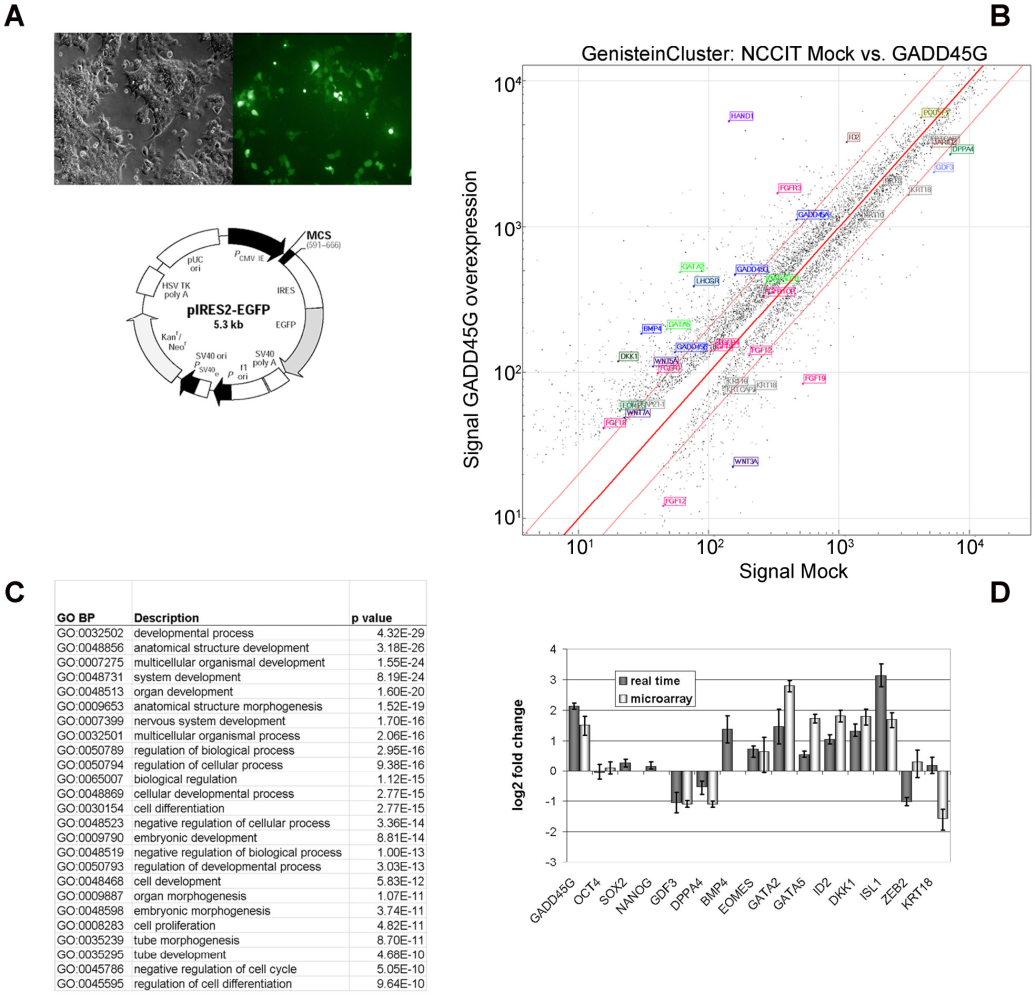
Figure 3.22 Expression of key pluripotency associated genes after induction with Genistein. (A) Real-Time PCR showing upregulated expression of *GADD45A* and *G*, and a drastic down-regulation of *NANOG*. (B) Western-blot showing down regulation of *NANOG*. (-) non-treated DMSO control, (+) Genistein treated NCCIT cells.

3.3 Transient overexpression of GADD45G in NCCIT cells

As the transcriptional level of *GADD45G* increases significantly (more than 2-fold) upon differentiation of ESC and EC cells as a result of ablating OCT4 function [5,8], the hypothesis was that activation of GADD45G activity would induce loss of self-renewal and hence differentiation of the cells with a concomitant decrease in the expression of OCT4. As Genistein induced upregulation of *GADD45G* was not specific enough and to test this hypothesis, the *GADD45G* coding sequence was cloned into the pIRES2-eGFP vector and used for transfection for NCCIT cells. The relative amount of eGFP positive cells was used as a control for transfection efficiency as good quality antibodies are currently unavailable and showed approximately 70% of transfected cells after 48 h (Fig. 3.23A). RNA was isolated two days post-transfection, and microarray based gene expression analysis carried out (Fig. 3.23B).

Though morphological changes could not be observed, transcriptional analysis revealed 531 genes with induced expression of 2-fold and higher. Functional annotation analysis revealed a significant enrichment for genes involved in developmental and differentiation processes (Fig. 3.23B, C, D), especially for

the neuronal lineage. An established differentiation method for EC cells is the addition of retinoic acid to the cells. In this case, the observation that EC cells can be driven to the neuronal lineage was supported by the observed transcriptional changes [31,32]. A selection of genes was chosen for independent confirmation of expression levels using real-time-PCR. An up-regulation of differentiation associated marker genes was noted, *BMP4, HAND1, EOMES, ID2, GATA4, GATA5, ISL1* and *MSX1* (Fig. 3.23D). Interestingly, MSX1 and MSX2 are known BMP4 downstream target genes [33]. Both genes were highly up-regulated upon OCT4 knockdown in ES and EC cells [5,8]. ISL1 is a LIM-homeobox containing gene important for developmental and regulatory function in islet, neural, and cardiac tissue [34].
Although over-expressing GADD45G in NCCIT cells induced up-regulated expression of genes associated with differentiation processes, this was not accompanied by a change in the mRNA or protein levels of OCT4, NANOG and SOX2 at the time point analyzed. However, down-regulation of pluripotency associated genes such as *ZEB2, GDF3* and *DPPA4* was observed (Fig. 3.23D).



GO BP	Description	p value
GO:0032502	developmental process	4.32E-29
GO:0048856	anatomical structure development	3.18E-26
GO:0007275	multicellular organismal development	1.55E-24
GO:0048731	system development	8.19E-24
GO:0048513	organ development	1.60E-20
GO:0009653	anatomical structure morphogenesis	1.52E-19
GO:0007399	nervous system development	1.70E-16
GO:0032501	multicellular organismal process	2.06E-16
GO:0050789	regulation of biological process	2.95E-16
GO:0050794	regulation of cellular process	9.38E-16
GO:0065007	biological regulation	1.12E-15
GO:0048869	cellular developmental process	2.77E-15
GO:0030154	cell differentiation	2.77E-15
GO:0048523	negative regulation of cellular process	3.36E-14
GO:0009790	embryonic development	8.81E-14
GO:0048519	negative regulation of biological process	1.00E-13
GO:0050793	regulation of developmental process	3.03E-13
GO:0048468	cell development	5.83E-12
GO:0009887	organ morphogenesis	1.07E-11
GO:0048598	embryonic morphogenesis	3.74E-11
GO:0008283	cell proliferation	4.82E-11
GO:0035239	tube morphogenesis	8.70E-11
GO:0035295	tube development	4.68E-10
GO:0045786	negative regulation of cell cycle	5.05E-10
GO:0045595	regulation of cell differentiation	9.64E-10

Figure 3.23 Over-expressing GADD45G in NCCIT cells

A: Presence of GFP expression 48h post-transfection (left) compared to the phase-contrast image of the cells. The map of the vector used is presented below.

B: Scatter plot comparing the transcriptomes of GADD45G transfected cells against cells transfected with the wild-type vector. GADD45G-mediated induction of transcription factors such as *HAND1* (purple), *GATA4* (green), and *ID2* (brown) depicted in boxes.

C: Table listing the most significant GO:biological processes related to the up-regulated (>2-fold) genes.

D: Real time PCR validation of a selection target genes (*NANOG, SOX2* and *BMP4* were below detection score 0.01)

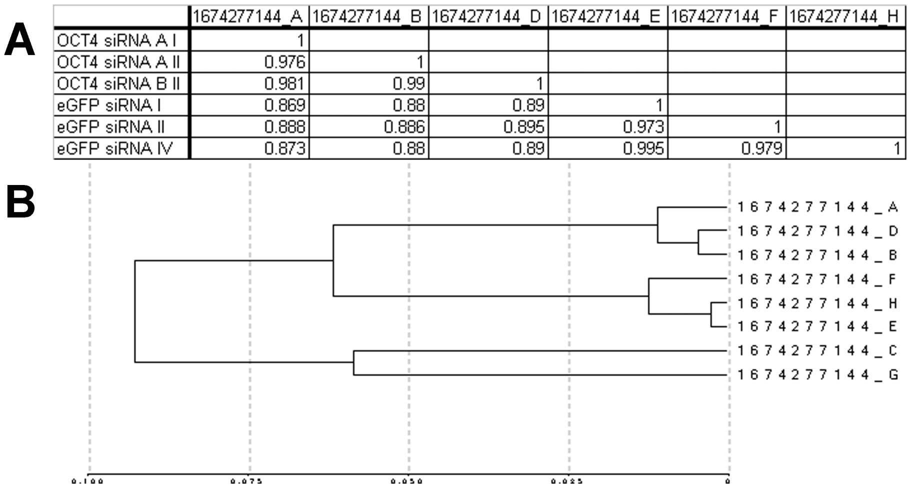
This result raises the possibility that GADD45G activates transcription of differentiation inducing transcription factors independent of the OCT4, SOX2 and NANOG circuitry. Alternatively, it could be that the increased activity of

GADD45G induces rapid suppression of OCT4, SOX2 and NANOG function via disrupting posttranslational modifications or protein-protein interaction required for sustaining the self-renewal circuitry. This action probably takes place long before the reduction of the mRNA and protein levels of OCT4, SOX2 and NANOG at least at the time point analyzed.

In summary, the indirect induction of GADD45G leads to an altered morphological phenotype which is likely to be based on the knockdown of OCT4 and NANOG at the protein level. This doesn't seem to be the direct effect of the GADD45G upregulation as the overexpression of GADD45G could not show these effects. However, analyzing the upregulated genes clearly show a link to differentiation inducing processes, which are independent from OCT4 and not associated with morphological changes up to five days posttransfection.

3.4 Differences in the transcriptional levels between human embryonal stem cells and human embryonal cancer cells

The aim was to get an indication if significant transcriptional differences related to functional annotation classes like Kegg or GO exist between the two cell culture systems. To exclude platform independent differences, this comparison was exclusively based on Illumina arrays. Thus, four RNA samples from undifferentiated H1 cells transfected with mock siRNA against GFP and four RNA samples from H1 cells, 72h after siRNA induced OCT4 knockdown (obtained from Babaie) were hybridized on the Illumina V2 arrays, which cover 22177 genes. Data reproducibility was verified by applying sample correlation analysis and clustering analysis for the different samples. As expected the clustering between the mock controls and the knockdown samples showed a better correlation for the biological replicates (0.98-0.99) than the comparison of the knockdown with the mock treated cells (0.87-0.89). The samples 1674277144_C and G (OCT4 siRNA I and eGFP siRNA III) did not cluster with their corresponding biological replicates and were thus excluded for further analysis (Figure 3.24).



	1674277144_A	1674277144_B	1674277144_D	1674277144_E	1674277144_F	1674277144_H
OCT4 siRNA A I	1					
OCT4 siRNA A II	0.976	1				
OCT4 siRNA B II	0.981	0.99	1			
eGFP siRNA I	0.869	0.88	0.89	1		
eGFP siRNA II	0.888	0.886	0.895	0.973	1	
eGFP siRNA IV	0.873	0.88	0.89	0.995	0.979	1

Figure 3.24 Clustering of samples and correlation factors for mouse and human in vivo experiments. The figure shows the sample clustering and the corresponding correlation coefficients derived from whole genome gene expression analyses for OCT4 knockdowns in human H1 ES cells.

Normalized data were analyzed for significant (Illumina: detection >0.99 for at least one group and p-value <0.05) changes in gene expression. The total number of target genes for all experiments are shown in Table 3.6.

total genes on chip	**22177**
genes with pval < 0.01 and diffscore < 0.05	**4843**
upregulated > 2.0	**1101**
upregulated > 1.5	**2030**
downregulated < 0.66	**1592**
downregulated < 0.5	**549**
not expressed in eGFP control (pval > 0.2)	**126**
not expressed in siRNA OCT4 (pval > 0.2)	**31**

Table 3.6 Summary of differential expressed genes in OCT4 knockdowns compared to siRNA GFP mock controls.

In total 22% of all genes on the Illumina chip were differential expressed based on the specifics of the Illumina software. On average more genes were upregulated. The genes obtained from this analysis were subsequently used as one data set for the construction of an interactive database (discussed in 4.3) and those genes expressed in the eGFP control were used to define the genes expressed in H1 cells.

3.4.1 Comparison with other published datasets

As expression (defined here as the difference in transcriptional levels) comparison of only two cell lines would introduce cell specific artefacts, the Illumina microarray data for the ES cell line H1 (see above) and the EC cell line NCCIT [38] were combined with Illumina microarray data from Josephson and colleagues, who compared the expression profiles between the human EC cell line 2102Ep and the human ES cell line BG03 [154]. The EC cell lines show significant functional differences as the 2102Ep is nullipotent and the NCCIT cells still have the capacity for differentiation into different lineages.

To get an idea if there exists a general difference between the two cell systems, the H1 and the BG03 were grouped and the 2102Ep and Ntera2 were grouped. As the data from Josephson was based on the older Illumina V1 version, instead of comparing the signal intensities, the detection scores were compared, thus including the background effects of the different arrays. In this case differential expression was defined as:

- Detection p-values <0.05 for NCCIT and 2102EP and detection p-values > 0.20 for H1 and BG03 for genes expressed exclusively in EC cells.
- There are no other transcription variants which are expressed in both hES and hEC cells.

This still arbitrary threshold for the p-values enabled the comparison on highly expressed genes and genes only slightly or none expressed (based on previously conducted real time experiments). These conditions led to the discovery of 25 unique expressed genes in EC cells but not expressed in ES cells and 111 unique expressed genes in ES cells but none in EC cells (see supplemental material).

3.4.2 Functional annotation analysis of differential expressed genes

The annotated gene symbols for each set of differentially expressed genes were used for functional annotation analysis, using DAVID and g:Profiler for acquiring REACTOME entries. Table 3.7 lists the most significant hits for genes expressed only in ES cells, based on a p-value of 0.01 and a benjamini p-value of 0.05. No REACTOME hits were detected with this query.

Term	pValue	Benjamini
GO:0032501~multicellular organismal process	5.40E-07	1.14E-05
GO:0048731~system development	4.76E-08	4.02E-05
GO:0048731~system development	1.74E-08	9.13E-05
GO:0032502~developmental process	1.66E-05	1.74E-04
GO:0048856~anatomical structure development	1.46E-06	2.38E-04
GO:0048513~organ development	8.54E-07	3.60E-04
GO:0022610~biological adhesion	7.38E-05	5.17E-04
GO:0007275~multicellular organismal development	7.94E-06	6.47E-04
GO:0032501~multicellular organismal process	5.40E-07	9.46E-04
GO:0048513~organ development	3.91E-07	0.001
GO:0048856~anatomical structure development	1.12E-06	0.001
GO:0048513~organ development	1.45E-06	0.002
signal	7.65E-06	0.004
BP00102:Signal transduction	1.97E-05	0.004
GO:0007155~cell adhesion	8.44E-05	0.004
GO:0007275~multicellular organismal development	6.16E-06	0.006
glycoprotein	6.41E-06	0.006
BP00124:Cell adhesion	6.75E-05	0.007
GO:0032502~developmental process	1.66E-05	0.014
GO:0044421~extracellular region part	5.39E-05	0.023
MF00016:Signaling molecule	1.12E-04	0.026
GO:0005576~extracellular region	3.13E-05	0.026
GO:0043062~extracellular structure organization and biogenesis	1.42E-04	0.039
GO:0009605~response to external stimulus	0.0011	0.044
GO:0022610~biological adhesion	7.38E-05	0.047
GO:0007155~cell adhesion	7.38E-05	0.047
cell adhesion	1.45E-04	0.050

Table 3.7 List of significant SP, PANTHER and GOTERMS for genes expressed in H1 and BG03 and not expressed in NCCIT and 2102Ep based on p-values by Student-T test and Benjamini.

Generally, genes involved in developmental processes were significantly expressed in ES cells, compared to EC cells.

Interestingly, PANTHER, as well as GO and SP detected an enrichment of cell adhesion associated genes, which are shown in Table 3.8.

SYMBOL	Gene Name
NRP2	NEUROPILIN 2
PCDHB13	PROTOCADHERIN BETA 13
SNAI1	SNAIL HOMOLOG 1 (DROSOPHILA)
COL8A2	COLLAGEN, TYPE VIII, ALPHA 2
JAM2	JUNCTIONAL ADHESION MOLECULE 2
TNFAIP6	TUMOR NECROSIS FACTOR, ALPHA-INDUCED PROTEIN 6
NRXN1	NEUREXIN 1
CNTN1	CONTACTIN 1
ALCAM	ACTIVATED LEUKOCYTE CELL ADHESION MOLECULE
NCAM1	NEURAL CELL ADHESION MOLECULE 1
COL6A3	COLLAGEN, TYPE VI, ALPHA 3
CYR61	CYSTEINE-RICH, ANGIOGENIC INDUCER, 61
COL12A1	COLLAGEN, TYPE XII, ALPHA 1
EGFR	EPIDERMAL GROWTH FACTOR RECEPTOR (ERYTHROBLASTIC LEUKEMIA VIRAL (V-ERB-B) ONCOGENE HOMOLOG, AVIAN)

Table 3.8 Genes involved in cell adhesion processes according to GOTERM BP GO:0007155

These genes might contribute in explaining the morphological differences between the colony forming ES cells, who need to be seeded on mouse embryonic fibroblast feeder layers or matrigel compared to human EC cells, who will remain in an undifferentiated state without feeder layers or matrigel and will not grow in a colony like fashion.

For the 25 unique expressed genes in EC cells, no specific enrichments could be detected, by GO, SP, PANTHER, Kegg and REACTOME annotations, negating a relationship between specifically expressed genes and known functional categories.

3.5 ChIP-seq data analysis of OCT4 targets in NCCIT cells

ChIP-chip studies are limited in their design as they can only detect enrichments depending on the tiling array. To detect OCT4 binding sites in a more unbiased way and to get the first indications of OCT4 binding sites in non five prime proximal regions, a ChIP-seq was performed with three biological replicates of NCCIT cells, which were merged for each the control DNA (Input) and the OCT4-enriched DNA (ChIP). Before sequencing, enrichment was quality tested for the Nanog and Sox2 promoters.

3.5.1 Data quality control

As a quality control for data consistency, Pearson correlation coefficients comparing individual ChIP lanes were calculated (0.61-0.75), the two input lanes (0.60), and the total ChIP reads against the total Input reads (0.76). Figure 3.25 A shows a scatter plot comparing genome-wide sequence coverage of two individual ChIP lanes (lane 081212_s2 with 4170527 Mio, and lane 081212_s1 with 4179779 Mio reads). The corresponding Pearson correlation coefficient is 0.75. Figure 3.25 B shows the same data in log2 scale. Figure 3.25 C shows an analogous scatter plot comparing the total sets of ChIP and Input reads (Pearson correlation coefficient 0.76). An enrichment of ChIP reads (x-axis) can be observed by visual inspection of the scatter plot in log2 scale (Figure 3.25 D).

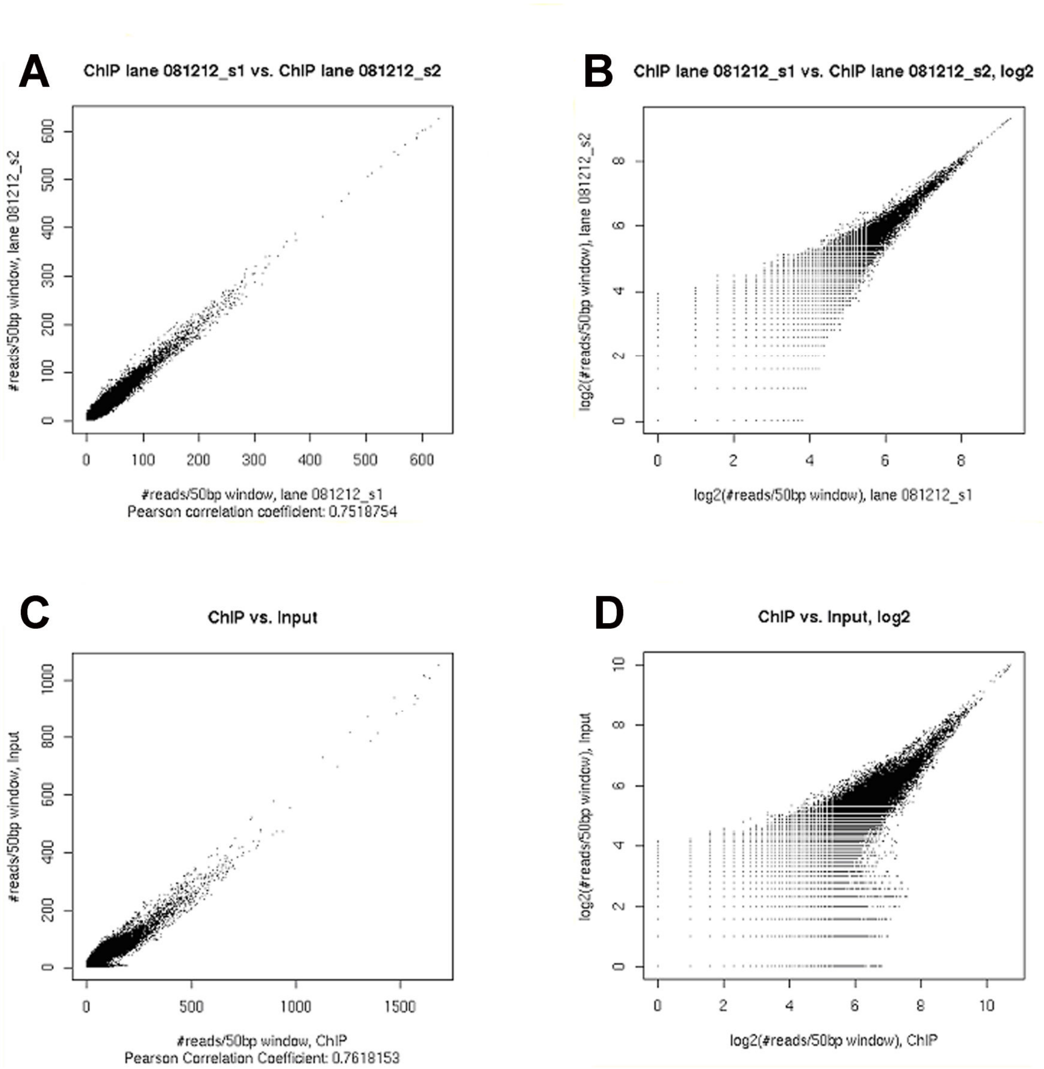


Figure 3.25 Quality control

A Comparison of genome-wide sequence coverage of two individual ChIP lanes (081212_s2 with 4170527 Mio, and 081212_s1 with 4,179,779 Mio distinct high-quality reads). The reads were extended to a length of 500bp (250bp bandwidth) and the genome was divided into windows of size 50bp. Each data point belongs to a window and the x axis show the number of reads that overlap the window. B log2 representation of A. C Comparison of genome-wide sequence coverage of the total ChIP reads against the total Input reads. D log2 representation of C.

3.5.2 Saturation analysis

Figure 3.26 shows the results of a saturation analysis performed to estimate the required number of ChIP reads necessary for obtaining an adequate and reproducible coverage of immunoprecipitated DNA fragments within the ChIP samples. The red line in Figure 3.26 A shows a successively increase of data accordance with increasing number of added ChIP reads. Comparing two distinct random sets of 5.6 Mio extended reads assembled from the total available set of ChIP reads a Pearson correlation coefficient of 0.77 is reached. In comparison to the saturation of ChIP reads, the blue line in Figure 3.26 A shows the results of an analogous saturation analysis for Input reads. With two distinct sets of 3.7 Mio Input reads, relatively lower data accordance is achieved (Pearson correlation coefficient of 0.62) compared to the same number of ChIP reads (0.7). One explanation that the comparable higher number of Input reads required for obtaining a similar reproducible coverage of DNA fragments is due to the enrichment of OCT4 binding site containing DNA fragments within the ChIP sample.

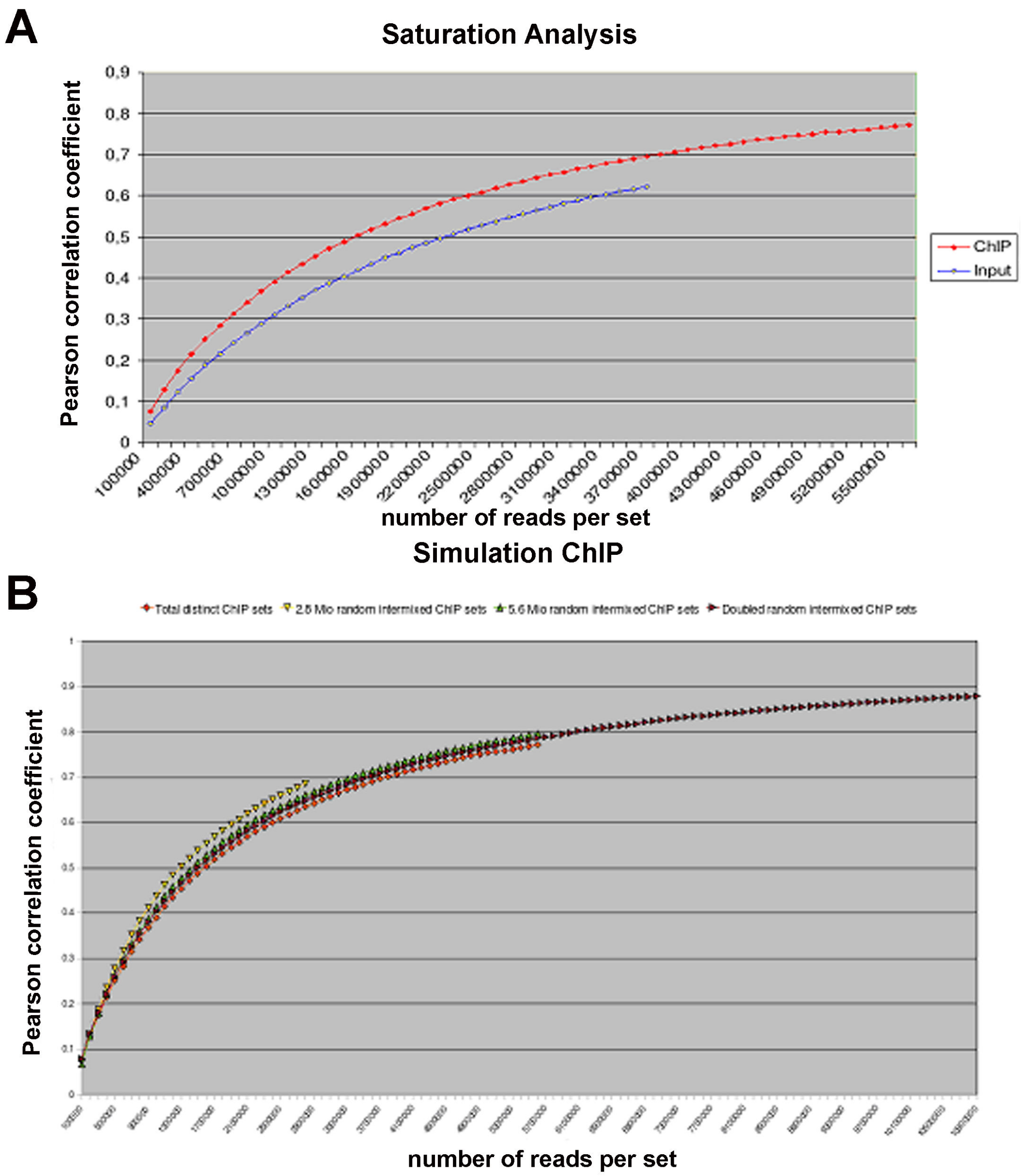


Figure 3.26 – Saturation analysis

A Saturation analysis of genome wide ChIP (red line) and Input (blue line) reads. Starting with two distinct sets of 100000 random selected and extended ChIP reads (250bp bandwidth) distributed over genome wide 50bp windows, a successively increase of data accordance can be observed in both cases. Comparing two distinct random sets of 5.6 Mio extended ChIP reads, a Pearson correlation coefficient of 0.77 is reached. B Estimated saturation for the total set of ChIP reads. The light red line represents the original saturation as given in a. The dark red line represents the estimated curve as deduced by the doubled random intermixed ChIP set. The yellow and the green lines represent simulations based on only subsets of the original set of ChIP reads.

Because the saturation analysis can be performed in an unbiased way on independent sets only, the total number of reads that can be utilized for this analysis is half of the original set. In order to estimate the correlation coefficient for our total ChIP data, the saturation analysis on an artificially duplicated set of the available total ChIP reads were performed. Figure 3.26 B shows the results of the original simulation study ("Total distinct ChIP sets") and of the simulation analysis based on the artificially doubled set of ChIP reads ("Doubled random intermixed ChIP reads"). Additionally, the saturation behaviour of artificially doubled read sets was tested by doubling only subsets of the available ChIP reads. Figure 3.26 B shows that the resulting correlations are overestimated for artificially doubled sub sets (see "2.8 Mio random intermixed ChIP sets" and "5.6 Mio random intermixed ChIP sets") compared to the original study. Nevertheless, a similar behaviour of the curve shapes can be observed and therefore, a correlation of 0.88 is estimated as an upper border for the total set of ChIP reads. In conclusion, the total set of available ChIP reads (~11.20 Mio) covers the existing immunoprecipitated DNA fragments in a reliable depth so that an experimental repetition with the same number of reads will give adequate similar results.

3.5.3 Genome wide distribution of sequencing reads

Table 3.9 summarizes the distribution of ChIP and Input reads in percentages over several gene associated regions as reported by CisGenome. In both cases, ~62% of all reads fall into intergenic and ~37% of all reads fall into intragenic regions.

Region	ChIP reads (%)	Input reads (%)
intergenic	62,87	62,41
intragenic	37,12	37,58
exon	1,77	1,89
intron	35,43	35,79
CDS	1,08	1,16
UTR	0,70	0,74
5'UTR	0,08	0,07
3'UTR	0,62	0,66
intergenic(<=1kb-from-gene)	1,07	1,08
TSSup1k	0,49	0,46
TESdown1k	0,58	0,62
intergenic(<=10kb-from-gene)	11,46	11,73
TSSup10k	6,38	6,47
TESdown10k	6,14	6,37
intergenic(<=100kb-from-gene)	54,94	55,62
TSSup100k	39,34	39,93
TESdown100k	38,95	39,57

Table 3.9 Location summary for ChIP and Input reads

In general, no obvious differences in occupancy of the examined regions like exons (1,77% ChIP and 1,89% Input reads), introns (35,43% ChIP and 35,79% Input reads), and -10kb promoters (6,38% ChIP and 6,47% Input reads) can be observed when considering the total distributions of ChIP and Input reads.

3.5.4 Genome wide distribution of binding regions

Genome wide enrichment analysis was performed using the two sample analysis module of CisGenome with varying parameters for the false discovery rate. Table 3.9 summarizes the location distribution of peaks identified by a varying FDR (see also Figure 3.27 B), arguing for a more scattered distribution of OCT4 binding sites rather than an accumulation in the five prime proximal promoter region.

Region	FDR 0.1 (%)	FDR 0.5 (%)	FDR 0.9 (%)
intergenic	69.13	64.00	61.57
intragenic	30.87	35.99	38.42
Exon	0.83	1.21	1.60
Intron	30.13	34.86	36.91
CDS	0.41	0.74	1.06
UTR	0.41	0.48	0.55
5'UTR	0.15	0.13	0.09
3'UTR	0.26	0.34	0.45
intergenic(<=1kb-from-gene)	1.10	1.00	0.97
TSSup1k	0.83	0.63	0.46
TESdown1k	0.26	0.37	0.52
intergenic(<=10kb-from-gene)	12.31	12.23	12.45
TSSup10k	6.86	7.22	7.07
TESdown10k	6.49	6.20	6.56
intergenic(<=100kb-from-gene)	58.17	60.12	59.28
TSSup100k	39.41	42.60	42.52
TESdown100k	45.44	44.35	42.88

Table 3.9 Location summary for peak regions identified by a varying FDR

Although the majority of peaks are located within genome wide intergenic regions in all three tested cases, its relative proportion decreases (69.13%, 64.00%, and 61.57%) with an increasing FDR (0.1, 0.5, and 0.9). Accordingly, the relative proportion of peaks located within intragenic regions increases (30.87%, 35.99%, and 38.42%). The largest fractions of intragenic peaks are located in introns (30.13%, 34.86%, and 36.91%) and the fraction of peaks located in conventionally analysed promoter regions (-10kb from the TSS) stays stable with varying FDR (6.86%, 7.22%, and 7.07%). Thus there seems to be a correlation between the decrease of false positives and an accumulation in intergenic regions.

3.5.5 Comparative analysis of gene associated binding regions

In order to compare the presented OCT4 ChIP-Seq data to previously published work, here a gene was considered as a putative direct target, if a peak exists within the -10kb to +2kb region around its TSS. This definition corresponds to promoter regions covered in a related study that analysed OCT4 binding sites in hESCs using tilling arrays [83]. As a high FDR means a high number of false positive detected peaks, for an FDR of 0.1, the closest genes correlating to the peaks were compared to the OCT4 targets, obtained by the ChIP-chip analysis with OCT4 explained above with a peak score of 0.5. Thus, comparing 1908 peaks with 497 peaks derived from the ChIP-chip, only 60 common genes could be identified. To evaluate the effect of the presence of octamer motifs within the peak sequences, only those peaks containing an octamer motif were compared. Comparing 161 ChIP-chip peaks, containing an octamer sequence and 228 ChIP-seq peaks containing an octamer site, revealed common peaks referring to 9 common genes, thus not increasing the percentage of overlapping genes. When comparing only those ChIP-seq peaks which were in the same region as used for the NimbleGen tiling array (96 genes) and comparing them to the 497 ChIP-chip targets, common peaks referred to 18 genes including *DPPA4, GDF3* and *PHC1* (see supplemental material).

Comparison with other published datasets

Table 3.10 shows the total number of identified peaks with respect to a varying false discovery rate (0.1, 0.5 and 0.9), putative OCT4 target genes, as well as overlaps to previously published OCT4 target genes [83] [137].

					Integration of motif mapping			
FDR	**Peaks**	**Annotated Genes**	**Overlap Boyer et al.**	**Overlap Jin et al.**	**Peaks with motif**	**Annotated Genes**	**Overlap Boyer et al.**	**Overlap Jin et al.**
FDR 0.1	1908	148	12	16	366	30	10	8
FDR 0.5	16591	1152	67	85	1881	139	29	21
FDR 0.9	93965	4987	179	267	6490	537	67	60

Table 3.10 Comparative analysis of identified peaks and putative OCT4 target genes

Considering a false discovery rate of 0.1, there are 1908 potential genome wide binding regions and 148 putative OCT4 target genes. From these target genes, 12 (8.11%) were also identified by Boyer et al. and the NCCIT derived OCT4 targets and 16 (10.81%) by Jin et al.. Setting the FDR to 0.5, there are 16591 potential genome wide binding regions that can be associated with 1152 genes. According to a higher number of putative OCT4 target genes, the total overlap to the two reference OCT4 target genes increases to 67 genes (5,82% Boyer et al.) and to 85 genes (7,38% Jin et al.). The overlap to the two reference OCT4 target genes increases to 179 (Boyer et al.) and to 267 genes (Jin et al.) when allowing a FDR of 0.9 at the expense of an increasing total number of genome wide peaks (93065) associated to 4987 putative OCT4 target genes (see target genes with motif sites in supplementary material). Peaks associated with a very high FDR (0.8-0.9) show a fold enrichment of 1.57-6 when the number of ChIP reads is compared against the number of Input reads that fall into the detected peak regions.

Thus, in total only very small overlaps could be observed when comparing peaks detected by the ChIP-seq approach with peaks detected in ChIP-chip approaches.

3.5.6 Motif mapping to binding regions

Table 3.10 also includes the results of mapping known OCT4 related motifs to the identified binding regions. Considering a FDR of 0.1, 366 (19.18%) of the original 1908 peaks contain a known OCT4 related motif. These 366 peaks can be associated with 30 putative OCT4 target genes (peak are located within the -10kb to +2kb range around a TSS). Although the total number of putative OCT4 target genes decreases by integrating the results of the motif mapping, the fraction of these target genes that were also identified by Boyer et al. increases to 33,33% (10 genes) and to 26.67% compared to the results from Jin et al. (8 genes). Lowering the stringency for the FDR results in a higher number of false positives and therefore, additional evidence for true binding sites is even more required: from the 16591 peaks identified by a FDR of 0.5, there are 1881 (11,34%) that contain an OCT4 related motif and subsequently there remain 139 sequence based confirmed putative OCT4 target genes. Allowing a

FDR of 0.9, there remain 6490 genome wide enriched regions (6,97% of all detected peaks) that can be associated with known OCT4 related motifs.

Conservation of OCT4 binding regions

The total set of identified putative OCT4 binding sites (FDR<=0.9) where analyzed for their conservation scores using the CisGenome software. The histograms in Figure 3.28 visualize the distributions of mean conservation scores for all peak regions with a conservation score>0 distinguishing between peaks associated with different FDRs (0.1, 0.5, and 0.9). It can be observed, that the frequency of peak regions is scattered over the full bandwidth of conservation scores when considering only peaks with a FDR<=0.1 (see Figure 3.28 A). By increasing the FDR, the majority of peak regions tend to assemble in the lower range of conservation scores (see Figure 3.28 C and 3.28 E).

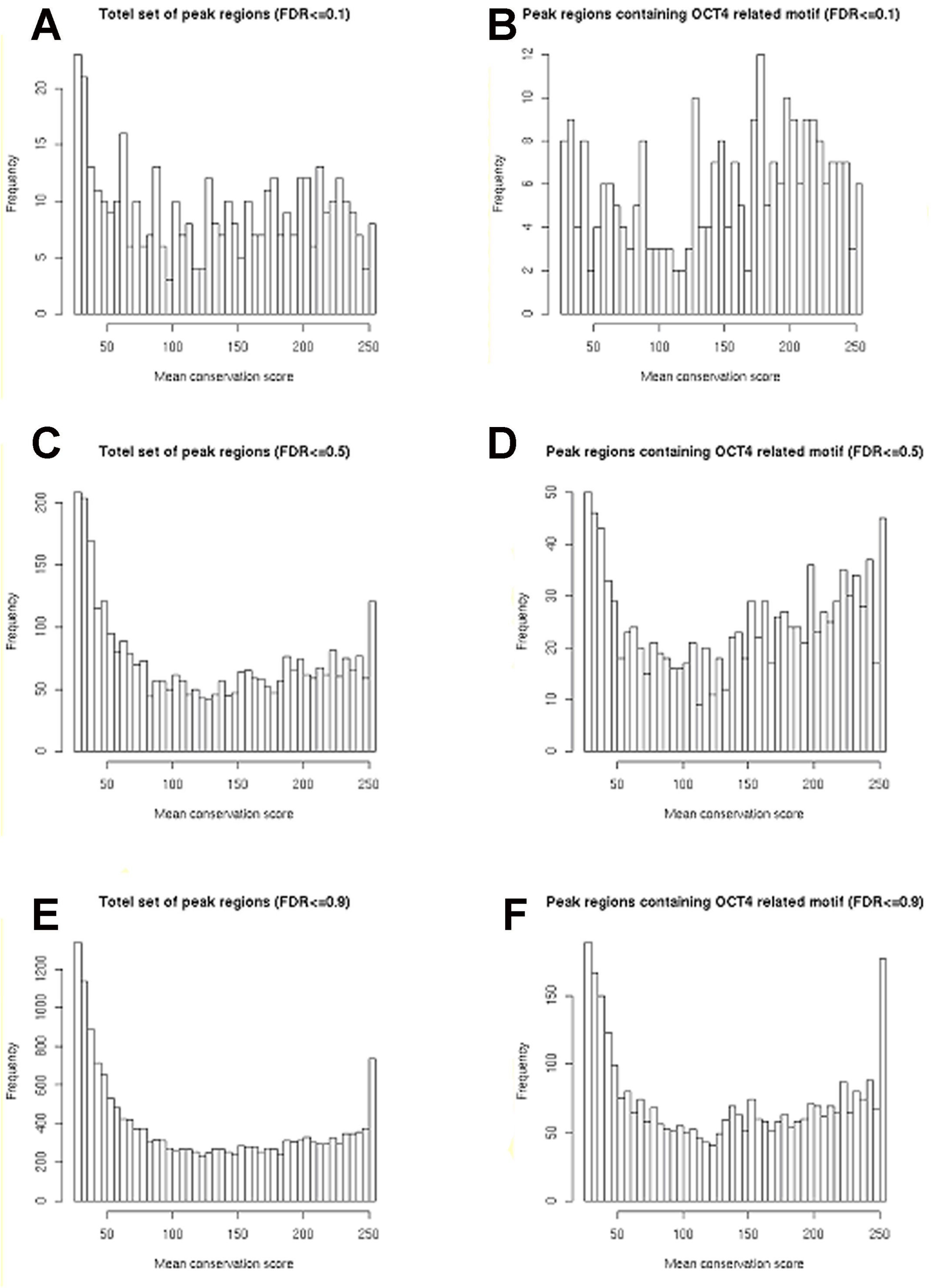


Figure 3.28 Conservation of peak regions

Histograms of mean conservation scores for A) all peak regions associated with a FDR≤0.1, B) all peak regions associated with a FDR≤0.1 that contain a known OCT4 related motif. C) all peak regions associated with a FDR≤0.5, D) all peak regions associated with a FDR≤0.5 that contain a known OCT4 related motif, E) all peak regions

associated with a FDR≤0.9, and F) all peak regions associated with a FDR≤0.5 that contain a known OCT4 related motif. Histograms include all peak regions that have a mean conservations score >0.

By selecting only OCT4 related motif containing peak regions, the majority of peaks assemble in a high range of conservation scores (see Figure 3.28 B). Although this effect is most clear for peaks associated with a FDR<=0.1, an analogous trend was observed when including peaks with higher FDRs (see Figure 3.28 D and 3.28 F).

Distribution of distances between OCT4 binding regions and TSSs

In order to identify direct target genes of a transcription factor, it is necessary to associate ChIP-Seq derived peaks and genes. For this, each peak is connected to the closest TSS (Transcription start site). The histograms in Figure 3.29 show frequencies of distances for the peak sets obtained by varying FDR (0.1, 0.5, and 0.9) and varying maximal distance (10kb, 100kb, and 1000kb) between peaks and TSSs.

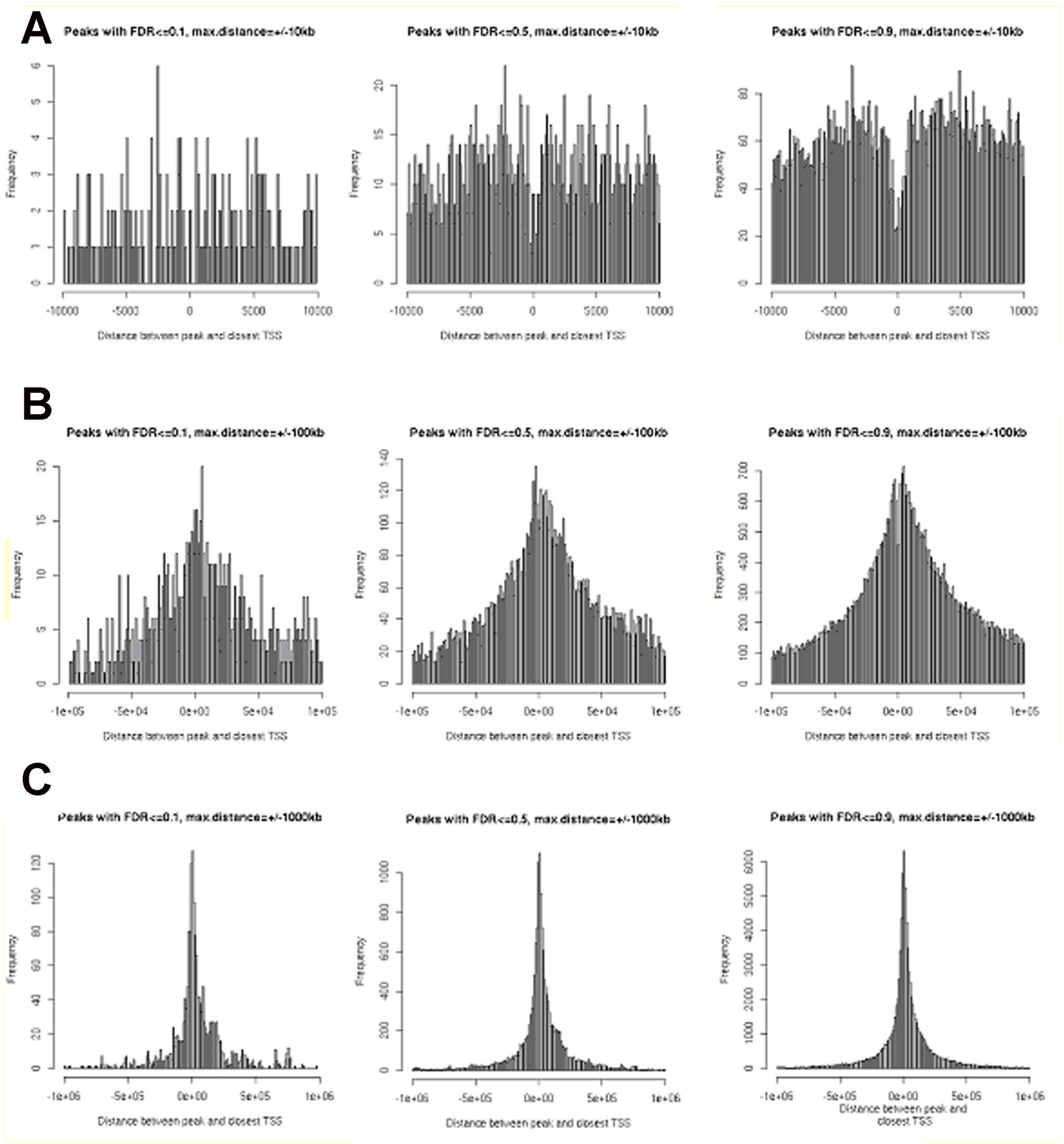


Figure 3.29 Distances between peaks and closest transcription start sites
Histograms of distances between peaks and transcription start sites with respect to a varying maximal distance (+/-10kb, +/-100kb, and +/-1000kb). a) Peaks associated with a FDR≤0.1. b) Peaks associated with a FDR≤0.5. c) Peaks associated with a FDR≤0.9.

At distance 0 the TSS is located and the positive regions (to the right) correspond to peaks located downstream of the TSS. Figure 3.29 A shows that the frequencies of distances vary over the range of -10kb to +10kb around the TSSs with no obvious main center. By including less-stringent peaks (FDR<=0.5 and FDR<=0.9), it can be seen that peaks located very close to TSSs (≤|1000bp|) become more underrepresented. The shape of frequencies of peak distances become more normally distributed when allowing for distances up to 100kb (see Figures 3.29 B and 3.29 C). By considering only highly specific peaks (FDR<=0.1) with a maximal distance of +/-100kb or +/-1000kb to the closest TSS respectively, it can be seen that distances are not totally symmetric distributed around the TSSs but there may be an enrichment of peaks in the range of up to 250kb downstream (see Figures 3.29 B and 3.29 C). Transcripts have a length of up to 2304kb, and 95.6% of all transcripts are shorter than 250kb. Therefore, an enrichment of peaks downstream of the TSS may originate from direct binding at gene associated introns or exons. Considering a FDR≤0.1 and a maximal distance of +/-250kb between peaks and TSSs, there are 1106 putative OCT4 target genes (6079 for a FDR≤0.5, and 13145 for a FDR≤0.9).

3.5.7 Functional regulation of OCT4 target genes

Although ChIP-Seq analysis allows for detecting putative direct target genes, no statement about transcriptional dependencies influencing gene expression can be deduced. Based on an RNAi mediated depletion of OCT4 function in hECCs followed by microarray analysis, 1946 genes were identified which show significantly altered expression 96h after the RNAi treatment. Of these, 528 genes were down and 1418 genes were up regulated. The overlap between functional regulation and direct TF binding can be tested with respect to a varying FDR and to a varying maximal distance between peaks and TSS. As an example, there are 134 genes showing altered expression after the RNAi mediated OCT4 knock down (42 down and 92 up regulated) and have an OCT4 binding site identified by a peak associated to a FDR≤0.1 within a range of +/-250kb around their TSS. Analogous, there are 589 functional regulated direct OCT4 target genes when setting the FDR to 0.5.

3.5.8 Functional enrichment analysis

For genes containing a peak with an FDR < 0.1

As the overlap of OCT4 target genes of the ChIP-seq approach presented here to previously performed ChIP-chip experiments was only 5 - 10%, it was tested if on a functional level, comparing the biological processes would reveal some similarities. For this reason, only those genes associated with an enriched region with an FDR<0.1 were considered. Thus 829 annotated genes were used as a query for the functional annotation software DAVID [145]. Table 3.11 shows the top ranked results, with a p-value as well as a benjamini p-value < 0.01

Term	PValue	Benjamini
GO:0032501~multicellular organismal process	2.42E-18	6.35E-15
GO:0007275~multicellular organismal development	1.31E-18	6.91E-15
GO:0032502~developmental process	1.70E-17	2.97E-14
GO:0048731~system development	2.99E-14	3.92E-11
GO:0048856~anatomical structure development	1.98E-13	2.08E-10
GO:0050789~regulation of biological process	9.71E-12	8.50E-09
GO:0065007~biological regulation	3.48E-11	2.61E-08
GO:0048513~organ development	1.21E-10	7.94E-08
GO:0009653~anatomical structure morphogenesis	2.82E-09	1.64E-06
GO:0009887~organ morphogenesis	1.14E-08	5.97E-06
GO:0030154~cell differentiation	1.34E-08	6.39E-06
GO:0048869~cellular developmental process	1.34E-08	6.39E-06
GO:0007399~nervous system development	3.70E-08	1.49E-05
GO:0048518~positive regulation of biological process	4.58E-08	1.72E-05
GO:0007154~cell communication	6.27E-08	2.20E-05
GO:0050794~regulation of cellular process	7.33E-08	2.41E-05
GO:0007267~cell-cell signalling	4.17E-07	1.29E-04
GO:0048468~cell development	7.07E-07	2.06E-04
GO:0048522~positive regulation of cellular process	4.54E-06	0.001
GO:0007165~signal transduction	5.97E-06	0.001
GO:0007167~enzyme linked receptor protein signalling pathway	1.16E-05	0.002
GO:0035295~tube development	1.27E-05	0.003
GO:0045165~cell fate commitment	1.91E-05	0.004
GO:0001505~regulation of neurotransmitter levels	2.06E-05	0.004
GO:0048519~negative regulation of biological process	2.78E-05	0.005

Table 3.11 Functional enrichment analysis of potential OCT4 target genes with peaks containing a FDR <0.1. Only those functional annotation terms having a P value below 0.01 (for both Student t test and Benjamini) are shown.

The main cluster of enriched ontologies contained developmental processes with a focus of nervous system development. When compared to the finally 497 OCT4 binding sites, obtained by the ChIP-chip analysis, developmental processes were a minor cluster with transcriptional regulation, homeobox genes and specific embryonal development being the main clusters. Thus it seems that the target sites obtained by ChIP-seq reveal more general developmental and differentiation inducing pathways, whereas the OCT4 ChIP-chip target sites are more specifically related to processes involved in transcriptional regulation.

For OCT4 target genes, showing significant altered expression after OCT4 RNAi knockdown.

Functionally regulated direct OCT4 target genes can be selected with respect to a variety of parameters based on the results of the presented study. In order to select a highly confirmed set of putative targets, the total set of identified peaks (FDR≤0.9) was chosen to exclude as many false negatives as possible and further selected those that have a mean conservation score ≥ 50, and contain a known OCT4 related binding site. The remaining 2667 peaks can be associated to 1737 genes by allowing a maximal distance of +/- 250kb between peak and TSS. Filtering this gene set further by selecting only those that showed significant altered expression after the RNAi knockdown. The final set of 203 highly confirmed functional direct OCT4 target genes were tested for enriched gene ontology's and pathways. Table 3.12 lists all pathways and gene ontologies enriched with a both a pValue and a Benjamini pvalue≤0.01.

Term	PValue	Benjamini
GO:0007275~multicellular organismal development	2.05E-20	1.08E-16
GO:0032502~developmental process	2.67E-19	7.02E-16
GO:0048731~system development	6.20E-17	1.94E-13
GO:0048856~anatomical structure development	3.94E-16	5.83E-13
GO:0032501~multicellular organismal process	1.68E-14	1.76E-11
GO:0048513~organ development	5.74E-12	5.03E-09
GO:0009653~anatomical structure morphogenesis	3.72E-10	2.79E-07
GO:0050794~regulation of cellular process	2.07E-09	1.36E-06
GO:0007399~nervous system development	4.62E-09	2.70E-06
GO:0050789~regulation of biological process	7.22E-09	3.79E-06
GO:0065007~biological regulation	1.80E-08	8.61E-06
GO:0048519~negative regulation of biological process	4.84E-08	2.12E-05
GO:0048523~negative regulation of cellular process	5.45E-08	2.20E-05
GO:0030154~cell differentiation	1.70E-07	5.97E-05
GO:0048869~cellular developmental process	1.70E-07	5.97E-05
hsa04510:Focal adhesion	8.38E-07	1.68E-04
GO:0009887~organ morphogenesis	1.36E-06	4.46E-04
GO:0006357~regulation of transcription from RNA polymerase II promoter	9.20E-06	0.002
GO:0048468~cell development	1.18E-05	0.003
GO:0022008~neurogenesis	1.46E-05	0.004
GO:0007507~heart development	1.64E-05	0.004
GO:0006366~transcription from RNA polymerase II promoter	2.72E-05	0.006
GO:0008283~cell proliferation	4.04E-05	0.009

Table 3.12 Functional enrichment analysis of highly confirmed OCT4 target genes. Only those functional annotation terms having a P value below 0.01 (for both Student t test and Benjamini) are shown.

A main cluster of enriched gene ontology's is connected to development especially, nervous system and heart development. For lower Benjamini P-values, OCT4 regulated direct target genes are enriched for functions connected to neuronal differentiation and development, heart, cartilage, skeletal development, and embryonic limb morphogenesis. Similar to the Boyer data of OCT4 bound regions in human ES cells, another cluster consists of transcriptional regulation. Interestingly, the KEGG pathway Focal Adhesion is enriched in this dataset. But unlike the comparison between EC and ES cells, where genes involved in cell adhesion were enriched (seen above) not collagens, tight junction adhesion molecules or EGFR were detected but genes involved in the signalling process like PI3K and Akt/PKB (See Figure 3.30).

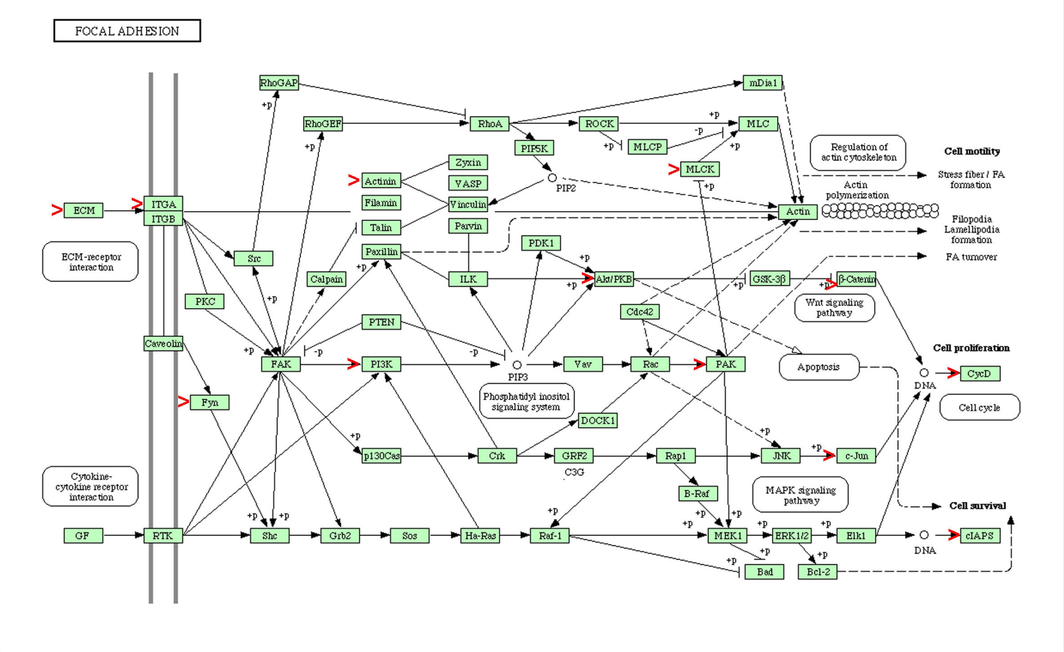


Figure 3.30 KEGG Pathway Focal Adhesion, showing genes (marked by red >) contained in highly confirmed functional direct OCT4 target genes.

Another minor cluster of enriched gene ontology's consisted of cell proliferation processes, reflecting possible changes in proliferation pathways after the induction of OCT4 ablated differentiation.

3.5.9 Summary

For the ChIP-seq analysis the three biological replicates, used for the ChIP-Chip hybridization were pooled for the control and the enriched DNA. The first steps in the analysis showed that the 11.2 Mio ChIP reads obtained, covered the existing immunoprecipitated DNA fragments in a reliable depth. The distribution of peaks was surprisingly not enriched in the proximal promoter regions but rather distributed to intergenic regions. When comparing the ChIP-seq peak regions with identified peak regions of ChIP-chip experiments, the overlap was only very small. Nonetheless, when performing a functional annotation, genes involved in diverse developmental processes were highly enriched, speaking against a randomized enrichment due to an amplification bias. Highly enriched regions in the ChIP-chip experiment like the OCT4 binding site for *NANOG* were correlated with an FDR of only 0.5 (See Figure 3.31). However, using the CisGenome intrinsic model for FDR calculation, this binding site was related to 21 reads in the IP channel compared to 0 reads in the Input channel, so obviously those peaks, correlated with a low amount of reads in the Input channel were associated with high FDR values.

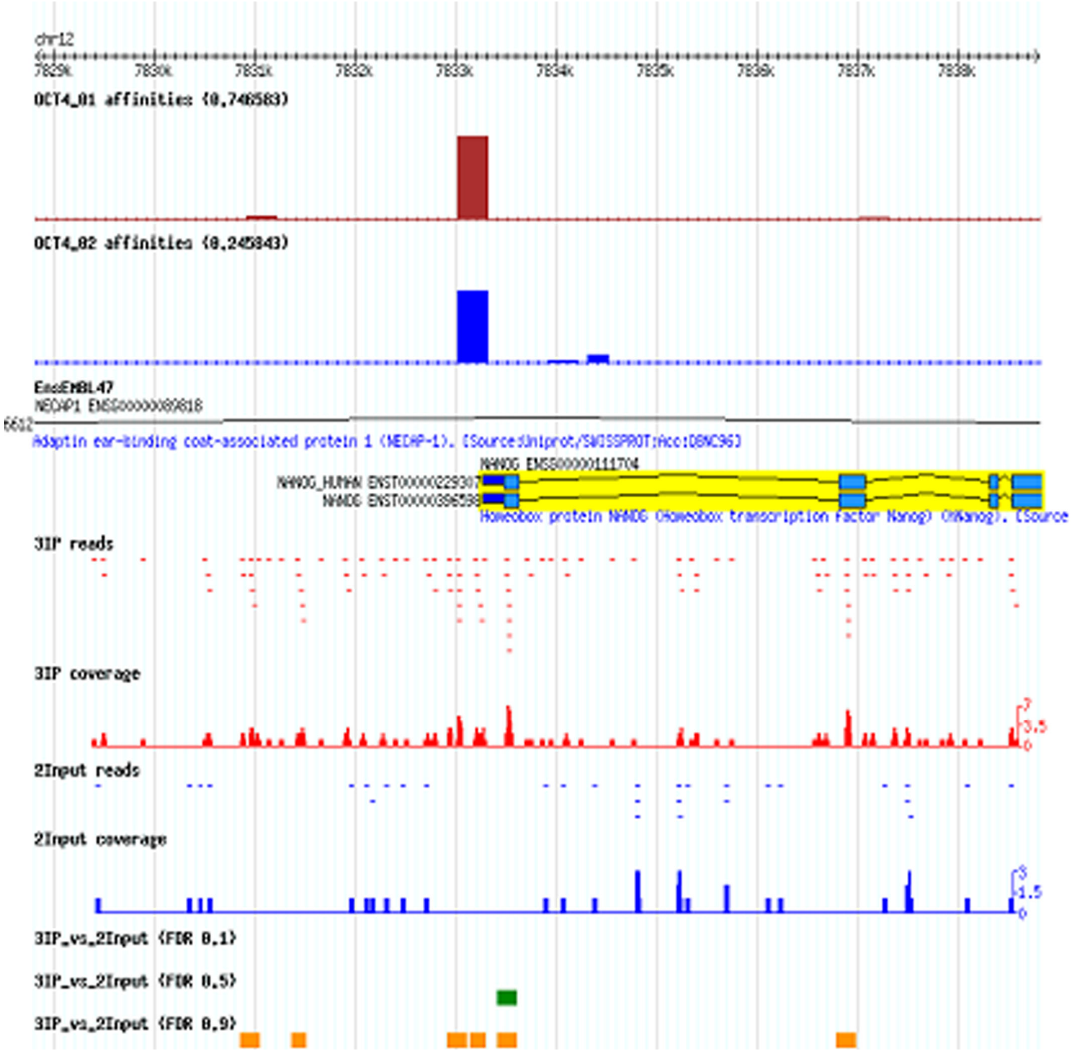


Figure 3.31 Genome browser view of the OCT4 binding site in the 5 prime proximal promoter of *NANOG*. Shown are the affinity values for one OCT4 motif, identified by PASTA [155], related to the oct-sox motif, validated by Rodda et al. [65]. Furthermore the number of ChIP-seq reads for the Input and the IP channel are shown.

For subsequent analysis steps, these cases need to be included for a more refined statistical model.

On the other hand, a plethora of new OCT4 binding regions could be discovered in intragenic, intergenic and 3 prime regions. However, further functional confirmation of these putative binding sites by band shift assays and luciferase reporter assays were beyond the scope of this study and are needed to obtain a functionally validated target set.

4 Discussion

4.1 General considerations regarding the ChIP technique

Combining siRNA or esiRNA induced knockdowns of the gene of interest with the potential binding sites of the gene products is a powerful method for deciphering gene regulatory networks.

With regards to deepening the understanding of ES cell specific pathways, such an approach has been applied using a genome-wide RNAi screen which leads to the identification of a new transcriptional module required for self-renewal. This module implicates over 100 genes in ES cell self-renewal, and illustrates the power of RNAi and forward genetics for the systematic study of self-renewal in mouse ES cells [76]. Since such an approach is still missing for human ES cell models, the basis of this work was to detect OCT4 specific regulatory nodes in human EC cell models, similar to human ES cells. Thus an OCT4 specific polyclonal antibody has been used. It would have been also possible to overexpress the protein with a tag (e.g. biotin tag [156], FLAG tag [157] or TAP tag [158]) and to enrich via the tag. These approaches, however, are not recommended, as the necessary over-expression leads to a disturbance of the normal physiological protein concentration. Therefore the results obtained using tagged proteins may not reflect accurately the normal binding behavior, even more true for transcription factors, which can form multimeric complexes.

4.2 ChIP-chip results compared to literature

ChIP based studies on the transcription factor OCT4 have been carried out by others [83,137]. However, none of these studies compared the peak regions identified using different detection programs. As been demonstrated in this work, using the online available programs MAC2 [134] and TAMALPAIS [135] and an in-house developed algorithm for peak discovery, the overlap between the single programs is below 50%, meaning that a substantial proportion of potential binding sites would be lost if one depends on one algorithm. Given that a lowered specificity and sensitivity after a random based PCR amplification have to be taken into account, TAMALPAIS and MAC2 are still the best algorithms for true positive prediction, although they would not achieve AUC (Area Under rock Curve) values beyond 0.7, using ROC-like curves (receiver operating characteristic curves) for diluted spike ins [146]. ROC-like curves plot sensitivity vs. filtering fraction at every threshold. On ROC curve True Positive Rate is plotted against False Positive Rate calculated at each cut-off [159]. To compare two different methods usually area under these curves is computed. A random method would have an area equal to 0.5 and a perfect method would have and area equal to 1. True positive peaks might be represented by different complex peak shapes, which one algorithm alone would not detect and thus the approach presented here, combining different programs in a rank based score, potentially leads to a more complete target list.

As shown before, NCCIT cells are a useful tool for investigating pathways involved in stemness and differentiation [38]. One question of this work was in how far NCCIT specific OCT4 enriched regions could be compared to human ES specific binding sites, identified by Boyer et al. [83] and in other human EC cell lines such as NTERA2 [137]. The overlap reported here is below 10%. This is based on different platforms and different peak finding programs used. This study shows that different programs can lead to a different set of target genes. Additionally, specific differences based on a different binding pattern for each specific cell line cannot be ruled out. Finally, the comparisons are obtained only for a selected promoter region, for which there is evidence that most binding events occur [83] but nonetheless reveal potential functional binding events associated to non-proximal promoter specific regions. Indeed cell-type specific pathways correlated to OCT4 binding sites between EC and ES cells could not been detected. To shed light on these questions, a comparison based on ChIP-seq, using the same tools for the data analysis would be needed.

Nonetheless, with regards to the common targets identified in this work, key stem cell markers like NANOG, OCT4, SOX2, HESX1, other homeodomain containing proteins like NKX2-2, SIX1, HOXB4, LHX5, transcription factors like ZIC4, SP8 and enzymes like DUSP6, PPP2R3A, which are potential candidates for either retaining pluripotency or inducing differentiation pathways not only in an ES cell specific model, were discovered as potential binding sites of OCT4.

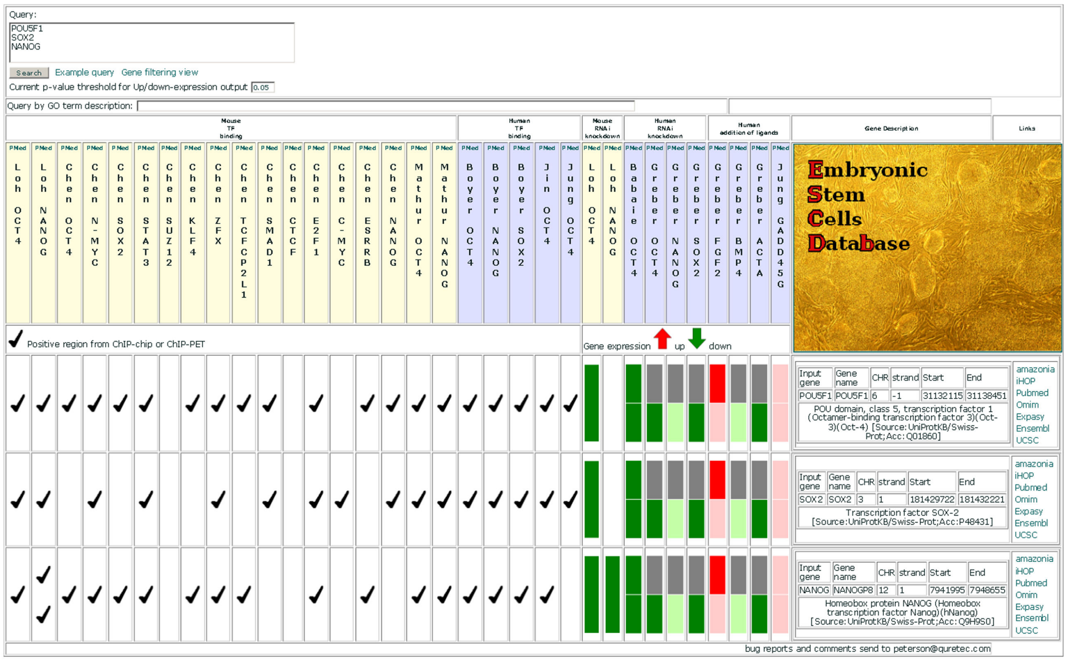
Performing a functional annotation with genes correlating with identified peaks in their promoter region for NCCIT cells, within the most stringent clusters were genes, containing homeodomains (See Table 3.14). With regards to the upregulated homeodomain containing proteins it is noteworthy to mention those genes which were also upregulated upon OCT4 ablation - *MSX1/2* and *ISL1*.

MSX1 and 2 are known BMP4 downstream targets [160]. An upregulation of BMP4 could be confirmed and both genes were strongly upregulated upon BMP4 stimulation in hES cells. In our case the upregulation also seemed to be independent of the overexpression of *PHOX2B*, formerly reported to upregulate *MSX1* gene expression [161].

ISL1 is a LIM-homeobox gene important for developmental and regulatory function in islet, neural, and cardiac tissue [162].

4.3 Data integration in the form of an Embryonic Stem Cell database

We are in an era of high-throughput functional genomics and systems biology-driven research where large datasets are usually needed and provided as supplementary tables in most publications. Though useful, such tables in isolation are of limited use for making cross-references across other related datasets. Furthermore, as similar approaches have recently been adopted in constructing the HaemAtlas which serves as a reference library for gene expression in human blood cells and as a resource for identifying key genes with roles in blood cell function [163], a specialized database has been developed, which enables rapid and convenient access and comparisons between published datasets related to embryonic stem cell biology to help overcome this shortfall. In order to facilitate the construction of this database, previously published datasets were gathered together with ChIP-on-chip using OCT4 and the NCCIT cell line to establish the Embryonic Stem Cell Database (http://biit.cs.ut.ee/escd/, see Fig. 4.1 for an example).



Query:
POU5F1
SOX2
NANOG

Search Example query Gene filtering view

Current p-value threshold for Up/down-expression output 0.05

Query by GO term description:

✓ Positive region from ChIP-chip or ChIP-PET

Gene expression up down

	Mouse TF binding																		Human TF binding					Mouse RNAi knockdown		Human RNAi knockdown				Human addition of ligands				Gene Description	Links
	PMed Loh OCT4	PMed Loh NANOG	PMed Chen OCT4	PMed Chen N-MYC	PMed Chen SOX2	PMed Chen STAT3	PMed Chen SUZ12	PMed Chen KLF4	PMed Chen ZFX	PMed Chen TCFCP2L1	PMed Chen SMAD1	PMed Chen CTCF	PMed Chen E2F1	PMed Chen C-MYC	PMed Chen ESRRB	PMed Chen NANOG	PMed Mathur OCT4	PMed Mathur NANOG	PMed Boyer OCT4	PMed Boyer NANOG	PMed Boyer SOX2	PMed Jin OCT4	PMed Jung OCT4	PMed Loh OCT4	PMed Loh NANOG	PMed Babaie OCT4	PMed Grebber OCT4	PMed Grebber NANOG	PMed Grebber SOX2	PMed Grebber FGF2	PMed Grebber BMP4	PMed Grebber ACTA	PMed Jung GADD45G	Embryonic Stem Cells Database	
POU5F1	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	🟩		🟩 🟩	⬛ 🟩	⬛ 🟩	⬛ 🟩	🟥 🟥	⬛ 🟩	⬛ 🟩	🟥 🟥	Input gene: POU5F1; Gene name: POU5F1; CHR: 6; strand: -1; Start: 31132115; End: 31138451 POU domain, class 5, transcription factor 1 (Octamer-binding transcription factor 3)(Oct-3)(Oct-4) [Source:UniProtKB/Swiss-Prot;Acc:Q01860]	amazonia iHOP Pubmed Omim Expasy Ensembl UCSC
SOX2	✓	✓		✓		✓			✓		✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	🟩		🟩 🟩	⬛ 🟩	⬛ 🟩	⬛ 🟩	🟥 🟥	⬛ 🟩	⬛ 🟩	🟥 🟥	Input gene: SOX2; Gene name: SOX2; CHR: 3; strand: 1; Start: 181429722; End: 181432221 Transcription factor SOX-2 [Source:UniProtKB/Swiss-Prot;Acc:P48431]	amazonia iHOP Pubmed Omim Expasy Ensembl UCSC
NANOG	✓	✓ ✓	✓	✓	✓	✓		✓	✓	✓			✓		✓		✓	✓	✓	✓	✓	✓		🟩	🟩	🟩 🟩	⬛ 🟩	⬛ 🟩	⬛ 🟩	🟥 🟥	⬛ 🟩	⬛ 🟩	🟥 🟥	Input gene: NANOG; Gene name: NANOGP8; CHR: 12; strand: 1; Start: 7941995; End: 7948655 Homeobox protein NANOG (Homeobox transcription factor Nanog)(hNanog) [Source:UniProtKB/Swiss-Prot;Acc:Q9H9S0]	amazonia iHOP Pubmed Omim Expasy Ensembl UCSC

bug reports and comments send to peterson@quretec.com

Figure 4.1 Example of the user interface for the constructed integrated database, three queries have been added (OCT4, SOX2 and NANOG). Potential binding sites for several transcription factors and expression changes induced by different perturbation changes can be seen at one glance.

The new database provides easy access to transcription factor binding data together with various perturbation experiments. ESCDb gathers mainly two types of data – chromatin immunoprecipitation array-based data on transcription factor targets and gene specific knockdown of pluripotency associated factors (OCT4, SOX2 and NANOG) as well as growth factor (FGF2) withdrawal and cytokine (BMB4 and ACTIVIN A) stimulation of human ES cells. Data for mouse and human embryonal stem cells have been collected, and complemented with data from embryonic stem cell experiments data from human embryonal carcinoma cells (NCCIT and NTERA2).

ESCDb offers a summarized view of multiple pluripotency related datasets. Individual genes are described as a row in the output table. A colour-scheme helps to illustrate the potential regulatory relations between genes. In the gene-expression datasets often more than one probe-set represent a gene and we treat each individual probe-set individually. The same order of probe-sets in the output table was kept for easier comparisons between probe-sets in all available datasets. Further details are given in numerical form when a given cell of the table is pointed with a cursor. The database can be queried with any widely used gene or protein identifier or Gene Ontology terms.

The current version of the database comprises gene expression data from 18 mouse transcription factor-targeting experiments for 14 known factors [83,164,165] (OCT4, SOX2, NANOG, N-MYC, C-MYC, STAT3, SUZ12, KLF4, ZFX, TCFCP2L1, SMAD1, CTCF, E2F1, ESRRB), 5 human transcription factor binding experiments [83,137] for the 3 main pluripotency regulators (OCT4, SOX2, NANOG), 2 mouse ES cell knock-down experiments for Oct4 and Nanog [150] and 8 perturbation experiments (including knockdowns of OCT4, SOX2 and NANOG in EC cells and overexpression of GADD45G in EC cells), BMP4 and ACTIVIN A stimulated hES cells and FGF2 withdrawal from hES cells during culture [38,86].

4.4 Different modules of OCT4 binding

Particularly in the case of OCT4, where a transcription factor could be part of several complexes with different factors interacting directly with DNA as proposed by Stuart Orkin [166], these sequences enriched in a ChIP-chip experiment could be a complex mixture of sequences which contain motifs for the profiled transcription factor and/or various interacting proteins.

In the following, six different modes of OCT4 binding (termed modules) will be discussed (see Figure 4.2):

Module 1: OCT4 and SOX2 will bind as a heterodimer co-operatively to their cis-elements in the respective promoter regions.

Module 2: Only OCT4 will bind to its cis-elements in the respective promoter regions.

Module 3: Only SOX2 will bind to its cis-elements in the respective promoter regions.

Module 4: Neither OCT4 nor SOX2 will bind to its cis-elements in the respective promoter regions.

Module 5: OCT4 will bind specifically to a PORE motif, forming a dimer.

Module 6: OCT4 will bind specifically to a MORE motif, forming a dimer.

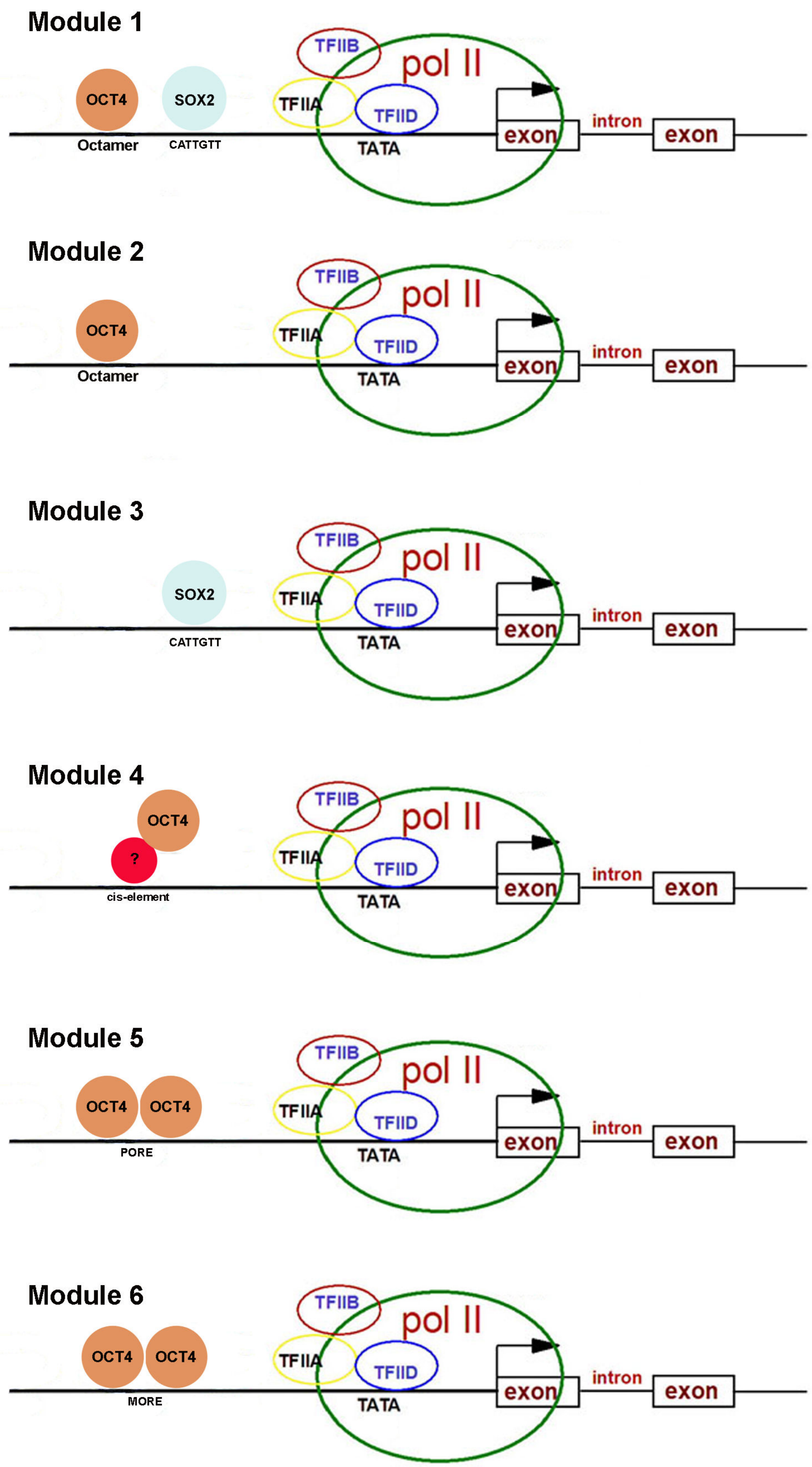


Figure 4.2 Six different modules of OCT4 binding modes, which are specific for the five prime proximal promoter regions. (Adopted from M. W. King)

The HMG containing transcription factor SOX2, is known to form a heterodimer with OCT4 which results in a protein-protein-DNA complex required for transcriptional regulation of genes such as *UTF1, FBX15, SOX2* and *NANOG* [60,82,167-169]. Based on the plurality of interactions between HMG and POU class proteins and the co-evolution of HMG/POU DNA target sequences, this interaction is thought to be a fundamental mechanism for the control of gene expression involved in developmental processes [61]. Furthermore, as shown for the *FGF4* promoter, the distance between the binding recognition sites of SOX2 and OCT4 seem to be crucial for synergistic activation. For enriched regions in NCCIT cells that contain both motifs the tendency to have a closer distance between each other is independent of strand orientation (Fig 4.3).

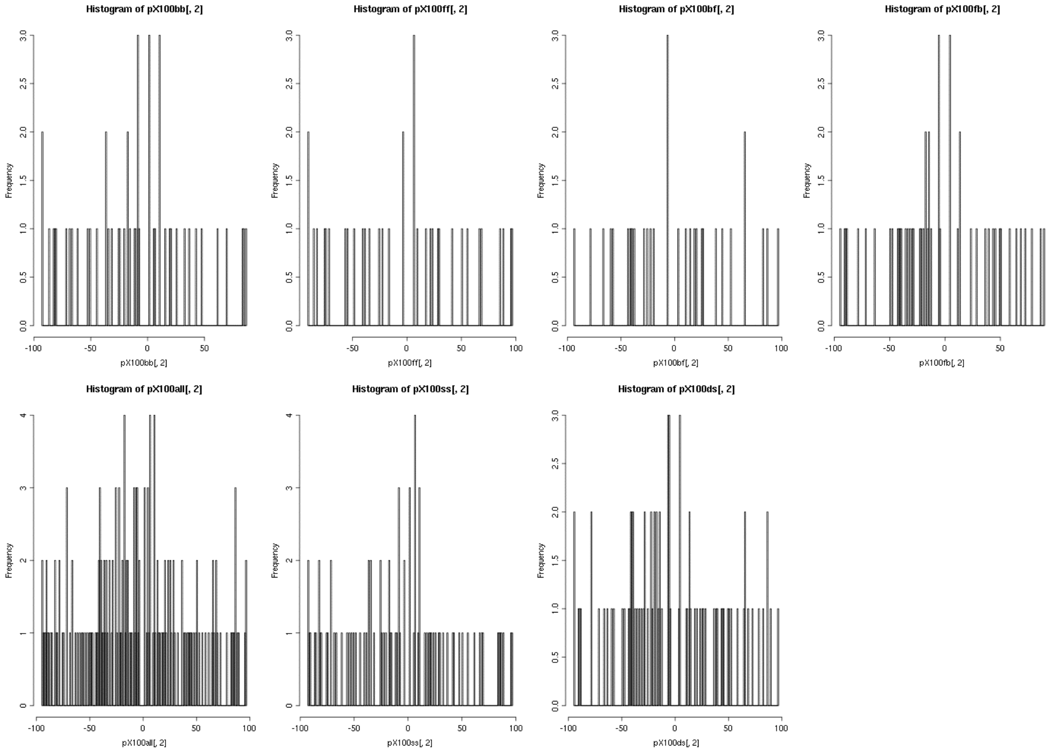


Fig. 4.3 OCT4 and SOX2 distances are depicted on the histogram. First row shows each strand configuration separately (e.g. pX100bf means motifs were chosen from only the highest scoring part of the peak, max window 100bp and OCT4 motif was found on backward strand while SOX2 motif was found on forward strand.

Another question is if the close proximity of the binding recognition sites of SOX2 and OCT4 is a pre-requisite for the proper assembly of functional activation complexes. The results suggest that there is no such correlation. This is based on the unveiling of 6 distinct modules of OCT4-regulated gene regulatory networks with genes within or between each module having distinct distances between the SOX2 and OCT4 binding motifs or even not having a SOX2 motif adjacent to that of OCT4.

Based on these results, it seems that the SOX2-OCT4 motif or the close proximity of both motifs is not required for the majority of OCT4 regulated target genes. For these genes, octamer motifs might be more displaced from our peak regions and hint at protein-chromatin interactions, bringing different chromatin regions into close proximity.

Boyer and colleagues have shown that approximately half of the promoter regions discovered by ChIP-chip analysis [83], occupied by OCT4 were also bound by SOX2 in human ES cells. In the analysis with human EC cells using the in silico-derived SOX2 motif for target identification instead of peak regions unveiled 108 SOX2-motif related putative binding sites out of 497 total binding sites and 161 binding sites linked to an OCT4 motif. However, this is only a fraction of the putative SOX2 binding sites identified in hES cells, thus suggesting distance related effects and/or other SOX2 motifs not discovered with our analysis. Additionally, one has to bear in mind that all thresholds defined for the OCT4 and SOX2 Poly Weight Matrices (PWMs) are arbitrarily set and therefore can only provide a prediction for a bona fide functional binding event, hence further experimental validation will be needed.

To identify binding modules, where the octamer element is not present, the 497 target genes for the presence of PORE or MORE motifs as an addition to target genes defined by module 4 were also screened. Four putative target genes harbouring a PORE motif (module 5) and 10 target genes, which contained a MORE motif (module 6) were identified.

Using a previously identified MORE (CTGCATATGCAT) motif within the *BMP4* promoter, Kang et al. [77] verified an interaction between Oct4 and this genomic region and showed using mouse ES cells subjected to ionizing radiation that Oct4 occupancy was induced by stress. Based on these observations, it is tempting to speculate that a subset of OCT4 targets harbouring the MORE motif might be associated with the modulation of stress responses.

Taken together, a concept of different direct and indirect OCT4 binding patterns is provided, depending on associated OCT4 related transcription factor binding sites. A similar approach applied by Segal and colleagues has been applied, to identify regulatory modules and their condition-specific regulators in yeast [170]. However, there was a difference in that the screen was started with potential transcription factor occupancy in relation to the presence of their specific binding sites.

There are many questions, which still remain unanswered. Which cellular constraints will lead to OCT4 differential regulated subsets of genes, which might belong to one of the 6 modules? As a provocative thought, is there an OCT4 regulatory module specific for maintaining the self-renewal circuitry, or specific for suppression of the induction of differentiation to distinct cell lineages by the recruitment of co-activators or repressors to the OCT4 transcriptional complex. In response to these questions, hypothetical schemes (Fig. 4.4) which are based on the *de novo* motif discovery analysis performed on the OCT4 indirect target genes were presented, postulated to be regulated under module 4.

As illustrated in Fig. 4.4 A and B, OCT4 might form a distinct or even the same complex with TCF3 and REST to maintain positive-gene regulatory networks supporting self-renewal. Interestingly both genes are highly expressed in undifferentiated ES and EC cells and their expression declines upon differentiation. Furthermore, TCF3 has been assigned as an integral component of an interconnected autoregulatory loop, where OCT4, SOX2, NANOG and TCF3 occupy each and their own promoters in maintaining the self-renewal circuitry in embryonic stem cells [171]. REST, a transcriptional co-repressor has been shown in mouse ES cells to selectively repress transcription of a subset of neuronal genes [172].

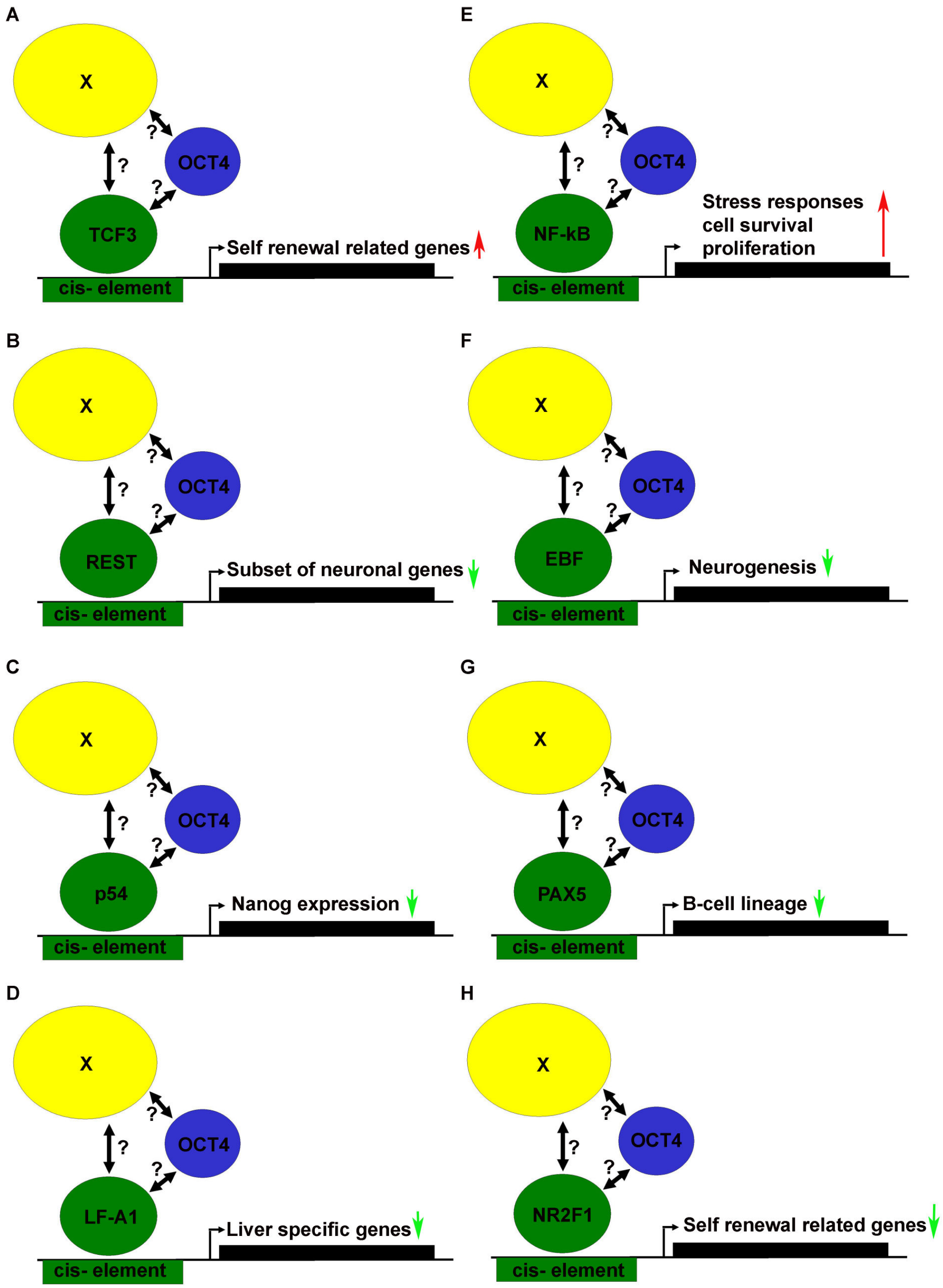


Figure 4.4 Hypothetical model based on module 4 of how OCT4 could be involved in regulating its target genes via non-direct DNA binding. OCT4 might be recruited by a mediator complex (X), which has additional affinity for the discussed transcription factors (A – H). Alternatively, there might be a direct interaction between OCT4 and the transcription factor(s) (indicated by '?'), which might then potentially bind to the identified in silico cis elements. Arrows: Red- induction and green- repression of transcription of the respective target genes.

Another protein complex that might promote self-renewal is composed of OCT4 and NFKB1 (Fig. 4.4 C) in positively regulating gene networks in response to stress signals to activate cell survival and proliferation pathways [173].

Furthermore, the regulatory schemes depicted in Fig. 4.4 D-G, represents scenarios where the OCT4-bound complex might sustain self-renewal by inhibiting the differentiation inducing activities of transcription factors such as TP53 [174], LF-A1 [175], EBF [176], PAX5 [177] and NR2F1 [178]. Unfortunately, experiments to test and confirm these hypotheses are beyond the scope of this study. However, evidence of indirect binding has been supported by an independent study by Gordan et al. [179]. They applied the hypothesis of indirect binding to yeast ChIP-chip data of 139 transcription factors (TFs). Their method revealed that only 48% of the data could be explained by direct binding of the profiled TFs, while 16% could be explained by indirect binding. In addition to the approach presented here, Gordan used *in vivo* nucleosome positioning. However they reported only a slight improve in the detection of indirect TFs and nucleosome data are not yet available for human EC or human ES cells. In addition, they suggested indirect TF-DNA interaction when the motif of the profiled TF was not significantly enriched in the ChIP-chip data. This was not the case for the motif we recovered for OCT4, but still around 66% of the enriched sequences did not contain OCT4 motifs, and one of the hypothesis of this study is that these sequences might still be valid candidates for a profiling of indirect TFs.

As a precautionary note, the possibility that the OCT4-regulatory modules described here are just the tip of the iceberg cannot be excluded and that with the adoption of an unbiased screen of OCT4 targets using ChIP-seq will reveal the complex nature of the self-renewal-gene regulatory network under the control of OCT4. A precedent for this is the identification in mouse ES cells of an extended network for pluripotency [167] and also indications that OCT4 can also bind to chimeric combinations of OCT4 half sites [180].

4.5 USP44 is a potential cell cycle regulator, controlled by OCT4

In this study, a highly conserved binding site for OCT4 was discovered in the proximal promoter region of the ubiquitin specific protease USP44. With respect to characterised potential downstream targets of OCT4, it was intriguing to speculate a possible direct regulation of USP44, an important regulator of the spindle checkpoint. A highly conserved OCT4 binding site was uncovered within its proximal promoter and a significant decrease of the transcript level in OCT4 knockdown experiments [38,45]. Furthermore, screening the online hESC expression atlas Amazonia [16], a significant decrease of this transcript upon embryoid body-based differentiation was uncovered, and the level remain low in various somatic tissues. Based on these findings it can be proposed that USP44 is a positive regulator of self-renewal in EC as well as ES cells and that this regulation is mediated by its prominent role in regulating the spindle checkpoint during the cell cycle [151]. Further experiments, like luciferase assays need to be performed to further validate USP44 as a direct OCT4 target. Concerning a functional role of USP44, no direct change of the key pluripotency factors OCT4, SOX2 and NANOG, arguing for a positive feed forward loop, and no morphological differences could be observed after partial ablation. Either way the knockdown was not efficient enough or a phenotype might be only observed after perturbations, e.g. differentiation of the cells with retinoic acid.

4.6 Genistein induces the upregulation of GADD45G and has an effect on the expression of key pluripotency markers

Previous experimental work addressing the effects of Genistein on cell proliferation and differentiation were performed using prolonged-cultured, transformed cell lines. These earlier findings, though informative, have short comings with respect to the genomic integrity of the cells used for these analyses. Genistein applied to low passage cultured cells has a noticeable effect on the transcription of common key regulators of cell-cycle progression. In terms of the mechanism(s) of action of Genistein, NFkB-mediated repression of *GADD45A* and *GADD45G* expression has been shown to be essential for cancer cell survival [181]. Furthermore, *GADD45A* expression has been shown to be induced by Genistein treatment of human prostate cancer cell lines [182]. To test if Genistein also imparts similar effects in other cancer cells, the embryonal carcinoma cell line (NCCIT) has been used, which has properties of cancer cells as well as pluripotent cells [26,38]. GADD45G and GADD45A are regulators of the cell-cycle at the G2/M transition [152] and act as tumor suppressors [153]. In this study, the direct effect on the gene expression of GADD45G and GADD45A could be shown. This result was confirmed by Oki et al. [182]. Furthermore, GADD45G has been shown to be a negatively regulated, direct downstream target of OCT4 [38,45,83]. Indeed, Genistein treatment of NCCIT cells led to the induction of *GADD45A* and *GADD45G* expression. Additionally, a reduction in *NANOG* transcription was noticed but not that of *POU5F1* and *SOX2*. A reduced level of NANOG could not be linked to a differentiation phenotype, but rather to reduced proliferation in NCCIT cells [38]. As shown before, down-regulation of OCT4 leads to the down-regulation of NANOG. The observed decrease in the transcript level of NANOG might be a downstream effect of Genistein-induced depletion of OCT4 protein [65].

Furthermore, a decrease in OCT4 and NANOG protein was detected. A speculation is that Genistein treatment might indirectly down-regulate *POU5F1* expression, possibly mediated by the up-regulated expression of *GADD45G*.

One mechanism to reduce OCT4 levels is the Ubiquitination of OCT4 by the HECT domain E3 ubiquitin ligase WWP2, thereby promoting its subsequent degradation by the 26S proteasome [183]. However by microarray analysis with Genistein treated cells against DMSO mock treated cells, instead of an upregulation of WWP2, a nearly 2 fold downregulation was observed.

4.7 GADD45G induces the upregulation of differentiation related genes

GADD45G is a potential tumour suppressor protein involved in cell cycle control. A functional octamer site for OCT4 could be identified and the array data indicates the upregulation of this gene upon siRNA induced OCT4 knockdown [45]. Furthermore, there are indications in mouse ES cells that the transcription level of *GADD45G* increases significantly upon differentiation stimuli [184]. It is tempting to speculate that OCT4 could have an impact on cell cycle regulation via GADD45G downstream cascades as it has been shown before that it is essential for the G2/M arrest and induction of apoptosis signalling, through the dissociation of the cyclinB/Cdc2 complex [185,186]. The influence of octamer sites in *GADD45G* expression have been explored recently, showing the functional importance along with NF-Y sites. In these cases OCT1 has bee shown to bind to the octamer site [187]. Since both OCT1 and OCT4 recognise the same binding site it would be interesting to hypothesize that the upregulation of GADD45G by OCT4 downregulation is promoted by a competitive effect in that OCT4 is replaced by OCT1, thus recruiting HAT-containing co-activator complexes as Oct-1 and NF-Y can interact with them. Since the focus of our study was the pattern of OCT4 linked motifs, competition events by OCT1 cannot ruled out as has been proposed before as an alternative model for regulation [188]. Furthermore there is indication that inhibition of NF-Y function leads to defects in ES cell proliferation correlating with accumulated cells at the G1/S transition of the cell cycle [189]. Further validation would be needed to support this theory.

Interesting to note was the correlation of a transient *GADD45G* upregulation in NCCIT cells with the upregulation of many specific differentiation associated genes according to GO terms. Among differentiation pathways was the development of the neuronal lineage dominant, consistent with the observation that hEC and mEC cells could be driven to the neuronal pathway upon the differentiation stimuli retinoic acid [190,191].

Finally, the role of GADD45G in cell cycle could be strengthened as cell cycle related genes were significantly enriched in all upregulated genes (p-value of 6.28e-07 by using g:Profiler), more precisely those genes connected to the negative regulation or arrest of cell cycle. Additionally interesting to note is the fact that within this group only genes involved in the progression from G1 to the S-phase could be found. A limited G1 phase is characteristic for undifferentiated ESCs, as could be supported before with OCT4 knockdowns in H1 cells [45].

4.8 Differences between human EC and human ES cells

There are still many problems associated with the culture of human ES cells, starting from a time consuming and challenging cell culture, further implying patent problems for some established cell lines and finally encompassing tedious governmental regulations. Human EC cells, on the other hand have none of these disadvantages. Nonetheless they share a number of similarities with hES cells, including a common set of cell surface proteins and a similarity in the global expression profile and are thus in principle useful as reference material for the hESC research [38,154].

However, yet there are obvious differences, as hEC cells are not needed to be grown with the addition of FGF2 on feeder cells. Furthermore they are not growing in colonies like hES cells but more homogenously. These differences might be partly explained by chromosomal aberrations, much more dominant in hEC cells compared to hES cells. And, contributing to their heritage, they express germ cell markers and possess a certain resistance to spontaneous differentiation, compared to hES cells.

In this study, two human EC cell lines, NCCIT and 2102Ep were compared with two human ES cell lines, BG03 and H1. The aim was to identify differentially expressed genes and describe them functionally. Thus, cell adhesion specific genes, enriched significantly and specifically in the hES cell lines investigated were detected. The following genes, identified by this approach, show an important role in development:

SNAI1
COL8A2
CNTN1
ALCAM
NCAM

For example *SNAI1* deletion results in embryonic lethality due to multiple vascular defects [192]. Targeted disruption of the *COL8A2* gene in mice can lead to anterior segment abnormalities in the eye [193]. Compton et al. reported that loss of contactin-1 from the neuromuscular junction where *CNTN1* is expressed, impairs communication or adhesion between nerve and muscle which results in the severe myopathic phenotype. *ALCAM* expression has been identified as a marker for isolating cardiomyocytes from differentiating cultures of hESCs [194]. *NCAM* is a cell adhesion macromolecule, which is known to play a critical role in development of the nervous system [195]. It is tempting to speculate that some of these genes might be important for developmental patterns, missing in EC cells. Moreover specific differences in adhesion genes might be a partial explanation for the different morphological phenotypes between hES and hEC cells. Further validation e.g. by immunofluorescence, showing the knockdown on protein levels are needed to confirm this hypothesis.

4.9 ChIP-seq discovered OCT4 binding sites

For the ChIP-seq the three biological samples used for the ChIP-chip were pooled for each the immunoprecipitated and the control DNA fraction. In total more than 11 million reads were obtained. For peak discovery, the CisGenome software was used, which applies a conditional binomial model to identify enriched regions [138]. Thus, 1908 unique mapped peaks were identified with an FDR of 0.1. Compared to a similar study which used a ChIP-seq approach with the transcription factor STAT1, this was a small amount as they discovered 41,582 peaks for an FDR of 0.001, starting from 15.1 million uniquely mapped reads in interferon-gamma stimulated HeLa S3 cells [119].

Furthermore, concerning the global peak distribution, surprisingly a focus of the discovered peaks around 1 Kbp from the transcription start site (TSS), as had been reported before by using ChIP-chip with OCT4, SOX2 and NANOG [83] could not been confirmed. Only at a resolution of 100 Kbp, the shape of frequencies of peak distances becomes more normally distributed around the TSS. This result is also in contrast to the study of Robertson et al. [119] as they could show that the majority of the peaks were detected from –500 bp to 500 bp in relation to the transcription start site. In contrast, another study investigating the genomic distribution of OCT4 in mouse for the chromosome 19, using CHIP-chip assays reported that most of the binding sites obtained were discovered in intragenic regions (38,9%) but still 7,20% could be mapped to the

5´proximal region (-10000 bp – 0 bp relative to the TSS) [165]. This result could be verified by this study (6,38% mapped to the 5´proximal region and 37,37% mapped to the intragenic region), arguing for a broader genomic location of OCT4 in relation to the TSS, compared to STAT1.

Nonetheless, the functional annotation analysis of peak regions for an FDR of 0.1 contained still main clusters of different development categories, also discovered by ChIP-chip analysis. Additionally cell communication and signaling pathways were significantly enriched, arguing against a randomized, biased peak selection, potentially introduced by amplification cycles. Only Transcription factor regulation groups, detected by the ChIP-chip OCT4 targets, could not be detected in the ChIP-seq targets.

5 Conclusion

In the present study, new binding modes of the transcription factor OCT4 have been postulated and new links to the cell cycle have been established. Relations with cell differentiation processes were uncovered by the putative direct OCT4 target GADD45G. Furthermore, OCT4 binding sites were studied on a global, unbiased level, using ChIP-seq analysis. The studies were done in human embryonal carcinoma (EC) cells (NCCIT) and compared with human embryonal stem (ES) cells (H1 and H9). A common set of OCT4 binding sites between these cell lines has been uncovered, using ChIP-Chip experiments and expression arrays. For the purpose of revealing enriched binding sites, new techniques have been established for a more unbiased target screen.

Furthermore, all published OCT4 ChIP large scale experiments have been used and been connected to microarray datasets, obtained from NCCIT cells and H9 cells. Giving the scientific community access to an online accessible graphical interface, which is updated constantly. The information in this database is related to OCT4 connected and regulated networks increasing the meager understanding of pluripotency in embryonic stem cells and embryonal carcinoma cells, but also differentiation processes operative in different cell systems.

In this era of high-throughput functional genomics and systems biology-driven research, which necessitates large datasets, there is a dire need for data integration platforms. To facilitate this, the datasets, presented in this work have been integrated along with existing related datasets from both human and mouse ES and EC cells to generate an Embryonic Stem Cell Data Base (ESCDb) that allows rapid and convenient access and comparisons between published datasets related to embryonic stem cell biology. This study will aid in increasing the meager understanding of pluripotency in ES, EC, iPS and cancer cells. In this context some indications have been found that hES cells seem to express a different set of cell attachment genes compared to hEC cells. Further validation experiments on protein level would be needed to validate this finding.

Applying OCT4 ChIP-seq, for the first time a global view of OCT4 binding sites was obtained. In this study we noticed that the majority of OCT4 binding sites were scattered in the intergenic region rather than focusing at the TSS. This might be one reason why the overlap to existing ChIP-chip studies was relatively small.

Finally, this study focused on the DNA level of the transcriptional regulation. Further studies on the protein level, which include possible post-translational mechanisms and determine in more detail the relationship between the protein levels of OCT4, SOX2 and other possible direct interacting factors of OCT4 and the variability of binding modes are required for a more accurate prediction by which mechanisms OCT4 regulates the cell fate of an undifferentiated pluripotent cell.

References

1. Hogan BL (1999) Morphogenesis. Cell 96: 225-233.
2. Pelton TA, Bettess MD, Lake J, Rathjen J, Rathjen PD (1998) Developmental complexity of early mammalian pluripotent cell populations in vivo and in vitro. Reprod Fertil Dev 10: 535-549.
3. Gilbert SF (2000) Developmental biology. (Sunderland, MA: Sinauer Associates).
4. Odorico JS, Kaufman DS, Thomson JA (2001) Multilineage differentiation from human embryonic stem cell lines. Stem Cells 19: 193-204.
5. Johnson MH, Maro B, Takeichi M (1986) The role of cell adhesion in the synchronization and orientation of polarization in 8-cell mouse blastomeres. J Embryol Exp Morphol 93: 239-255.
6. Fong CY, Bongso A, Ng SC, Kumar J, Trounson A, et al. (1998) Blastocyst transfer after enzymatic treatment of the zona pellucida: improving in-vitro fertilization and understanding implantation. Hum Reprod 13: 2926-2932.
7. Beddington RS, Robertson EJ (1999) Axis development and early asymmetry in mammals. Cell 96: 195-209.
8. Jones JM, Thomson JA (2000) Human embryonic stem cell technology. Semin Reprod Med 18: 219-223.
9. Nichols J, Zevnik B, Anastassiadis K, Niwa H, Klewe-Nebenius D, et al. (1998) Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. Cell 95: 379-391.
10. Tanaka S, Kunath T, Hadjantonakis AK, Nagy A, Rossant J (1998) Promotion of trophoblast stem cell proliferation by FGF4. Science 282: 2072-2075.
11. Kunath T, Strumpf, D., Rossant, J., and Tanaka, S., Marshak, D.R., Gardner, D.K., and Gottlieb, D. (2001) Trophoblast stem cells. eds Cold Spring Harbor Laboratory Press, : 267–288.
12. Damjanov I (1993) Pathogenesis of testicular germ cell tumours. Eur Urol 23: 2-5; discussion 6-7.
13. Damjanov I (1993) Teratocarcinoma: neoplastic lessons about normal embryogenesis. Int J Dev Biol 37: 39-46.
14. Stevens LC (1967) The biology of teratomas. Adv Morphog 6: 1-31.
15. Mostofi FK SI (1998) WHO International histological classification of tumours: histological typing of testis tumours. 2nd ed Berlin: Springer-Verlag.
16. Assou S, Le Carrour T, Tondeur S, Strom S, Gabelle A, et al. (2007) A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. Stem Cells 25: 961-973.
17. Illmensee K, Mintz B (1976) Totipotency and normal differentiation of single teratocarcinoma cells cloned by injection into blastocysts. Proc Natl Acad Sci U S A 73: 549-553.
18. Mintz B, Illmensee K (1975) Normal genetically mosaic mice produced from malignant teratocarcinoma cells. Proc Natl Acad Sci U S A 72: 3585-3589.
19. Andrews PW, Bronson DL, Benham F, Strickland S, Knowles BB (1980) A comparative study of eight cell lines derived from human testicular teratocarcinoma. Int J Cancer 26: 269-280.
20. Wang N, Trend B, Bronson DL, Fraley EE (1980) Nonrandom abnormalities in chromosome 1 in human testicular cancers. Cancer Res 40: 796-802.
21. Andrews PW, Casper J, Damjanov I, Duggan-Keen M, Giwercman A, et al. (1996) Comparative analysis of cell surface antigens expressed by cell lines derived from human germ cell tumours. Int J Cancer 66: 806-816.

22. Andrews PW, Damjanov I, Berends J, Kumpf S, Zappavigna V, et al. (1994) Inhibition of proliferation and induction of differentiation of pluripotent human embryonal carcinoma cells by osteogenic protein-1 (or bone morphogenetic protein-7). Lab Invest 71: 243-251.
23. Andrews PW (1984) Retinoic acid induces neuronal differentiation of a cloned human embryonal carcinoma cell line in vitro. Dev Biol 103: 285-293.
24. Damjanov I, Horvat B, Gibas Z (1993) Retinoic acid-induced differentiation of the developmentally pluripotent human germ cell tumor-derived cell line, NCCIT. Lab Invest 68: 220-232.
25. Henderson JK, Draper JS, Baillie HS, Fishel S, Thomson JA, et al. (2002) Preimplantation human embryos and embryonic stem cells show comparable expression of stage-specific embryonic antigens. Stem Cells 20: 329-337.
26. Andrews PW, Matin MM, Bahrami AR, Damjanov I, Gokhale P, et al. (2005) Embryonic stem (ES) cells and embryonal carcinoma (EC) cells: opposite sides of the same coin. Biochem Soc Trans 33: 1526-1530.
27. Andrews PW (2006) The selfish stem cell. Nat Biotechnol 24: 325-326.
28. Draper JS, Smith K, Gokhale P, Moore HD, Maltby E, et al. (2004) Recurrent gain of chromosomes 17q and 12 in cultured human embryonic stem cells. Nat Biotechnol 22: 53-54.
29. Sperger JM, Chen X, Draper JS, Antosiewicz JE, Chon CH, et al. (2003) Gene expression patterns in human embryonic stem cells and human pluripotent germ cell tumors. Proc Natl Acad Sci U S A 100: 13350-13355.
30. Evans MJ, Kaufman MH (1981) Establishment in culture of pluripotential cells from mouse embryos. Nature 292: 154-156.
31. Martin GR (1981) Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proc Natl Acad Sci U S A 78: 7634-7638.
32. Robertson EJ, Kaufman, M.H., Bradley, A.,Evans, M. (1983) Isolation, Properties and Karyotype Analysis of Pluripotent (EK) Cell Lines from Normal and Parthenogenetic Embryos. Cold Spring Harb Conf Cell Prolif 10: 647-663.
33. Thomson JA, Kalishman J, Golos TG, Durning M, Harris CP, et al. (1995) Isolation of a primate embryonic stem cell line. Proc Natl Acad Sci U S A 92: 7844-7848.
34. Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, et al. (1998) Embryonic stem cell lines derived from human blastocysts. Science 282: 1145-1147.
35. Reubinoff BE, Pera MF, Fong CY, Trounson A, Bongso A (2000) Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. Nat Biotechnol 18: 399-404.
36. Assady S, Maor G, Amit M, Itskovitz-Eldor J, Skorecki KL, et al. (2001) Insulin production by human embryonic stem cells. Diabetes 50: 1691-1697.
37. Amit M, Carpenter MK, Inokuma MS, Chiu CP, Harris CP, et al. (2000) Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture. Dev Biol 227: 271-278.
38. Greber B, Lehrach H, Adjaye J (2007) Silencing of core transcription factors in human EC cells highlights the importance of autocrine FGF signaling for self-renewal. BMC Dev Biol 7: 46.
39. Scholer HR (1991) Octamania: the POU factors in murine development. Trends Genet 7: 323-329.
40. Scholer HR, Ruppert S, Suzuki N, Chowdhury K, Gruss P (1990) New type of POU domain in germ line-specific protein Oct-4. Nature 344: 435-439.

41. Adjaye J, Huntriss J, Herwig R, BenKahla A, Brink TC, et al. (2005) Primary differentiation in the human blastocyst: comparative molecular portraits of inner cell mass and trophectoderm cells. Stem Cells 23: 1514-1525.
42. Niwa H (2001) Molecular mechanism to maintain stem cell renewal of ES cells. Cell Struct Funct 26: 137-148.
43. Niwa H, Miyazaki J, Smith AG (2000) Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nat Genet 24: 372-376.
44. Pesce M, Scholer HR (2001) Oct-4: Gatekeeper in the beginnings of mammalian development. Stem Cells 19: 271-278.
45. Babaie Y, Herwig R, Greber B, Brink TC, Wruck W, et al. (2007) Analysis of oct4-dependent transcriptional networks regulating self-renewal and pluripotency in human embryonic stem cells. Stem Cells 25: 500-510.
46. Cheng L (2004) Establishing a germ cell origin for metastatic tumors using OCT4 immunohistochemistry. Cancer 101: 2006-2010.
47. Cheng L, Thomas A, Roth LM, Zheng W, Michael H, et al. (2004) OCT4: a novel biomarker for dysgerminoma of the ovary. Am J Surg Pathol 28: 1341-1346.
48. Clark AT, Bodnar MS, Fox M, Rodriquez RT, Abeyta MJ, et al. (2004) Spontaneous differentiation of germ cells from human embryonic stem cells in vitro. Hum Mol Genet 13: 727-739.
49. Gidekel S, Pizov G, Bergman Y, Pikarsky E (2003) Oct-3/4 is a dose-dependent oncogenic fate determinant. Cancer Cell 4: 361-370.
50. Jones TD, Ulbright TM, Eble JN, Baldridge LA, Cheng L (2004) OCT4 staining in testicular tumors: a sensitive and specific marker for seminoma and embryonal carcinoma. Am J Surg Pathol 28: 935-940.
51. Jones TD, Ulbright TM, Eble JN, Cheng L (2004) OCT4: A sensitive and specific biomarker for intratubular germ cell neoplasia of the testis. Clin Cancer Res 10: 8544-8547.
52. Looijenga LH, Stoop H, de Leeuw HP, de Gouveia Brazao CA, Gillis AJ, et al. (2003) POU5F1 (OCT3/4) identifies cells with pluripotent potential in human germ cell tumors. Cancer Res 63: 2244-2250.
53. Goto T, Adjaye J, Rodeck CH, Monk M (1999) Identification of genes expressed in human primordial germ cells at the time of entry of the female germ line into meiosis. Mol Hum Reprod 5: 851-860.
54. Kehler J, Tolkunova E, Koschorz B, Pesce M, Gentile L, et al. (2004) Oct4 is required for primordial germ cell survival. EMBO Rep 5: 1078-1083.
55. Jin T, Branch DR, Zhang X, Qi S, Youngson B, et al. (1999) Examination of POU homeobox gene expression in human breast cancer cells. Int J Cancer 81: 104-112.
56. Steingart RA, Heldenberg E, Pinhasov A, Brenneman DE, Fridkin M, et al. (2002) A vasoactive intestinal peptide receptor analog alters the expression of homeobox genes. Life Sci 71: 2543-2552.
57. Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, et al. (2008) An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nat Genet 40: 499-507.
58. Schoenhals M, Kassambara A, De Vos J, Hose D, Moreaux J, et al. (2009) Embryonic stem cell markers expression in cancers. Biochem Biophys Res Commun 383: 157-162.

59. Botquin V, Hess H, Fuhrmann G, Anastassiadis C, Gross MK, et al. (1998) New POU dimer configuration mediates antagonistic control of an osteopontin preimplantation enhancer by Oct-4 and Sox-2. Genes Dev 12: 2073-2090.
60. Nishimoto M, Fukushima A, Okuda A, Muramatsu M (1999) 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. Mol Cell Biol 19: 5453-5465.
61. Ambrosetti DC, Basilico C, Dailey L (1997) Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. Mol Cell Biol 17: 6321-6329.
62. Fong H, Hohenstein KA, Donovan PJ (2008) Regulation of self-renewal and pluripotency by Sox2 in human embryonic stem cells. Stem Cells 26: 1931-1938.
63. Li J, Pan G, Cui K, Liu Y, Xu S, et al. (2007) A dominant-negative form of mouse SOX2 induces trophectoderm differentiation and progressive polyploidy in mouse embryonic stem cells. J Biol Chem 282: 19481-19492.
64. Chew JL, Loh YH, Zhang W, Chen X, Tam WL, et al. (2005) Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. Mol Cell Biol 25: 6031-6046.
65. Rodda DJ, Chew JL, Lim LH, Loh YH, Wang B, et al. (2005) Transcriptional regulation of nanog by OCT4 and SOX2. J Biol Chem 280: 24731-24737.
66. Card DA, Hebbar PB, Li L, Trotter KW, Komatsu Y, et al. (2008) Oct4/Sox2-regulated miR-302 targets cyclin D1 in human embryonic stem cells. Mol Cell Biol 28: 6426-6438.
67. Becker KA, Ghule PN, Therrien JA, Lian JB, Stein JL, et al. (2006) Self-renewal of human embryonic stem cells is supported by a shortened G1 cell cycle phase. J Cell Physiol 209: 883-893.
68. Fluckiger AC, Marcy G, Marchand M, Negre D, Cosset FL, et al. (2006) Cell cycle features of primate embryonic stem cells. Stem Cells 24: 547-556.
69. Savatier P, Huang S, Szekely L, Wiman KG, Samarut J (1994) Contrasting patterns of retinoblastoma protein expression in mouse embryonic stem cells and embryonic fibroblasts. Oncogene 9: 809-818.
70. Tomilin A, Remenyi A, Lins K, Bak H, Leidel S, et al. (2000) Synergism with the coactivator OBF-1 (OCA-B, BOB-1) is mediated by a specific POU dimer configuration. Cell 103: 853-864.
71. Herr W, Cleary MA (1995) The POU domain: versatility in transcriptional regulation by a flexible two-in-one DNA-binding domain. Genes Dev 9: 1679-1693.
72. Kemler I, Schaffner W (1990) Octamer transcription factors and the cell type-specificity of immunoglobulin gene expression. FASEB J 4: 1444-1449.
73. LeBowitz JH, Clerc RG, Brenowitz M, Sharp PA (1989) The Oct-2 protein binds cooperatively to adjacent octamer sites. Genes Dev 3: 1625-1638.
74. Poellinger L, Roeder RG (1989) Octamer transcription factors 1 and 2 each bind to two different functional elements in the immunoglobulin heavy-chain promoter. Mol Cell Biol 9: 747-756.
75. Remenyi A, Lins K, Nissen LJ, Reinbold R, Scholer HR, et al. (2003) Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. Genes Dev 17: 2048-2059.

76. Saxe JP, Tomilin A, Scholer HR, Plath K, Huang J (2009) Post-translational regulation of Oct4 transcriptional activity. PLoS ONE 4: e4467.

77. Kang J, Gemberling M, Nakamura M, Whitby FG, Handa H, et al. (2009) A general mechanism for transcription regulation by Oct1 and Oct4 in response to genotoxic and oxidative stress. Genes Dev 23: 208-222.

78. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, et al. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298: 799-804.

79. Odoms K, Shanley TP, Wong HR (2004) Short-term modulation of interleukin-1beta signaling by hyperoxia: uncoupling of IkappaB kinase activation and NF-kappaB-dependent gene expression. Am J Physiol Lung Cell Mol Physiol 286: L554-562.

80. Avilion AA, Nicolis SK, Pevny LH, Perez L, Vivian N, et al. (2003) Multipotent cell lineages in early mouse development depend on SOX2 function. Genes Dev 17: 126-140.

81. Chambers I, Colby D, Robertson M, Nichols J, Lee S, et al. (2003) Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell 113: 643-655.

82. Tokuzawa Y, Kaiho E, Maruyama M, Takahashi K, Mitsui K, et al. (2003) Fbx15 is a novel target of Oct3/4 but is dispensable for embryonic stem cell self-renewal and mouse development. Mol Cell Biol 23: 2699-2708.

83. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, et al. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122: 947-956.

84. Boyer LA, Mathur D, Jaenisch R (2006) Molecular control of pluripotency. Curr Opin Genet Dev 16: 455-462.

85. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, et al. (2005) A high-resolution map of active promoters in the human genome. Nature 436: 876-880.

86. Greber B, Lehrach H, Adjaye J (2008) Control of early fate decisions in human ES cells by distinct states of TGFss pathway activity. Stem Cells Dev.

87. Xu RH, Sampsell-Barron TL, Gu F, Root S, Peck RM, et al. (2008) NANOG is a direct target of TGFbeta/activin-mediated SMAD signaling in human ESCs. Cell Stem Cell 3: 196-206.

88. Suzuki A, Raya A, Kawakami Y, Morita M, Matsui T, et al. (2006) Nanog binds to Smad1 and blocks bone morphogenetic protein-induced differentiation of embryonic stem cells. Proc Natl Acad Sci U S A 103: 10294-10299.

89. Smith AG, Heath JK, Donaldson DD, Wong GG, Moreau J, et al. (1988) Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. Nature 336: 688-690.

90. Nordhoff V, Hubner K, Bauer A, Orlova I, Malapetsa A, et al. (2001) Comparative analysis of human, bovine, and murine Oct-4 upstream promoter sequences. Mamm Genome 12: 309-317.

91. Pesce M, Scholer HR (2000) Oct-4: control of totipotency and germline determination. Mol Reprod Dev 55: 452-457.

92. Summerton JE (2007) Morpholino, siRNA, and S-DNA compared: impact of structure and mechanism of action on off-target effects and sequence specificity. Curr Top Med Chem 7: 651-660.

93. Summerton J, Weller D (1997) Morpholino antisense oligomers: design, preparation, and properties. Antisense Nucleic Acid Drug Dev 7: 187-195.

94. Capecchi MR (1989) The new mouse genetics: altering the genome by gene targeting. Trends Genet 5: 70-76.

95. Fire AZ (2007) Gene silencing by double-stranded RNA (Nobel Lecture). Angew Chem Int Ed Engl 46: 6966-6984.

96. Mello CC (2007) Return to the RNAi world: rethinking gene expression and evolution (Nobel Lecture). Angew Chem Int Ed Engl 46: 6985-6994.

97. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, et al. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391: 806-811.

98. Montgomery MK, Xu S, Fire A (1998) RNA as a target of double-stranded RNA-mediated genetic interference in Caenorhabditis elegans. Proc Natl Acad Sci U S A 95: 15502-15507.

99. Echeverri CJ, Beachy PA, Baum B, Boutros M, Buchholz F, et al. (2006) Minimizing the risk of reporting false positives in large-scale RNAi screens. Nat Methods 3: 777-779.

100. Hutvagner G, Zamore PD (2002) RNAi: nature abhors a double-strand. Curr Opin Genet Dev 12: 225-232.

101. Sharp PA (2001) RNA interference--2001. Genes Dev 15: 485-490.

102. Waterhouse PM, Wang MB, Lough T (2001) Gene silencing as an adaptive defence against viruses. Nature 411: 834-842.

103. Bernstein E, Caudy AA, Hammond SM, Hannon GJ (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature 409: 363-366.

104. Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, et al. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature 411: 494-498.

105. Ketting RF, Fischer SE, Bernstein E, Sijen T, Hannon GJ, et al. (2001) Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. Genes Dev 15: 2654-2659.

106. Chendrimada TP, Gregory RI, Kumaraswamy E, Norman J, Cooch N, et al. (2005) TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. Nature 436: 740-744.

107. Haase AD, Jaskiewicz L, Zhang H, Laine S, Sack R, et al. (2005) TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing. EMBO Rep 6: 961-967.

108. Carmell MA, Xuan Z, Zhang MQ, Hannon GJ (2002) The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. Genes Dev 16: 2733-2742.

109. Chiu YL, Rana TM (2002) RNAi in human cells: basic structural and functional features of small interfering RNA. Mol Cell 10: 549-561.

110. Kim K, Lee YS, Carthew RW (2007) Conversion of pre-RISC to holo-RISC by Ago2 during assembly of RNAi complexes. Rna 13: 22-29.

111. Orlando V (2000) Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. Trends Biochem Sci 25: 99-104.

112. Kurdistani SK, Grunstein M (2003) In vivo protein-protein and protein-DNA crosslinking for genomewide binding microarray. Methods 31: 90-95.

113. Lieb JD, Liu X, Botstein D, Brown PO (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet 28: 327-334.

114. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, et al. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409: 533-538.

115. Bohlander SK, Espinosa R, 3rd, Le Beau MM, Rowley JD, Diaz MO (1992) A method for the rapid sequence-independent amplification of microdissected chromosomal material. Genomics 13: 1322-1324.
116. Mueller PR, Wold B (1989) In vivo footprinting of a muscle specific enhancer by ligation mediated PCR. Science 246: 780-786.
117. Liu CL, Schreiber SL, Bernstein BE (2003) Development and validation of a T7 based linear amplification for genomic DNA. BMC Genomics 4: 19.
118. Kato M, Hata N, Banerjee N, Futcher B, Zhang MQ (2004) Identifying combinatorial regulation of transcription factors and binding motifs. Genome Biol 5: R56.
119. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4: 651-657.
120. Bentley DR (2006) Whole-genome re-sequencing. Curr Opin Genet Dev 16: 545-552.
121. Brierley MM, Fish EN (2005) Stats: multifaceted regulators of transcription. J Interferon Cytokine Res 25: 733-744.
122. Schroder K, Hertzog PJ, Ravasi T, Hume DA (2004) Interferon-gamma: an overview of signals, mechanisms and functions. J Leukoc Biol 75: 163-189.
123. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497-1502.
124. Buck MJ, Nobel AB, Lieb JD (2005) ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. Genome Biol 6: R97.
125. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, et al. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell 116: 499-509.
126. Gottardo R, Li W, Johnson WE, Liu XS (2008) A flexible and powerful bayesian hierarchical model for ChIP-Chip experiments. Biometrics 64: 468-478.
127. Keles S, van der Laan MJ, Dudoit S, Cawley SE (2006) Multiple testing methods for ChIP-Chip high density oligonucleotide array data. J Comput Biol 13: 579-613.
128. Hochberg Y, Benjamini Y (1990) More powerful procedures for multiple significance testing. Stat Med 9: 811-818.
129. Li W, Meyer CA, Liu XS (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. Bioinformatics 21 Suppl 1: i274-282.
130. Ji H, Wong WH (2005) TileMap: create chromosomal map of tiling array hybridizations. Bioinformatics 21: 3629-3636.
131. Efron B, Tibshirani R (2002) Empirical bayes methods and false discovery rates for microarrays. Genet Epidemiol 23: 70-86.
132. Zheng M, Barrera LO, Ren B, Wu YN (2007) ChIP-chip: data, model, and analysis. Biometrics 63: 787-796.
133. Sun W, Buck MJ, Patel M, Davis IJ (2009) Improved ChIP-chip analysis by a mixture model approach. BMC Bioinformatics 10: 173.
134. Song JS, Johnson WE, Zhu X, Zhang X, Li W, et al. (2007) Model-based analysis of two-color arrays (MA2C). Genome Biol 8: R178.
135. Bieda M, Xu X, Singer MA, Green R, Farnham PJ (2006) Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. Genome Res 16: 595-605.

136. Sandve GK, Drablos F (2006) A survey of motif discovery methods in an integrated framework. Biol Direct 1: 11.

137. Jin VX, O'Geen H, Iyengar S, Green R, Farnham PJ (2007) Identification of an OCT4 and SRY regulatory module using integrated computational and experimental genomics approaches. Genome Res 17: 807-817.

138. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 26: 1293-1300.

139. Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, et al. (2006) Dissecting self-renewal in stem cells with RNA interference. Nature 442: 533-538.

140. Kuhn K, Baker SC, Chudin E, Lieu MH, Oeser S, et al. (2004) A novel, high-performance random array platform for quantitative gene expression profiling. Genome Res 14: 2347-2356.

141. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

142. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, et al. (2009) The UCSC Genome Browser Database: update 2009. Nucleic Acids Res 37: D755-761.

143. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851-1858.

144. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374-378.

145. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4: P3.

146. Johnson DS, Li W, Gordon DB, Bhattacharjee A, Curry B, et al. (2008) Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. Genome Res 18: 393-403.

147. Reimand J, Kull M, Peterson H, Hansen J, Vilo J (2007) g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res 35: W193-200.

148. Houldsworth J, Heath SC, Bosl GJ, Studer L, Chaganti RS (2002) Expression profiling of lineage differentiation in pluripotential human embryonal carcinoma cells. Cell Growth Differ 13: 257-264.

149. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37: D412-416.

150. Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, et al. (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet 38: 431-440.

151. Stegmeier F, Rape M, Draviam VM, Nalepa G, Sowa ME, et al. (2007) Anaphase initiation is regulated by antagonistic ubiquitination and deubiquitination activities. Nature 446: 876-881.

152. Sun L, Gong R, Wan B, Huang X, Wu C, et al. (2003) GADD45gamma, down-regulated in 65% hepatocellular carcinoma (HCC) from 23 chinese patients, inhibits cell growth and induces cell cycle G2/M arrest for hepatoma Hep-G2 cell lines. Mol Biol Rep 30: 249-253.

153. Ying J, Srivastava G, Hsieh WS, Gao Z, Murray P, et al. (2005) The stress-responsive gene GADD45G is a functional tumor suppressor, with its response

to environmental stresses frequently disrupted epigenetically in multiple tumors. Clin Cancer Res 11: 6442-6449.

154. Josephson R, Ording CJ, Liu Y, Shin S, Lakshmipathy U, et al. (2007) Qualification of embryonal carcinoma 2102Ep as a reference for human embryonic stem cell research. Stem Cells 25: 437-446.

155. Roider HG, Manke T, O'Keeffe S, Vingron M, Haas SA (2009) PASTAA: identifying transcription factors associated with sets of co-regulated genes. Bioinformatics 25: 435-442.

156. Viens A, Mechold U, Lehrmann H, Harel-Bellan A, Ogryzko V (2004) Use of protein biotinylation in vivo for chromatin immunoprecipitation. Anal Biochem 325: 68-76.

157. Jiang H, Daniels PJ, Andrews GK (2003) Putative zinc-sensing zinc fingers of metal-response element-binding transcription factor-1 stabilize a metal-dependent chromatin complex on the endogenous metallothionein-I promoter. J Biol Chem 278: 30394-30402.

158. Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, et al. (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods 24: 218-229.

159. Metz CE (1978) Basic principles of ROC analysis. Semin Nucl Med 8: 283-298.

160. Chen YH, Ishii M, Sucov HM, Maxson RE, Jr. (2008) Msx1 and Msx2 are required for endothelial-mesenchymal transformation of the atrioventricular cushions and patterning of the atrioventricular myocardium. BMC Dev Biol 8: 75.

161. Revet I, Huizenga G, Chan A, Koster J, Volckmann R, et al. (2008) The MSX1 homeobox transcription factor is a downstream target of PHOX2B and activates the Delta-Notch pathway in neuroblastoma. Exp Cell Res 314: 707-719.

162. Li H, Heilbronn LK, Hu D, Poynten AM, Blackburn MA, et al. (2008) Islet-1: a potentially important role for an islet cell gene in visceral fat. Obesity (Silver Spring) 16: 356-362.

163. Watkins NA, Gusnanto A, de Bono B, De S, Miranda-Saavedra D, et al. (2009) A HaemAtlas: characterizing gene expression in differentiated human blood cells. Blood 113: e1-9.

164. Chen X, Xu H, Yuan P, Fang F, Huss M, et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133: 1106-1117.

165. Mathur D, Danford TW, Boyer LA, Young RA, Gifford DK, et al. (2008) Analysis of the mouse embryonic stem cell regulatory networks obtained by ChIP-chip and ChIP-PET. Genome Biol 9: R126.

166. Wang JL, Rao S, Chu JL, Shen XH, Levasseur DN, et al. (2006) A protein interaction network for pluripotency of embryonic stem cells. Nature 444: 364-368.

167. Kuroda T, Tada M, Kubota H, Kimura H, Hatano SY, et al. (2005) Octamer and Sox elements are required for transcriptional cis regulation of Nanog gene expression. Mol Cell Biol 25: 2475-2485.

168. Nishimoto M, Miyagi S, Yamagishi T, Sakaguchi T, Niwa H, et al. (2005) Oct-3/4 maintains the proliferative embryonic stem cell state via specific binding to a variant octamer sequence in the regulatory region of the UTF1 locus. Mol Cell Biol 25: 5084-5094.

169. Tomioka M, Nishimoto M, Miyagi S, Katayanagi T, Fukui N, et al. (2002) Identification of Sox-2 regulatory region which is under the control of Oct-3/4-Sox-2 complex. Nucleic Acids Res 30: 3202-3213.

170. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34: 166-176.

171. Cole MF, Johnstone SE, Newman JJ, Kagey MH, Young RA (2008) Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. Genes Dev 22: 746-755.

172. Jorgensen HF, Terry A, Beretta C, Pereira CF, Leleu M, et al. (2009) REST selectively represses a subset of RE1-containing neuronal genes in mouse embryonic stem cells. Development 136: 715-721.

173. Pahl HL (1999) Activators and target genes of Rel/NF-kappaB transcription factors. Oncogene 18: 6853-6866.

174. Lin T, Chao C, Saito S, Mazur SJ, Murphy ME, et al. (2005) p53 induces differentiation of mouse embryonic stem cells by suppressing Nanog expression. Nat Cell Biol 7: 165-171.

175. Ramji DP, Tadros MH, Hardon EM, Cortese R (1991) The transcription factor LF-A1 interacts with a bipartite recognition sequence in the promoter regions of several liver-specific genes. Nucleic Acids Res 19: 1139-1146.

176. Garcia-Dominguez M, Poquet C, Garel S, Charnay P (2003) Ebf gene function is required for coupling neuronal differentiation and cell cycle exit. Development 130: 6013-6025.

177. Cotta CV, Zhang Z, Kim HG, Klug CA (2003) Pax5 determines B- versus T-cell fate and does not block early myeloid-lineage development. Blood 101: 4342-4346.

178. Ben-Shushan E, Sharir H, Pikarsky E, Bergman Y (1995) A dynamic balance between ARP-1/COUP-TFII, EAR-3/COUP-TFI, and retinoic acid receptor:retinoid X receptor heterodimers regulates Oct-3/4 expression in embryonal carcinoma cells. Mol Cell Biol 15: 1034-1048.

179. Gordan R, Hartemink AJ, Bulyk ML (2009) Distinguishing direct versus indirect transcription factor-DNA interactions. Genome Res.

180. Tantin D, Gemberling M, Callister C, Fairbrother W (2008) High-throughput Biochemical Analysis of in-vivo Location Data Reveals Novel Classes of POU5F1(Oct4)/DNA complexes. Genome Res.

181. Zerbini LF, Czibere A, Wang Y, Correa RG, Otu H, et al. (2006) A novel pathway involving melanoma differentiation associated gene-7/interleukin-24 mediates nonsteroidal anti-inflammatory drug-induced apoptosis and growth arrest of cancer cells. Cancer Res 66: 11922-11931.

182. Oki T, Sowa Y, Hirose T, Takagaki N, Horinaka M, et al. (2004) Genistein induces Gadd45 gene and G2/M cell cycle arrest in the DU145 human prostate cancer cell line. FEBS Lett 577: 55-59.

183. Xu H, Wang W, Li C, Yu H, Yang A, et al. (2009) WWP2 promotes degradation of transcription factor OCT4 in human embryonic stem cells. Cell Res 19: 561-573.

184. Sharov AA, Masui S, Sharova LV, Piao Y, Aiba K, et al. (2008) Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. BMC Genomics 9: 269.

185. Regenbrecht CR, Jung M, Lehrach H, Adjaye J (2008) The molecular basis of genistein-induced mitotic arrest and exit of self-renewal in embryonal carcinoma and primary cancer cell lines. BMC Med Genomics 1: 49.
186. Taylor WR, Stark GR (2001) Regulation of the G2/M transition by p53. Oncogene 20: 1803-1815.
187. Campanero MR, Herrero A, Calvo V (2008) The histone deacetylase inhibitor trichostatin A induces GADD45 gamma expression via Oct and NF-Y binding sites. Oncogene 27: 1263-1272.
188. Smith AE, Ford KG (2005) Use of altered-specificity binding Oct-4 suggests an absence of pluripotent cell-specific cofactor usage. Nucleic Acids Res 33: 6011-6023.
189. Grskovic M, Chaivorapol C, Gaspar-Maia A, Li H, Ramalho-Santos M (2007) Systematic identification of cis-regulatory sequences active in mouse and human embryonic stem cells. PLoS Genet 3: e145.
190. Serra M, Leite SB, Brito C, Costa J, Carrondo MJ, et al. (2007) Novel culture strategy for human stem cell proliferation and neuronal differentiation. J Neurosci Res 85: 3557-3566.
191. Xia C, Wang C, Zhang K, Qian C, Jing N (2007) Induction of a high population of neural stem cells with anterior neuroectoderm characters from epiblast-like P19 embryonic carcinoma cells. Differentiation 75: 912-927.
192. Lomeli H, Starling C, Gridley T (2009) Epiblast-specific Snai1 deletion results in embryonic lethality due to multiple vascular defects. BMC Res Notes 2: 22.
193. Hopfer U, Fukai N, Hopfer H, Wolf G, Joyce N, et al. (2005) Targeted disruption of Col8a1 and Col8a2 genes in mice leads to anterior segment abnormalities in the eye. Faseb J 19: 1232-1244.
194. Rust W, Balakrishnan T, Zweigerdt R (2009) Cardiomyocyte enrichment from human embryonic stem cell cultures by selection of ALCAM surface expression. Regen Med 4: 225-237.
195. Bisaz R, Conboy L, Sandi C (2009) Learning under stress: a role for the neural cell adhesion molecule NCAM. Neurobiol Learn Mem 91: 333-342.

Publications

A data integration approach to mapping OCT4 regulated transcriptional networks operative in embryonic stem cells and embryonal carcinoma cells
Marc Jung, Hedi Peterson, Lukas Chavez, Pascal Kahlem, Hans Lehrach, Jaak Vilo and James Adjaye
PLOS ONE, submitted

Full Genome analysis of OCT4 binding sites in human embryonal carcinoma cells using ChIP-Seq
Marc Jung, Lukas Chavez, Holger Klein, Hans Lehrach, Bernd Timmermann, Ralf Herwig, and James Adjaye
In preparation

The molecular basis of phytoestrogene Genistein induced exit of self-renewal and proliferation in embryonal carcinoma and primary cancer cell lines
Regenbrecht CRA*, Jung M*, Lehrach H and Adjaye J
*Authors contributed equally to the study.
BMC Med Genomics, 2008

The origins of human embryonic stem cells: a biological conundrum.
Brink TC, Sudheer S, Janke D, Jagodzinska J, Jung M, Adjaye J.
Cells Tissues Organs. 2008

Embryonale Stammzellen
Selbsterneuerung und Pluripotenz in humanen embryonalen Stammzellen
Greber B, Jung M, Adjaye J
Biospektrum, 2007

Appendix

5.1.1 Solutions, Buffers and Media

Antibiotics (1000x)	50 mg/ml Ampicillin; 30 mg/ml Kanamycin
	34 mg/ml Chloramphenicol
LB medium	Bacto-tryptone 10 g
	Bacto-yeast extract 5g
	NaCl 10 g
	pH adjusted to 7.2; autoclaved
LB agar	
LB medium	15 g/l Bacto agar
	2x YT
	16 g Bacto-tryptone
	10 g Bacto-yeast extract
	5 g NaCl
	H2O added to 1 l; autoclaved
6x DNA loading buffer	0.2% Bromophenol blue
	60% Glycerol
	60 mM EDTA
TE buffer	10 mM Tris-HCl
	1 mM EDTA; pH 8.0
PBS	instamed PBS Dulbecco w/o Mg2+ , Ca2+
Blocking buffer	
1x TBS	
3% BSA	

4x Laemmli buffer 100ml	8g SDS
	40 ml Glycerin
	40 ml 0.6M Tris pH 6.8
	80 mg Bromophenol blue
	H2O added to 80ml
	20 ml β-Mercapto ethanol

Ponceau-S staining solution	0.2% Ponceau
	3% Trichloroacetic acid

Glycerol stocks plasmids clones were prepared in 25% glycerol and stored at –80°C.

5.1.2 Buffers for SDS-PAGE gel electrophoresis

Resolving Buffer

1.5M Tris-HCl pH8.8:	180g Tris base (121g/mol)
	ad 900ml dH_2O
	adjust pH8.8 with 37% HCl (approx. 26ml)
	ad 1000ml dH_2O

Stacking Buffer

0.5M Tris-HCl pH6.8:	60g Tris base (121g/mol)
	ad 900ml dH_2O
	adjust pH6.8 with 37% HCl (approx. 47ml)
	ad 1000ml dH_2O

3x Loading Buffer (Sample Buffer)

3x SDS-PAGE SB:	9.375ml Stacking Buffer
	17.2ml 87% glycerol
	15ml 10% SDS
	some grains Bromophenol blue
	ad 47.5ml dH_2O

	store at RT	
working solution:	950µl (47.5ml) 3x Loading Buffer	
	50µl (2.5ml) beta-Mercapto-ethanol	
	store at –20℃	
10x Running Buffer		
10x SDS-PAGE RB:	250mM Tris base (121g/mol)	→ 30.3g
	1.92M Glycine (75g/mol)	→ 144.1g
	100ml 10% SDS	
	ad 1000ml dH$_2$O	

5.1.3 Buffers for western blotting

10x Transfer buffer:	250mM Tris base (121g/mol)	→ 30.3g
	1.92M Glycine (75g/mol)	→ 144.1g
	ad 1000ml dH$_2$O	
1x Transfer buffer:	100ml 10x Transfer buffer	
	200ml Methanol	
	ad 1000ml dH$_2$O	
1M Tris-HCl pH7.6:	120g Tris base (121g/mol)	
	ad 900ml dH$_2$O	
	adjust pH7.6 with 37% HCl	
	ad 1000ml dH$_2$O	
1x TBS:	8g Sodium Chloride	
	20ml 1M Tris-Hcl pH7.6	
	ad 1000ml dH$_2$O	

1x TBST:	1000 ml 1x TBS 1 ml Tween 20
Blocking solution:	3% milk powder → 1.5g ad 50ml 1x TBST

5.1.4 Cells, Vectors and antibodies

NCCIT	ATCC
HEK293T	MPI, Berlin
pIRES2-IRES-GFP	INVITROGEN, Germany
pLKO.1-puro	SIGMA, Germany
Anti-Oct3/4 (N-19), sc-8628-x	Santa Cruz, USA
Anti-Oct3/4 (H134), sc-9081-x	Santa Cruz, USA
Anti-Sox2 (Y17), sc-17320	Santa Cruz, USA
Anti hNanog, AF1997	R&D SYSTEMS, USA
Rabbit Anti-Goat IgG, 401504	CALBIOCHEM, USA
ECL^{TM} Anti-mouse IgG, NA9310V	GE Healthcare, USA

5.1.5 Equipment and Reagents

5.1.5.1 Equipment

Phase lock Gel (Heavy) 1.5/2ml tubes	Eppendorf, Germany
Neubauer Counting Chamber	Carl Roth, Germany
Dounce homogenizer	Dounce, USA
Branson 250	Branson Ultrasonics, USA
Branson Tip	Branson Ultrasonics, USA
Agarose gel electrophoresis equipment	Amersham, UK
SDS-PAGE gel electrophoresis equipment	Eppendorf, Germany
Nanodrop Spectrophotometer	Nanodrop, USA
Thermocycler PTC100,	MJ Research Inc, USA
Thermomixer	Eppendorf, Germany
ABI Prism 7700	Applied Biosystems, USA
Microscopes	Carl Zeiss AG, Germany

5.1.5.2 Reagents used in Chromatin immunoprecipitation

37% Formaldehyde, mol. biol. grade	Sigma, USA

Glycogen mol. biol. grade	Roche, Germany
Dynabeads Protein G	Invitrogen, USA
Dynabeads Protein A	Invitrogen, USA
PMSF	Sigma, USA
Protease Inhibitor Cocktail, Complete mini EDTA free	Roche, Germany
Proteinase K	Sigma, USA
Triton X-100	Sigma, USA
Glycine	Merck, Germany
DTT	Sigma-Aldrich, USA
PBS	Sigma-Aldrich, USA
Na-desoxycholate	Merck, Germany
NaCl	Merck, Germany
SDS	Sigma-Aldrich, USA
Tris	Sigma-Aldrich, USA
EDTA	Sigma-Aldrich, USA
LiCl	Merck, Germany
RNase	Sigma-Aldrich, USA
Glycogen	Roche, Germany
Proteinase K	Sigma-Aldrich, USA
Chromatography Water	Merck, Germany
Phenol/Chloroform/Isoamylalcohol (25:24:1)	Roth, Germany
Chloroform	Merck, Germany
Ethanol p.a.	Merck, Germany
PMSF	Sigma-Aldrich, USA
HEPES	Sigma-Aldrich, USA
MgCl2	Merck, Germany
KCl	Merck, Germany
NaN3	Merck, Germany
EDTA	Merck, Germany
EGTA	Merck, Germany
Nonidet P-40 IGEPAL	Sigma-Aldrich, USA

5.1.5.3 Reagents for linear DNA amplification

Sequenase buffer	Amersham, UK
BSA mol. boil grade 500µg/ml	Biolabs
DTT 0.1M, RNase free	Promega, USA
dNTPs	MPI Berlin
Taq polymerase	MPI Berlin
Sequenase T7 DNA Polymerase Version 2.0, 13U/µl	Amersham, UK

Wizard SV Gel and PCR clean-up System	Promega, USA
Trishydrochlorid	Merck, Germany
KCl	Merck, Germany
Tween20, nuclease free	Sigma-Aldrich, USA
MgCl2	Merck, Germany

5.1.5.4 Software

ABI PRISM 7900HT Sequence detection System	Applied Biosystems, USA
Primer Express	Applied Biosystems, USA
SDS 2.1 software	Applied Biosystems, USA
BeadStudio 1.0	Illumina, USA
AxioVision 6.4	Zeiss, Germany

5.1.5.5 Other Reagents

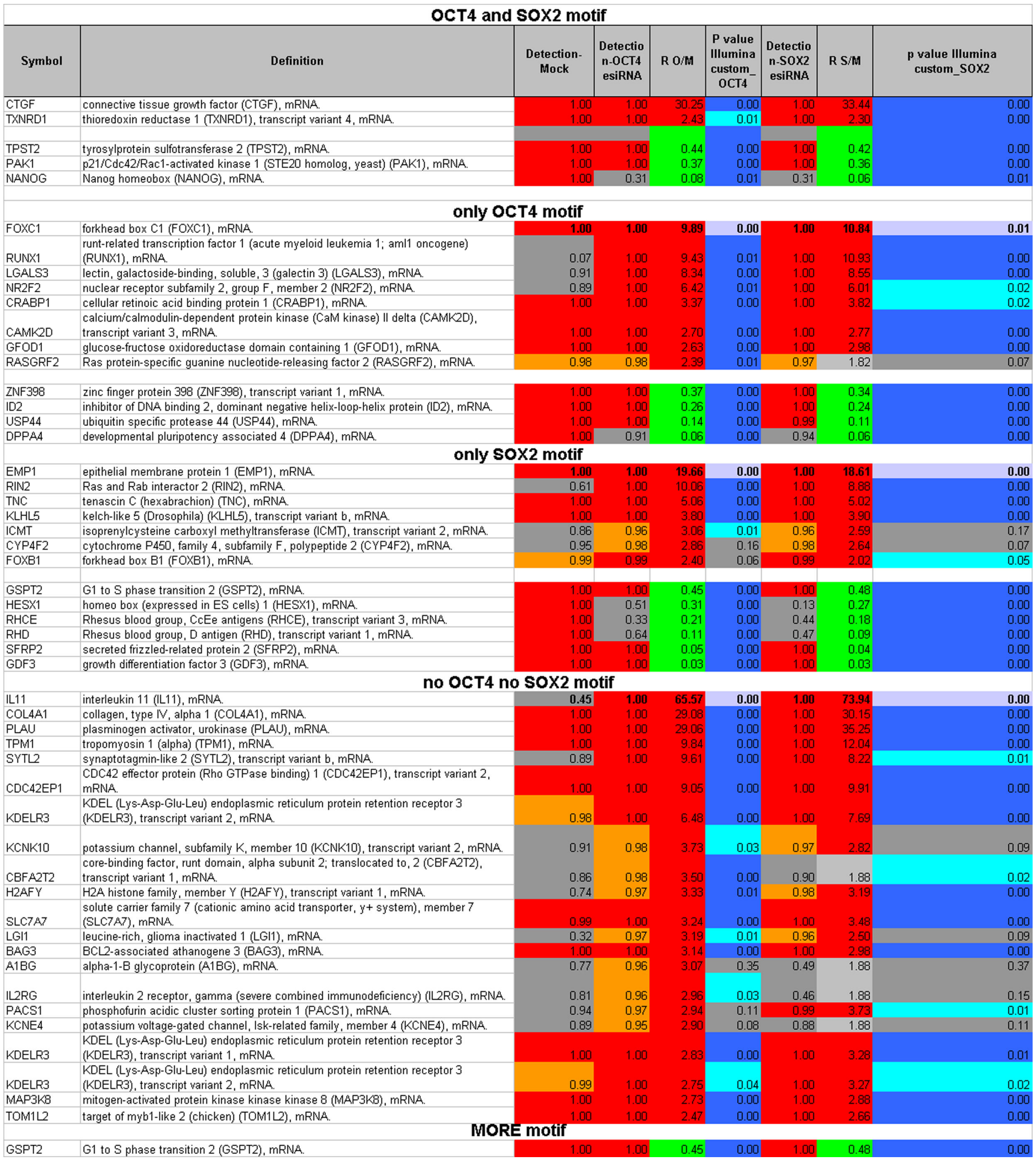
1 kb marker DNA Ladder	New England BioLabs, USA
100 bp marker DNA Ladder	New England BioLabs, USA
30% Hydrogen Peroxyde	Sigma, USA
384 clear well optical reaction plates	Applied Biosystems
Agarose	Bio&Sell, Germany
M-MLV reverse transcriptase	Promega, USA
Benzonase	Roche, Germany
DNase	Promega, USA
dNTPs	Amersham, UK
DTT 0.1M, RNase free	Promega, USA
ECL Advance detection	Amersham, UK
Ethidiumbromide solution	Sigma, USA
Optical adhesive covers	Applied Biosystems
Protein Marker: Precision Plus Protein Standard	Biorad, USA
RNase Away	Roth, Germany
RNasin Ribonuclease inhibitor	Promega, USA
RQ1 RNase-free DNase	Promega, USA
SybrGreen Master Mix	ABgene
Transfer Membrane for Western Blots	Millipore
Trizol	Invitrogen, USA
Vectashield DAPI mounting	Vector Labs, USA
BioMAx XAR film	Kodak, Germany

S

5.1.6 Supplemental material

Term	PValue	Benjamini
GO:0007275~multicellular organismal development	4.38E-08	2.30E-04
GO:0032502~developmental process	6.43E-07	0.001688603
GO:0048856~anatomical structure development	7.75E-06	0.013486922
GO:0032501~multicellular organismal process	3.19E-05	0.041059899
GO:0030154~cell differentiation	3.45E-04	0.202840828
GO:0048869~cellular developmental process	3.45E-04	0.202840828
GO:0045165~cell fate commitment	3.94E-04	0.205685528
GO:0048731~system development	2.44E-04	0.226516822
GO:0009653~anatomical structure morphogenesis	2.97E-04	0.229176342
GO:0003677~DNA binding	3.77E-04	0.237825024
GO:0043565~sequence-specific DNA binding	3.37E-04	0.276469364
GO:0003700~transcription factor activity	2.51E-04	0.303519215
GO:0043283~biopolymer metabolic process	7.57E-04	0.328183354
GO:0030528~transcription regulator activity	1.42E-04	0.336335082
GO:0050794~regulation of cellular process	0.0011072	0.410897733
GO:0009790~embryonic development	0.001289	0.431483768
GO:0050789~regulation of biological process	0.0025237	0.639849939
GO:0032774~RNA biosynthetic process	0.0038292	0.694470743
GO:0006355~regulation of transcription, DNA-dependent	0.0032066	0.70040288
GO:0048513~organ development	0.0034894	0.706050388
GO:0006351~transcription, DNA-dependent	0.0037947	0.713053688
GO:0016070~RNA metabolic process	0.0048135	0.755463947
GO:0045449~regulation of transcription	0.0052502	0.766747244
GO:0019219~regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	0.0062074	0.805196738
GO:0065007~biological regulation	0.0070742	0.816485677
GO:0006350~transcription	0.0068645	0.821533564
GO:0007399~nervous system development	0.0075844	0.824329006
GO:0010468~regulation of gene expression	0.0081532	0.833403711

Table X.0.1 Functional characterization of the core 31 target genes common between NCCIT, H9 and NTERA2 cells, according to GO terms.



Symbol	Definition	Detection-Mock	Detection-OCT4 esiRNA	R O/M	P value Illumina custom_OCT4	Detection-SOX2 esiRNA	R S/M	p value Illumina custom_SOX2
OCT4 and SOX2 motif								
CTGF	connective tissue growth factor (CTGF), mRNA.	1.00	1.00	30.25	0.00	1.00	33.44	0.00
TXNRD1	thioredoxin reductase 1 (TXNRD1), transcript variant 4, mRNA.	1.00	1.00	2.43	0.01	1.00	2.30	0.00
TPST2	tyrosylprotein sulfotransferase 2 (TPST2), mRNA.	1.00	1.00	0.44	0.00	1.00	0.42	0.00
PAK1	p21/Cdc42/Rac1-activated kinase 1 (STE20 homolog, yeast) (PAK1), mRNA.	1.00	1.00	0.37	0.00	1.00	0.36	0.00
NANOG	Nanog homeobox (NANOG), mRNA.	1.00	0.31	0.08	0.01	0.31	0.06	0.01
only OCT4 motif								
FOXC1	forkhead box C1 (FOXC1), mRNA.	**1.00**	**1.00**	**9.89**	**0.00**	**1.00**	**10.84**	**0.01**
RUNX1	runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene) (RUNX1), mRNA.	0.07	1.00	9.43	0.01	1.00	10.93	0.00
LGALS3	lectin, galactoside-binding, soluble, 3 (galectin 3) (LGALS3), mRNA.	0.91	1.00	8.34	0.00	1.00	8.55	0.00
NR2F2	nuclear receptor subfamily 2, group F, member 2 (NR2F2), mRNA.	0.89	1.00	6.42	0.01	1.00	6.01	0.02
CRABP1	cellular retinoic acid binding protein 1 (CRABP1), mRNA.	1.00	1.00	3.37	0.00	1.00	3.82	0.02
CAMK2D	calcium/calmodulin-dependent protein kinase (CaM kinase) II delta (CAMK2D), transcript variant 3, mRNA.	1.00	1.00	2.70	0.00	1.00	2.77	0.00
GFOD1	glucose-fructose oxidoreductase domain containing 1 (GFOD1), mRNA.	1.00	1.00	2.63	0.00	1.00	2.98	0.00
RASGRF2	Ras protein-specific guanine nucleotide-releasing factor 2 (RASGRF2), mRNA.	0.98	0.98	2.39	0.01	0.97	1.82	0.07
ZNF398	zinc finger protein 398 (ZNF398), transcript variant 1, mRNA.	1.00	1.00	0.37	0.00	1.00	0.34	0.00
ID2	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein (ID2), mRNA.	1.00	1.00	0.26	0.00	1.00	0.24	0.00
USP44	ubiquitin specific protease 44 (USP44), mRNA.	1.00	1.00	0.14	0.00	0.99	0.11	0.00
DPPA4	developmental pluripotency associated 4 (DPPA4), mRNA.	1.00	0.91	0.06	0.00	0.94	0.06	0.00
only SOX2 motif								
EMP1	epithelial membrane protein 1 (EMP1), mRNA.	**1.00**	**1.00**	**19.66**	**0.00**	**1.00**	**18.61**	**0.00**
RIN2	Ras and Rab interactor 2 (RIN2), mRNA.	0.61	1.00	10.06	0.00	1.00	8.88	0.00
TNC	tenascin C (hexabrachion) (TNC), mRNA.	1.00	1.00	5.06	0.00	1.00	5.02	0.00
KLHL5	kelch-like 5 (Drosophila) (KLHL5), transcript variant b, mRNA.	1.00	1.00	3.80	0.00	1.00	3.90	0.00
ICMT	isoprenylcysteine carboxyl methyltransferase (ICMT), transcript variant 2, mRNA.	0.86	0.96	3.06	0.01	0.96	2.59	0.17
CYP4F2	cytochrome P450, family 4, subfamily F, polypeptide 2 (CYP4F2), mRNA.	0.95	0.98	2.86	0.16	0.98	2.64	0.07
FOXB1	forkhead box B1 (FOXB1), mRNA.	0.99	0.99	2.40	0.06	0.99	2.02	0.05
GSPT2	G1 to S phase transition 2 (GSPT2), mRNA.	1.00	1.00	0.45	0.00	1.00	0.48	0.00
HESX1	homeo box (expressed in ES cells) 1 (HESX1), mRNA.	1.00	0.51	0.31	0.00	0.13	0.27	0.00
RHCE	Rhesus blood group, CcEe antigens (RHCE), transcript variant 3, mRNA.	1.00	0.33	0.21	0.00	0.44	0.18	0.00
RHD	Rhesus blood group, D antigen (RHD), transcript variant 1, mRNA.	1.00	0.64	0.11	0.00	0.47	0.09	0.00
SFRP2	secreted frizzled-related protein 2 (SFRP2), mRNA.	1.00	1.00	0.05	0.00	1.00	0.04	0.00
GDF3	growth differentiation factor 3 (GDF3), mRNA.	1.00	1.00	0.03	0.00	1.00	0.03	0.00
no OCT4 no SOX2 motif								
IL11	interleukin 11 (IL11), mRNA.	**0.45**	**1.00**	**65.57**	**0.00**	**1.00**	**73.94**	**0.00**
COL4A1	collagen, type IV, alpha 1 (COL4A1), mRNA.	1.00	1.00	29.08	0.00	1.00	30.15	0.00
PLAU	plasminogen activator, urokinase (PLAU), mRNA.	1.00	1.00	29.06	0.00	1.00	35.25	0.00
TPM1	tropomyosin 1 (alpha) (TPM1), mRNA.	1.00	1.00	9.84	0.00	1.00	12.04	0.00
SYTL2	synaptotagmin-like 2 (SYTL2), transcript variant b, mRNA.	0.89	1.00	9.61	0.00	1.00	8.22	0.01
CDC42EP1	CDC42 effector protein (Rho GTPase binding) 1 (CDC42EP1), transcript variant 2, mRNA.	1.00	1.00	9.05	0.00	1.00	9.91	0.00
KDELR3	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 3 (KDELR3), transcript variant 2, mRNA.	0.98	1.00	6.48	0.00	1.00	7.69	0.00
KCNK10	potassium channel, subfamily K, member 10 (KCNK10), transcript variant 2, mRNA.	0.91	0.98	3.73	0.03	0.97	2.82	0.09
CBFA2T2	core-binding factor, runt domain, alpha subunit 2; translocated to, 2 (CBFA2T2), transcript variant 1, mRNA.	0.86	0.98	3.50	0.00	0.90	1.88	0.02
H2AFY	H2A histone family, member Y (H2AFY), transcript variant 1, mRNA.	0.74	0.97	3.33	0.01	0.98	3.19	0.00
SLC7A7	solute carrier family 7 (cationic amino acid transporter, y+ system), member 7 (SLC7A7), mRNA.	0.99	1.00	3.24	0.00	1.00	3.48	0.00
LGI1	leucine-rich, glioma inactivated 1 (LGI1), mRNA.	0.32	0.97	3.19	0.01	0.96	2.50	0.09
BAG3	BCL2-associated athanogene 3 (BAG3), mRNA.	1.00	1.00	3.14	0.00	1.00	2.98	0.00
A1BG	alpha-1-B glycoprotein (A1BG), mRNA.	0.77	0.96	3.07	0.35	0.49	1.88	0.37
IL2RG	interleukin 2 receptor, gamma (severe combined immunodeficiency) (IL2RG), mRNA.	0.81	0.96	2.96	0.03	0.46	1.88	0.15
PACS1	phosphofurin acidic cluster sorting protein 1 (PACS1), mRNA.	0.94	0.97	2.94	0.11	0.99	3.73	0.01
KCNE4	potassium voltage-gated channel, Isk-related family, member 4 (KCNE4), mRNA.	0.89	0.95	2.90	0.08	0.88	1.88	0.11
KDELR3	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 3 (KDELR3), transcript variant 1, mRNA.	1.00	1.00	2.83	0.00	1.00	3.28	0.01
KDELR3	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 3 (KDELR3), transcript variant 2, mRNA.	0.99	1.00	2.75	0.04	1.00	3.27	0.02
MAP3K8	mitogen-activated protein kinase kinase kinase 8 (MAP3K8), mRNA.	1.00	1.00	2.73	0.00	1.00	2.88	0.00
TOM1L2	target of myb1-like 2 (chicken) (TOM1L2), mRNA.	1.00	1.00	2.47	0.00	1.00	2.66	0.00
MORE motif								
GSPT2	G1 to S phase transition 2 (GSPT2), mRNA.	1.00	1.00	0.45	0.00	1.00	0.48	0.00

Table X.0.2 List of differential expressed genes in NCCIT cells upon OCT4 ablation combined with ChIP-chip OCT4 targets in relative to different binding modules

HGNC symbol	Description
GATA6	Transcription factor GATA-6 (GATA-binding factor 6) [Source:UniProtKB/Swiss-Prot;Acc:Q92908]
JAM2	Junctional adhesion molecule B Precursor (JAM-B)(Junctional adhesion molecule 2)(Vascular endothelial junction-associated molecule)(VE-JAM)(CD322 antigen) [Source:UniProtKB/Swiss-Prot;Acc:P57087]
C6orf25	Protein G6b Precursor [Source:UniProtKB/Swiss-Prot;Acc:O95866]
KIAA1328	Uncharacterised protein KIAA1328 [Source:UniProtKB/Swiss-Prot;Acc:Q86T90]
SCT	Secretin Precursor [Source:UniProtKB/Swiss-Prot;Acc:P09683]
IFNAR1	Interferon-alpha/beta receptor alpha chain Precursor (IFN-alpha-REC) [Source:UniProtKB/Swiss-Prot;Acc:P17181]
DOCK5	Dedicator of cytokinesis protein 5 [Source:UniProtKB/Swiss-Prot;Acc:Q9H7D0]
BCL2	Apoptosis regulator Bcl-2 [Source:UniProtKB/Swiss-Prot;Acc:P10415]
SERPINB4	Serpin B4 (Squamous cell carcinoma antigen 2)(SCCA-2)(Leupin) [Source:UniProtKB/Swiss-

	Prot;Acc:P48594]
BCL2L14	Apoptosis facilitator Bcl-2-like protein 14 (Bcl2-L-14)(Apoptosis regulator Bcl-G) [Source:UniProtKB/Swiss-Prot;Acc:Q9BZR8]
C21orf88	Putative uncharacterised protein C21orf88 [Source:UniProtKB/Swiss-Prot;Acc:P59052]
TNFSF11	Tumour necrosis factor ligand superfamily member 11 (Receptor activator of nuclear factor kappa B ligand)(RANKL)(TNF-related activation-induced cytokine)(TRANCE)(Osteoprotegerin ligand)(OPGL)(Osteoclast differentiation factor)(ODF)(CD254 antigen) [Contains Tumour necrosis factor ligand superfamily member 11, membrane form;Tumor necrosis factor ligand superfamily member 11, soluble form] [Source:UniProtKB/Swiss-Prot;Acc:O14788]
CCKBR	Gastrin/cholecystokinin type B receptor (CCK-B receptor)(CCK-BR)(Cholecystokinin-2 receptor)(CCK2-R) [Source:UniProtKB/Swiss-Prot;Acc:P32239]
SLC1A1	Excitatory amino acid transporter 3 (Sodium-dependent glutamate/aspartate transporter 3)(Excitatory amino-acid carrier 1)(Neuronal and epithelial glutamate transporter)(Solute carrier family 1 member 1) [Source:UniProtKB/Swiss-Prot;Acc:P43005]
FOXQ1	Forkhead box protein Q1 (Hepatocyte nuclear factor 3 forkhead homolog 1)(HNF-3/forkhead-like protein 1)(HFH-1) [Source:UniProtKB/Swiss-Prot;Acc:Q9C009]
SQRDL	Sulfide:quinone oxidoreductase, mitochondrial Precursor (EC 1.-.-.-) [Source:UniProtKB/Swiss-Prot;Acc:Q9Y6N5]
LMO7	LIM domain only protein 7 (LOMP)(F-box only protein 20) [Source:UniProtKB/Swiss-Prot;Acc:Q8WWI1]
AFP	Alpha-fetoprotein Precursor (Alpha-1-fetoprotein)(Alpha-fetoglobulin) [Source:UniProtKB/Swiss-Prot;Acc:P02771]
DUSP18	Dual specificity protein phosphatase 18 (EC 3.1.3.48)(EC 3.1.3.16)(Low molecular weight dual specificity phosphatase 20) [Source:UniProtKB/Swiss-Prot;Acc:Q8NEJ0]
FOXA2	Hepatocyte nuclear factor 3-beta (HNF-3-beta)(HNF-3B)(Forkhead box protein A2) [Source:UniProtKB/Swiss-Prot;Acc:Q9Y261]
EDN1	Endothelin-1 Precursor (Preproendothelin-1)(PPET1) [Contains Endothelin-1(ET-1);Big endothelin-1] [Source:UniProtKB/Swiss-Prot;Acc:P05305]
CNTN1	Contactin-1 Precursor (Neural cell surface protein F3)(Glycoprotein gp135) [Source:UniProtKB/Swiss-Prot;Acc:Q12860]
TIMP3	Metalloproteinase inhibitor 3 Precursor (Tissue inhibitor of metalloproteinases 3)(TIMP-3)(Protein MIG-5) [Source:UniProtKB/Swiss-Prot;Acc:P35625]
SPARCL1	SPARC-like protein 1 Precursor (High endothelial venule protein)(Hevin)(MAST 9) [Source:UniProtKB/Swiss-Prot;Acc:Q14515]
HIST1H3G	Histone H3.1 (H3/a)(H3/b)(H3/c)(H3/d)(H3/f)(H3/h)(H3/i)(H3/j)(H3/k)(H3/l) [Source:UniProtKB/Swiss-Prot;Acc:P68431]
LOX	Protein-lysine 6-oxidase Precursor (EC 1.4.3.13)(Lysyl oxidase) [Source:UniProtKB/Swiss-Prot;Acc:P28300]
BTN2A2	Butyrophilin subfamily 2 member A2 Precursor [Source:UniProtKB/Swiss-Prot;Acc:Q8WVV5]
LMO2	Rhombotin-2 (LIM domain only protein 2)(Cysteine-rich protein TTG-2)(T-cell translocation protein 2) [Source:UniProtKB/Swiss-Prot;Acc:P25791]
HS3ST2	Heparan sulfate glucosamine 3-O-sulfotransferase 2 (EC 2.8.2.29)(Heparan sulfate D-glucosaminyl 3-O-sulfotransferase 2)(Heparan sulfate 3-O-sulfotransferase 2)(h3-OST-2) [Source:UniProtKB/Swiss-Prot;Acc:Q9Y278]
UGT8	2-hydroxyacylsphingosine 1-beta-galactosyltransferase Precursor (EC 2.4.1.45)(UDP-galactose-ceramide galactosyltransferase)(Ceramide UDP-galactosyltransferase)(Cerebroside synthase) [Source:UniProtKB/Swiss-Prot;Acc:Q16880]
AMPH	Amphiphysin [Source:UniProtKB/Swiss-Prot;Acc:P49418]
KCNH8	Potassium voltage-gated channel subfamily H member 8 (Voltage-gated potassium channel subunit Kv12.1)(Ether-a-go-go-like potassium channel 3)(ELK channel 3)(ELK3)(ELK1)(hElk1) [Source:UniProtKB/Swiss-Prot;Acc:Q96L42]
MYL7	Myosin regulatory light chain 2, atrial isoform (Myosin light chain 2a)(MLC-2a)(MLC2a)(Myosin regulatory light chain 7) [Source:UniProtKB/Swiss-Prot;Acc:Q01449]
PCDHB13	Protocadherin beta-13 Precursor (PCDH-beta-13) [Source:UniProtKB/Swiss-Prot;Acc:Q9Y5F0]
EGFR	Epidermal growth factor receptor Precursor (EC 2.7.10.1)(Receptor tyrosine-protein kinase ErbB-1) [Source:UniProtKB/Swiss-Prot;Acc:P00533]
SLC15A3	Solute carrier family 15 member 3 (Peptide transporter 3)(Peptide/histidine transporter 2)(Osteoclast transporter) [Source:UniProtKB/Swiss-Prot;Acc:Q8IY34]
STAC	SH3 and cysteine-rich domain-containing protein (SRC homology 3 and cysteine-rich domain protein) [Source:UniProtKB/Swiss-Prot;Acc:Q99469]
NPPB	Natriuretic peptides B Precursor (Gamma-brain natriuretic peptide) [Contains Brain natriuretic peptide 32(BNP-32)(BNP(1-32));BNP(1-30);BNP(1-29);BNP(1-28);BNP(2-31);BNP(3-32);BNP(3-30);BNP(3-29);BNP(4-32);BNP(4-31);BNP(4-30);BNP(4-29);BNP(4-27);BNP(5-32);BNP(5-31);BNP(5-29)] [Source:UniProtKB/Swiss-Prot;Acc:P16860]
CALN1	Calneuron-1 (Calcium-binding protein CaBP8) [Source:UniProtKB/Swiss-Prot;Acc:Q9BXU9]
PTGIS	Prostacyclin synthase (EC 5.3.99.4)(Prostaglandin I2 synthase) [Source:UniProtKB/Swiss-Prot;Acc:Q16647]
SNAI1	Zinc finger protein SNAI1 (Protein snail homolog 1)(Protein sna) [Source:UniProtKB/Swiss-Prot;Acc:O95863]
DOK5	Docking protein 5 (Downstream of tyrosine kinase 5)(Protein dok-5)(IRS6) [Source:UniProtKB/Swiss-Prot;Acc:Q9P104]
RNF128	E3 ubiquitin-protein ligase RNF128 Precursor (EC 6.3.2.-)(RING finger protein 128)(Gene related to anergy in lymphocytes protein) [Source:UniProtKB/Swiss-Prot;Acc:Q8TEB7]
GTPBP5	GTP-binding protein 5 (Protein obg homolog 1)(ObgH1) [Source:UniProtKB/Swiss-Prot;Acc:Q9H4K7]
LIN7A	Lin-7 homolog A (Lin-7A)(hLin-7)(Mammalian lin-seven protein 1)(MALS-1)(Vertebrate lin-7 homolog

	1)(Veli-1 protein)(Tax interaction protein 33)(TIP-33) [Source:UniProtKB/Swiss-Prot;Acc:O14910]
KCNK17	Potassium channel subfamily K member 17 (TWIK-related alkaline pH-activated K(+) channel 2)(2P domain potassium channel Talk-2)(TWIK-related acid-sensitive K(+) channel 4)(TASK-4) [Source:UniProtKB/Swiss-Prot;Acc:Q96T54]
C5	Complement C5 Precursor (C3 and PZP-like alpha-2-macroglobulin domain-containing protein 4) [Contains Complement C5 beta chain;Complement C5 alpha chain;C5a anaphylatoxin;Complement C5 alpha' chain] [Source:UniProtKB/Swiss-Prot;Acc:P01031]
MYL4	Myosin light chain 4 (Myosin light chain 1, embryonic muscle/atrial isoform)(Myosin light chain alkali, GT-1 isoform) [Source:UniProtKB/Swiss-Prot;Acc:P12829]
NRXN1	Neurexin-1-alpha Precursor (Neurexin I-alpha) [Source:UniProtKB/Swiss-Prot;Acc:Q9ULB1]
ANKRD1	Ankyrin repeat domain-containing protein 1 (Cardiac ankyrin repeat protein)(Cytokine-inducible nuclear protein)(C-193) [Source:UniProtKB/Swiss-Prot;Acc:Q15327]
COL1A1	Collagen alpha-1(I) chain Precursor (Alpha-1 type I collagen) [Source:UniProtKB/Swiss-Prot;Acc:P02452]
MCF2	Proto-oncogene DBL (Proto-oncogene MCF-2) [Contains MCF2-transforming protein;DBL-transforming protein] [Source:UniProtKB/Swiss-Prot;Acc:P10911]
MYEOV	Myeloma-overexpressed gene protein (Oncogene in multiple myeloma) [Source:UniProtKB/Swiss-Prot;Acc:Q96EZ4]
ZCWPW1	Zinc finger CW-type PWWP domain protein 1 [Source:UniProtKB/Swiss-Prot;Acc:Q9H0M4]
PLA2G7	Platelet-activating factor acetylhydrolase Precursor (PAF acetylhydrolase)(EC 3.1.1.47)(PAF 2-acylhydrolase)(LDL-associated phospholipase A2)(LDL-PLA(2))(2-acetyl-1-alkylglycerophosphocholine esterase)(1-alkyl-2-acetylglycerophosphocholine esterase) [Source:UniProtKB/Swiss-Prot;Acc:Q13093]
EPO	Erythropoietin Precursor (Epoetin) [Source:UniProtKB/Swiss-Prot;Acc:P01588]
PIP5KL1	Phosphatidylinositol-4-phosphate 5-kinase-like protein 1 (PtdIns(4)P-5-kinase-like protein 1)(PI(4)P 5-kinase-like protein 1)(EC 2.7.1.68) [Source:UniProtKB/Swiss-Prot;Acc:Q5T9C9]
TMLHE	Trimethyllysine dioxygenase, mitochondrial Precursor (EC 1.14.11.8)(Epsilon-trimethyllysine 2-oxoglutarate dioxygenase)(TML-alpha-ketoglutarate dioxygenase)(TML dioxygenase)(TMLD)(TML hydroxylase) [Source:UniProtKB/Swiss-Prot;Acc:Q9NVH6]
COL12A1	Collagen alpha-1(XII) chain Precursor [Source:UniProtKB/Swiss-Prot;Acc:Q99715]
SLMAP	Sarcolemmal membrane-associated protein (Sarcolemmal-associated protein) [Source:UniProtKB/Swiss-Prot;Acc:Q14BN4]
HPD	4-hydroxyphenylpyruvate dioxygenase (EC 1.13.11.27)(4-hydroxyphenylpyruvic acid oxidase)(HPPDase)(4HPPD)(HPD) [Source:UniProtKB/Swiss-Prot;Acc:P32754]
COL8A2	Collagen alpha-2(VIII) chain Precursor (Endothelial collagen) [Source:UniProtKB/Swiss-Prot;Acc:P25067]
PLA2G4C	Cytosolic phospholipase A2 gamma Precursor (cPLA2-gamma)(EC 3.1.1.4)(Phospholipase A2 group IVC) [Source:UniProtKB/Swiss-Prot;Acc:Q9UP65]
C6orf163	Uncharacterised protein C6orf163 [Source:UniProtKB/Swiss-Prot;Acc:Q5TEZ5]
GUCY1A2	Guanylate cyclase soluble subunit alpha-2 (GCS-alpha-2)(EC 4.6.1.2) [Source:UniProtKB/Swiss-Prot;Acc:P33402]
MANEA	Glycoprotein endo-alpha-1,2-mannosidase (Endo-alpha mannosidase)(Endomannosidase)(hEndo)(EC 3.2.1.130)(Mandaselin) [Source:UniProtKB/Swiss-Prot;Acc:Q5SRI9]
HACE1	E3 ubiquitin-protein ligase HACE1 (EC 6.3.2.-)(HECT domain and ankyrin repeat-containing E3 ubiquitin-protein ligase 1) [Source:UniProtKB/Swiss-Prot;Acc:Q8IYU2]
BVES	Blood vessel epicardial substance (hBVES)(Popeye domain-containing protein 1)(Popeye protein 1) [Source:UniProtKB/Swiss-Prot;Acc:Q8NE79]
NCAM1	Neural cell adhesion molecule 1 Precursor (NCAM-1)(N-CAM-1)(CD56 antigen) [Source:UniProtKB/Swiss-Prot;Acc:P13591]
ALCAM	CD166 antigen Precursor (Activated leukocyte cell adhesion molecule)(CD166 antigen) [Source:UniProtKB/Swiss-Prot;Acc:Q13740]
MOXD1	DBH-like monooxygenase protein 1 Precursor (EC 1.14.17.-)(Monooxygenase X) [Source:UniProtKB/Swiss-Prot;Acc:Q6UVY6]
MAP3K5	Mitogen-activated protein kinase kinase kinase 5 (EC 2.7.11.25)(MAPK/ERK kinase kinase 5)(MEK kinase 5)(MEKK 5)(Apoptosis signal-regulating kinase 1)(ASK-1) [Source:UniProtKB/Swiss-Prot;Acc:Q99683]
MGLL	Monoglyceride lipase (MGL)(EC 3.1.1.23)(Lysophospholipase homolog)(Lysophospholipase-like)(HU-K5) [Source:UniProtKB/Swiss-Prot;Acc:Q99685]
CPNE4	Copine-4 (Copine IV)(Copine-8) [Source:UniProtKB/Swiss-Prot;Acc:Q96A23]
SLCO2A1	Solute carrier organic anion transporter family member 2A1 (Solute carrier family 21 member 2)(Prostaglandin transporter)(PGT) [Source:UniProtKB/Swiss-Prot;Acc:Q92959]
EPHB1	Ephrin type-B receptor 1 Precursor (EC 2.7.10.1)(Tyrosine-protein kinase receptor EPH-2)(NET)(HEK6)(ELK) [Source:UniProtKB/Swiss-Prot;Acc:P54762]
SYNJ2	Synaptojanin-2 (EC 3.1.3.36)(Synaptic inositol-1,4,5-trisphosphate 5-phosphatase 2) [Source:UniProtKB/Swiss-Prot;Acc:O15056]
UNC93A	Protein unc-93 homolog A (Protein UNC-93A)(HmUNC-93A) [Source:UniProtKB/Swiss-Prot;Acc:Q86WB7]
MCOLN2	Mucolipin-2 [Source:UniProtKB/Swiss-Prot;Acc:Q8IZK6]
CYR61	Protein CYR61 Precursor (Cysteine-rich angiogenic inducer 61)(Insulin-like growth factor-binding protein 10)(Protein GIG1) [Source:UniProtKB/Swiss-Prot;Acc:O00622]
TNFAIP6	Tumor necrosis factor-inducible gene 6 protein Precursor (TNF-stimulated gene 6 protein)(TSG-6)(Tumor necrosis factor, alpha-induced protein 6)(Hyaluronate-binding protein) [Source:UniProtKB/Swiss-Prot;Acc:P98066]
CYBRD1	Cytochrome b reductase 1 (EC 1.-.-.-)(Duodenal cytochrome b)(Ferric-chelate reductase 3)

	[Source:UniProtKB/Swiss-Prot;Acc:Q53TN4]
MAN1A2	Mannosyl-oligosaccharide 1,2-alpha-mannosidase IB (EC 3.2.1.113)(Processing alpha-1,2-mannosidase IB)(Alpha-1,2-mannosidase IB)(Mannosidase alpha class 1A member 2) [Source:UniProtKB/Swiss-Prot;Acc:O60476]
PGAP1	GPI inositol-deacylase (EC 3.1.-.-)(Post-GPI attachment to proteins factor 1)(hPGAP1) [Source:UniProtKB/Swiss-Prot;Acc:Q75T13]
AOX1	Aldehyde oxidase (EC 1.2.3.1) [Source:UniProtKB/Swiss-Prot;Acc:Q06278]
NRP2	Neuropilin-2 Precursor (Vascular endothelial cell growth factor 165 receptor 2) [Source:UniProtKB/Swiss-Prot;Acc:O60462]
DNER	Delta and Notch-like epidermal growth factor-related receptor Precursor [Source:UniProtKB/Swiss-Prot;Acc:Q8NFT8]
LHX4	LIM/homeobox protein Lhx4 (LIM homeobox protein 4) [Source:UniProtKB/Swiss-Prot;Acc:Q969G2]
NPL	N-acetylneuraminate lyase (NALase)(EC 4.1.3.3)(N-acetylneuraminic acid aldolase)(N-acetylneuraminate pyruvate-lyase)(Sialic acid lyase)(Sialate lyase)(Sialate-pyruvate lyase)(Sialic acid aldolase) [Source:UniProtKB/Swiss-Prot;Acc:Q9BXD5]
COL6A3	Collagen alpha-3(VI) chain Precursor [Source:UniProtKB/Swiss-Prot;Acc:P12111]
APOBEC3B	Probable DNA dC->dU-editing enzyme APOBEC-3B (EC 3.5.4.-)(Phorbolin-1-related protein)(Phorbolin-2/3) [Source:UniProtKB/Swiss-Prot;Acc:Q9UH17]
ELF3	ETS-related transcription factor Elf-3 (E74-like factor 3)(Epithelium-specific Ets transcription factor 1)(ESE-1)(Epithelium-restricted Ets protein ESX)(Epithelial-restricted with serine box) [Source:UniProtKB/Swiss-Prot;Acc:P78545]

Table X.0.3 Annotated genes based on HUGO symbols, which are expressed only in H1 and BG03 hES cells compared to NCCIT and 2102Ep hEC cells.

HGNC symbol	Description
DAZ4	Putative uncharacterised protein DKFZp666C074 [Source:UniProtKB/TrEMBL;Acc:Q658T2]
TPTE	Putative tyrosine-protein phosphatase TPTE (EC 3.1.3.48)(Transmembrane phosphatase with tensin homology)(Tumor antigen BJ-HCC-5)(Cancer/testis antigen 44)(CT44) [Source:UniProtKB/Swiss-Prot;Acc:P56180]
GRK4	G protein-coupled receptor kinase 4 (EC 2.7.11.16)(G protein-coupled receptor kinase GRK4)(ITI1) [Source:UniProtKB/Swiss-Prot;Acc:P32298]
HMX1	Homeobox protein HMX1 (Homeobox protein H6) [Source:UniProtKB/Swiss-Prot;Acc:Q9NP08]
HLA-DQB1	HLA class II histocompatibility antigen, DQ(3) beta chain Precursor (Clone II-102) [Source:UniProtKB/Swiss-Prot;Acc:P01920]
USP6	Ubiquitin carboxyl-terminal hydrolase 6 (EC 3.1.2.15)(Ubiquitin thioesterase 6)(Ubiquitin-specific-processing protease 6)(Deubiquitinating enzyme 6)(Proto-oncogene TRE-2) [Source:UniProtKB/Swiss-Prot;Acc:P35125]
LMO1	Rhombotin-1 (LIM domain only protein 1)(Cysteine-rich protein TTG-1)(T-cell translocation protein 1) [Source:UniProtKB/Swiss-Prot;Acc:P25800]
EMR1	EGF-like module-containing mucin-like hormone receptor-like 1 Precursor (Cell surface glycoprotein EMR1)(EMR1 hormone receptor) [Source:UniProtKB/Swiss-Prot;Acc:Q14246]
FOLH1	Glutamate carboxypeptidase 2 (EC 3.4.17.21)(Glutamate carboxypeptidase II)(Membrane glutamate carboxypeptidase)(mGCP)(N-acetylated-alpha-linked acidic dipeptidase I)(NAALADase I)(Pteroylpoly-gamma-glutamate carboxypeptidase)(Folylpoly-gamma-glutamate carboxypeptidase)(FGCP)(Folate hydrolase 1)(Prostate-specific membrane antigen)(PSMA)(PSM) [Source:UniProtKB/Swiss-Prot;Acc:Q04609]
PCDHB12	Protocadherin beta-12 Precursor (PCDH-beta-12) [Source:UniProtKB/Swiss-Prot;Acc:Q9Y5F1]
WFDC3	WAP four-disulfide core domain protein 3 Precursor (Putative protease inhibitor WAP14) [Source:UniProtKB/Swiss-Prot;Acc:Q8IUB2]
TP53TG3	TP53-target gene 3 protein (TP53-inducible gene 3 protein) [Source:UniProtKB/Swiss-Prot;Acc:Q9ULZ0]
NKX2-5	Homeobox protein Nkx-2.5 (Homeobox protein NK-2 homolog E)(Cardiac-specific homeobox)(Homeobox protein CSX) [Source:UniProtKB/Swiss-Prot;Acc:P52952]
MLN	Promotilin Precursor [Contains Motilin;Motilin-associated peptide(MAP)] [Source:UniProtKB/Swiss-Prot;Acc:P12872]
HTR2C	5-hydroxytryptamine receptor 2C (5-HTR2C)(5-HT-2C)(5-HT2C)(5HT-1C)(Serotonin receptor 2C) [Source:UniProtKB/Swiss-Prot;Acc:P28335]
IL13RA2	Interleukin-13 receptor alpha-2 Precursor (IL-13 receptor alpha-2)(IL-13R-alpha-2)(IL-13RA-2)(Interleukin-13-binding protein)(CD213a2 antigen) [Source:UniProtKB/Swiss-Prot;Acc:Q14627]
PLA2G2A	Phospholipase A2, membrane associated Precursor (EC 3.1.1.4)(Phosphatidylcholine 2-acylhydrolase)(Group IIA phospholipase A2)(GIIC sPLA2)(Non-pancreatic secretory phospholipase A2)(NPS-PLA2) [Source:UniProtKB/Swiss-Prot;Acc:P14555]
ZNF229	Zinc finger protein 229 [Source:UniProtKB/Swiss-Prot;Acc:Q9UJW7]
LIM2	lens intrinsic membrane protein 2, 19kDa isoform 1 [Source:RefSeq peptide;Acc:NP_085915]
ACRV1	Acrosomal protein SP-10 Precursor (Acrosomal vesicle protein 1) [Source:UniProtKB/Swiss-Prot;Acc:P26436]
XCL1	Lymphotactin Precursor (C motif chemokine 1)(Cytokine SCM-1)(ATAC)(Lymphotaxin)(SCM-1-alpha)(Small-inducible cytokine C1)(XC chemokine ligand 1) [Source:UniProtKB/Swiss-Prot;Acc:P47992]

Table X.0.4 Annotated genes based on HUGO symbols, which are expressed only in NCCIT and 2102Ep hEC cells compared to H1 and BG03 hES cells.

HGNC symbol	Description
SNRPN	Small nuclear ribonucleoprotein-associated protein N (snRNP-N)(Sm protein N)(Sm-N)(SmN)(Sm-D)(Tissue-specific-splicing protein) [Source:UniProtKB/Swiss-Prot;Acc:P63162]
GALNT8	Probable polypeptide N-acetylgalactosaminyltransferase 8 (EC 2.4.1.41)(Polypeptide GalNAc transferase 8)(pp-GaNTase 8)(GalNAc-T8)(Protein-UDP acetylgalactosaminyltransferase 8)(UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase 8) [Source:UniProtKB/Swiss-Prot;Acc:Q9NY28]
ATPBD4	ATP-binding domain-containing protein 4 [Source:UniProtKB/Swiss-Prot;Acc:Q7L8W6]
GDF3	Growth/differentiation factor 3 Precursor (GDF-3) [Source:UniProtKB/Swiss-Prot;Acc:Q9NR23]
PHC1	Polyhomeotic-like protein 1 (hPH1)(Early development regulatory protein 1) [Source:UniProtKB/Swiss-Prot;Acc:P78364]
PCSK1	Neuroendocrine convertase 1 Precursor (NEC 1)(EC 3.4.21.93)(Prohormone convertase 1)(Proprotein convertase 1)(PC1) [Source:UniProtKB/Swiss-Prot;Acc:P29120]
CSF2RB	Cytokine receptor common subunit beta Precursor (GM-CSF/IL-3/IL-5 receptor common beta-chain)(CDw131)(CD131 antigen) [Source:UniProtKB/Swiss-Prot;Acc:P32927]
CYP4F2	Leukotriene-B(4) omega-hydroxylase 1 (EC 1.14.13.30)(Cytochrome P450 4F2)(CYPIVF2)(Leukotriene-B(4) 20-monooxygenase 1)(Cytochrome P450-LTB-omega) [Source:UniProtKB/Swiss-Prot;Acc:P78329]
MAN2C1	Alpha-mannosidase 2C1 (EC 3.2.1.24)(Alpha-D-mannoside mannohydrolase)(Mannosidase alpha class 2C member 1)(Alpha mannosidase 6A8B) [Source:UniProtKB/Swiss-Prot;Acc:Q9NTJ4]
AMN	Protein amnionless Precursor [Source:UniProtKB/Swiss-Prot;Acc:Q9BXJ7]
EBF1	Transcription factor COE1 (O/E-1)(OE-1)(Early B-cell factor) [Source:UniProtKB/Swiss-Prot;Acc:Q9UH73]
FLRT1	Leucine-rich repeat transmembrane protein FLRT1 Precursor (Fibronectin-like domain-containing leucine-rich transmembrane protein 1) [Source:UniProtKB/Swiss-Prot;Acc:Q9NZU1]
FRS2	Fibroblast growth factor receptor substrate 2 (FGFR substrate 2)(Suc1-associated neurotrophic factor target 1)(SNT-1)(FGFR-signaling adaptor SNT) [Source:UniProtKB/Swiss-Prot;Acc:Q8WU20]
CLLU1	Chronic lymphocytic leukemia up-regulated protein 1 [Source:UniProtKB/Swiss-Prot;Acc:Q5K131]
ANKRD40	Ankyrin repeat domain-containing protein 40 [Source:UniProtKB/Swiss-Prot;Acc:Q6AI12]
DPPA4	Developmental pluripotency-associated protein 4 [Source:UniProtKB/Swiss-Prot;Acc:Q7L190]
GIMAP5	GTPase IMAP family member 5 (Immunity-associated nucleotide 4-like 1 protein)(Immunity-associated protein 3)(IAN-5) [Source:UniProtKB/Swiss-Prot;Acc:Q96F15]
FLAD1	FAD synthetase (EC 2.7.7.2)(FMN adenylyltransferase)(FAD pyrophosphorylase)(Flavin adenine dinucleotide synthetase) [Includes Molybdenum cofactor biosynthesis protein-like region;FAD synthetase region] [Source:UniProtKB/Swiss-Prot;Acc:Q8NFF5]

Table X.0.5 Common target sites of OCT4, comparing ChIP-Chip and ChIP-seq.

HGNC symbol	Description
SNRPN	Small nuclear ribonucleoprotein-associated protein N (snRNP-N)(Sm protein N)(Sm-N)(SmN)(Sm-D)(Tissue-specific-splicing protein) [Source:UniProtKB/Swiss-Prot;Acc:P63162]
GALNT8	Probable polypeptide N-acetylgalactosaminyltransferase 8 (EC 2.4.1.41)(Polypeptide GalNAc transferase 8)(pp-GaNTase 8)(GalNAc-T8)(Protein-UDP acetylgalactosaminyltransferase 8)(UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase 8) [Source:UniProtKB/Swiss-Prot;Acc:Q9NY28]
C1R	Complement C1r subcomponent Precursor (EC 3.4.21.41)(Complement component 1, r subcomponent) [Contains Complement C1r subcomponent heavy chain;Complement C1r subcomponent light chain] [Source:UniProtKB/Swiss-Prot;Acc:P00736]
MYST3	Histone acetyltransferase MYST3 (MYST protein 3)(EC 2.3.1.48)(EC 2.3.1.-)(MOZ, YBF2/SAS3, SAS2 and TIP60 protein 3)(Runt-related transcription factor-binding protein 2)(Monocytic leukemia zinc finger protein)(Zinc finger protein 220) [Source:UniProtKB/Swiss-Prot;Acc:Q92794]
YAF2	YY1-associated factor 2 [Source:UniProtKB/Swiss-Prot;Acc:Q8IY57]
FUS	RNA-binding protein FUS (Oncogene FUS)(Oncogene TLS)(Translocated in liposarcoma protein)(POMp75)(75 kDa DNA-pairing protein) [Source:UniProtKB/Swiss-Prot;Acc:P35637]
MAN2C1	Alpha-mannosidase 2C1 (EC 3.2.1.24)(Alpha-D-mannoside mannohydrolase)(Mannosidase alpha class 2C member 1)(Alpha mannosidase 6A8B) [Source:UniProtKB/Swiss-Prot;Acc:Q9NTJ4]
SIN3A	Paired amphipathic helix protein Sin3a (Transcriptional corepressor Sin3a)(Histone deacetylase complex subunit Sin3a) [Source:UniProtKB/Swiss-Prot;Acc:Q96ST3]
HMG20A	High mobility group protein 20A (HMG box-containing protein 20A)(HMG domain-containing protein HMGX1)(HMG domain-containing protein 1) [Source:UniProtKB/Swiss-Prot;Acc:Q9NP66]
HOXB13	Homeobox protein Hox-B13 [Source:UniProtKB/Swiss-Prot;Acc:Q92826]
KIAA1919	Sodium-dependent glucose transporter 1 [Source:UniProtKB/Swiss-Prot;Acc:Q5TF39]
OLFML3	Olfactomedin-like protein 3 Precursor (HNOEL-iso)(hOLF44) [Source:UniProtKB/Swiss-Prot;Acc:Q9NRN5]

Table X.0.6 Common putative OCT4 target genes between the ChIP-seq experiment and H9 ChIP-Chip targets [83], containing an OCT4 motif.

HGNC symbol	Description
SNRPN	Small nuclear ribonucleoprotein-associated protein N (snRNP-N)(Sm protein N)(Sm-N)(SmN)(Sm-D)(Tissue-specific-splicing protein) [Source:UniProtKB/Swiss-Prot;Acc:P63162]
GALNT8	Probable polypeptide N-acetylgalactosaminyltransferase 8 (EC 2.4.1.41)(Polypeptide GalNAc transferase 8)(pp-GaNTase 8)(GalNAc-T8)(Protein-UDP acetylgalactosaminyltransferase 8)(UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase 8) [Source:UniProtKB/Swiss-Prot;Acc:Q9NY28]
C1R	Complement C1r subcomponent Precursor (EC 3.4.21.41)(Complement component 1, r subcomponent) [Contains Complement C1r subcomponent heavy chain;Complement C1r subcomponent light chain] [Source:UniProtKB/Swiss-Prot;Acc:P00736]
PHC1	Polyhomeotic-like protein 1 (hPH1)(Early development regulatory protein 1) [Source:UniProtKB/Swiss-Prot;Acc:P78364]
MAN2C1	Alpha-mannosidase 2C1 (EC 3.2.1.24)(Alpha-D-mannoside mannohydrolase)(Mannosidase alpha class 2C member 1)(Alpha mannosidase 6A8B) [Source:UniProtKB/Swiss-Prot;Acc:Q9NTJ4]
FBXO40	F-box only protein 40 (Muscle disease-related protein) [Source:UniProtKB/Swiss-Prot;Acc:Q9UH90]
FLAD1	FAD synthetase (EC 2.7.7.2)(FMN adenylyltransferase)(FAD pyrophosphorylase)(Flavin adenine dinucleotide synthetase) [Includes Molybdenum cofactor biosynthesis protein-like region;FAD synthetase region] [Source:UniProtKB/Swiss-Prot;Acc:Q8NFF5]
ZNF238	Zinc finger protein 238 (Transcriptional repressor RP58)(58 kDa repressor protein)(Zinc finger protein C2H2-171)(Translin-associated zinc finger protein 1)(TAZ-1)(Zinc finger and BTB domain-containing protein 18) [Source:UniProtKB/Swiss-Prot;Acc:Q99592]

Table X.0.7 Common putative OCT4 target genes between the ChIP-seq experiment and NTERA2 ChIP-Chip targets [137], containing an OCT4 motif.