Analysis of SOX1 regulation in stem cell and cancerous cell lines

Azaz Ahmad, MSc
School of Medicine
The University of Nottingham
September 2016

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

Abstract

The SOX family of transcription factors are well-known regulators of diverse cellular events during development. SOX1, which belongs to the SOXB1 sub-family, is a key regulator of neural stem cell fate and a known specific marker of the neuroectoderm lineage. SOX1 plays an important role in early embryonic and postnatal CNS development. Recently, several studies have implicated *SOX1* as a tumour suppressor gene in different cancer types. Conversely, *SOX1* has also been reported to act as an oncogene in a prostate cancer model. In order to better understand *SOX1* gene regulation, this project set out to gain a deeper insight into the regulation of *SOX1* in the context of stem cells and cancer, and to identify potential regulatory mechanisms that can significantly regulate its function.

Initially, *SOX1* gene expression and its promoter DNA methylation pattern was analysed in a range of cancer cell lines to establish whether *SOX1* epigenetic silencing was consistently found in cancer lines. Differential SOX1 expression across the analysed cancer cell lines suggests differential regulation of SOX1 in cancer, accompanied by cancer type dependent epigenetic silencing of SOX1 by DNA methylation.

The second part of the study focused on the characterisation of the structure and expression of a newly identified *SOX1* overlapping transcript (*SOX1*-OT), using RT-PCR and 5'5'RACE techniques. The *SOX1*-OT genomic locus was found evolutionary conserved across different species. *SOX1*-OT expression was further analysed in a human neuroprogenitor cell line across different time points of neural differentiation, highlighting its possible role in neural differentiation. Furthermore, the *SOX1*-OT gene expression profile was matched with *SOX1* gene expression in a panel of different stem cell and cancerous cell lines. The co-expression profiles of *SOX1-OT* and SOX1 in stem cells and carcinogenesis indicated towards a potential role of *SOX1-OT* regulating SOX1 gene expression.

Finally, a comprehensive bioinformatics analysis was performed to investigate evidence of SOX1 post translational modifications (PTMs). *In silico* prediction of phosphorylation, acetylation and sumoylation sites support SOX1 PTMs. The predicted PTMs within different SOX1 protein domains may affect its function through altering its DNA binding activities, cellular localisation and interaction with partner proteins.

In conclusion, *SOX1* expression in different stem and cancer cell lines is likely to be regulated by promoter DNA methylation, a long non coding RNA (*SOX1*-OT) and its function by different types of PTMs. These regulatory features may in the future advance the understanding of the *SOX1* transcription regulatory network in stem cell developmental processes and its role in cancer development.

Acknowledgement

First of all, I would like to thank my supervisors, Dr Virginie Sottile and Dr Cristina Tufarelli, for giving me the opportunity to carry out this project. It was their continuous guidance, encouragement and support during my studies that enabled me to complete the project on time.

I am thankful to my internal assessor Dr Raheela khan for the guidance and taking out her valuable time to read my reports throughout the project. Special thanks go to Dr Stephanie Strohbuecker, for her useful suggestions and feedbacks throughout the project and her invaluable help and support in the lab.

I am thankful to Dr Pamela Collier, Dr Shelanah Salih and Ujjal Bose for their assistance in the lab work. Thanks to the lab members of Sottile lab and STEM, and to those who took me on induction and training. I am also grateful to Caroline Reffin for her assistance in the administrative work.

I would like to thank the Vice Chancellor's Scholarship for Research Excellence (International) from the University of Nottingham for funding this project and the School of Medicine for their support.

Finally, I am especially thankful to my family, friends and wife Dr Nageen Naseer for their support and faith in me.

Table of Content

List of Figures

List of Tables

Abbreviations

<	less than
>	greater than
ActB	actin, beta
ADC	adenocarcinoma
AFB	Animal free blocker
AGNA	Anit-glial nuclear antibody
APP	Abridged Anchor Primer
APP	Amyloid precursor protein
AUAP	Abridged Universal Amplification Primer
bFGF	basic fibroblast growth factor
bHLH	basic helix-loop-helix
BLAST	BASIC local Alingment search tool
BRCA	breast cancer susceptibility gene
C	Cytosine
CDS	Coding DNA Sequence
ChIP	Chromatin immunoprecipitation
CNS	central nerveous system
CO2	Carbondioxide
CSC	cancer stem cell
DAVID	Database for Annotation, Visualization and Integrated Discovery
DG	Dentate Gyrus
DJ-1	Parkinsonism associated deglycase-1
DMEM	Dulbecco´s modified Eagle medium
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNMT	DNA cytosine methyltransferases
ECR	Evolutionary conserved region
EGF	Epidermal growth factor
EGFR	epidermal growth factor receptor
EMT	Epithelial mesenchymal transition
ENCODE	Encyclopaedia of DNA elements
ESC	embryonic stem cell
FCS	fetal calf serum
FGF	Fibroblast growth factor

G	guanine
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
GATA-1	GATA-binding factor 1
GFP	green fluorescence protein
GI	General
GO	Gene Ontology
GSP	Gene specific primer
HCC	Hepatocellular Carcinoma
HMG	high-mobility-group
HOTAIR	HOX Transcript Antisense RNA
HPRT1	Hypoxanthine Phosphoribosyltransferase 1
i.e.	that is (id est)
IIP456	the Migration and Invasion-inhibitory protein
LEMS	Lambert-Eaton myasthenic syndrome
LincRNA	long intervening non-coding RNA
LMXIA	LIM homeobox transcription factor 1, alpha
lncRNA	long non-coding RNA
LV	Lateral Ventricle
MALAT-1	Metastasis Associated Lung Adenocarcinoma Transcript 1
MgCl2	Magnesium Chloride
miRNA	microRNA
MRG	Multiple reference genes
MSC	mesenchymal stem cell
MYC	V-Myc Avian Myelocytomatosis Viral Oncogene Homolog
ncbi	National Centre of Biotechnology Information
ncRNA	non-coding RNA
NSC	neural stem cell
O-GlcNAc	β-linked N-acetylglucosamine
Pax	Paired box protein
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PCR	Polymerase chain reaction
PDSM	phosphorylation dependent sumoylation motif
PFA	Paraformaldehyde
PFA	paraformaldehyde
PRC2	Polycomb Repressive Complex 2
PSA	Prostate specific Antigen
PTM	Post translational modification
qPCR	quantitative PCR
RACE	RAPID AMPLIFICATION OF cDNA ENDS
RIPA	Radio immuno precipitation assay
RNA	Ribonucleic acid

RT	room temperature
RT-PCR	reverese transcriptase PCR
SCLC	Small cell lung cancer
SDS	Sodium dodecyl sulfate
SDW	sterile destilled water
SENP	Sentrin/SUMO-specific proteases
SFRP	Secreted frizzled-related protein 1
siRNA	small interfering RNA
SNP	single nucleatide polymorphism
Sox	Sry-related high-mobility-group box
SOX1	Sry-BOX 1
SOX1-OT	SOX1 overlapping Transcript
STAT3	signal transducer and activator of transcription
STR	Short Tandem Repeats
SUMO	small ubiquitin-like modifier
TdT	terminal deoxynucleotidyl transferase
TF	transcription factor
TSS	Transcription start site
TTE	Transcription Termination End
Ubc	ubiquitin-conjugating
UCSC	University of California, Santa Cruz
UTR	untranslated region
Ywhaz	tryptophan 5-monooxygenase activation protein Zeta

1 Chapter 01

Literature Review

SOX1 is a transcription factor which is mainly involved in early embryonic development and neural cell fate determination [1-3]. During embryogenesis, SOX1 plays a key role in neural induction and is expressed in ectodermal cells committed to the neural fate. SOX1 is one of the earliest established markers of the neuroectoderm lineage [1]. Postnatal expression of Sox1 is confined into adult CNS, within two stem cell populations of the subventricular zone and dentate gyrus of the hippocampus [4, 5]. SOX1 express in neural stem cell (NSC) which keep it in undifferentiated state, but upon continuous expression it leads to neurogenesis [4, 5]. The expression of *SOX1* transcript has recently been reported in several cancer types [6-9], suggesting *SOX1* as a tumour suppressor gene [6, 10, 11]. For example, in nasopharyngeal carcinoma (NPC) ectopic expression of *SOX1* has been shown to inhibit cell proliferation and invasive ability of the tumour [6, 10, 11]. Conversely, *SOX1* has also been reported to act as a oncogene, which was found expressed in more invasive cell populations of prostate cancer as compared to non-invasive cells to promote tumour invasion [12]. Due to the increasing evidence of *SOX1* involvement in cancer, it is important to understand *SOX1* gene regulation in the context of cancer and analyse different co-factors regulating its expression.

1.1 SOX genes family of transcription factors

The identification of mammalian testis determining factor SRY (Sex determining region Y) present on the Y chromosome has led to the discovery of the *SOX* gene family encoding transcription factors [13, 14]. SOX family proteins are present throughout the animal kingdom, they are known to regulate diverse developmental processes such as embryonic development and cell fate determination [15]. SOX proteins are closely related to SRY region through their HMG-box DNA binding domain [16]. SOX proteins with 50% or higher sequence similarity to that of SRY HMG domain are referred as SOX proteins (SRY related HMG box). So far, 20 different SOX proteins have been identified in human and mouse, while two SOX like genes have been identified in unicellular choanoflagellate named *Monosiga brevicollis* which suggest origin of SOX proteins that exist before multicellularity [17]. SOX family of proteins has been divided into subfamilies from A to H, based upon 80% amino acid sequence similarity within or outside of HMG domain [16, 18]. Among them, SOXB group has been divided into sub group SOXB1 (SOX1, SOX2 and SOX3), and SOXB2 (SOX14 and SOX21) based on amino acid similarity within HMG domain and its immediate C-proximal domain (Figure 1-1).

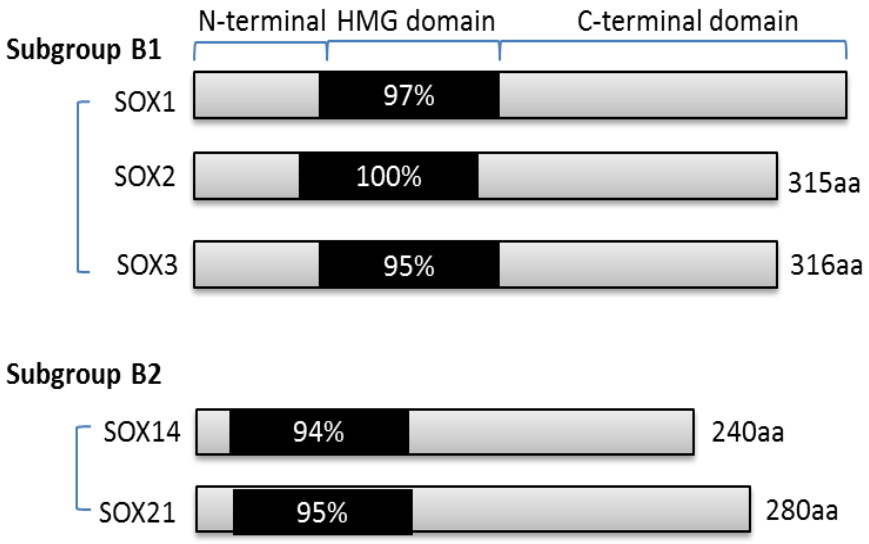


Figure 1-1 **SOXB group of protein structure and sequence similarity:** Structure comparison between the SOXB family proteins showing different domains of the SOXB group of proteins. Percentage sequence identity within HMG domain is shown in comparison with the SOX2 HMG domain. This image is adopted from [19].

All SOX family proteins recognize essentially the same DNA binding motif through their HMG domain and bind within the minor groove of DNA, which brings conformational changes to the DNA structure [20]. During embryonic development, SOX proteins act as transcription factors by switching on and off the transcription of development-related genes [21], Their functions are highly dependent on cell type and promoter context, and they also exhibit functional redundancy among each other [15].

Interestingly, some of the SOX family factors are involved in the reprogramming of differentiated cells into somatic or pluripotent stem cells [17]. In addition to these important functions, recent evidence has documented that SOX family also plays an important role in adult homeostasis and tissue regeneration [17]. Interestingly, altered expression of *SOX* genes has been reported in human cancer, causing these

genes to be extensively studied in an effort to determine their functional role in disease [17].

1.1.1 SOXB1 subfamily

The SOXB1 sub family of SOX transcription factors contains SOX1, SOX2, and SOX3 which are mainly expressed in neural tissues and are established regulators of cell fate decisions during early development [13]. SOXB1 subfamily is evolutionary conserved across different species indicating its importance in diverse developmental processes [13]. SOXB1 group share more than 90% amino acid sequence similarity in the HMG-DNA binding domain, and they exhibits functional redundancy where they are co-expressed. In neural progenitor cells, SOX2 plays a major role in maintaining pools of neural progenitors, whereas loss of its expression leads to cell-cycle exit and onset of neural differentiation [17, 22]. It has been reported that due to functional redundancy among the SOXB1 group, the phenotype elicited by inhibition of SOX2 expression can be rescued by co-expression of SOX1 [23]. The role of SOXB1 factors in maintaining neural progenitor's identity and their function as transcription regulators remain poorly understood [24].

1.1.2 *SOX1*

SOX1 is an intronless gene (Figure 1-2) that encodes for a transcription factor regulating the transcription of development-related genes [2, 3, 25]; it plays an important role in embryonic development and stem cell differentiation. *SOX1* has been reported as a key regulator of neural stem cell fate and is established as one of the earliest markers in neural

differentiation [1, 25]. Its expression is also maintained in the adult human CNS in specific areas such as lateral ventricles[4], dentate gyrus[5] and cerebellum[26]. Mouse SOX1 has also been reported in mouse lens development, where the SOX1 protein interacts with the γ-Crystallin gene and regulates its expression, which is essential for normal lens development in mammals [13, 27].

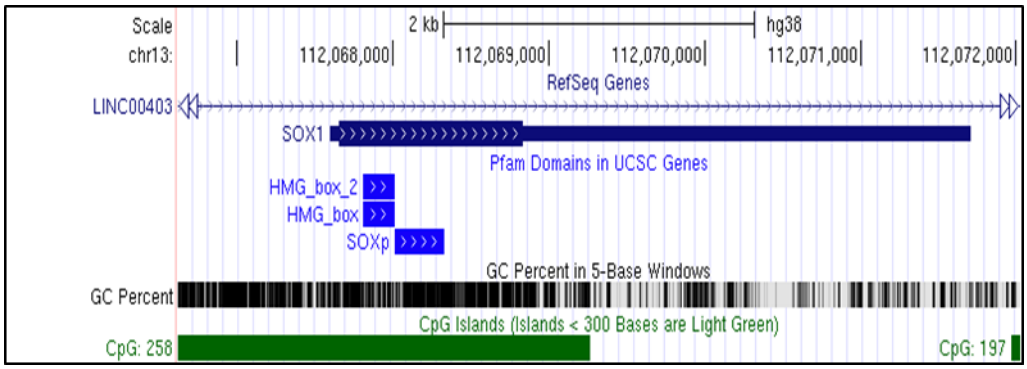


Figure 1-2: Genomic structure of *SOX1* Gene: Taken from UCSC genome browser (http://genome.ucsc.edu) [28] showing *SOX1* gene genomic co-ordinates, along with CpG islands and GC content percentage present at this genomic region are also shown.

Considered as the earliest marker for the neuroectoderm lineage, the role of *SOX1* in neural lineage commitment and differentiation is still elusive. Some studies have reported the function of *SOX1* is to determine neural stem cell fate [29, 30] [1], while others suggested that SOX1 keeps neural stem cells in an undifferentiated state by blocking neurogenesis [22, 31]. A dual function of SOX1 considering both scenarios has been suggested whereby SOX1 initially keeps neural stem cell (NSC) progenitors in an undifferentiated state to maintain progenitor pools, but upon continued expression leads to neural differentiation [4, 5].

1.2 Neural Stem cells (NSCs) development

Stem cells are characterized by their self-renewal property and potential to differentiate into multiple cell lineages [32]. Embryonic stem cells give rise to all embryonic lineages, while somatic stem cells are considered to give rise to specific cell lineages within a tissue. Development of the central nervous system (CNS) starts with neural stem cells (NSCs) differentiation. During NSC differentiation, the NSCs give rise to transit amplifying progenitors which then subsequently differentiate into three different types of lineage restricted mature cells (neurons, astrocytes and oligodendrocytes) at different time points in development (Figure 1-3).

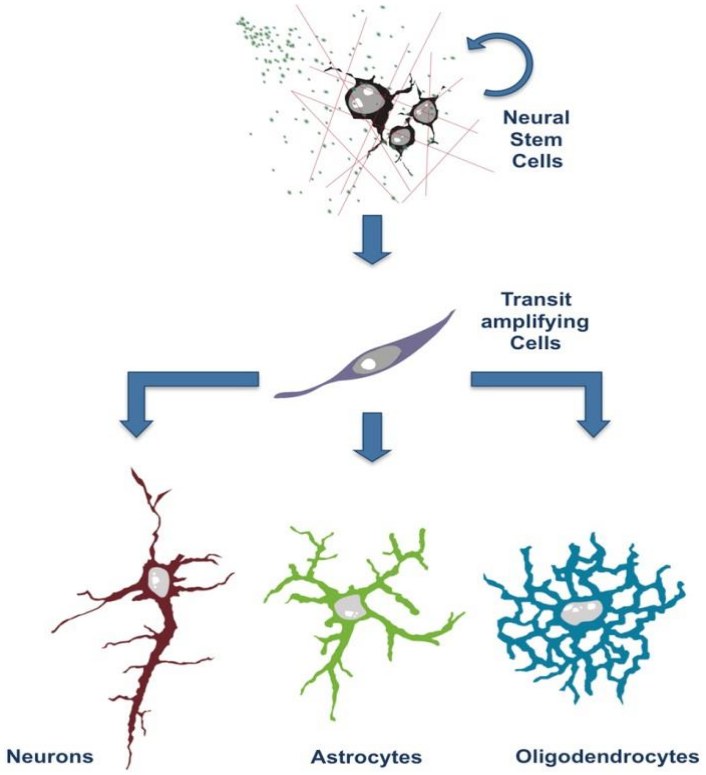


Figure 1-3 **Neural Stem cell self-renewal and differentiation:** During neurogenesis the tri-potent NSCs give rise to transit amplifying progenitors which then subsequently differentiate into lineage restricted three different types of mature cells. This image has been taken from [33]

During mammalian CNS development, NSCs are under the influence of multiple signalling pathways to produced different cell types as shown in the (Figure 1-3). These pathways interact with each other in a dynamic way to determine the fate of NSCs [34]. Their interaction is highly specific and co-ordinated according to the time, space and intensity to produce specific cell types and prevent inappropriate cell formation [34]. These signalling pathways modulate different transcription factors including SOX1, which in turn act upon neural fate associated genes to regulate their expression in order to determine NSC fate [29, 34]. SOX1 as a transcription factor and marker of NSCs is known to play direct role in NSCs fate determination and differentiation [29]. It has been reported that SOX1 interacts through multiple independent pathways to promote neurogenesis [29]. Interaction of SOX1 and its role in these signalling pathways that regulate neurogenesis has been discussed in details in the section 1.2.1

1.2.1 Role of SOX1 in NSCs fate determination

The mechanism involved in NSC development and fate determination is complex and largely unknown, which may involve different factors such as environmental influence, epigenetic modification and transcription factors [34, 35]. The major signalling pathways which cause the cascade of reactions inside the cell to modify gene expression of target genes are specified in a time and space-related manner, and any abnormal changes to these will result in abnormal development [36, 37][2]. SOX1 has been

known to interact with different developmental related pathways for NSCs fate determination such as Notch, Wnt and Stat3 signalling pathways.

1.2.1.1 Notch signalling

Notch Signalling is an evolutionary conserved intercellular signalling mechanism involved in cell processes such as development, cell differentiation, proliferation, apoptosis, epithelial to mesenchymal transition (EMT) and angiogenesis [38] [39]. There are four single pass transmembrane receptors for Notch signalling (Notch1-4) and five structurally similar Notch ligands (Delta1, 3, 4 and Jagged1-2). Notch signalling is initiated by cell-to-cell communication; cells expressing a notch receptor bind to adjacent cell ligand, which results in many downstream processes of cell development and growth [40]. During CNS development, Notch signalling plays a key role in maintaining the neuronal progenitor pool by blocking neurogenesis [34]. Notch signalling induces expression of target genes Hairy and enhancer split 3 (Hes3), Sonic hedgehog (Shh) and the transcription factor STAT3 which promote the survival of neural stem cells [39]. Neurogenesis is believed to be regulated by interplay between transcriptional factors SOX1-3 and proneural proteins [22]. Notch signalling mainly represses the expression of pro-neural genes to maintain the characteristics of neural progenitor, whereas the transcription factor SOX1 counteracts neurogenesis downstream of pro-neural factors by suppression of cell cycle exit to block neurogenesis [31]. SOX1 binds to the Hes1 promoter and supresses Hes1 transcription which is a potential repressor of neurogenesis, thus

attenuating Notch signalling [29]. SOX1 also up-regulates proneural bHLH protein neurogenin1 which promotes cell cycle exit. This net effect of SOX1 induces neuronal cell fate and exit of the cell from cell cycle [29]. Notch signalling is considered fundamental to development while aberrant notch signalling has been suggested to be involved in a wide variety of human cancer [41]. Better understanding of the *SOX1* role in this pathway can possibly add to better understanding of Notch signalling mechanism in diseases like cancer.

1.2.1.2 Wnt signalling

The Wnt Signalling pathway has important roles in various events during cell development. The Wnt protein has been assumed to act as a stem cell growth factor by promoting stem cell maintenance and proliferation [42]. Wnt signalling regulates different development-related genes through accumulation of β-Catenin in the cytoplasm which then translocates to the nucleus. Inside the nucleus, β-Catenin forms complexes with T-cell factors (TCFs) and/or lymphocyte enhance factors (LEFs) to start transcription of target genes [43]. In the absence of Wnt signalling, phosphorylation of β-Catenin occurs within the cytoplasm due to interaction of Glycogen synthase kinase (GSK-3B), Axin, Conductin and Adenomateous polyposis coli (APC). Phosphorylation cause degradation of β-Catenin and as a consequences results in repression of target genes within the nucleus [34, 44]. Direct interaction of SOX1 with β-Catenin has been reported previously, SOX1 suppresses β-Catenin mediated TCF signalling by interacting with β-Catenin, which attenuates the Wnt signalling pathway.

An interaction between β-Catenin and SOX1 is considered important as it maintains a balance between proliferation and differentiation of neural progenitor cells [29]. Studies have found that SOX1 could regulate TCF-responsive transcriptional activity to inhibit Wnt downstream genes [6] Wnt signalling role in cancer development has been extensively studied [36], and defects in the level of Wnt ligands or altered activity of Wnt protein can result in accumulation of tumour by increased cell proliferation and premature stem cell differentiation [36]. Recently, in different cancer types it has been found that loss of SOX1 expression lead to aberrant activation of Wnt signalling pathway that cause cancer development [6]. Control of Wnt signalling levels is important because inappropriate increased activity of β-Catenin can lead to cancer. SOX proteins have been found to act as modulators for Wnt signalling by interaction with β-Catenin/TCFs to facilitate target gene selection, as Wnt signalling targets different genes in different cell types [45, 46].

1.2.1.3 STAT3 pathway

The STAT (Signal transducer and activator of transcription) family of proteins are transcription factors which have important roles in transcription regulation of developmentally important genes [47]. These proteins are present in inactive form in the cytoplasm and are activated upon phosphorylation by tyrosine kinases in response to extracellular signalling. The phosphorylated STAT proteins translocate into the nucleus and trigger transcription of target genes [47] STAT3 a member of STAT proteins family has important role in processes like cell growth and

apoptosis [34] STAT3 has been known to promote NSC proliferation by blocking neurogenesis and favouring astrocyte differentiation in the presence of Notch and Bone morphogenic (BMP) signalling [34, 48, 49]. In cancer studies, loss or continuous active STAT3 signalling has been reported with negative effects on cell growth and development [48, 49]. The role of SOX1 and its interaction with STAT3 has been found in prostate cancer cells. In-vitro SOX1 has been found to interact directly with STAT3, loss of SOX1 expression lead to decreased DNA binding activity of STAT3 and in-vitro loss of invasiveness in prostate cancer cell line (52). SOX1 expression is high in aggressive prostate cancer cells which also express high levels of STAT3 suggesting the importance of SOX1 interaction with STAT3. Any alteration in the levels of either protein can affect STAT3 pathway which could possibly lead to cancer [50, 51].

1.3 Epigenetics

Epigenetics was first defined by Waddington as a "causal interaction of genes with their products (proteins), which bring phenotype into being" [52]. Later its definition changed with identification of its implication within many biological processes. Now epigenetics is defined as the study of heritable modification in gene expression that occurs without altering the DNA sequence. Epigenetic changes in gene expression are set at early embryogenesis and inherited through cell division [53].

It is now well established that gene regulation requires interaction between genome and protein complexes within the nucleus. Inside the nucleus the genome is organised into a highly compact and dynamic structure which influences genes transcriptional status, i.e. whether a

gene is active or silent [54]. During cell division, this precise organisation is maintained by specific mechanisms such as epigenetic regulation and gene expression to ensure the integrity of the cell [54, 55]. Chromosomes have transcriptionally active regions called “Euchromatin” and transcriptionally inactive regions called “Heterochromatin” [55]. Chromatin structure is highly dynamic, and changes its spatial position within different compartments of the nucleus to help regulate gene expression [54]. DNA methylation and histone modifications, the most studied epigenetic processes, both influence chromatin structure, which ultimately impact on regulation of gene expression [56].

Epigenetic regulation has an important role in regulating different developmental processes including early embryogenesis, neural development, X-chromosome inactivation and genomic imprinting [57-60]. Cells of multicellular organism are genetically the same but differ in structure and function, allowing cells to function differentially in the context of different tissues and environments [61]. Epigenetic regulation by means of DNA methylation and histone modification has an important role in developmental processes by regulating gene expression within the nucleus, and any alteration to the proteins which regulate these processes could lead to disease like cancer [61]. Research in the past decade has established that epigenetic modifications play important roles in the maintenance of proper cognitive brain function, while their dysregulation may result in devastating consequences. For example, neurological disorders such as Alzheimers, Huntingtons and Rett syndrome in which altered histone modifications and/or DNA methylation of different

developmental genes lead to abnormal cognitive functions [62]. Epigenetic modifications are also potentially reversible, which may allow pharmacological intervention to reverse the state of a disease. For example, Different histone deacetylase (HDAC) inhibitors are in preclinical trials for the treatment of Alzheimer and Huntington disease [62].

1.3.1 DNA methylation

DNA methylation involves modification of DNA without changes in its genetic code [63]; this involves transfer of a methyl (-CH_3) group from *S*-adenosyl-L-methionine to the carbon atom at position 5 in the Cytosine base within a DNA strand [64]. In mammals, cytosine methylation mostly occurs at cytosine nucleotides that lie next to guanine referred to as CpG dinucleotides (5'-CG-3') [65]. Groups of CpG dinucleotides that occur in clusters spanning at least 200bp in length, with a G+C content more than 50%, are called CpG Islands [66]. CpG Islands are mostly present at gene promoter and exonic regions; promoter methylation of CpG islands has been linked to important developmental processes [66, 67].

DNA methylation has an essential role in normal embryonic development, regulation of gene expression and in events such as gene silencing, genomic imprinting and X-chromosome inactivation [68-70]. For example, X-chromosome inactivation in mammals, one of the X chromosomes becomes heavily methylated, highly compacted and silenced, thus ensuring an equal amount of gene expression from this chromosome in male and females [71]. DNA methylation works closely together with histone modification and chromatin remodelling complexes to control and

regulate gene expression [72]. Chromatin structures influence the accessibility of DNA to transcription factors and RNA polymerase complexes to regulate transcription of genes [72, 73].

DNA methylation occurs throughtout the genome with tissue and differentiation stage specific patterns; any disruption to normal tissue patterns of DNA methylation has been proposed as a hallmark of cancer [74]. Abberrant DNA methylation of the promoter region of a gene is a mechanism that can for example inactivate a gene that suppressess tumourogenesis [75]. Abberant DNA methylation patterns reported in cancer have described genome-wide hypomethylation mainly within the repetitive region of the genome [76] and hypermethylation of CpG Islands [77]. DNA hypermethylation of CpG Islands at the promoter region of a gene is associated with silencing of gene expression [78, 79]. In the case of tumour suppressor genes (TSGs), which have important roles in processes such as cell cycle, differentiation and apoptosis, silencing by DNA hypermethylation in the promoter region results in suppression of gene expression and causes loss of function, which eventually results in pathogenesis of cancer [80-82]. For example in retinoblastoma, expression of the Rb gene which is a tumour suppressor gene that controls cell-cycle, is frequently inactivated by DNA hypermethylation in its promoter region thus promoting the development of retinoblastoma [80]. Oncogenes such as cMYC and H-RAS have been found activated in cervical cancer and breast tumour due to DNA hypomethylation at promoter region of these genes, leading to oncogenic expression [79, 83, 84].

DNA methylation is maintained by the DNA methyltransferase DNMT1, which prefers hemi-methylated DNA and known as maintenance methyltransferases in mammals. While DNMT3a and DNMT3b are specific for de novo methylation and thus responsible for establishment of methylation [85] [86]. In general, DNA methylation patterns are established during early embryogenesis by the DNMT3 enzymes and maintained throughout development by DNMT1 [85]. DNA methyltransferases DNMT3a and DNMT3b are essential for de novo methylation and mammalian development as combined knock-out in mice results in earlier embryonic lethality [87].

Until recently, It was believed that DNA methylation is an irreversible modification but then a remarkable discovery of ten-eleven translocation protein 1 (TET1) which potentially de-methylate DNA by modifying 5mC (5-methylCytosine) was made [88, 89]. DNA demethylation is considered as equally an important event as DNA methylation. TET1 belongs to the family of three proteins (TET1, TET2 and TET3) that promote DNA demethylation by converting 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) [89]. TET enzymes play role in important processes like transcriptional regulation and reprogramming of DNA methylation [90] [89]. DNA demethylation can be passive or active, passive DNA demethylation occur through loss of 5mC during rounds of DNA replication while active DNA demethylation is mediated by the activity of the TET enzymes [89].

It has been known that dysregulation of epigenetic marks is hallmark of cancer global hypo-methylation has been observed in cancer together

with hyper-methylation of specific genes, such as tumour suppressor gene [91]. Therefore, precise regulation of DNA methylation and demethylation is important for normal cellular development [89]. Better understanding of these epigenetic regulations will help define better therapeutic strategies for somatic cell reprogramming, regenerative medicine and cancer treatment.

1.3.2 Alteration of *SOX1* DNA methylation in cancer

In recent years reports have been accumulating which suggest a potential role of *SOX1* in cancer, probably due to a combination of readily availability of reagents for *SOX1*, its high profile as a stem cell regulator and the development of genomic approaches. As a result *SOX1* expression has been reported in several cancer types including hepatocellular carcinoma (HCC), cervical, prostate and ovarian as described in detail below [6, 8, 10, 12].

1.3.2.1 SOX1 methylation in Hepatocellular Carcinoma (HCC)

Hepatocellular Carcinoma (HCC) is a common type of liver cancer with many associated risk factors associated. The molecular mechanisms involved in the development of HCC are still unclear; both genetic as well epigenetic factors are considered to be involved with the development of HCC [6]. DNA methylation at promoter region has been associated with the inactivation of TSG which causes cancer [92, 93]. There is significant correlation between *SOX1* down regulation and methylation of its promoter in HCC [94]. *SOX1* has been found to act as tumour suppressor gene in HCC by antagonising Wnt/β-Catenin pathway [6]. DNA

hypermethylation within the promoter region of *SOX1* causes loss of function which can lead to aberrant activation of Wnt signalling pathways that could result in progression of HCC [6]. Induced overexpression of *SOX1* in HCC cell lines has been found to inhibit cancer phenotype while reports on knockdown analysis of mouse SOX1 showed partial restoration of a cancer phenotype [6]. Based on these findings, it has been suggested that loss of *SOX1* expression through promoter hypermathylation may be an early event in carcinogenesis making it an interesting candidate for early detection of HCC [6].

1.3.2.2 *SOX1 methylation in prostate cancer*

In prostate cancer, Serum PSA (Prostate Specific Antigen) test is mainly used for detection and follow-up, but due to problems in specificity of the PSA test, which often leads to over-diagnosing and hence over-treatment, the search for a specific biomarker for prostate cancer is still ongoing [95]. Epigenetically modified loci, especially DNA methylation of CpG at promoter regions, have been linked with down-regulation of tumour suppressor genes in many cancers [6, 96] including prostate cancer, whereby different developmental related genes including *SOX1* are differentially methylated in prostate cancer cell populations. *SOX1* gene promoter region in the non-invasive population of prostate cells were found to be hypermethylated. It was also noted that *SOX1* expression in non-invasive populations of prostate cells was significantly lower than that observed in an invasive population, suggesting a possible role for *SOX1* to act as biomarker in prostate cancer [97]. Interestingly, *SOX1* has

also been found to interact with *STAT3* pathway and both *SOX1* and STAT3 are expressed at higher level in more aggressive metastatic prostate cancer. STAT3 and *SOX1* are transcription factors and are key genes regulating the progression of prostate cancer [50].

1.3.2.3 *SOX1 methylation in cervical cancer*

Cervical cancer in women is one of the leading causes of cancer death worldwide and it is diagnosed as one in ten cancers in women [98]. Aberrant DNA methylation which has been found to contribute to development of cancer has a potential in cervical cancer screening [99, 100]. In one study on cervical cancer searching for a biomarker with high sensitivity and specificity, a CpG Island microarray was performed which identified six different genes (SOX1, PAX1, LMX1A, NKX6-1, WT1 and ONECUT1) which were differentially methylated, including *SOX1*, in squamous cell carcinoma (SCC) of the uterine cervix [100]. They are all developmental related genes and were found more frequently methylated in SCC tissues [100]. This analysis of DNA methylation has been suggested as promising approach for the identification of tumour suppressor genes and identification of novel biomarkers for cervical cancer [100]. Therefore, it could be suggested that *SOX1* can act as biomarker even though its function remains unknown in cervical cancer.

1.3.2.4 *SOX1 methylation in ovarian cancer*

In ovarian cancer, similar observations of DNA hypermethylation in promoter regions of tumour suppressor genes have been reported to lead to the pathogenesis of ovarian carcinoma [82, 101]. Differential DNA

methylation profiles have been found in high grade carcinoma as compared with low grade carcinoma [8, 50, 101].

Promoters of different developmental genes such as Secreted frizzled receptor proteins 1 (*SFRP1*), *SOX1*, paired box gene 1 (*PAX1*) and LIM homeobox transcription factor 1 alpha (*LMX1A*) gene, are commonly found hypermethylated in variety of cancer types [8, 102-104]. Among these genes, combined methylation of *SOX1*, *PAX1* and *SFRP1* genes has the best sensitivity and specificity for detecting ovarian cancer, while *SOX1*, *PAX1* and *LAMX1A* are development-related genes that have higher methylation rate in malignant ovarian tumour than in non-malignant tumour tissues, suggesting a possible role in ovarian cancer [8].

1.3.2.5 *SOX1 methylation in Non-small cell lung cancer (NSCLC)*

In Non-small cell lung cancer (NSCLC) *SOX1* has been reported to be abnormally methylated and its co-methylation along with other genes has been linked with squamous cell carcinoma [9]. Function of *SOX1* in lung cancer is still elusive and needs further exploration. In early stages of NSCLC, *SOX1* methylation frequency has been found significantly higher than normal lung disease control. Co-methylation of *SOX1*, SIX6 (SIX homeobox 6) and RARB (retinoic acid receptor, β) has been associated with adenosquamous carcinoma (ADC) [9]. Previously, it has been reported that long term exposure of cisplatin promotes methylation of *SOX1* in ovarian cancer [105]. Cisplatin is an anti-cancer drug, widely used in non-small cell lung cancer (NSCLC) therapy but after exposure to cisplatin, treated cells creates resistance to it and make it ineffective [106].

SOX1 hypermethylation in HCC and NSCLC has been already reported [6, 9], based on these finding research has been done and it was find out that inactivation of *SOX1* by promoter hypermethylation is responsible for cisplatin resistance in NSCLC. Additionally, silencing of *SOX1* by promoter methylation enhances autophagy induced by cisplatin resistance in NSCLC [106]. Similarly, loss of SOX1 expression increases metastatic abilities in NSCL through epithelial to mesenchymal transition (EMT) [106].

1.3.3 SOX1 as a promising cancer biomarker

Since the discovery that alterations in DNA methylation are associated with aberrant gene regulation that could lead to cancer development, [79, 107] enormous research has been done to analyse differentially methylated regions that can help identify potential biomarkers for early cancer detection [64]. In ovarian cancer, DNA methylation has been suggested as a potential prognostic factor and promising biomarker for its early detection. It has been found that DNA hypermethylation of developmental related genes such as SRY-box 1 (*SOX1*), paired box gene 1 (*PAX1*) and LIM homeobox transcription factor 1 alpha (*LMX1A*) play an important role in tumorigenesis and progression of ovarian cancer [8]. Genetic mutations or a single nucleotide polymorphism (SNP) can also be used as a molecular biomarkers for ovarian cancer screening like mutation in TSGs (BRCA1 and BRCA2), however the mutation rate in ovarian cancer patients suggest these are not suitable as molecular biomarkers [108]. For early detection of cervical cancer, DNA methylation analysis of certain development-related genes such as *SOX1*, *PAX1* and

LMX1A has been evaluated as methylation biomarkers, but due to its moderate specificity and sensitivity it is not available in clinical practice [65]. Therefore, there is a need for methylation biomarkers with high sensitivity and sufficient specificity in order to detect cervical cancer at an early stage for a treatment to be successful. In prostate cancer, SOX1 has been reported to be involved in the progression of prostate cancer and is one of the epigenetically regulated targets in prostate cancer invasion [50]. SOX1 expression in prostate cancer cells makes it an attractive potential methylation biomarker and could be helpful to differentiate between more aggressive invasive cells population in prostate cancer.

SOX1 antibodies have been found in patients with paraneoplastic neuropathy like Lambert-Eaton myasthenic syndrome (LEMS) and nonparaneoplastic neuropathy [109]. LEMS is an autoimmune disorder of neuromuscular junctions which causes muscle weakness and autonomic dysfunction and is followed by diagnosis of small cell lung carcinoma (SCLC) in more than 50% of LEMS patients [110, 111]. Previously, Graus F et al had identified anti-glial nuclear antibody (AGNA) in LEMS and SCLC patient serum, which shows a characteristic nuclear staining of the bergmann glia cell in purkinje cell layer of the cerebellum [112]. AGNA was present in 43% of patients with LEMS and SCLC. AGNA reactivity was found widely expressed in the developing nervous system. A fetal brain library was screened with AGNA positive sera and it was found that the antibodies produced were against the antigen SOX1 [113].

Presence of SOX1 antibodies has been proposed as specific serological marker for detection of SCLC-LEMS [111]. In another study, SOX1

antibodies have been found to predict SCLC in cerebellar ataxia patients with a specificity of 100% and sensitivity of 49% [114]. Recently, it has been suggested that neuronal marker antibodies such as SOX1 in CNS paraneoplastic syndromes only indicate the presence of underlying tumour rather than a cause [114]. The significance of *SOX1* as a marker in immune responses during autoimmune diseases is still elusive, and further understanding of *SOX1* regulation in control and disease contexts will help to refine screening methods for detection of lung cancer [111].

1.4 Long non-coding RNAs (lncRNAs)

During the transcription of protein coding genes, genetic information is transcribed from DNA into coding RNA transcripts; these are then exported to the cytosol and translated into proteins. Studies over the last several decades have shown that only a small proportion of the genome is transcribed into protein coding RNA transcripts whilst a large amounts of RNA transcripts do not code for protein and are called non-coding RNAs (ncRNAs) [43-45]. Encyclopedia of DNA Elements (ENCODE) is a project that has revolutionised the field of genomics with the aim to identify all functional elements in the human genome. ENCODE has found that 0.1% genes within the human genome showed evidence of protein expression, suggesting that the majority of RNA transcripts are non-coding [115, 116]. RNA transcripts that do not code for protein are called non-coding RNAs (ncRNAs). It was found that the level of ncRNAs transcription is four times higher than protein coding RNAs [46]. The ncRNAs are divided into different classes containing many types of short non-coding RNAs (<200nt) and long non-coding RNAs or lncRNAs (>200nt). LncRNAs are

found in sense or anti-sense direction to protein coding genes, or within introns of protein-coding genes. Long non-coding RNAs which are found in intergenic regions of the genome (between two genes) are referred to as long intergenic non-coding RNAs (LincRNAs). Similar to mRNAs, majority of lncRNAs are transcribed by RNA polymerase II, polyadenylated and can show complex splicing patterns [117].

The lncRNAs which were mainly considered as the result of transcriptional noise are the least well studied ncRNAs. In mammals, thousands of lncRNAs have recently been described, and suggested to play a role in a several biological processes including transcriptional, post-transcriptional modifications chromatin organization and epigenetic regulations. However, the biological significance and function of the vast majority of these transcripts remain unclear [47, 48].

The field of cancer research has also recently turned its attention to non-coding transcripts. With the help of genome wide studies, it has been revealed that more than 80% of cancer associated genetic variations (SNPs) are located within the non-coding region of the human genome and only small number are found in protein coding regions [117, 118]. While the role of long non-coding RNA in cancer is still emerging, studies have reported aberrant expression of lncRNA in many cancer types, demonstrating both oncogenic and tumour suppressor roles in tumorigenesis [119]. LncRNA function through various mechanisms to play key role in cell growth, proliferation and differentiation. Any perturbation in the lncRNAs expression may contribute to the several processes related to carcinogenesis, including cell growth and

proliferation [117]. Some lncRNAs significantly contribute in molecular pathways in cancer, such as cell proliferation, tumour suppression evasion, cancer angiogenesis, anti-apoptosis and metastasis [120]. It has been found that overexpression of oncogenic lncRNAs results in cancer development through chromatin looping and distal engagement with the androgen receptor, antisense gene regulation, alternative splicing, and impeding DNA repair [121].

HOTAIR is the first long intervening non-coding RNA (lincRNA) found to be involved in cancer development [122]. HOTAIR is known as an epigenetic regulator that regulate genes involved in different cellular pathways by interacting with Polycomb Repressive Complex 2 (PRC2) [123]. Its expression is dysregulated in different cancer types such as breast cancer and Hepatocellular carcinoma (HCC) [122]. In primary breast cancer, expression of HOTAIR is highly upregulated and linked with metastasis and poor survival rate [122].

Another lncRNA called MALAT-1 is highly evolutionary conserved in mammals indicating its potential important function. MALAT-1 has been identified in different cellular processes like alternative splicing, nuclear organization, epigenetic modulating of gene expression [124]. Studies have shown that MALAT-1 is complicated in various pathological processes, including cancer [124, 125]. It is known as prognostic marker in lung cancer metastasis [126] and function as a critical regulator of metastasis-associated genes [127]. MALAT-1 exact mode of action in different physiological and pathological conditions still need to be explore [126]. Recent studies have indicated potential role for lncRNAs in cancer

and the molecular mechanisms through which they may play a role in cancer development largely remains unclear [117].

1.5 Post-Translational Modification of protein

Proteins perform vast variety of biological functions in a living organism such as catalyzing cell metabolism, DNA replication, cell signaling, cellular transport, etc. During protein biosynthesis, each mRNA is translated into a polypeptide chain of amino acid in a manner specified by the encoded gene. Subsequently, proteins are usually folded into specified three-dimensional structures that determine their activity. During or after protein synthesis, the amino acid residues in a protein can be modified by post-translational modification (PTM), which alters the structure, stability, localization and activity of the protein and ultimately change its function. Therefore, to understand the function of a protein it is important to know about possible post-translational modifications [128]. There are many types of PTMs that can occur to a protein, often including proteolytic cleavage events or covalent modifications at specific amino acid residues, such as addition of phosphoryl, sumoyl, acetyl, glycosyl, methyl or other groups. Sumoylation, acetylation and phosphorylation are the most essential post-translational modifications because of their important role in the cellular processes including gene expression regulation, signalling pathways and intracellular transport. The more studied PTMs are briefly described below.

1.5.1 Phosphorylation

Phosphorylation is one of the most studied post-translational modifications (PTMs). This modification is caused by the transfer of a phosphate group from adenosine triphosphate (ATP) to the acceptor residue of an organic molecule to generate adenosine diphosphate (ADP) and the organic molecule carrying phosphorylated acceptor residue. This transfer is performed by protein kinases that can display specificity for individual residues such as serine, threonine or tyrosine [128]. Phosphorylation event is a reversible biochemical reaction that can be reversed by the action of enzymes called phosphatases [128]. This reversible phosphorylation event is essential for the regulation of cellular processes such as metabolism, proliferation, differentiation and apoptosis [128, 129].

Transcription factor proteins are involved in a wide variety of cellular processes [130]. Phosphorylation regulates different aspects of transcription factor function, including cellular localization, protein stability, protein-protein interactions and DNA binding activities [131]. For example, SOX2, a member of SOXB1 transcription factors, is a master regulator of embryonic stem cell (ESC). Phosphorylation-based regulation of SOX2 has been reported in the literature; it has been shown that protein kinase B (Akt) phosphorylates SOX2 at position Thr118 that enhanced its transcriptional activities in ESC [132]. SOX2 plays key role during reprogramming of somatic cell into induced pluripotent cell [133]. It has been demonstrated that mouse SOX2 phosphorylation by Cdk proteins

promotes the establishment of pluripotent state during reprogramming [134].

Many human diseases are the consequences of abnormal phosphorylation including neurodegenerative diseases and cancer (122), For example, In Alzheimers's disease, abnormal hyper-phosphorylation of tau protein is responsible for misfolding and aggregation that lead to the pathogenesis of Alzheimer's disease [135]. In human cancer, loss of cyclin D1 phosphorylation at C-terminal residue (Threonine 286) inhibits its nuclear export signal which significantly increased its oncogenic potential [136].

1.5.2 The O-GlcNAc Modification

Glycosylation is the enzymatic process in which glycan (saccharide chain) is attached to the protein and is another important PTM. Protein glycosylation is the most abundant and diverse form of modification, which occurs to at least 50% of all mammalian proteins [137]. There are different types of glycosylation; most common are N-linked and O-linked glycosylation. In eukaryotes, most of the nuclear and cytoplasmic proteins are o-glycosylated by a-linked O-GlcNAc to a serine or threonine residues [138]. Others types of o-glycosylations are O-fucosylation, O-mannosylation, and O-glucosylation that are of functionally high relevance during early embryonic development and for vital physiological function of proteins [137]. O-GlcNAc modification is the attachment of O-linked N-acetylglucosamine (O-GlcNAc) to serine or threonine residues [138]. Like phosphorylation, O-GlcNAc is highly dynamic, with rapid cycling in

response to cellular signals. O-GlcNAc glycosylation is considered reciprocal to phosphorylation during the cell cycle, cell stimulation, and/or cell growth. Thus, if phosphorylation occurs, O-GlcNAc does not, and vice versa [137, 138]. This reciprocal modification serves as a nutrient/stress sensor to modulate signalling, transcription, and cytoskeletal functions. O-GlcNAc has been also reported in the aetiology of different chronic diseases like diabetes and neurodegenerative disorders [139]. For example, In hyperglycemia increase expression of O-GlcNAc proteins within the insulin signalling pathway contribute to insulin resistance [140]. In neurodegenerative disorders like Alzheimer's disease, under normal brain condition the proteins involved in the pathology of the disease are O-GlcNAcylated such as tau, neurofilaments, beta-amyloid precursor protein, and synaptosomal proteins. It has been proposed that due to hypoglycaemia within the brain may reduce O-GlcNAcylation of tau protein that leads to hyperphosphorylation, and as a result causes tangle formation and neuronal death [140]. It has been reported that O-GlcNAc regulates signalling and transcriptional processes related to cancer, and is also involved in the trafficking of cell adhesion molecules necessary for metastasis [139]. For example, O-GlcNAcylation of β-Catenin regulates its nuclear localization and transcriptional activity. It also leads to loss of E-cadherin which is a cell adhesion molecule critically important to mechanism underlying metastasis of cancer cells [141]. Many oncogenic proteins and tumour suppressor proteins are also regulated by O-GlcNAc modification. Due to known role of phosphorylation in cancer development it has been speculated that the extensive cross talk between

phosphorylation and O-linked glycosylation could have a function in cancer [142].

1.5.3 Sumoylation

Sumoylation is a type of a post-translational modification (PTM) in which a small ubiquitin-like modifier (SUMO) protein covalently binds to a lysine (K) amino acid in a protein substrate and regulates its functional properties [143]. Sumoylation is a highly dynamic and ubiquitous PTM, which occurs mostly within the nucleus and that regulates many cellular processes such as transcription, chromatin remodelling, nucleo-cytoplasmic transport and cell signalling [144]. For example, Sumoylation of transcription factors P3 led to the transcription repression by establishment of compacted repressive chromatin with characteristics of compacted heterochromatin [145]. Smad4, a factor which play role in TGF-βsignal transduction pathways requires sumoylation to transport into the nucleus where it activate or repress transcription of other transcription factors [146]. Techniques used to identify sumoylated proteins include immunoprecipitation, an in vitro sumoylation assay, and gel shift mobility assays [147]. Studies have suggested that sumoylation occurs on specific lysine (K) residue within the canonical consensus motif Ψ-K-X-E, (ψ, a hydrophobic amino acid, such as A, I, L, M, P, F, V or W; X, any amino acid residue) [148]. However, 52% of the reported sumoylation sites do not contain the predicted consensus sequence [144].

Other modifications, such as phosphorylation, may regulate sumoylation of a substrate both positively and negatively [149]. A specific motif

ΨKxExxSP known as PDSM (Phosphorylation dependent Sumoylation motif) regulate phosphorylation-dependent sumoylation of numerous mutually unrelated transcriptional regulators [149]. PDSM has been described to have a SUMO consensus site followed by two amino-acid and then a phosphorylated residue preceding a Proline amino acid [149]. This highly conserved PDSM has been found in several human proteins and its orthologs [149].PDSM is conserved in many proteins families including some of the human SOX family of proteins, SOX3 protein which is a member of SOXB1 sub-family contains PDSM at position 374-381 (VKSEpSsp) [149]. In human SOX2, modification like phosphorylation dependent sumoylation at similar conserved motif has been previously reported to regulate transcriptional activities of SOX2 [150]. In mouse SOX2, Tsuruzoe et al. have found sumoylation of lysine (K) at similar conserved motif that negatively regulate SOX2 transcriptional activity through impairing the DNA binding site *[151]*.

Sumoylation is known to play important roles in DNA damage repair and maintaining genome integrity, in some cases any abnormal changes to sumoylation system causes a defect in maintaining homeostasis which might hint to cancer development [152]. Studies have described sumoylation of proteins involved in neurodegenerative diseases such as Huntington's disease (huntingtin), Parkinson's disease (tau, α-synuclein, DJ-1) and Alzheimer's disease (tau, APP). Sumoylation is also known to regulate different tumour suppressor proteins such as p53, pRB (retinoblastoma protein), p63, p73, and Mdm2 (murine double minute 2) [153]. Overexpression of SUMO conjugating enzymes (Ubc9) and SUMO

proteases (SENP1 and SENP5) have been identified in many cancer types including osteosarcoma, colon, and prostate cancer [154]. These evidences suggest that sumoylation needs to be tightly regulated to prevent tumourigenesis [154].

1.6 Hypothesis and aims of the project

Transcriptional regulation of the *SOX1* gene in normal and cancer development is unclear, and little is known about factors that regulate its expression. Recently, *SOX1* has been reported in different cancer types as a tumour suppressor gene, while contrary to this, *SOX1* has been also found as an oncogene in prostate cancer progression [10, 12]. Interestingly, *SOX1* lies within an intron of a long-non-coding RNA transcript called *SOX1* overlapping transcript (*SOX1-OT*) whose structure is largely unknown. Little is known about the transcriptional regulation of human *SOX1* and *SOX1-OT* or whether there is any regulatory relationship between *SOX1* and *SOX1-OT*. Taken together this information led to the following working hypothesis:

SOX1 possesses a complex regulatory network and its function as a tumour suppressor or oncogenes depends upon different regulatory mechanisms.

The overall goal of this project was to address this hypothesis by gaining deeper insights into the regulation of *SOX1* in the context of stem cells and cancer, and to identify potential regulatory factors and mechanisms that might regulate its function. To achieve this, three specific aims were developed.

Aim 1: Expression and DNA methylation profile of *SOX1* in stem cells and cancer cell lines (Chapter-3).

SOX1 gene expression and its promoter DNA methylation pattern was probed in a wide range of stem cells and cancer cell lines in order to find if *SOX1* gene regulation through promoter DNA methylation is a common

regulatory mechanism and whether its gene expression in a panel of cancer cell lines is in any way associated with tumourigenesis.

Aim2: Structure and expression profile of *SOX1-OT* and its relationship to *SOX1* expression (Chapter-4).

Two specific aims were addressed:

Specific aim 2.1: Structural Characterisation of *SOX1-OT*

Structure of SOX1-OT and its different isoforms were characterised in a neural stem cell line (ReN) across different time points of differentiation.

Specific aim 2.2: Expression profile of *SOX1* and *SOX1-OT* in stem cells and cancer cell lines

SOX1 and *SOX1-OT* expression profile was analysed in a panel of different stem cells and cancerous cell lines in order to identify any possible co-relation between their relative expressions in these experimental contexts.

Aim 3: SOX1 protein expression and identification of possible post-translational modifications (Chapter-5).

SOX1 belongs to a family of transcription factors which plays fundamental roles in embryonic and CNS development [1-3]. There have been few reports about SOX1 protein expression in cancer. Therefore, SOX1 protein expression in different cancer types was investigated to see whether SOX1 gene expression translate into its protein. Additionally, Online Bioinformatics databases were used to predict potential post translational modifications for SOX1 that might be important to SOX1 function as transcription factor.

2 Chapter 02

Methods and Materials

2.1 Chemicals and Reagents

Reagents used in this study were purchased from Life Technologies™ unless otherwise stated. Most of the cell lines were kindly provided by Dr. Virginie Sottile and Dr Cristina Tufarelli. ReN cells pellets were kindly provided by Dr Stephanie Strohbuecker. Mouse brain tissues and frozen section were kindly provided by Dr Shelanah Salih and Dr Virginie Sottile. Breast carcinoma cells pellets were kindly provided by Dr Pamela Collier from the Lab of Dr Anna Grabowska, School of Medicine, The University of Nottingham. Human cell lines used in this study are shown in the *Table 1*. Cells from heterozygous Sox1-GFP mice tissues [155] labelled as +/- indicating that they carry both the wild-type Sox1 and the GFP reporter gene.

2.2 Cell culture

2.2.1 Cell culture in standard medium

The cell lines that cultured were NTera, hMSCs, HeLa, SH-SY5Y, HOS, CaCo2 and MCF7. NTera Cells were kindly gifted by C. Allegrucci, SH-SY5Y cells were kindly gifted by Prof E. Billet (NTU), HOS cells were kindly gifted by F. Rose.. The cell lines CaCo2 and MCF7 were kindly provided by Cristina Tufarelli and their STR profiling was done by the donor. HeLa cells were already available from a previous project, provide by the McWhir lab. hMSC cells were commercially bought for the project. All of

the cells were brought up on receiving them and cell pellets were prepared for the future experiments.

The cell lines used in this study NTera, hMSCs, HeLa, SH-SY5Y, HOS, CaCo2 and MCF7 were grown in a standard MSC medium, which is made up of Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal calf serum (FCS), 1% L-Glutamate, 1% Non-essential amino acids and 0.5% Penicillin/ Streptomycin. Cells were incubated in a humidified incubator in an atmosphere of 5% CO_2 at 37˚C.

2.2.2 Passaging of cell lines

Protocol for all the cell lines was same unless otherwise stated. Cell lines NTera (passage 60 or P+60), hMSCs (P+43), HeLa (P+4), SH-SY5Y (P+34), HOS (P+68) and MCF7 (P+4) were passaged when they reached 90% confluency except for CaCo2 (P+6), when reached 75% confluency, washed with PBS and treated for up to 5 minutes except for CaCo2, which was up to 10 minutes with 500µl 0.05% trypsin/EDTA per T25 flask. To deactivate the trypsin 5mL of medium was added into the flask, cells were gently mixed and transferred into the centrifuge tube. Cells solution was centrifuged at 200g for 5 min. After centrifugation, cell pellet was formed at the bottom and the supernatant was aspirated. Cell pellet was re-suspended in a 2ml of medium. Cell suspension aliquot (1mL each) was added into new T25 flasks already containing 4mL of fresh medium. All the cell lines were routinely passaged after two days, except for CaCo2 which was very slow to grow were passaged after 3 or 4 days.

2.2.3 Cryopreservation of cells

Cells were cryopreserved in an ice cold freezing mix, which is made up of 60% MSC medium, 20% FCS and 20% Dimethyl sulfoxide (DMSO). Trypsinised cells were washed in a PBS and centrifuged for 5min at 12,000x rpm. After aspirating the supernatant, the cells were re-suspended in a 1:1 solution of MSC medium and freezing mix and quickly transferred and kept at -80°C. To thaw cryopreserved cells, tubes were quickly transferred into the cell culture hood and thawed in a waterbath, centrifuged for 5min at 12,000x rpm. Supernatant was removed and re-suspended in 5ml of MSC medium and transferred to a T25 flask.

Table 1 Cell lines used in the experimental study.

No.	Cell line	Tissue	Cell type	Ref
1	Ntera2	Testis	Pluripotent-embryonal carcinoma	[156]
2	hMSCs	Bone marrow	Mesenchymal progenitors	[157]
3	ReN cells	Brain	Neural progenitors	[158]
4	HeLa	Cervix	Adenocarcinoma	[159]
5	SH-SY5Y	Bone marrow	Neuroblastoma	[160]
6	HOS	Bone	Osteosarcoma	[161]
7	CaCo2	Colon	Colorectal Adenocarcinoma	[162]
8	MCF7	Mammary gland	Adenocarcinoma	[163]
9	HuES7	Embryo	Human embryonic stem cell	[164]
10	MCF10A	Mammary gland	Immortalized epithelial cell line	[165]
11	MRC5	Lung	Fibroblast (Normal)	[166]
12	HCT116	Colon	Colorectal Carcinoma	[167]
13	MDA-MB-361	Mammary gland	Adenocarcinoma	[168]
14	MDA-MB-231	Mammary gland	Adenocarcinoma	[168]
15	Hs578T	Mammary gland	Carcinoma	[169]
16	T47D	Mammary gland	Ductal carcinoma	[170]

2.2.4 Neural treatment

Neural treatment was used for SH-SY5Y cells. Cells were grown in Neurobasal medium, containing 25ml DMEM-F12 supplemented with 1ml

B27 (50x), 0.5ml N2 (100x), 0.5% Penicillin/Streptomycin, Heparin (5 µg/ml) and the growth factors 20ng/ml bFGF and 20ng/ml EGF) for 6 days along with control cells kept under basic growth medium (section, 2.2.1) and labelled as SH-SY5Y-neuro and SH-SY5Y-control respectively.

2.3 Molecular Biology

2.3.1 Harvesting Cells for RNA and DNA extraction

For RNA/DNA extraction, cell monolayers were washed with PBS, detached with 0.05% trypsin/EDTA and pelleted for 5 min at 12,000x rpm, cell pellets were stored at -80°C with or without TRI® Reagent (Sigma-Aldrich, UK) for RNA and DNA extraction respectively.

2.3.2 Genomic DNA extraction

Genomic DNA from different cell lines was extracted by using Quick-gDNA (Miniprep) kit (Zymo Research, UK). The frozen cell pellet was resuspended in genomic lysis buffer, vortexed for 3 seconds and incubated at room temperature for 10min. The mixture was transfer to a Zymo-Spin™ column and centrifuge at 10,000 x g for 1min. Flow through was discarded and Zymo-Spin™ column was transfer into new collection tube, followed by centrifuge at 10,000 x g for 1min. 500µl of g-DNA Wash Buffer was to the Zymo-Spin™ column and centrifuge at 10,000 x g for 1min. Zymo-Spin™ column was transfer into the clean microcentrifuge tube. 50uL of autoclaved distilled water was added to the Zymo-Spin™ column and stand for 5min at room temperature. DNA was eluted by centrifugation at top speed for 30 seconds. The DNA concentration was

determined based on the OD260 nm using the Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies, Rockland, Delaware, USA).

2.3.3 RNA Extraction and purification

Impure RNA can have negative effects on downstream processes. Therefore, commercially available kits from different companies were tested as shown in the Table 2, in order to purify the extracted RNA from different impurities which could be potential inhibitors for downstream processes like reverse transcription cDNA synthesis and real time PCR. Extracted RNA was run on 2% Agarose gel to analyse the purity of the RNA and check for genomic DNA contamination.

Table 2 Different commercially available kits were used to compare for RNA extraction yields.

Tri-Reagent	Sigma, Cat# T9424
RNA Clean and Concentrator	Zymo research, Cat# R1015
One step PCR inhibitor removal kit	Zymo research, Cat# D6030

2.3.4 Polymerase chain reaction (PCR)

PCR reactions were set up in a thermal cycler (Multigene Model: TC9600-G, Labnet International, Inc.). 20µl volume PCR master mix contained 14.15µl SDW, 2µl 10xPCR buffer, 0.6µl MgCl2 (50mM), 0.6 µl dNTP (2.5mM, Invitrogen), 0.15µl Platinum® Taq DNA Polymerase (Invitrogen, USA), 1.5µl Primer mix (10pmol/µl) and 1µl DNA (50-200ng/µl) unless otherwise stated. For PCR amplification of *SOX1-OT*, PCR mix was prepared containing 2xDMSO buffer (32mM Ammonium sulphate, 134mM Tris-HCl, 20mM β-Mercaptoethanol and 20% DMSO. PCR amplification was mostly kept to 40 cycles, unless otherwise stated. PCR reaction was

step up as following: an initial denaturation for 5 minutes at 95°C; then denaturation step for 30 seconds at 95°C, followed by 45 seconds annealing temperature (primer specific) and an extension at 72°C for 60 seconds; the extension step dependent upon the length of the amplification product, which is approximately 1 min for 1000 bps. All reactions contained PCR negative control in which there was sterile distal water added instead of DNA template. The PCR products were analysed by agarose gel electrophoresis

2.3.5 Primer optimization

Primer pairs for SOX1 and SOX1-OT were designed by using Primer-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi?LINK_LOC=BlastHomeAd), Primer3 (http://frodo.wi.mit.edu/) and Oligocalc tools (for checking primer dimers and hairpin loops, available at (http://www.basic.northwestern.edu/biotools/oligocalc.html), with sequences available on the NCBI (National Centre of Biotechnology Information) website (http://www.ncbi.nlm.nih.gov/). The primers were purchased from Eurofins MWG Operon (Ebersberg, Germany). To find optimised annealing temperature for primers, temperature gradient PCR in a range of 40-65°C was performed.

2.3.6 DNase Treatment

Extracted RNA samples were subjected to DNAse I treatment in order to avoid genomic DNA contamination. DNase-I, Amplification grade kit (Invitrogen, USA) was used according to the manufacturer's protocol using 1U/µl of DNase-I for each 1µg of RNA at 25°C for 20min.

2.3.7 Clean-Up after PCR, RT-PCR or Enzyme digestion

For clean-up PCR products, cDNA after reverse transcription PCR or restriction enzyme digestion, MinElute® PCR purification kit (Qiagen, UK) was used according to the manufacturer's protocols, with alteration: To elute DNA, 10µl sterile distal water (SDW) was added to the column and kept at a temperature of 70°C for 5 minutes followed by centrifugation.

Agarose gel extraction of PCR products, the DNA fragment was cut out from a gel and extracted by using QIAquick Gel Extraction Kit (Qiagen, UK), according to the manufacturer's protocol.

2.3.8 cDNA synthesis and Reverse Transcription PCR

After DNase-I treatment, RNA samples were converted to cDNA by reverse transcription (cDNA synthesis). 2µg RNA were used to synthesize cDNA by reverse transcription using 200 units/µl of SuperScript® III Reverse Transcriptase (Invitrogen, USA) in 30µL of total reaction volume, including 100pmol/µl of random 15mer (pentadecamer [171]) primers (MWG Biotech, Germany), 0.5mM dNTP and 0.1mM DTT. All steps were performed according to the manufacturer's protocol. cDNA samples obtained by RT-PCR were cleaned up by using MinElute PCR purification kit (Qiagen, UK) and stored at -20°C.

2.3.9 Reverse Transcription PCR primer pairs for SOX1

Different human *SOX1* primer pairs collected from published literature were tested for *SOX1* cDNA expression (Table 3) To test the specificity of the SOX1 primers, PCR amplification was performed on NTera cDNA and genomic DNA (positive control) and then products were run on 2%

agarose gel. This was followed by gel extraction and then purified PCR products were send for Sanger sequencing (SourceBioScience, Nottingham) (section, 2.5.4.6).

Table 3 Different SOX1 primer pairs collected from published literature were tested for detection of SOX1 mRNA transcript reversed transcribed into cDNA.

SOX1 Primers	T m	Product size	Sequence	Ref
ahS1_1	62	258bp	Up: 5'5'CCAATTGTTGGCATCTAGGTCT`3	[10, 100]
			Dn: 5'5'- GCACCACTACGACTTAGTCCG-3`	
ahS1_3	64	848bp	Up: 5'-TCACTTTCCTCCGCGTTGCTTCC-3'	[172]
			Dn: 5'-TGCCCTGGTCTTTGTCCTTCATCC-3'	
ahS1_4	59	54bp	Up: 5'5'-AAGGCAGGTCCAAGCACTTA-3`	[173]
			Dn: 5'5'-ACCCAAAAGAGCGGTAACAA-3`	
ahS1_5	60	201bp	Up: 5-CCTCCGTCCATCCTCTG-3	[174]
			Dn: 5-AAAGCATCAAACAACCTCAAG-3	
ahS1_6	60	468bp	Up: 5'5'-TACAGCCCCATCTCCAACTC-3`	[175]
			dn: 5'5'- GCTCCGACTTCACCAGAGAG-3	

2.3.10 Electrophoresis

Agarose gels (1-2% w/v) were prepared by mixing agarose powder in a required amount of 1x TAE buffer (Tris-acetate-EDTA) to bring into a desired percentage, heated up in a microwave to completely dissolve and melt the agarose, Ethidium-bromide (Sigma-Aldrich, UK) was added at a concentration 4ng/ml. The melted agarose was settled in a gel-casting tray for 20-30 minutes. 10-20µl of the PCR products were loaded onto the gel by mixing with 6µl loading buffer (3.5g Sucrose, 4 ml 10xTAE buffer, small quantity of bromphenol blue, filled up with sterile distilled water to 10 ml). For band sizes 2µl HyperLadder™1Kb (Bioline, UK) or HyperLadder™ 50bp (Bioline, UK) were used on gels. Gels were run at 65- 100V for about

60-120 minutes. Gels were visualized using the Luminescent Image Analyzer LAS-4000 (Fujifilm).

2.3.11 Quantitative-Real time PCR

For gene quantification by real time PCR, Two different assays SYBR Green fluorescent dye and TaqMan probe based methods were optimized in order to establish both assays for quantification of *SOX1* gene expression. qPCR was performed on Applied Biosystem Fast 7500. SYBR/TaqMan assays were performed in 20µL reaction volume containing SYBR Green/TaqMan® Gene Expression Master Mix (Applied Biosystems, UK), 1µL SYBR Green/TaqMan gene expression assay and distilled water as described below.

2.3.11.1 TaqMan qPCR assays

qPCR was performed on Applied Biosystem Fast 7500. TaqMan qPCR assays were performed in 20µL reaction volume containing 10µL TaqMan® Gene Expression Master Mix (Applied Biosystems, UK), 1µL TaqMan gene expression assay and 5µL distilled water. Serial dilutions (1:10) of NTera cDNA were performed to generate quantitative relative standard curve (from 0.001 to 100 ng) for *SOX1* (Invitrogen Ref: Hs01057642_s1,). Four reference genes available as TaqMan gene expression assays [Act-B (Invitrogen Ref: Hs99999903_m1,), YWHAZ (Invitrogen Ref: Hs03044281_g,), GAPDH (Invitrogen Ref: Hs02758991_g1,) and HPRT1 (Invitrogen Ref: Hs02800695_m1,)] were also used for normalization purposes.

2.3.11.2 SYBRGreen

For SYBR Green, Power SYBR® Green Master Mix (Applied Biosystems, UK) was used. Primer pairs used for *SOX1* (ahS1#5) see Table 3, and GAPDH as a housekeeping gene see [174]. PCR conditions were as follow: 50 numbers of cycles performed and each cycle had hold stage at 94^0C for 5min followed by denaturation step at 94^0C for 30sec and then annealing at specific primer temperatures for 45 sec, followed by extension at 72^0C for 1 min. At the end, Melt Curve analysis was performed on q-PCR product to measure the melting temperature of each q-PCR products and to see if amplification of non-specific or primer dimers has occurred, Melt curve was generated at 95^0C for 10min followed by 60^0C for 1min and then 95^0C for 5min. Product from q-PCR was also run on 2% agarose gel to look for any unspecific bands or primer dimers which can be amplified during the PCR amplification. *SOX1* primer pair ahS1#5 was optimised for SYBR Green assay, ahS1#5 was tested for different primer concentration, annealing temperature and cycle number to optimise the primer conditions for *SOX1* detection by SYBR Green Assay. Both forward and reverse primer pairs were compared at concentrations of 0.5µM, 0.3µM, 0.2µM and 0.1µM. Comparison of annealing temperatures ranges between 50^0C to 60^0C, while 40 and 50 cycles number were also compared.

2.3.12 Relative quantification of SOX1 gene expression by RT-qPCR

Relative quantification of *SOX1* gene expression was performed across all cancerous and normal cell lines by real time qPCR using the $2^{-\Delta\Delta Ct}$ method [176]. *SOX1* gene expression was quantified relative to the expression

level of the calibrator/normal sample ReN cells (human neural stem cells) and was normalised to multiple reference genes (*GAPDH, HPRT1,* and *YWHAZ*). In order to avoid expression variations in different samples, Geometric means values for the reference genes were normalised to a multiple reference genes geometric means values. Multiple reference genes geometric means were then used for normalisation of *SOX1* gene expression in a reference sample (ReN cells).

2.3.12.1 Statistical analysis

For relative gene quantification of *SOX1*, by using qPCR Ct values were normalized to the expression of three reference genes (GAPDH, HPRT1 and YWHAZ). Using $2^{-\Delta\Delta Ct}$ method, median fold changes (Δct) in relative gene expression of SOX1 were calculated in comparison to ReN cells undifferentiated at day 0. Relative fold changes of gene expression displayed with error bars represent + values of relative quantification is defined by the standard error of the Δct's values. For statistical analysis, One way ANOVA multiple comparison test was carried out with 95% confidence interval, Three technical replicates were used (n=3). Statistical software GraphPad Prism 6 was used for data analysis.

2.3.13 Detection of SOX1 gene transcript in a mouse MSCs (mMSC+hSOX1) transfected with human *SOX1* gene

In order to identify a positive control for human *SOX1* protein expression, mouse MSCs transfected cell line with human *SOX1* gene was tested for *SOX1* expression by using human *SOX1* primer (ahS1#5). Human *SOX1* was expressed under the promoter pCAG. The cell line provided by Dr. Virginie

Sottile was grown in a cell culture. Cell pellets were harvested for DNA/RNA and protein extraction subsequently. First of all normal PCR was performed on gDNA in order to detect human SOX1 gene in a mMSC+hSOX1 cell line, which is then followed by RT-PCR to look for cDNA expression.

2.4 Bacterial cultures and cloned DNA purification

2.4.1 Growth media and conditions

Growth media LB (25g LB broth (Miller) in 1 litre of SDW at 37°C) was used for culturing of bacteria (E.Coli, XL10 Gold® Ultracompetent cells, Stratagene, Australia). Antibiotic Ampicillin (Sigma) at a concentration 50µg/ml was added to LB media both solid (agar) and liquid; Solid media was prepared by adding 1.5% w/v bacto-agar (Oxoid), Ampicillin was added at a temperature below 50°C. Agar plates were streaked with IPTG/X-Gal Solution ChromoMax (Fisher Scientific).

2.4.2 Ligation

Ligation was performed by using T4 DNA ligase (Promega, UK) and pGEM-T® Easy Vector (Promega, UK). Total reaction volume was 10µl, containing 1µL pGEM-T® Easy Vector and 3µl DNA fragment (with 3:1 molar ratio of vector and insert DNA respectively), 1µL of T4 DNA ligase enzyme and 5µl 2x Rapid. The reaction mix was incubated for 45 minutes at room temperature, kept on ice and then followed by transformation.

2.4.3 Transformation

β-Mercaptoethanol was added to the cells thawed on ice (4µl per 100µl of cells), flicked and incubated on ice for 10 minutes. In order to transform

the ligated vector into XL10-Gold Ultracompetent cells (Stratagene, Australia), 30µl of competent cells were added into a tube already containing 2µl of the ligation mixture, incubated for 30min on ice, heat-shocked to the cells was given at 42°C for 30 seconds and chilled on ice for 2 minutes. This was followed by addition of 450µl SOC medium into the transformation mix samples and kept at 37°C in a shaking incubator for 60 minutes. LB plates were prepared that were supplemented with 30µg/ml Ampicillin and streaked with IPTG/X-Gal solution ChromoMax (Fisher Scientific) as selection agent for blue/white colonies. After incubation of transformation mix, 50µl and 100µl aliquots for each transformation were streaked on LB plates. Streaked plates were kept overnight at 37°C.

2.4.4 Plasmid DNA purification (Miniprep)

Bacterial colonies grown on LB plates, white colonies (containing insert) were picked up and inoculated in 5ml of LB medium containing 50µg/ml Ampicillin and incubated at 37°C overnight in a shaking incubator. After incubation, 1.5ml cultures were taken and pelleted down by centrifugation for 3 min at 9000rpm. The bacterial pellets were stored at -20°C until subsequent purification of plasmid DNA using the QuickLyse Mini Prep Kit (Qiagen, UK) according to the manufacturer's protocol.

2.5 DNA Methylation Analysis of SOX1 promoter region

SOX1 gene promoter region was analysed for DNA methylation status by bisulphite conversion of DNA. Bisulphite treatment involved treatment of DNA with bisulphite that converts unmethylated Cytosine into Uracil while methylated Cytosine remained unchanged. 500ng of genomic DNA

was subjected to bisulphite conversion using EZ DNA Methylation-Gold™ Kit (Zymo Research, UK) according to the manufacturer's protocol, with the alteration of using autoclaved water (~70°C) instead of Elution Buffer. Bisulphite conversion protocol followed as: 500ng/20µl DNA was added into a master mix containing 130µl CT conversion reagent, 300µl M-Dilution buffer, 50µl M-Dissolving buffer and 900µl autoclaved water. This was followed by incubation in a thermal cycler with the steps: 98°C for 10 min and 64°C for 2.5 hours. After incubation, DNA was loaded into a Zymo-Spin™ IC Column by adding 600µl M-Binding buffer. The DNA was washed once with M-Wash buffer. 200µl M-Desulphonation buffer was added to the Zymo-Spin™ IC column which was followed by an incubation at room temperature for 20 min. The Zymo-Spin™ IC column was washed twice with M-Wash buffer and the DNA was eluted with 10µl autoclaved water (~70°C). The bisulphite converted DNA obtained was immediately stored at -20°C for later use.PCR on bisulphite converted DNA

PCR amplification of bisulphite treated DNA was performed in total volume of 20ul, each sample contained 2µl 10xPCR buffer, 0.15µl Platinum® Taq DNA Polymerase (Invitrogen,), 1.6µl dNTP (2.5mM, Invitrogen), 0.6µl $MgCl_2$ (50mM), 0.8µl bisulphite primer mix (10pmol/µl) and 2µl of bisulphite converted DNA. PCR was performed for 40 cycles, after heating at 95°C for 10 min was followed by denaturation at 95°C for 30 sec, annealing at 55°C for 60 sec and extension at 72°C for 60 Sec which was followed by final 7min extension at 72°C. PCR products were analysed by electrophoresis on 2% agarose gel. For all samples, 10µl of the PCR product was run on agarose gel. Samples that were positive by showing

band on the agarose gel, the remaining 10µl of the PCR products were cleaned up by MinElute® PCR purification kit (Qiagen, UK), section, 2.3.7. Purified PCR product was then cloned into PGEM-T easy vector (section, 2.4).

2.5.1 Bisulphite Sequencing

After cloning and plasmid DNA purification (Minipreps) of bisulphite converted DNA, 100ng/µl of plasmid DNA was subsequently sequenced using the T7F or SP6 primer (SourceBioScience, Nottingham). The electropherograms of the sequences obtained were analysed for quality check on sequence alignment software BioEdit v.7.0.9[177] or FinchTv (http://www.geospiza.com/Products/finchtv.shtml).

2.5.2 Primer pairs used for bisulphite converted PCR amplification

Primer for bisulphite converted DNA was designed using Methyl Primer Express® software v1.0 (Applied Biosystems). Details of the SOX1 primer pair (hSOX1#3BSP) that was designed to amplify promoter region of SOX1 gene has shown below.

Table 4 SOX1 Bisulphite primers sequences

Primer	SOX1 bisulphite primers	Original sequence
Forward	GTTTTGTTAGAAGTTGTAGTTTT	GCCCTGCTAGAAGTTGCAGCCTC
Reverse	AAATAAAAAATTTCTCCTAATACACA	AGGTGGAAAGTTTCTCCTGATGCACA

2.5.3 RT- PCR Primer pairs designed for SOX1-OT:

Different RT-PCR primers were designed to amplify different regions of *SOX1-OT* (Figure 2-1). Primers sequences are shown in the Table 5.

Table 5 Primer pairs used for the *SOX1-OT* amplification.

No.	Sense (F) and anti-sense (R) primer	Sequence
1	F1	TGGAAGTTTCACTCAGCCGT
2	F2	CTTGGCATCTTCTTCCGAGCA
3	F2A	TGGGCAGGCAGGACTTCA
4	F4	ACCAGAGCCGAGGACTAAAC
5	F4a	GACCAGAGCCGAGGACTAAAC
6	F5a	ACCACTCCATTGCAGAAAAGC
7	F6	CCACCCGGTCCGGAATGA
8	F6a	GCAGAGCGTTAGGGGCG
9	F7	TCACTTATCTGGAAACCTGCGG
10	F7a	ATCTGGAAACCTGCGGTTGG
11	F11	ATGTGCAGGACTAAGGCGAC
12	F11a	CTGCGACCACCTACCATCAC
13	F12	ACCCAGGAAAAAGCTACGGG
14	R1	GATAATGACCCCCGGTTCCC
15	R2	GCATGGGCACGACTTGG
16	R3	TTGTTGGTTGCACTACCCCT
17	R4	GCACTACCCCTTCACATCCT
18	R4a	TTCACATCCTACCCCCTCCTT
19	R5	TCAATGTTTATTTCGACTTCCCG
20	R5a	TTATTTCGACTTCCCGGGGC
21	R6	TTACAGTTAGTTCCTCCTCCAGC
22	R6a	ACAGTTAGTTCCTCCTCCAGCTC
23	R7	CTGCGGATTGCAGCGAC
24	R7a	CGTTCGCTGCGGATTGC
25	R11	GTAGGTGGTCGCAGTGAGAG
26	R11a	CTTGCAACTTCCGTGACCAA
27	R12	GACCTCTGCATCCCCTCAAC

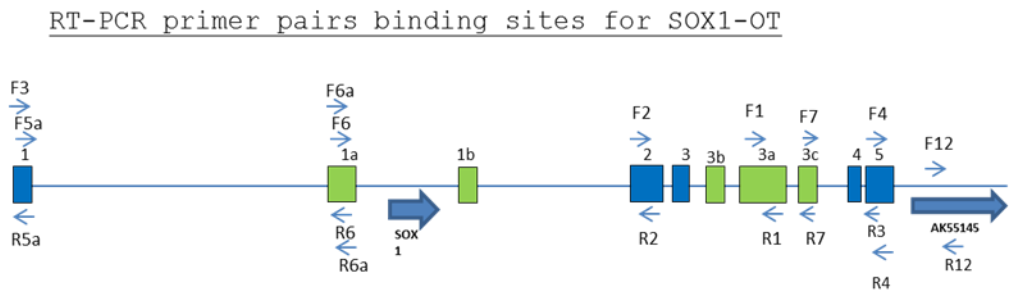


Figure 2-1: **RT-PCR Primer pairs binding sites around SOX1-OT**: Primer pairs used for *SOX1-OT* detection by RT-PCR amplification. Square boxes represent the exons; blue coloured boxes are RefSeq annotated exons while the green coloured ones are novel unannotated exons

identified by the RT-PCR and 5'RACE analysis. Orange coloured arrows represent protein coding genes.

2.5.4 5'RACE (RAPID AMPLIFICATION OF cDNA ENDS)

5'RACE experiments was performed by using the 5'RACE System for Rapid Amplification of cDNA Ends, version 2.0 (Invitrogen, UK). An overview of the 5'RACE process is illustrated in the Figure 2-2.

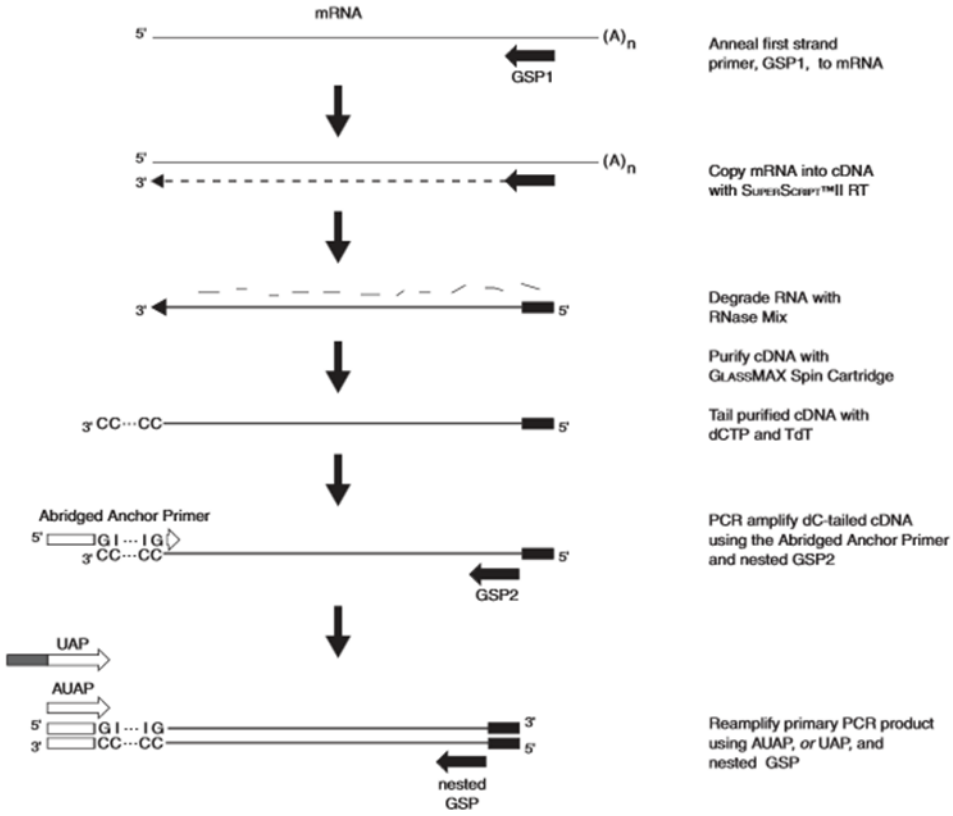


Figure 2-2: **Overview of the 5'RACE:** taken from Invitrogen protocol guide

2.5.4.1 First strand cDNA synthesis:

For the 5'RACE reaction, first strand cDNA synthesis was carried out in a 50µL reverse transcription reaction by using 2.5µg DNAase-I digested RNA (ReN cells-Day6), 200 units/µl of enzyme SuperScript® III Reverse Transcriptase (Invitrogen, UK), 2.5µM of sense *SOX1-OT* primers GSP1 (*GCACTACCCCTTCACATCCT*), GSP2 (*TTCACATCCTACCCCCTCCTT*), nested

GSP3 (*CATCCTACCCCCTCCTTTTGT*), and control primer cGSP1 (provided with a 5' RACE kit) with annealing temperature of 60°C were used for reverse transcription. All the steps were performed according to manufacturing protocol. After cDNA synthesis the reaction was cleaned up by using MiniElute PCR purification kit (Qiagen, UK), each sample was eluted two times with 11µL of elution buffer and then the two elutes were combined together.

2.5.4.2 dC tailing of cDNA:

MiniElute purified cDNA (ReN, D6) was used in the TdT-tailing reaction by using total 25µl reaction volume. A reaction mix containing 5x tailing buffer, 2mM dCTP and 10µl cDNA filled up with water to 24µl was incubated for 3min at 94°C and (immediately) chilled on ice for 1min. 1µl of TdT enzyme (provided with 5' RACE) was added to the reaction mix and incubated for 30min at 37°C. Finally, the TdT enzyme was heat inactivated for 10min at 65°C and placed on ice.

2.5.4.3 GI-Primary PCR amplification:

dC-tailed cDNA was amplified directly by PCR using 10µM Abridged Anchor Primer (AAP), 10µM nested GSP2 primer, 25mM MgCl2, 0.5µL Taq Polymerase (5units/µL) and 5µL dC-tailed cDNA in a total volume of 50µL reaction. After 35 PCR cycles, each consisting of a denaturation step at 94°C for 1min, an annealing step at 60°C for 1min and an extension step at 72°C for 2 min, the final extension was performed at 72°C for 7min. After the PCR amplification, the 5'RACE product(s) was (were) analysed by 2% agarose gel electrophoresis.

2.5.4.4 AUAP secondary PCR amplification:

Secondary PCR amplification was performed on 1µL of the primary PCR product by using Abridged Universal Amplification Primer (AUAP) and GSP3 primer nested to the GSP2 primer used for the primary PCR. Same reaction conditions were used as for GI-Primary PCR; refer to above section (2.5.4.3).

After the secondary PCR amplification, to check for the presence of target *SOX1-OT* variants, RT-PCR was performed on 1µL of secondary PCR product by using different *SOX1-OT* primer pairs from the Table 5.

2.5.4.5 5'RACE product gel purification and cloning into PGEM-T easy vector

After Secondary PCR amplification, 20µl of 5'RACE product was run on a 2% agarose gel with ethidium bromide staining (section, 2.3.10), multiple bands, as expected in case of different transcript variants, were excised and then purified using the MiniElute gel extraction kit (Qiagen, UK). Purified 5'RACE product(s) were then cloned into the PGEM-T easy vector (promega, UK). Formed colonies were tested with overnight liquid culture PCR (see protocol for cloning in section, 2.4*)* and RcoRI enzyme digestion (BioLabs, UK) in order to check for the presence of desired target. All the samples positive for the desired insert (5'RACE product) were send for Sanger sequencing (SourceBioscience, Nottingham).

2.5.4.6 Sequence analysis:

Samples received from the sequencing were first quality checked using the FinchTV software version 1.4.0. After quality check, all the samples were analysed for the presence of EcoRI cleavage sites (GAATTC), the abridged

primers were then removed from each sequence and the remaining sequence was aligned to the human genome (GRCh37/hg19) using the UCSC genome browser [178], images were then generated.

2.6 SOX1 protein analysis by western blot

Different cell lines and tissues were tested for SOX1 protein expression through western blot. Mouse brain tissues (Table 6) were collected post-mortem from animals by Dr Virginie Sottile following Schedule1 sacrifice.

2.6.1 Proteins extraction

. Tissues or cells (Table 1 and Table 6) were processed for total protein extraction. Samples were homogenised according to manufacturer's protocol in RIPA lysis buffer (Sigma Aldrich, UK) or HEPES lysis buffer (25mM Hepes, 134mM NaCl, 1% NP40, 0.1% SDS, 0.5% Sodium-deoxycholate, 5% Glycerol, 100mM Sodium Fluoride and 1mM EDTA), 1x phosphatase/protease inhibitors cocktail were added freshly to the buffer. Samples were followed by Dounce homogenization (for tissues only) and Sonication for 15cylces with 30sec ON/OFF at 4ºC. This was followed by centrifugation at 13,000 x rpm at 4ºC and supernatant obtained were stored at -80ºC. Samples were all the time kept on ice during processing.

Table 6 Listed are the different tissues samples processed for protein extraction.

Tissues	Type/Source	Labelled as
Adult mouse Brain	Wild type	mBrain WT
Adult mouse cerebellum	Heterozygous mouse sox1+/-GFP	mCB+/-
Adult mouse cerebellum	Homozygous mouse sox1+/-GFP	mCB-/-

2.6.2 Bradford Assay

To quantify unknown proteins concentration in each sample, a standard curve was generated through Bradford assay. A Bovine gamma globulin was used as a standard protein (Bio-Rad, UK) along with assay dye

reagent (Bio-Rad, UK). All steps were performed according to the manufacturer's protocols.

2.6.3 SDS polyacrylamide gel electrophoresis

Before loading, total proteins from cell lysates were mixed in 1:1 ratio with 2x Laemmli buffer (BioRad, UK) containing 65.8 mM Tris-HCl, pH 6.8, 2.1% SDS, 26.3% (w/v) glycerol, 0.01% bromophenol blue and 355mM β-mercaptoethanol, was heated at 95ºC for 5 min to denature. 50ug of proteins were loaded into the 10% pre-cast Tris-Glycine gels (Invitrogen, UK) including protein marker (Bio Rad, UK). Gel was allowed to run in Novex® Tris-Glycine SDS Running Buffer 10X (Invitrogen, UK) at 40mA, 125V constant for 90min using commercially available NuPAGE electrophoresis system from Invitrogen.

2.6.4 Electroblotting/Transfer

Proteins on gel were transferred to Nitrocellulose membrane (Invitrogen, UK), performing wet transfer in NuPAGE® Transfer Buffer 20X (Invitrogen, UK) by using commercially available XCell II™ Blot (Invitrogen, UK) Module running at 100mA 25V constant for 2hr. After the transfer membrane was soaked in a transfer buffer for 20min and then transferred protein on membrane was visualized by developing in Ponceau S solution (Sigma) for 10min. Ponceau S solution was removed by rinsing blots in TBS-T (TBS containing Tris 20mM, NaCl 500mM + 0.1%Tween20, pH: 7.4) for 2 min and then soaked in TBS-T for 30min without shaking. This was followed by blocking blots in 5% blocking solution made up in TBS-T for 2 hours at room temperature without

shaking. Membranes were incubated with primary antibody made up in 3% blocking solution (Table 7) and left overnight at 4˚C with slight shaking. 8x washes were performed with TBS-T each for 20 mins. Membranes were then incubated for 1.5hr at room temperature with Secondary antibody (Table 9) made up in 3% blocking solution against each primary antibody. After secondary antibody incubation 8x washes were performed with TBS-T. Blots were developed for chemoluminescent detection by using Immun-Star™ Alkaline Phosphatase Substrate and enhancer kit (Biorad, UK). Images were taken by chemi-doc system (Fuji Film-LAS 4000). Bands were analysed and quantified by Aida Image Analyzer V.4.15.

Table 7 Types of primary antibodies used for western blot are shown with details provided by the supplier.

Antibodies	Host	type	cat#	Supplier
Sox1 SC (C20)	Goat	Polyclonal	sc-17318	Santa Cruz
Sox1 SC (L20)	Goat	Polyclonal	sc-17317	Santa Cruz
sox1 mp	Rabbit	Polyclonal	AB5768	Miilipore
Sox1 abcam	Rabbit	Polyclonal	ab22572	Abcam
Sox1 cs	Rabbit	Polyclonal	4194	Cell signalling
Sox1 sig	Rabbit	Polyclonal	S8318	Sigma-Aldrich
Sox1 mAB	mouse	Monoclonal	MAB3369	R&D Systems
Sox1 abcam	rabbit	Monoclonal	ab109290	Abcam
Lamin B (C-20)	goat	Polyclonal	sc-6216	Santa Cruz
Anti-Actin-B	Mouse	Monoclonal	A5441	Sigma

2.6.5 Testing of different Blocking Solution to improve signal

Different blocking solutions were tested to see which one can give maximised signal with less background and to get rid of non-specific protein detection. Blocking solutions were 5% Fetal Calf Serum (FSC), 5%

Animal Free Blocker (AFB) (Vector), 5% dry Milk (Marvel), 5% Bovine Serum Albumin (BSA) all of them prepared in TBS-T.

2.7 Detection of SOX1 by Immunocytochemistry (ICC)

Immunocytochemistry (ICC) was performed on different cell lines and mouse brain tissues frozen section fixed with 4% paraformaldehyde (PFA) to check for SOX1 protein expression. Cells such as NTera hiMSCs, HOS, HeLa, MCF7, and CACO2 were grown on glass slides in MSC medium, kept inside in incubator at 37^0C with medium changed regularly. When cells became more than 90% confluent then they were fixed in a 4% PFA for 20 min followed by a wash with PBS. Slides were kept at 4^0C in a PBS for later used.

2.7.1 Growing & Fixation of Cells on a glass slide

Cells were grown on glass slides in MSC medium, kept in an incubator at 37^0C with medium changed regularly. When cells became more than 90% confluent then they were fixed in a 4% paraformaldehyde (PFA) for 20 min followed by a wash with PBS. Slides were kept at 4^0C in a PBS for later used.

2.7.2 Preparation of cells for immunostaining

To start Immunostaining, Slides containing cells were immersed in PBS for 5 min to remove any debris.

2.7.3 Antigen retrieval

Antigen retrieval treatment was performed in order to retrieve the hidden antigenic sites for the antibodies. First of all, Steamer filled with distilled

water was preheated to the boiling point. Citrate buffer (10mM Sodium Citrate, 0.05%Tween20, pH 6.0) was preheated in a plastic chamber for 5mins and then slides were incubated in it for 25mins. After incubation, slides were allowed to cool down at room temperature for 30mins and then washed 2X in PBS for 5 min each.

2.7.4 Endogenous peroxidase blocking step (3%H_2O_2)

Endogenous peroxidase blocking step was performed to reduce unspecific staining or background. To prepared peroxidase blocking solution 10mL of 30% H_2O_2 Solution was added to 90mL Methanol. This was followed by incubation of slides in the solution for 10mins and then washed in PBS twice for 5 minute each.

2.7.5 Blocking Step

Cells were blocked in a blocking solution made up of 0.1% Fetal Calf Serum in a PBT (Phosphate buffer Saline+ 0.1% Tween20) for 30mins to prevent unspecific staining.

2.7.6 Primary Antibody

Cells were incubated with 100µL of Primary antibody diluted in blocking solution covered with parafilm and left over night at 4^0C in a humidified chamber. A list of the primary antibodies used for immunostaining is shown in Table 8.

2.7.7 Secondary Antibody

After primary antibody incubation cells were washed with PBT 3x times each for 20mins to remove any unbound primary antibodies. Cells were

incubated with peroxidase conjugated secondary antibodies (Table 9) diluted in PBT for 1 hour at room temperature in a humidified chamber. After incubation, Slides were washed with PBT 3x times for 15 minute each.

2.7.8 Antigen labelling/Development of Slides

DAB (3, 3'-diaminobenzidine) peroxidase substrate kit (Vector Laboratories) was used to developed slides for 2-10min (depending upon brown colour intensity) according to the manufacturer's protocol, followed by washes with PBS 3x for 2min each. Finally, slides were mounted with (4`, 6-Diamidino-2-Phenylindole) Dapi-containing Vectashield (Vector Laboratories).

2.7.9 Image processing

Bright field images were taken using MicroPublisher colour camera 5.0 RTV (Canada) attached to Nikon Eclipse 90i fluorescent microscope using Velocity imaging software version 5.2.1 (2007).

Table 8 List of Primary Antibodies used for immunostaining and different concentration they were tested

Antibodies	Species Raised in	Dilution	Supplier, Cat No:
Sox1 R&D	Goat	1:100,	R&D system, AF3369
Sox1 SC-C20	Goat	1:50,1:100	Santa Cruz, sc-17318
Sox2	Rabbit	1:200,1:500	Active Motif, 39823

Table 9 List of Secondary Antibodies used with concentration and supplier details

Secondary Antibodies used	Dilution	Company, Cat no
HRP anti-goat	1:200	Vector Laboratories
HRP anti-Rabbit	1:200	Vector Laboratories

2.7.10 Retrieval of Human SOX1 protein Sequence

The UniProt database was used to retrieve the Human SOX1 protein sequence in FASTA format (UniProt ID: O00570) [179].

2.7.11 IBS Illustrator for Biological Sequences

The IBS webserver was used for schematic illustration of SOX1 protein sequence figures [180].

2.7.12 ScanProsite tools

ScanProsite which is one of the ProSite database tools, was used to scan for the protein motif XKSExSxxP for matches against the PROSITE collection of motifs. Parameters selected for the Input query were kept to default; the protein database UniProtKB was selected from the options, the search was carried out against human proteins only, and the default option of splice variants was excluded. The result obtained was matched against UniProt protein entries and was analysed by retrieving the Gene ontology (molecular) in the UniProt database. [181]

2.7.13 Multiple Sequence Alignment of SOX1 protein

Multiple sequence alignment (MSA) was performed using Human SOX1 protein sequence in FASTA format. NCBI protein BLAST[182] was used to identify homologous sequences for the Human SOX1 protein in the refseq collection database (refseq_protein)[183], the protein-protein BLAST (blastp) algorithm was selected in the available options. Protein BLAST retrieved homologous sequences with high sequence similarities, Species selected were Mouse (NP_033259), Chimpanzee (XP_016781055), Rabbit (XP_002724136), Whale (XP_004273645), Monkey (XP_017357713),

Mouse (NP_033259), Dog (XP_849239), and Chicken (NP_989664). Sequences in FASTA format were obtained and multiple sequence alignment was performed by using the MSA tool Clustal-Omega from the EMBL-EBI resources [184].

2.8 Post-translational modifications (PTMs) databases

Different online Bioinformatics tools and databases were searched to identify available evidence of potential PTM residues in the human SOX1 protein. Following are the PTM databases which were accessed and used for accumulating evidences of potential PTM residues.

2.8.1.1 PhosphoSitePlus® (PSP)

PhosphoSitePlus® is an online, highly interactive and continuously curated system biology resource for studying PTMs of proteins which are experimentally verified in the regulation of biological processes. PhosphoSitePlus® also provides coverage of commonly studied PTMs including acetylation, methylation, ubiquitination, and O-glycosylation. The PhosphoSitePlus® database was searched to identify already available experimentally verified data about SOX1 PTMs. The PhosphoSitePlus® resources can be accessed at http://www.phosphosite.org/homeAction.do. The search was carried out by protein name SOX1 and human SOX1 (ACC#O00570) was selected from the options given. [185]

2.8.1.2 CBS prediction servers

CBS (The Center for Biological Sequence analysis at the Technical University of Denmark) prediction servers provide sequence and

structure based prediction of eukaryotic protein phosphorylation sites. Following are CBS prediction servers that were accessed for computational analysis of the SOX1 Protein.

2.8.1.3 NetPhos 3.1 server

The NetPhos server predicts phosphorylation sites at serine, threonine and tyrosine residues in a eukaryotic protein using an ensemble of neural networks (a machine learning approach) [186]. NetPhos3.1 server was accessed at http://www.cbs.dtu.dk/services/NetPhos [186]**.** Search for prediction of phosphorylation within SOX1 was carried out by selecting options shown in the Figure 2-3 Error! Reference source not found..

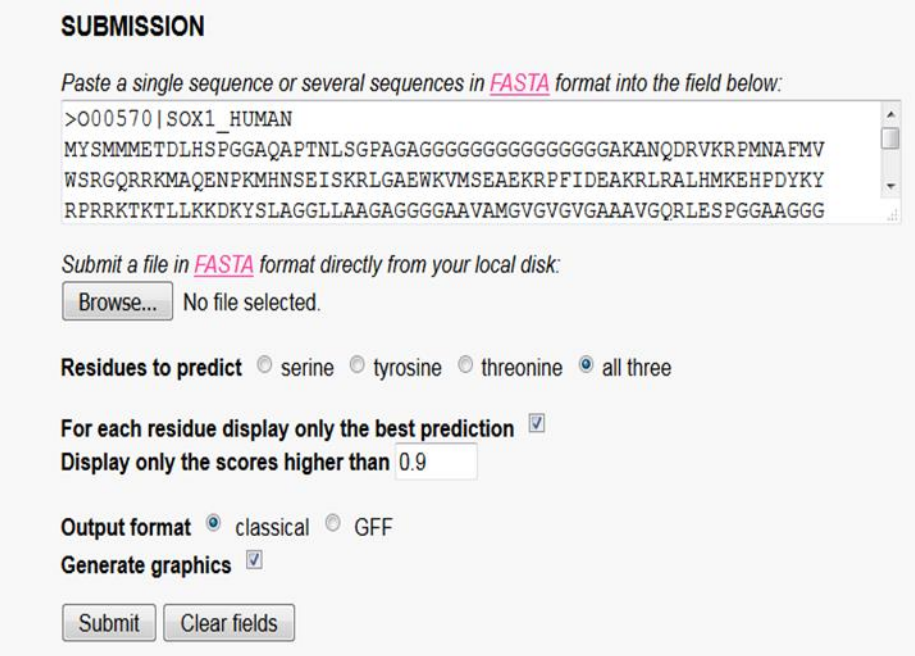


Figure 2-3 **Screenshots of NetPhos 3.1 server Input page**: The amino acid sequence of SOX1 in FASTA format was used to search for all serine, tyrosine and threonine sites, displaying only the best prediction for each residue with a prediction score of equal to or greater than 0.9 [186].

2.8.1.4 YinOYang1.2 server

This server predicts O-GlcNAc (O-glycosylation-N-Acetylation) attachment sites in eukaryotic protein sequences. Amino acid residues in a protein

with cross talk between phosphorylation and glycosylation are called YinOYang site. This server runs parallel with the NetPhos 3.1 server, to mark possible phosphorylated sites and hence identify "Yin-Yang" sites [187]. The weblink for YinOYang 1.2 Server is http://www.cbs.dtu.dk/services/YinOYang. The prediction of YinOYang sites was carried out by selecting option shown in the **Error! Reference source not found.**.

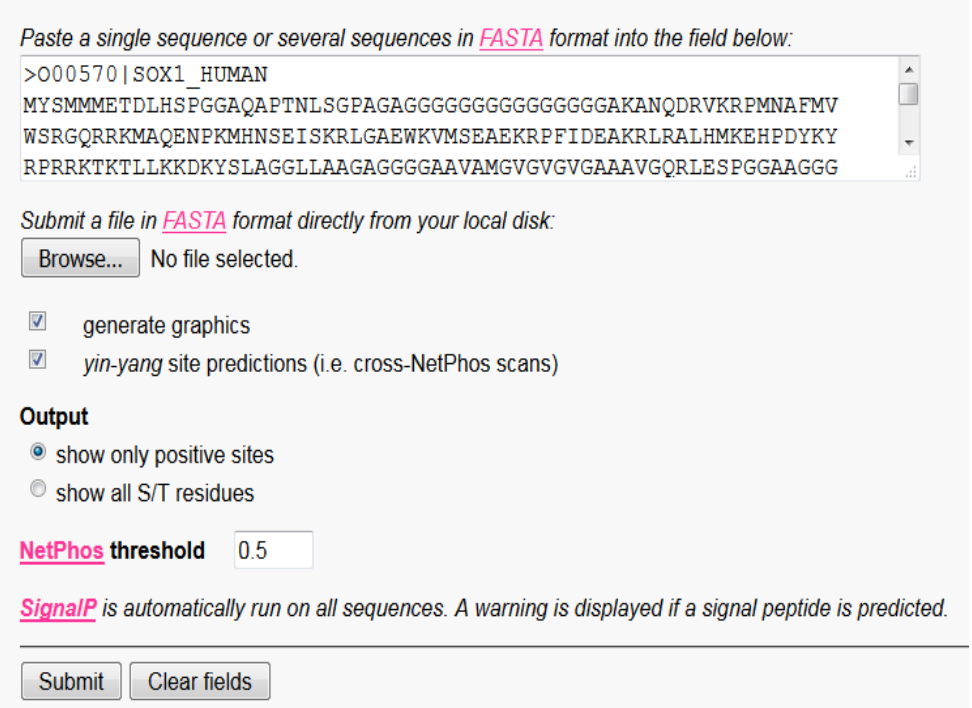


Figure 2-4 Screenshots of YinOYang 1.2 server Input page: To predict Yin-Yang sites in a SOX1, SOX1 amino acid sequence in a FASTA format was given in a search box. Yin-yang site predictions (i.e. cross-NetPhos scans) option was selected and rest left as per default settings. NetPhos threshold value 0.5 remains default as predictions are run in parallel from the NetPhos server [187].

2.8.1.5 Sumoylation prediction databases

GSP-SUMOv2 is a sumoylation prediction database, which is publically available at http://sumosp.biocuckoo.org or can be access through the ExPASy proteomics tools. GSP-SUMO database provide prediction of sumoylation sites and SUMO-interaction Motifs (SIMs) in a given protein sequence. The database has scientific literature, used to manually collect 983 sumoylation sites in 545 proteins and 151 known SIMs in 80 proteins as the non-redundant data sets. [148] GSP-SUMO database webserver was used for prediction of sumoylation sites or SIMs in a SOX1 protein sequence. Query input was SOX1 sequence in FASTA format, the rest of the option was kept as default. Another sumoylation prediction database JASSAv4 was also used for the prediction of sumoylation sites in SOX1 protein. JASSA, a Joint Analyser of sumoylation site and SIMs is designed to define the best sumoylation sites for experimental validation [188]. JASSA is freely accessible at http://www.jassa.fr. The query input was SOX1 protein in FASTA format and the rest of the option was kept as default.

2.8.1.6 DAVID Software analysis

DAVID is a database for Annotation, Visualization and Integrated Discovery (DAVID) that provides a comprehensive set of functional annotation tools for researchers to understand biological meaning behind large list of genes/proteins [189, 190]. DAVID software version 6.8 was used for the functional annotation clustering of the proteins list in the Table 13. DAVID version 6.8 was accessed through a web link (https://david.ncifcrf.gov). Analysis was started by keeping the setting

into default and selecting species *Homo sapiens,* Proteins list was provided in the list manager window, Result was generated by selecting functional annotation tool (Figure 2-5).

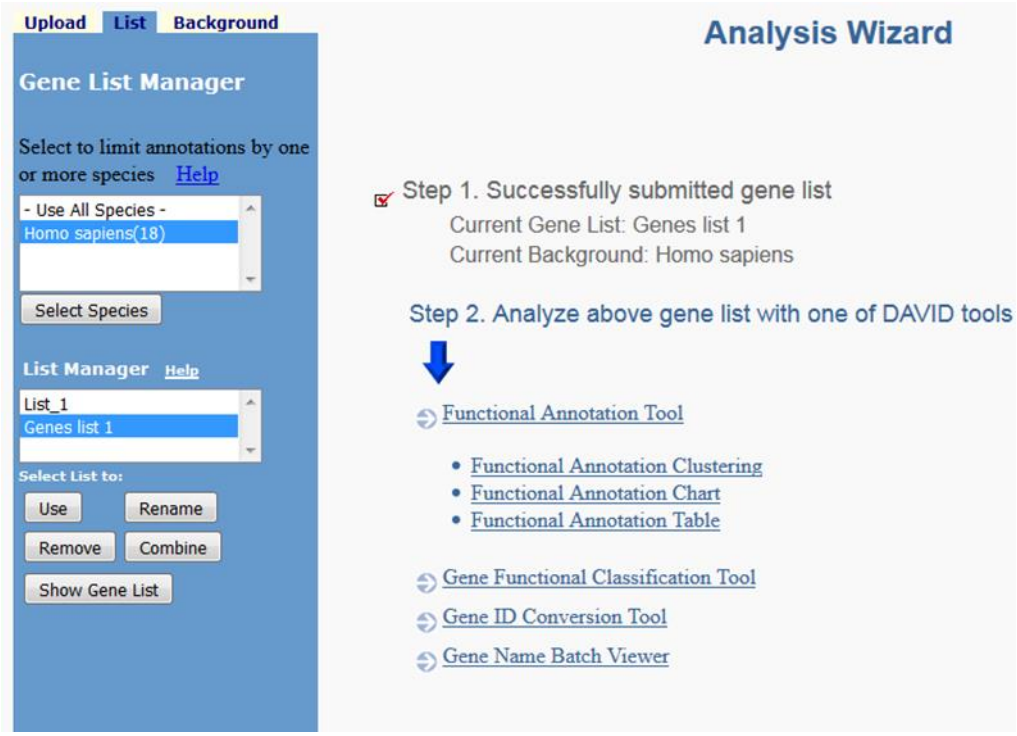


Figure 2-5 **DAVID SOFTWARE query page**: Analysis Wizard was used to generate result by providing desired proteins list.

3 Chapter 03

Analysis of SOX1 gene expression in neural stem cell and cancerous cell lines

3.1 Introduction

Recently, *SOX1* has been shown implicated in several cancer types, Accumulating evidence suggests *SOX1* act as a tumour suppressor gene in many cancers types and demonstrated that *SOX1 in vitro* inhibits tumour invasion [6, 10, 11]. In hepatocellular carcinoma and ovarian cancer, *SOX1* down regulation correlates with poor prognosis and tumour development [6, 8]. In prostate cancer, contrary to its anti-tumour role in other cancers, *SOX1* has been found to act as a oncogene, expressed in metastatic tissue and promote tumour invasion [12]. *SOX1* functions as a tumour suppressor or oncogenes depending upon differences in genetic background, signalling pathways and cellular context [10].

After identification of its role in cancer, *SOX1* has been suggested as a promising methylation biomarker for early detection of cancer [6, 8]. auto-antibodies to SOX1 are common in SCLC, and can serve as serological markers [114]. Therefore, due to the increasing reports of *SOX1* involvement in cancer development, it is important to understand *SOX1* gene regulation in control and disease contexts which will help to refine methods for the detection or treatment of cancer.

3.2 Results

The evidences reported above highlight the importance of SOX1 expression in different cancer types and its DNA methylation profile which has been suggested as promising biomarker for early detection of cancer. However, detailed analysis of SOX1 expression in different cancer types are lacking. Therefore, experiments were performed to analyse SOX1 gene expression in different cancerous and normal cell lines at the RNA and protein levels, In addition promoter DNA methylation pattern of SOX1 gene was also analysed in order to identify any co-relation with its gene expression.

3.2.1 Optimization of different *SOX1* primer pairs for Reverse Transcription PCR:

To study *SOX1* expression in different cell lines, a variety of primer pairs that amplify different region of *SOX1* gene were used as shown in the Figure 3-1**.** Amplified fragments were obtained with the different *SOX1* primer pairs are shown in the Figure 3-2**.** Sanger sequencing of the PCR products revealed that ahS1_2 is an unspecific primer pair that did not amplify *SOX1* while ahS1_3 failed to work. Primer pairs ahS1_1 and ahS1_5 were the only primers that amplified *SOX1* gene (Figure 3-2).

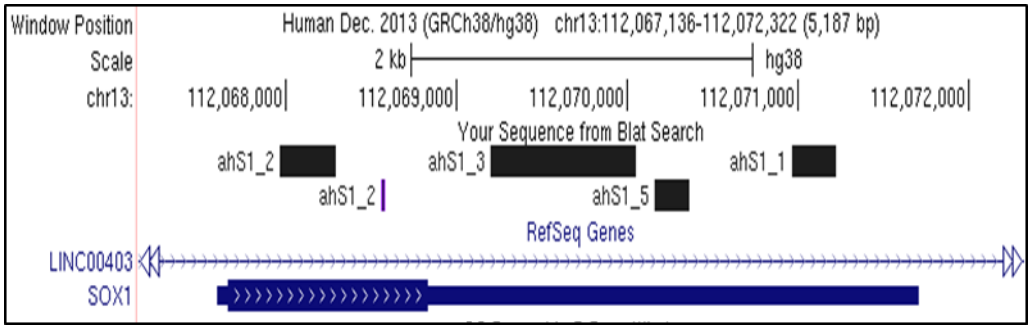


Figure 3-1: **RT-PCR SOX1 primer pairs binding sites**: UCSC browser (http://genome.ucsc.edu*) [178]* generated diagram for different *SOX1* primer pairs that amplify different regions of *SOX1* gene. The Black boxes above the *SOX1* gene (in blue) represent each primer position that bind to the *SOX1* gene and amplify it.

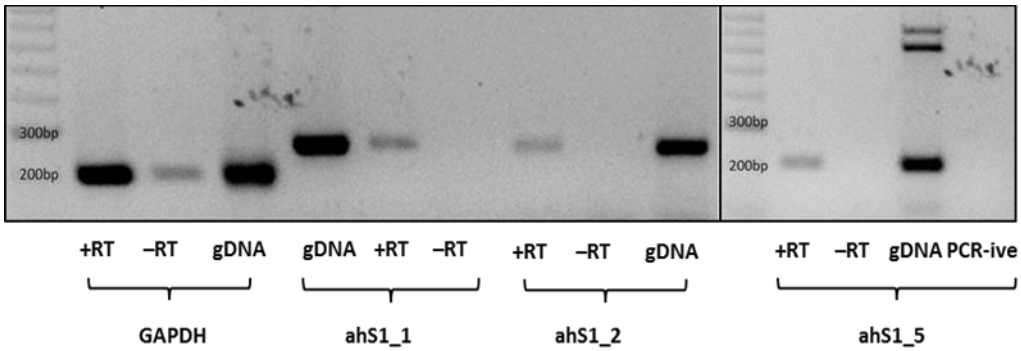


Figure 3-2 ***SOX1* primers optimization:** RT-PCR amplification products obtained through amplification of *SOX1* transcript in ReN cells by different primer pairs. Genomic DNA (NTera) was used as positive control for these primers and –RT (No cDNA) as a negative control. Amplified products for *SOX1* transcript can be seen under +RT (cDNA) lane. *GAPDH* was included as a reference gene. Ethidium-bromide stained 2 % Agarose gel was used.

3.2.2 Detection of *SOX1* gene transcript in a mouse MSCs (mMSC+hSOX1) transfected with human *SOX1* gene:

In order to find a positive control for human *SOX1* gene detection, the mMSCs+h*SOX1* transfected cell line, which carries a transgene for the human *SOX1* protein coding region (CDS), was tested for *SOX1* expression. PCR was initially performed with genomic DNA of the mMSCs+h*SOX1* cell line to check whether h*SOX1* gene was present. Two different *SOX1* primer pairs were used, the human specific (ahS1#5) and as a control the mouse specific (msox1#13). For the human specific primer pair (ahS1#5) band

was detected in the transfected mMSCs+hSOX1 suggesting presence of human *SOX1* gene as shown Figure 3-3, for this primer hiMSCs cell line was used positive control while mouse-NSC (mNSCs) was negative control. A mouse specific primer pair (msox1#13) was used as a control primer that only amplifies the mouse Sox1 gene (Figure 3-3). For the msox1#13 primer pair, mouse NSCs were used as a positive control and hiMSC was a negative control. There was a band in mMSC+hSOX1 sample for msox1#13 primer pair suggesting the amplification of endogenous mouse Sox1 gene copy in the transfected mMSC genome (Figure 3-3).

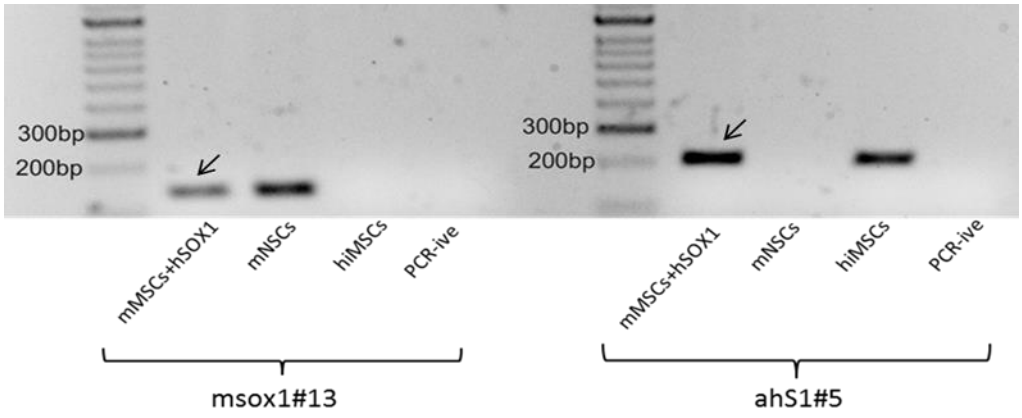


Figure 3-3: **Detection of human *SOX1* in the transfected mouse MSC (mMSC+hSOX1) cell line:** Genomic amplification of inserted Human SOX1 by primer pairs human (ahS1#5) and mouse (msox1#13), amplified product obtained are marked by arrows. Ethidium-bromide stained 2 % Agarose gel was used.

After the PCR test on genomic DNA samples, the mMSCs+hSOX1 cell line was then tested for mRNA expression of transfected human *SOX1* gene by RT-PCR shown in the Figure 3-4. Human specific primers (hsox1#6 and hsox1) did not show human *SOX1* mRNA expression in mMSCs+hSOX1 cell line Figure 3-4A, while the control mouse specific primers (msox1#13 and msox1#9) showed mRNA expression for Sox1 in the mMSCS+hSOX1 cell line (Figure 3-4B), under normal conditions mMSCs do not express Sox1

gene. These results suggest there is *SOX1* mRNA expression in the transfected cell line which is only detected by Sox1 mouse specific primer but not by *SOX1* human specific. Therefore, the transfected mMSCs_hSOX1 cell line was found not reliable to use as a human *SOX1* positive control.

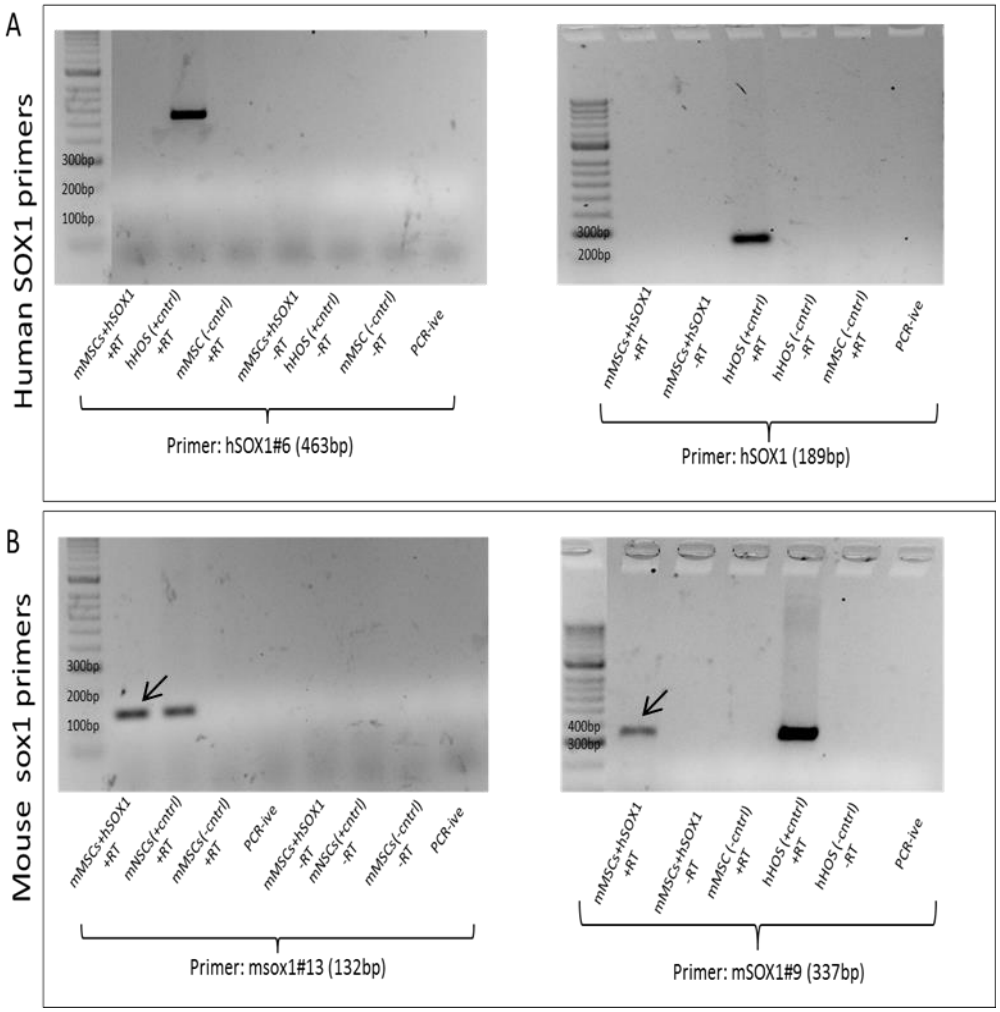


Figure 3-4: **RT-PCR detection on the cDNA of the transfected mouse MSC (mMSCs+hSOX1) cell lines:** (A) RT- PCR by using human primer pairs to amplify human *SOX1* gene in the mouse MSCs transfected cell line. (B) RT- PCR by using mouse primer pairs as a control for the human *SOX1* gene in the mouse MSCs transfected cell line.

3.2.3 Optimization and validation of real time qPCR target (*SOX1*) and reference genes assays:

Before gene quantification by real time PCR, It was important to optimise a method for accurate measurement of *SOX1* gene expression by either SYBR Green (Florence dye) or TaqMan (probe based method). Both methods were optimized as described below in details.

3.2.3.1 Optimization of SYBR Green assay for the real-time qPCR analysis:

In order to perform relative quantification of gene expression, the SYBR Green based method was used to generate standard curves for *SOX1* and the reference gene *GAPDH*. Amplification of serially diluted cDNA for the *SOX1* and *GAPDH* to generate standard curves are shown in the Figure 3-5.

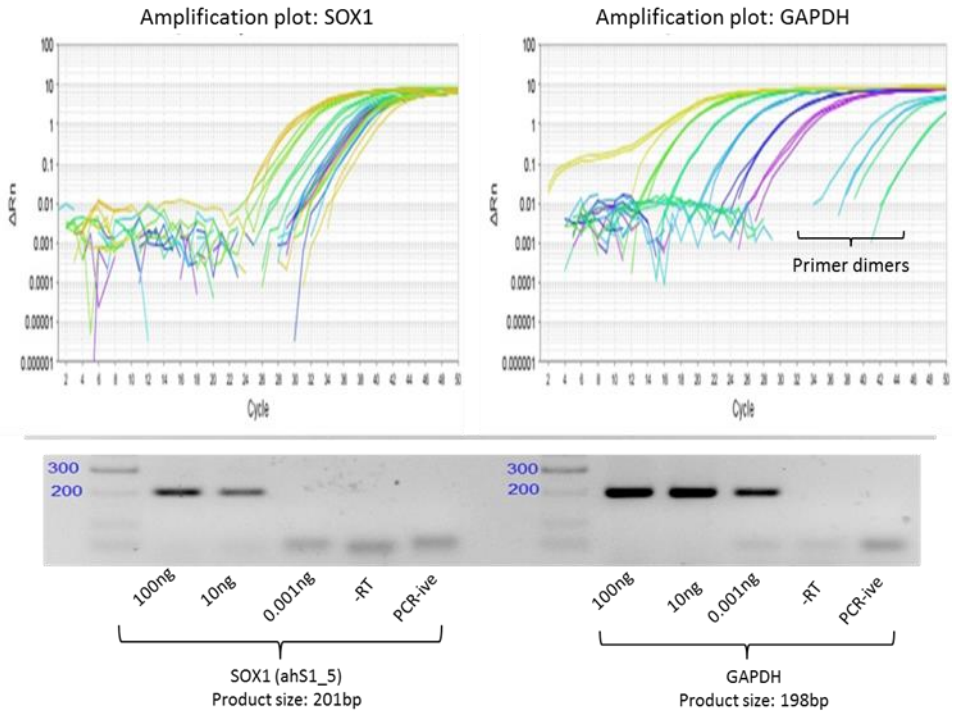


Figure 3-5: **Standard curve by SYBR Green assay**: Serially diluted cDNA amplification by SYBR Green assay in order to generate standard curve for *SOX1* and reference gene *GAPDH*, ΔRn is plotted against the cycle number (ΔRn is the magnitude of the signal generated by the given set of PCR conditions which is Rn minus baseline). After the q-PCR amplification, the

products were run on 2% agarose gel to test for the presence of primer dimers.

SOX1 standard curve showed wide variation between Ct values of each replicates and the slope of the curve was not straight. Primer dimers were formed during the PCR amplification which can contribute to the false fluorescence signal as SYBR Green dye binds to any double stranded structure. This was confirmed by running the product on gel as shown in the Figure 3-5 and melt curve at the end (Figure 3-6).

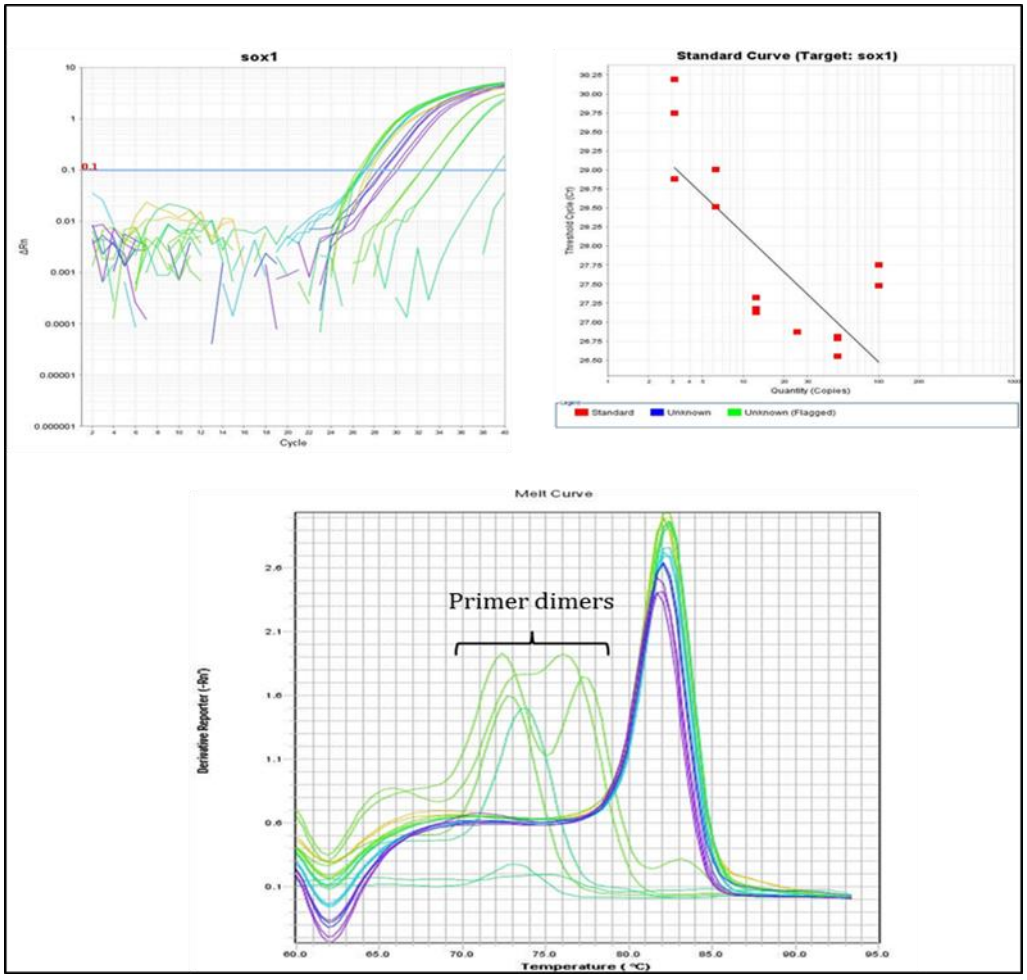


Figure 3-6: **Melt curve analysis of *SOX1* standard curve:** SYBR Green *SOX1* standard curve (HOS cDNA) and melt curve generated at the end.

Different parameters were adjusted for SYBR Green assay. Comparison between different cycle numbers (40 and 50) has shown that 40 cycles had less effect on the primer dimers see Figure 3-7**.** Different Primer

annealing temperatures tested showed that 60^{0}C was the best primer annealing temperature for *SOX1*. Different *SOX1* primer concentrations (0.5, 0.3. 0.2 and 0.1µM) showed that 0.1µM concentration for both *SOX1* forward and reverse primer gave less primer dimers without compromising the signal intensity (Figure 3-7).

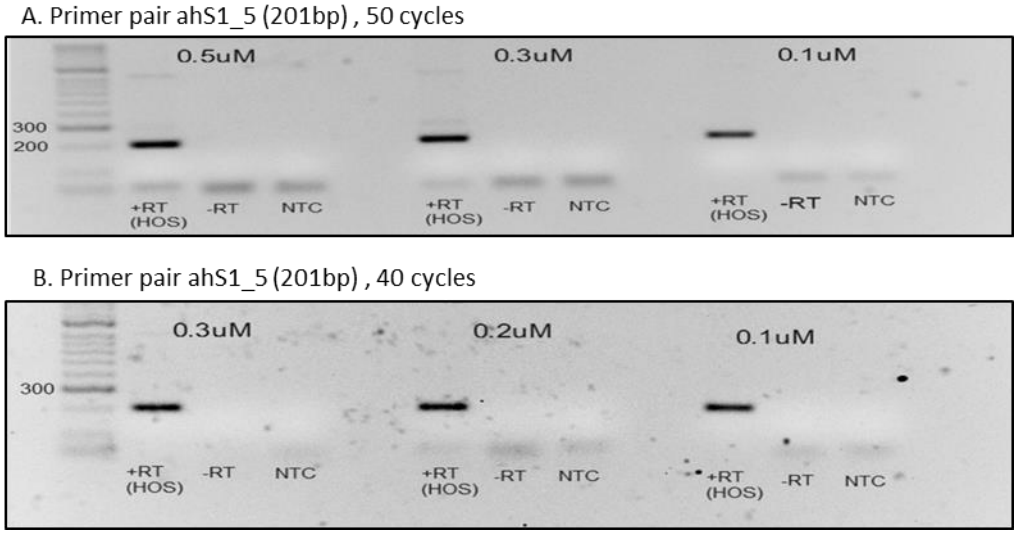


Figure 3-7:. **Primer concentration optimization for SYBR Green Assay:** (A) Different primer concentrations for the ahS1_5 forward and reverse primers with 50 cycles number PCR. (B) 40 cycles number PCR with different Primer Concentrations. SYBR Green PCR products were run on 2 % agarose gel.

3.2.3.2 Optimization of Probe-based TaqMan assay for the real time qPCR analysis:

TaqMan probe-based detection for *SOX1* gene expression analysis was optimised at different steps starting from RNA preparation till cDNA clean up. Previously generated *SOX1* standard curve for q-PCR had problems with very low efficiency and wide distribution of Ct values between technical replicates. This problem was tackled through many steps to remove PCR inhibitors either coming from RNA or cDNA preparations, which were causing difficulty to generate *SOX1* standard curve for the relative quantification of standard curves. This was the case with

reference genes as well, such as GAPDH which was tried multiple times to generate standard curves with acceptable PCR efficiency and R^2 value.

First of all, *SOX1* TaqMan assay was tested on genomic DNA extracted from HOS cell line which had been cleaned up using Qiagen Mini-Elute columns. It was found that the assay for the *SOX1* standard curve gave 100.9% PCR efficiency with R^2 value of 0.991 in the serial dilution between 120ng and 3.8ng of gDNA (1 in 2 dilutions) see Figure 3-8A. This was followed by cDNA clean up with MiniElute columns for *SOX1* and GAPDH genes (Figure 3-8B). It was found that the GAPDH standard curve was 99% efficient with R^2 value of 0.996, while the *SOX1* standard curve was 73.50% efficient with R^2 value of 0.905, within the same range see (Figure 3-8B). cDNA clean up by columns showed less variation in ct values for *SOX1* and improved standard curve with PCR efficiency of 73.50%, which is within acceptable range. Therefore, It was suggested that cDNA clean up by columns could improve the PCR results

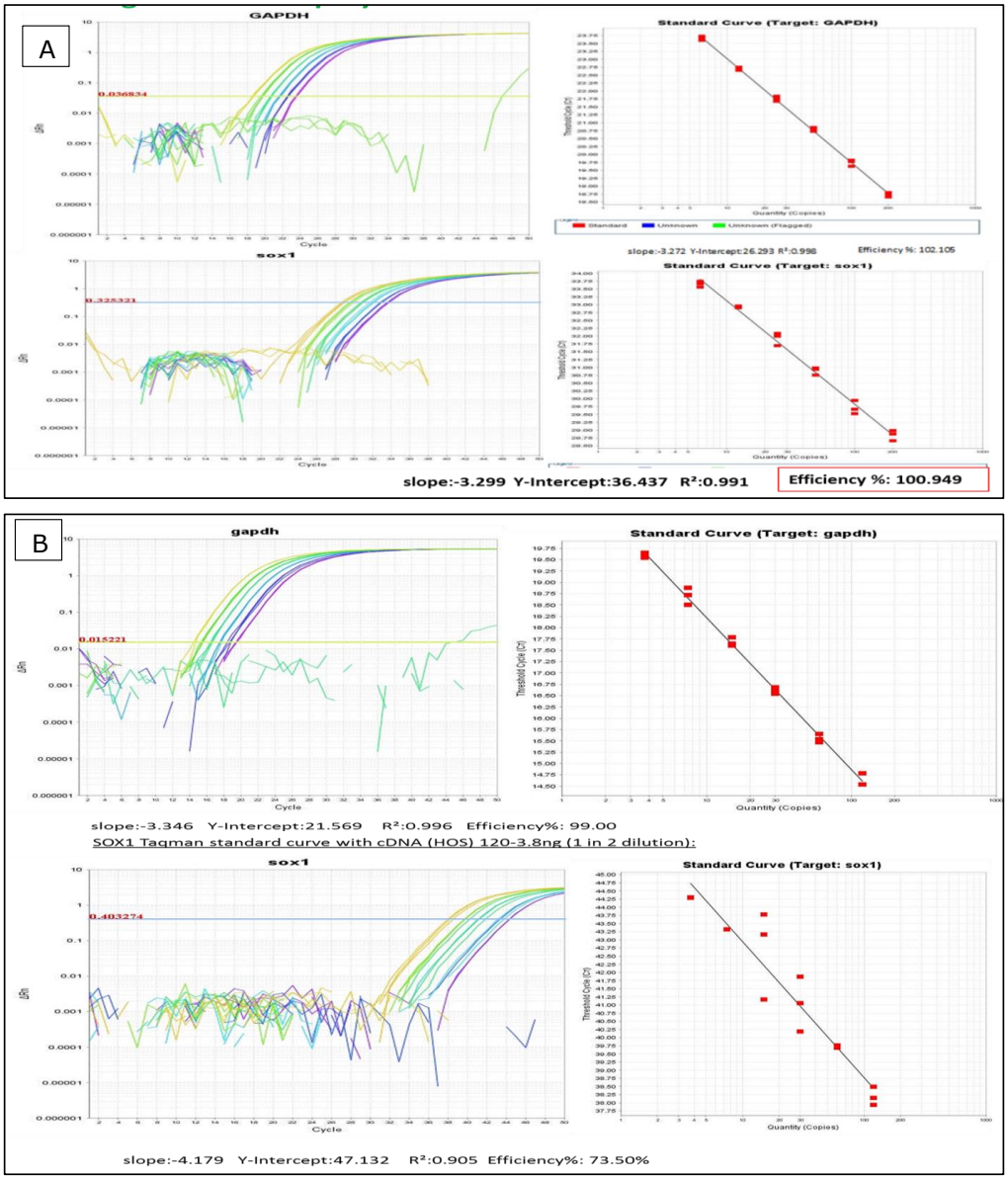


Figure 3-8: **TaqMan generation of Standard curve**: HOS gDNA (A) and cDNA (B) cleaned up by columns were used to generate TaqMan *SOX1* standard curves with reference gene GAPDH, cDNA starting quantity was 120ng diluted to the 3.75ng (1 in 2 dilution).

Before cDNA clean-up for qPCR there was still a need to improve *SOX1* assay by considering others parameters such as RNA preparation. Therefore, three different experiments were carried out to do a comparison between differently prepared cDNA samples, before generating TaqMan standard curves with each cDNA. Frist, *SOX1-Ia* was a cDNA prepared by only RNA clean up before cDNA synthesis. Secondly,

SOX1-Ib was a sample with both RNA and cDNA clean up and lastly, *SOX1-II* was a cDNA only cleaned up after cDNA synthesis. *SOX1* TaqMan PCR results showed that cDNA clean up by MiniElute column was the best option to consider for generating *SOX1* standard curves (Figure 3-9) for the relative quantification of *SOX1* gene expression in different cell lines.

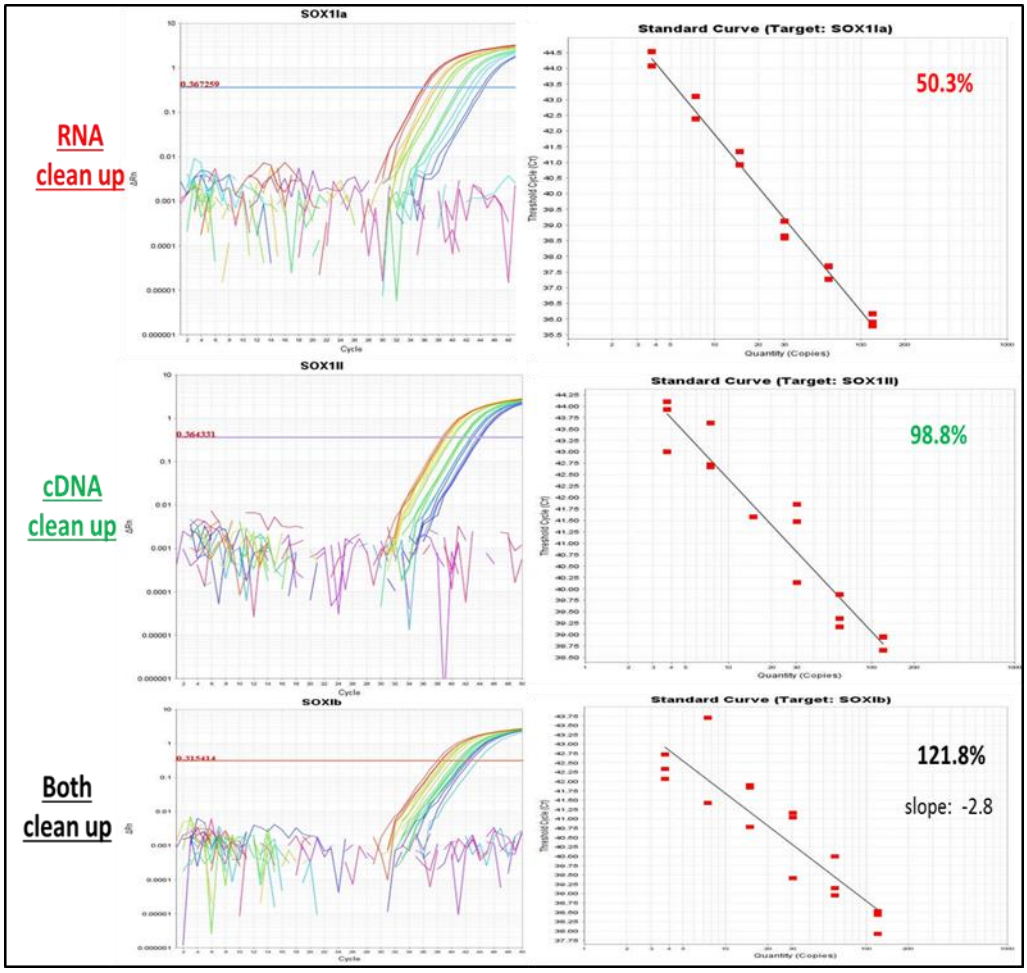


Figure 3-9: **RNA clean up experiments for real time qPCR**: Comparison between differently processed cDNA before TaqMan standard curve generation for the *SOX1* gene.

3.2.4 Relative quantification of *SOX1* gene expression in different cancerous and normal cell lines by qPCR:

3.2.4.1 Generation of Standard curves for *SOX1* and reference genes:

PCR assays validation was performed prior to quantification by qPCR for the target (*SOX1*) and reference genes. Four different reference genes (*GAPDH, HPRT1, YWHAZ* and *ACT-B*) were selected based on published data in literature. Standard curves were performed on target and all reference genes to validate PCR efficiencies for each assay and to find out optimal amount of starting cDNA for the samples to be quantified. Standard curves by serial dilution of cDNA (ReN cells) from 200ng till 6.25ng (1 in 2 dilutions) were generated as shown in Figure 3-10 and Figure 3-11. The R^2 value, slope and efficiency of the assay are shown in the Table 10. *GAPDH, HPRT1* and *YWHAZ* showed less variation in the Ct values and were in an acceptable range of efficiency with R^2 value close to 1.

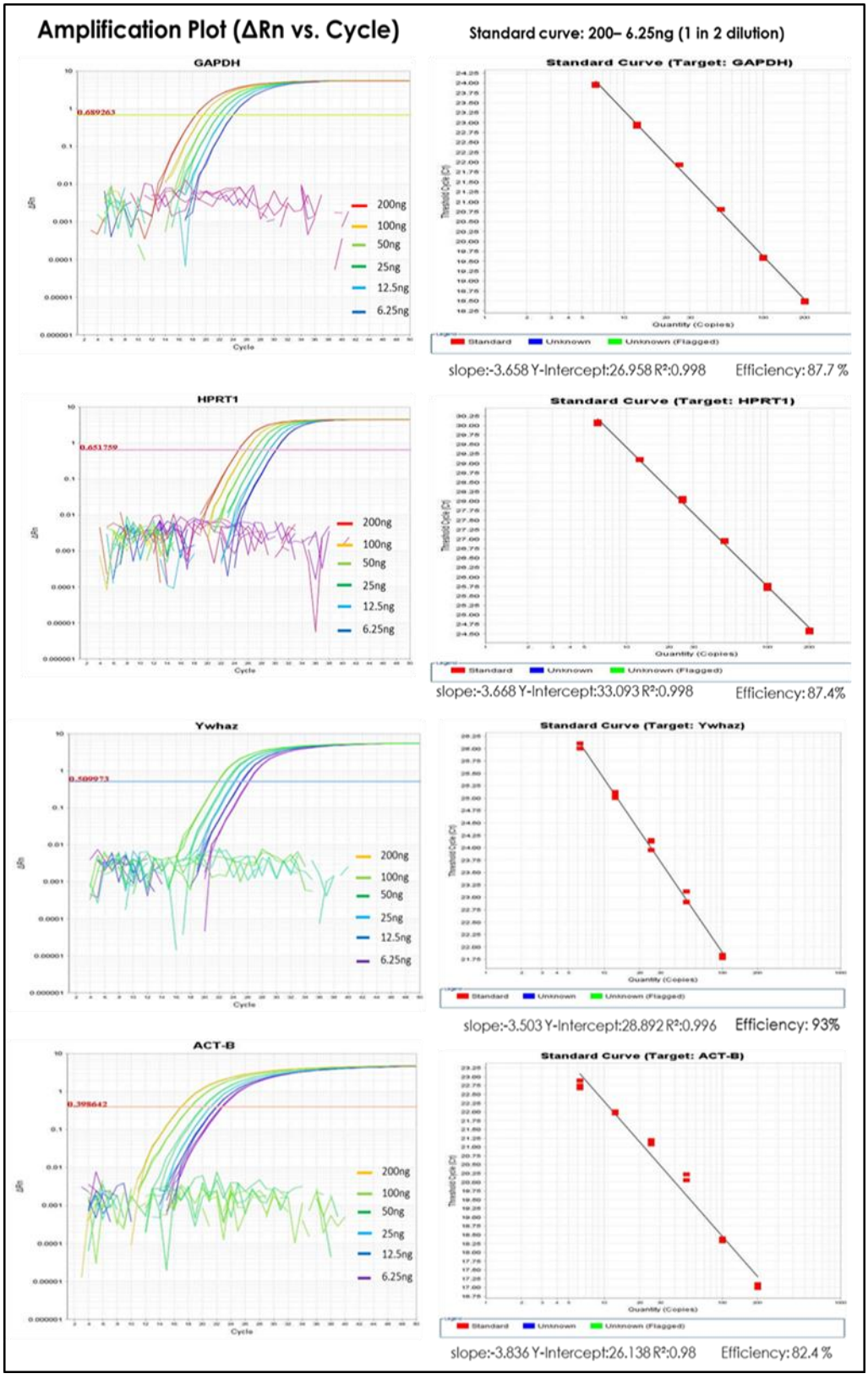


Figure 3-10: **TaqMan standard curves of reference genes**: Standard curves (TaqMan) for different reference genes showing amplification graph generated by 1 in 2 serial dilution of hNSCs starting from 200ng to 6.25ng.

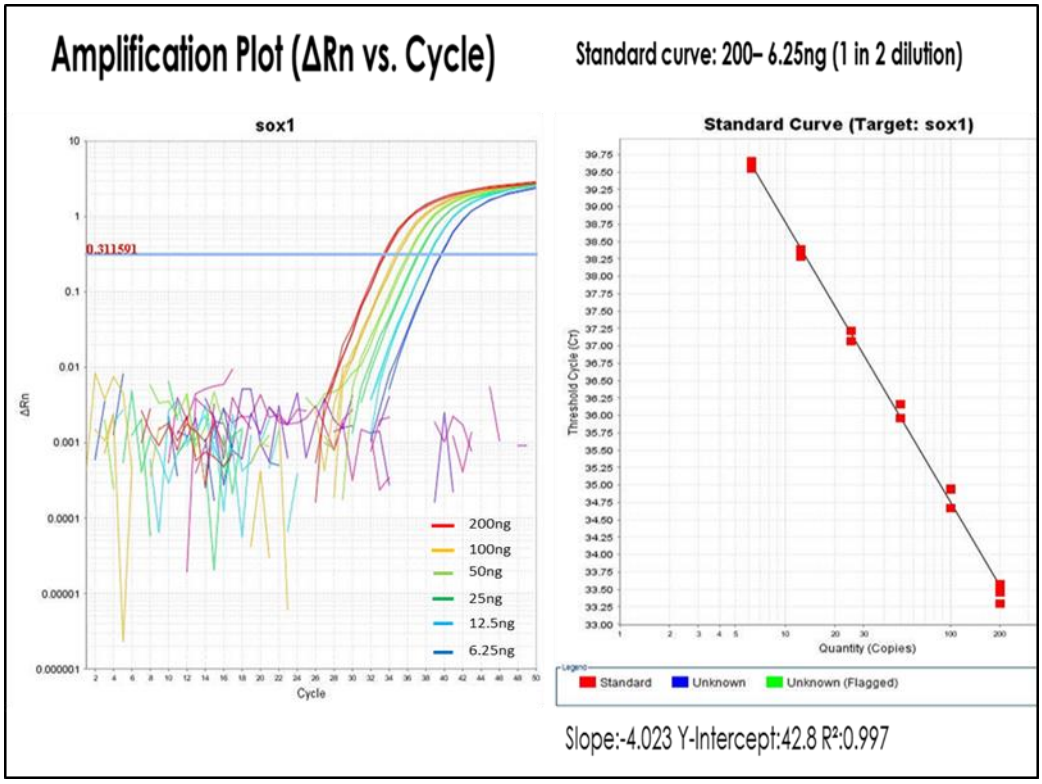


Figure 3-11: **TaqMan standard curves of *SOX1* gene gene:** *SOX1* (TaqMan) standard curves with amplification graph generated by 1 in 2 dilutions of hNSCs starting from 200ng to 6.25ng

Table 10: Slope, R^2 value and efficiencies of the standard curves for *SOX1* gene and different reference genes generated by TaqMan assay real time PCR

TaqMan Assays	Slope	R^2	Efficiencies
GAPDH	-3.658	0.998	87.7%
HPRT1	-3.668	0.998	87.4%
ACT-B	-3.836	0.980	82.4%
YWHAZ	-3.503	0.996	93%
SOX1	-4.023	0.997	77.3%

SOX1 standard curve (TaqMan) produced a straight curve with an R^2 value of 0.997 and showed a PCR efficiency of 77.3%. This was considered efficient for the generation of *SOX1* standard curve by a cDNA template (Table 10), and the efficiency not equal to 100% was attributed to low

expression of *SOX1* and to the fact that its Poisson distribution in a sample would contribute to the lower limit of detection of expression.

Three reference genes *GAPDH, HPRT1* and *YWHAZ* were selected as reference genes according to MIQE guideline with minimum information needed for publication [191] *ACT-B* standard curve showed high standard deviation between technical replicates, the slope of the curve was not straight enough (R= 0.98) and the efficiency was low therefore *ACT-B* was excluded from further studies.

3.2.4.2 RT-PCR for *SOX1*:

Before quantification of *SOX1* gene expression by qPCR, all of the cell lines were first analysed to detect *SOX1* gene transcript by reverse transcription PCR (RT-PCR). Results obtained are shown in the Figure 3-12**.** It was found that ReN cells (human neural stem cells) were positive for *SOX1* gene expression as expected. Bands were obtained for *SOX1* in NTera, MCF7, SH-SY5Y, T47D, HuES7 and HOS cell lines while no bands were obtained in HeLa, HCT116, MDA-MB-231, MDA-MB-361, Hs578T, MRC5 and CACO2 cell lines. Reference gene GAPDH was expressed across all cell lines confirming that the lack of *SOX1* expression was not due to failed reverse transcription (Figure 3-12).

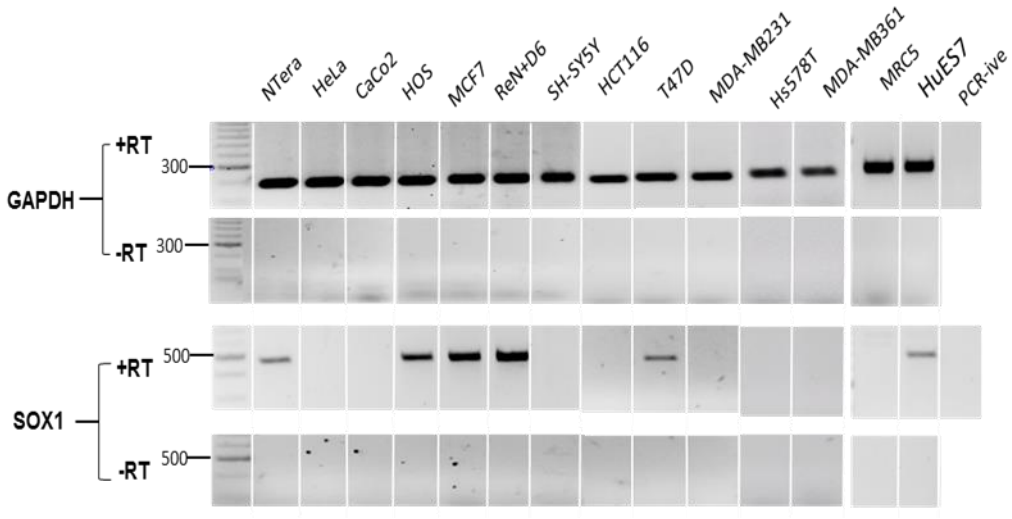


Figure 3-12: *SOX1* gene expression by RT-PCR: RT-PCR gel picture showing *SOX1* bands for +RT (cDNA) samples. GAPDH was used as reference gene for all the samples.

After RT-PCR analysis, relative quantification of *SOX1* gene expression was performed across all cancerous and normal cell lines by real time qPCR using the $2^{-\Delta\Delta Ct}$ method. *SOX1* gene expression was quantified relative to the expression level of the calibrator/normal sample ReN cells (human neural stem cells) and was normalised to multiple reference genes (*GAPDH, HPRT1,* and *YWHAZ*). Variable *SOX1* gene expression levels were found across the different cell lines tested (Figure 3-13B). Cancerous cell lines such as NTera, HOS and T47D were found to express *SOX1* at low level relative to the *SOX1* expression in ReN cells. MCF7 was found to express *SOX1* slightly less than one fold relatively to the *SOX1* expression in ReN cells. No *SOX1* expression was found in different cancerous cell lines such as CACO2, HCT116, Hs578T, HeLa, and MDA-MB361/231. No amplification was found in the –RT samples and negative controls confirming that there had not been any contamination with DNA. The calibrator sample (HOS) used for plate to plate variation showed no

drastic variation between the Ct values for each assay used in the experiment as shown in Figure 3-14.

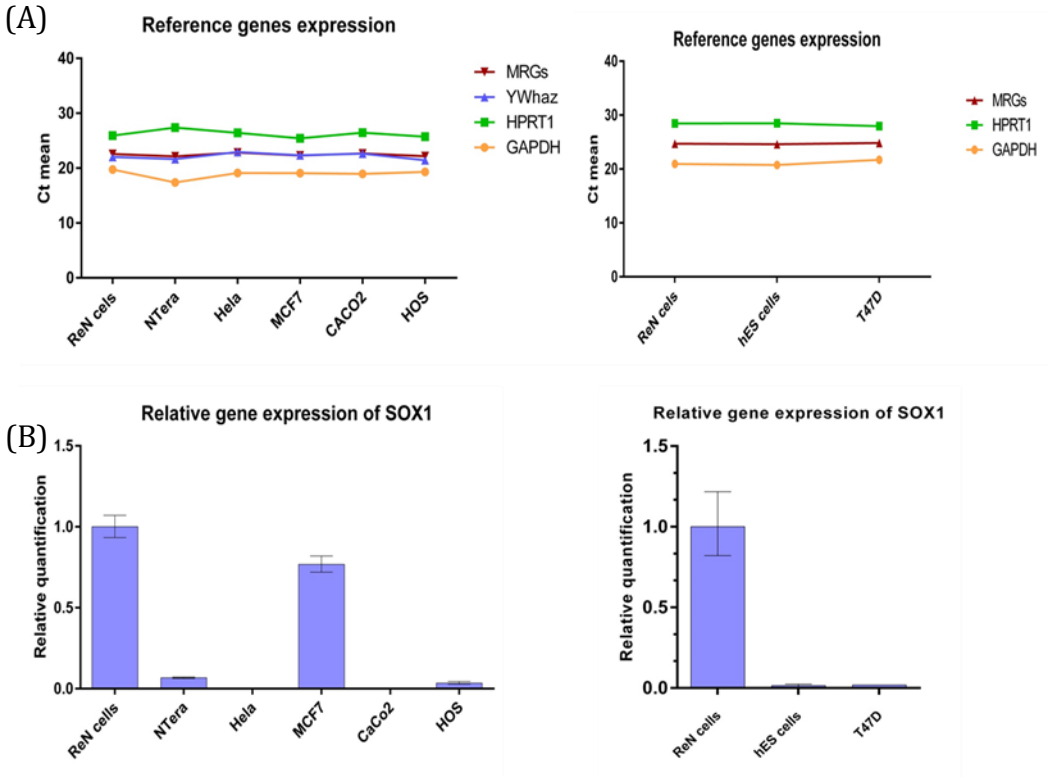


Figure 3-13: **Graph representation of relative gene expression of *SOX1* and reference genes** (A) Multiple reference genes (MRGs) Ct means across all the cell lines. (B) Relative quantification (RQ) of *SOX1* gene expression across different cancerous cell lines relative to *SOX1* expression in ReN cells, Error bars represent the standard error of the Δct's values.

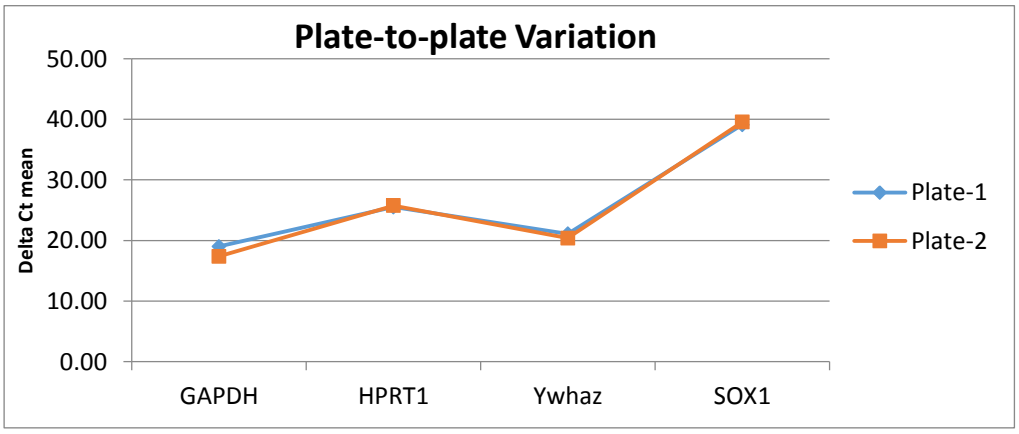


Figure 3-14 **Graphic representation of qPCR plate to plate variation:** Showing plate to plate variations for a calibrator sample (HOS) for each assay used

3.2.5 *SOX1* gene promoter DNA methylation pattern

To assess whether the level of expression was related to DNA methylation, DNA methylation pattern of the *SOX1* gene promoter was determined for different cancerous and normal cell lines. Direct sequencing of amplification products of PCR performed on bisulphite treated DNA from different cell lines was used to generate DNA methylation pattern of the *SOX1* promoter CpG Island (Figure 3-15). The region of the *SOX1* promoter analysed was found to be differentially methylated across different cell lines. Comparing *SOX1* promoter methylation status with its gene expression profile it was also found that *SOX1* promoter methylation correlates to its gene expression level. Cell lines such as HeLa, Hs578T, MDA-MB-361, CaCo2 and HCT116 and hMSCs which do not express *SOX1* were found to be highly methylated at promoter region (Figure 3-15). By contrast, cell lines such as ReN cells before (D0) and after neural differentiation (D6), HUES7, NTera, MCF7, T47D and HOS that do express *SOX1* were found to have relatively lower levels of methylation at the *SOX1* promoter than non-expressing cell lines (Figure 3-15). Therefore, it is likely that *SOX1* expression is regulated at the epigenetic level in these cancerous cell lines. Future studies will be required to determine the relationship of the differential methylation status of *SOX1* to the different cancerous phenotypes and medical outcome to establish its potential as a biomarker in cancer.

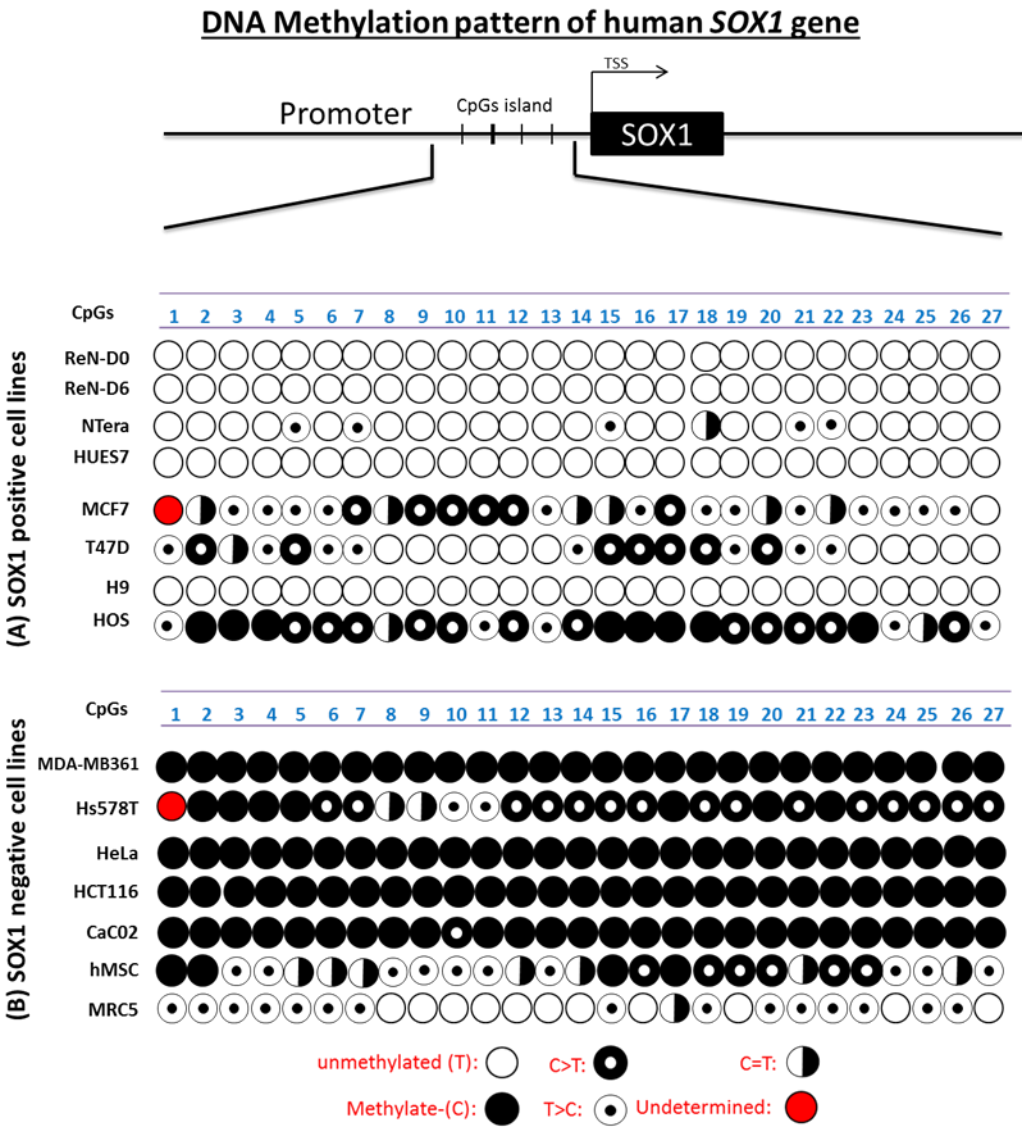


Figure 3-15: **Illustration of *SOX1* gene Promoter DNA methylation pattern obtained through direct sequencing:** (A) CpGs Island DNA methylation pattern of *SOX1* positive cell lines, Each CpGs represented as a circle. (B) CpGs Island DNA methylation pattern of *SOX1* negative cell lines

3.2.6 *SOX1* gene expression across different time points of human neural stem (ReN) cell differentiation.

SOX1 plays important role in neural differentiation and it is the earliest known marker for neural stem cell. To characterize the dynamics of *SOX1* expression in neural stem cell differentiation, *SOX1* gene expression was analysed on ReN cells RNA at day 0, 2, 4 and 6 of differentiation. ReN cell pellets provided by a lab member (Dr. Stephanie Strohbuecker) were processed for RNA preparation. First of all, gene specific endpoint RT-PCR was performed to check *SOX1* expression (Figure 3-16)**.** *GAPDH* was used as a ubiquitously expressed gene as a positive control for reverse transcription. *SOX1* was found to be expressed across all different time points of neural differentiation of ReN cells.

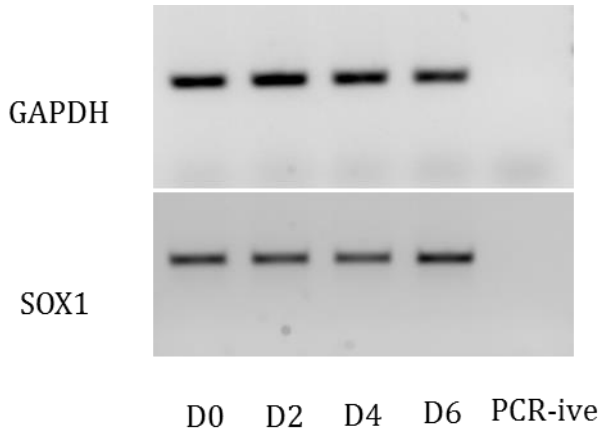


Figure 3-16: ***SOX1* gene expression across different time points of human neural stem (ReN) cell differentiation**: Gene specific RT-PCR across different time points of ReN cell differentiation for *SOX1* and reference gene GAPDH.

Quantitative real time PCR was then performed on the cDNA to study whether levels of *SOX1* RNA changed during differentiation. As previously shown (Figure 3-10, Figure 3-11), three reference genes were used (*GAPDH, HPRT1 and YWHAZ*) for quantification, Reference genes

expression was combined using geometric mean into multiple reference genes (MRGs) expression (Figure 3-17) which was then used for the quantification of *SOX1* RNA.

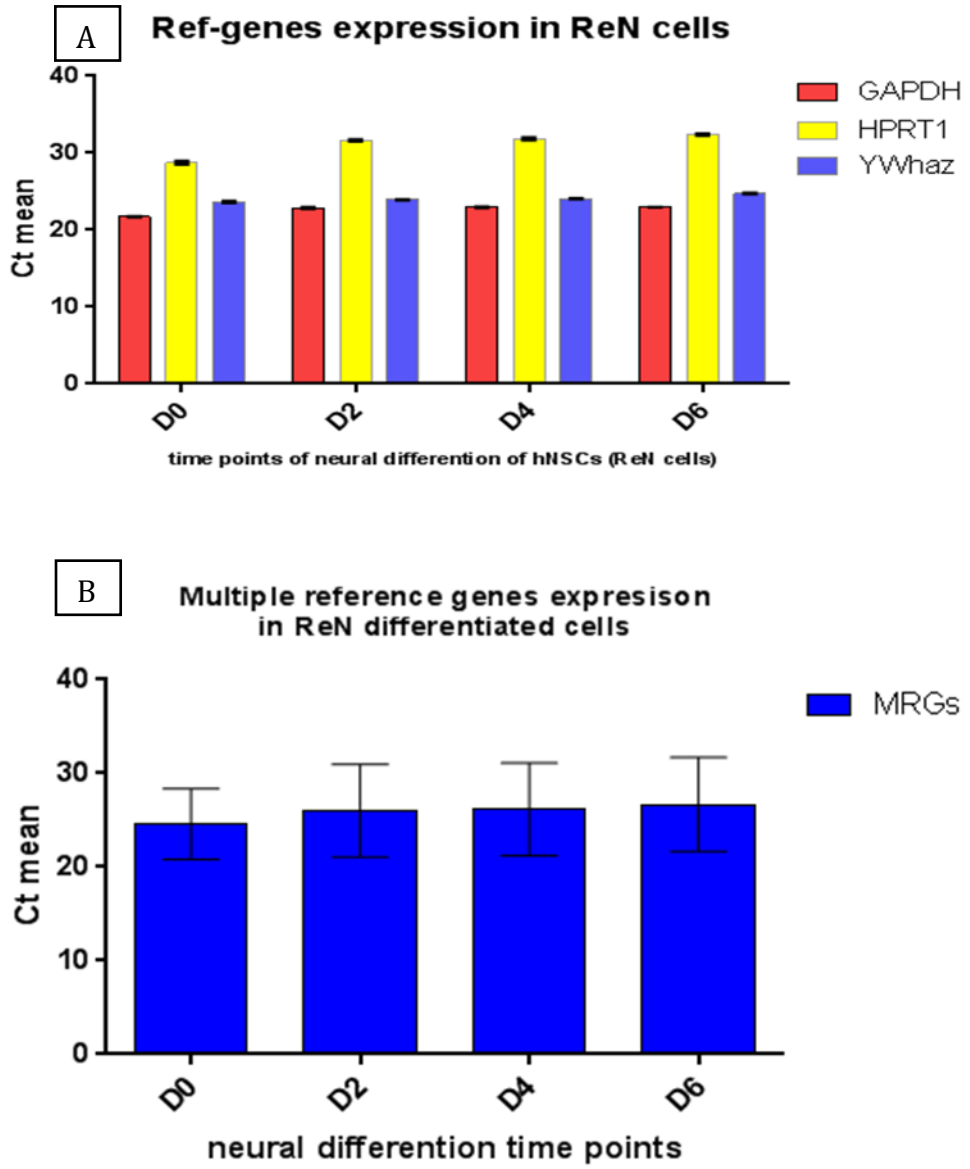


Figure 3-17: **Quantitative real time PCR gene expression for reference genes across different time points of ReN cell differentiation:** (day0, 2, 4 and 6) (A) Relative gene expression of three different reference genes (*GAPDH, HPRT1 and YWHAZ).* (B) Gene expression of three reference genes was normalized into multiple reference gene expression (MRGs). Error bars represent standard deviation (SD).

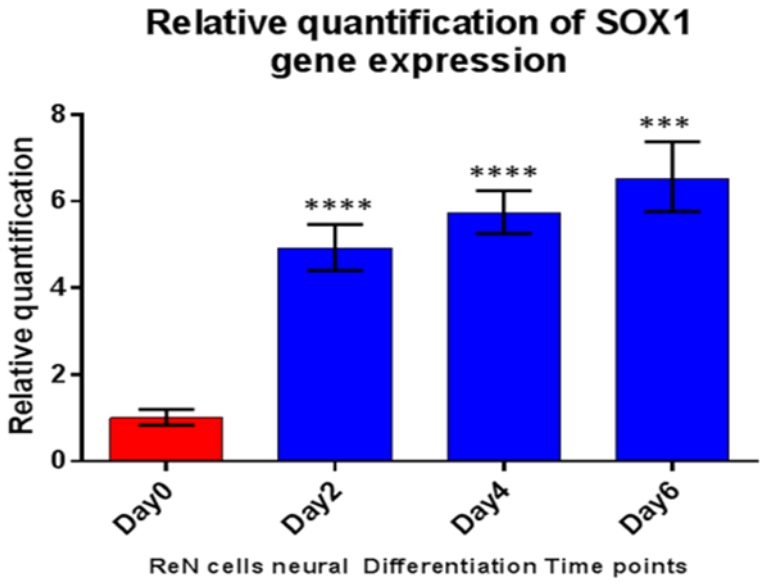


Figure 3-18 **Relative quantification of *SOX1* gene expression analysed by qPCR across different time-points of ReN cells differentiation**: (day 0, 2, 4 and 6). qPCR Ct values were normalized to three reference genes expression (*GAPDH, HPRT1 and YWHAZ*. Median fold changes (delta Ct) in relative gene expression of *SOX1* in comparison to ReN cells undifferentiated at day 0 ($2^{-\Delta\Delta Ct}$), Error bars represent the standard error of the Δct's values. Statistical test one way ANOVA was performed that showed *SOX1* was significantly UP-regulated at day 2, 4 and 6 in comparison to D0. , n=3, ****P value <0.0001, 95% confidence interval.

SOX1 gene expression at different time point of neural differentiation was quantified relative to MRGs and expressed relative to its gene expression at day0 of ReN cells. It was found that *SOX1* mRNA was significantly up regulated at day 2, 4 and 6 of neural differentiation compared to Day 0 (Figure 3-18).

3.2.7 Immunostaining approach to detect SOX1 signal in different human cell lines:

Immunostaining was performed using SOX1-SC antibody to detect SOX1 protein in human cell lines grown on glass slides. The experiment aim was to identify human SOX1 positive cell line that can be used as a positive control for future experiments. Human cell lines, HOS and MCF7 were tested for SOX1 protein expression as *SOX1* mRNA transcript was previously identified by qPCR in these cell lines (figure 3.13). HeLa and CaCo2 cell lines were included as a negative control for SOX1 protein expression on the basis of its qPCR results (figure 3.13) while hiMSCs under normal conditions is negative for SOX1 expression. Mouse cerebellum tissue highly express SOX1 and included as a positive control. Work in the lab previously performed by another PhD student demonstrated that, a SOX1 goat polyclonal antibody (from Santacruz, 'SC') was successful in detecting SOX1 in mouse brain tissues by immunofluorescence, therefore, mouse brain cerebellum tissue fixed on the glass slides were used as a mouse SOX1 positive control for the anti-SOX1 antibody.

Results from the immunostaining showed strong SOX1 positive staining in Bergmann glia cells located in the Purkinje cell layer (PCL) of mouse brain cerebellum tissue (Figure 3-19), indicating immunostaining experiment has worked. Human cell lines negative controls (HeLa, CaCo2 and hiMSCs) showed no SOX1 staining as expected. In the target cell lines HOS and MCF7 no SOX1 signal was detected (Figure 3-20) and therefore they cannot be used as a SOX1 positive control for future experiments. It has

been suggested that HOS and MCF7 cell lines which expressed *SOX1* mRNA transcript might not actually translated into SOX1 protein or its protein expression might be below the limit of detection by the technique used. It is therefore important to test these human cell lines for SOX1 protein expression with more sensitive technique like western blot, See section 3.2.8.

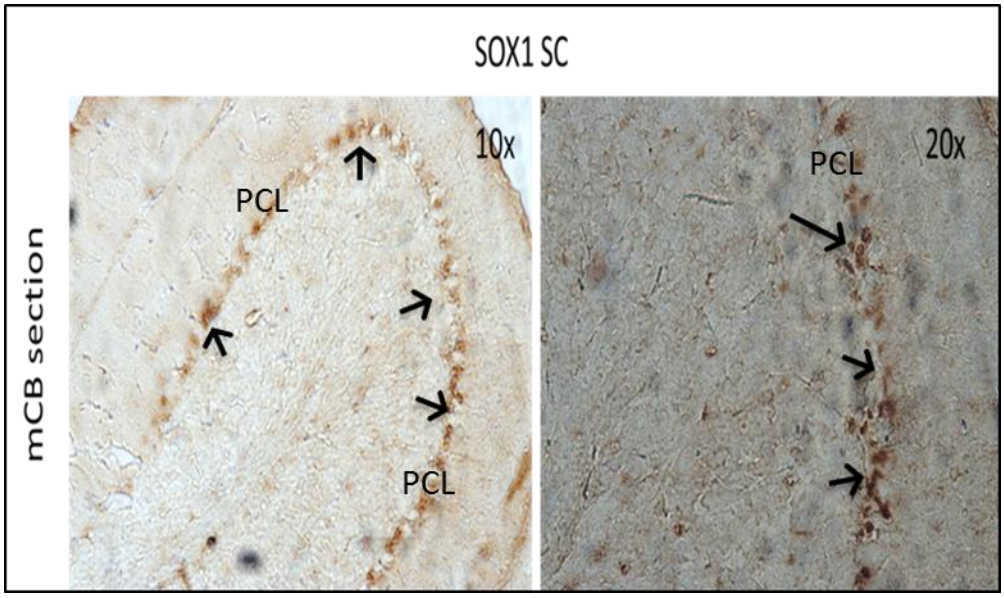


Figure 3-19: **Immunostaining images to detect SOX1 signal in mouse cerebellum**: Frozen section of cerebellum of adult wild type mouse fixed on glass slide, these sections were used as a positive control for immunostaining of Sox1 protein by using anti-SOX1 SC. Black arrows showing SOX1 positive staining in Bergmann glia cells located in the Purkinje cell layer (PCL).

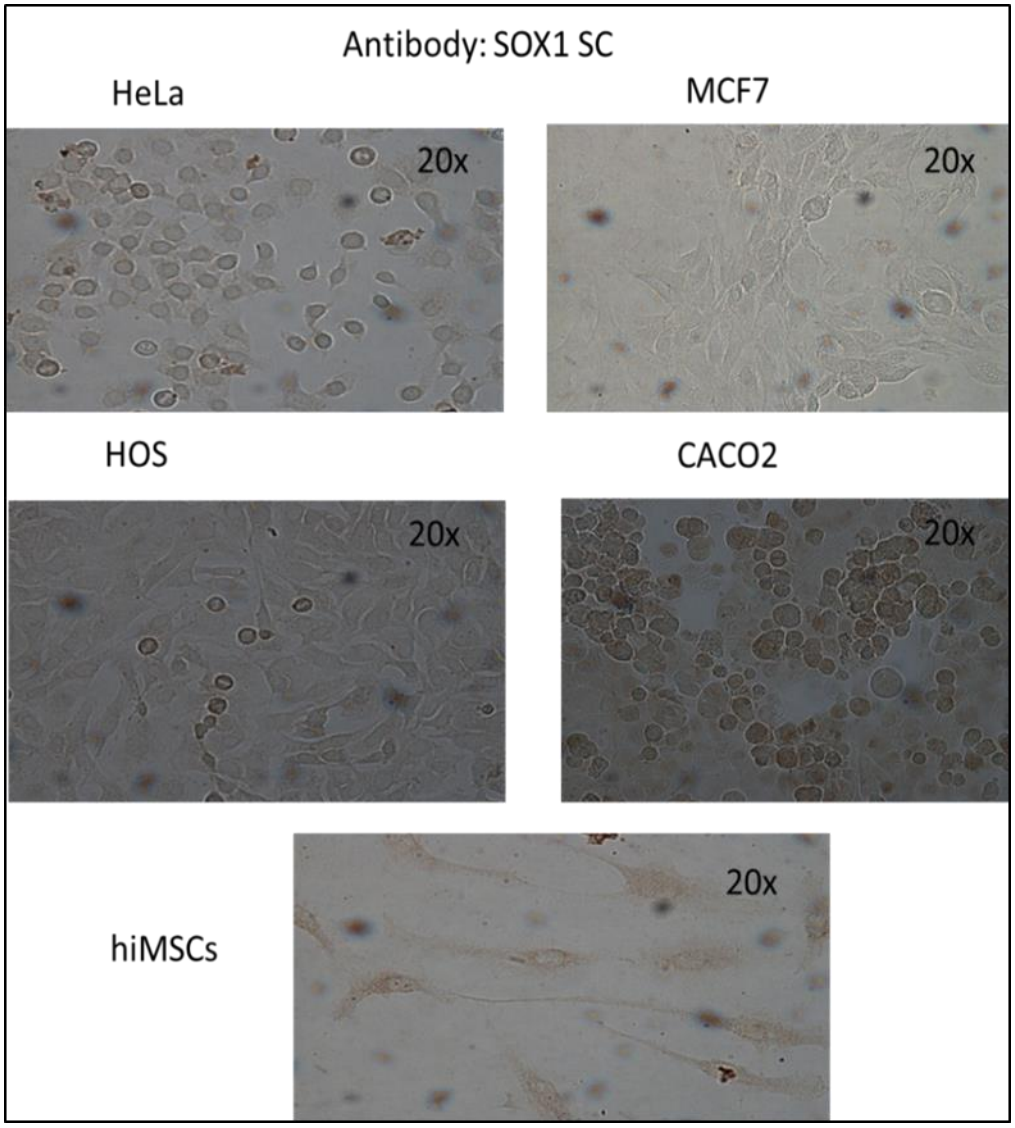


Figure 3-20 **Immunostaining images to detect SOX1 signal in different cell lines:** Immunostaining was performed for SOX1 protein signal on different cell lines grown on a glass slides in a cell culture by using SOX1 SC antibody.

3.2.8 Western Blot analysis of SOX1 protein:

Having established the presence of *SOX1* RNA in differentiating ReN cells and in several cell lines, western blot experiments were performed in order to evaluate the expression of the SOX1 protein in the different cell lines. It is very difficult to develop antibodies that specifically recognised SOX1 because SOX1 shares high sequence homogy with subfamily members SOX2 and SOX3 proteins, therefore different antibodies from different companies were tried to detect SOX1 protein.

SOX1 SC antibody was tested on human cells such as NTera with appropiate controls such as mouse cerebellum. The results obtained are shown in the Figure 3-21. The SOX1 SC antibody showed a band for the mouse cerebellum but failed to obtained a band for the NTera cell line. An anti-ActinB antibody was used as loading control, which showed uniform protein loading accros the three lanes. hiMSC cells were used as a negative control for SOX1, the unspecific bands in the hiMSC cells indicates non-specific cross reactivity of the antibody.

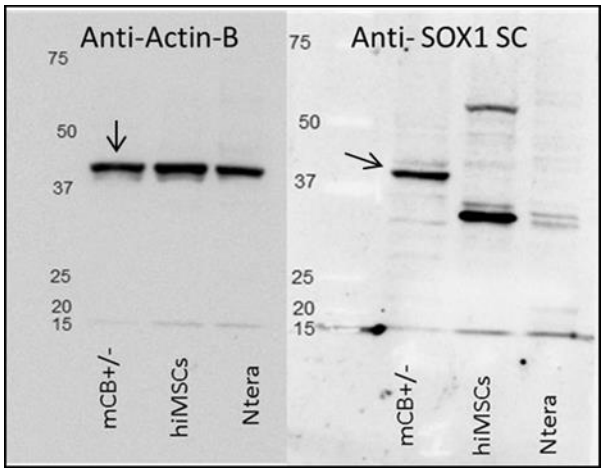


Figure 3-21: **Western Blot optimization for SOX1 protein detection:** Testing of SOX1 SC antibody on human NTera cell line, with mCB+/- used as a positive control for mouse Sox1 and hiMSCs as negative control for SOX1.

3.2.8.1 Validation of SOX1 antibody by using different mouse genotypes to compare Sox1 protein expression

SOX1 SC antibody (from Santa Cruz) was the only antibody which was working on the mouse tissues, but did not give signal on the cell lines (both mouse & human). In order to validate the specificity of our antibody SOX1 SC, total tissue lysates prepared from three different genotypes of adult mouse cerebellum tissue were used: 1) Wild type mouse-cerebellum (mCB WT), 2) Heterozygous mouse-cerebellum Sox1+/-Gfp (mCB+/-) and 3) Homozygous mouse-cerebellum Sox1-/-Gfp (mCB-/-). These three genotypes samples were compared for mouse SOX1 and GFP protein expression in order to validate the assumption that there should be no SOX1 expression in homozygous sample and Wild type should express SOX1 at higher level than heterozygous sample. This assumptions will be opposite for GFP expression as there should be no GFP in wild type and more GFP in homozygous than heterozygous samples. The results obtained for GFP expression was consistent with the assumption as expected (Figure 3-22 top panel). However, unexpedectly it was found that SOX1 SC produce a positive signal around 37KDa for all three genotypes (Figure 3-22 bottom panel).

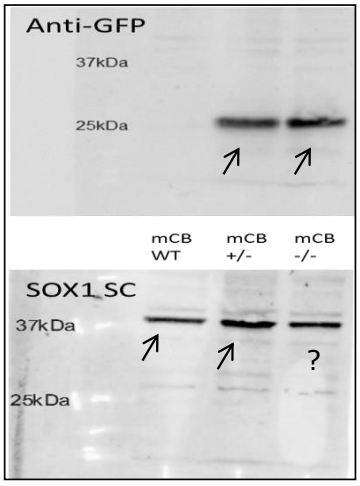


Figure 3-22 **Western blot for SOX1 detection in different mouse genotype samples:** Western blot membrane showing bands for GFP and SOX1 in different mouse genotype samples Mouse Cerebellum Wild type (mCB WT), mouse cerebellum heterozygous for Sox1+/- Gfp and mouse cerebellum homozygous for Sox1-/- Gfp

This data raised a question about the specificity of SOX1 SC antibody. It is suggested that the SOX1 peptide which has been used to raise this antibody might be in the regoin which shares sequence similarties with SOX2 or SOX3 proteins, as SOXB1 family protein shares more than 50% sequence identity. However, mouse SOX2 protein size is 34kDa while mouse SOX3 is 38kDa, which is quite close to the mouse SOX1 protein size and can be difficult to differentiate between the two size bands. Therefore it is suggested that exploitation of online bioinformatics tools for sequence alignment between SOX1 and other SOX family can be helpful to identify the candidate SOX proteins which share high sequence similarties with SOX1. It would be also useful to find regions that are less conserved between the homologous to be used as potential epitopes for the generation of specific antibodies for use in western blot.

3.2.8.2 Testing of different commercially available SOX1 antibodies

Different SOX1 antibodies cited in published studies were tested in a search for SOX1 detection in human cell lines but no satisfactory results were obtained so far. As shown in Figure 3-23, SOX1 mouse monoclonal antibody (R&D) was tried on both human cell lines and mouse tissues (as a control) but no specific bands were obtained. Lamin B was used as a loading control. GFP band in the mouse homozygous mCB+/- confirmed SOX1 expression along with GFP but so far no SOX1 specific protein band has been detected by this antibody (Figure 3-23).

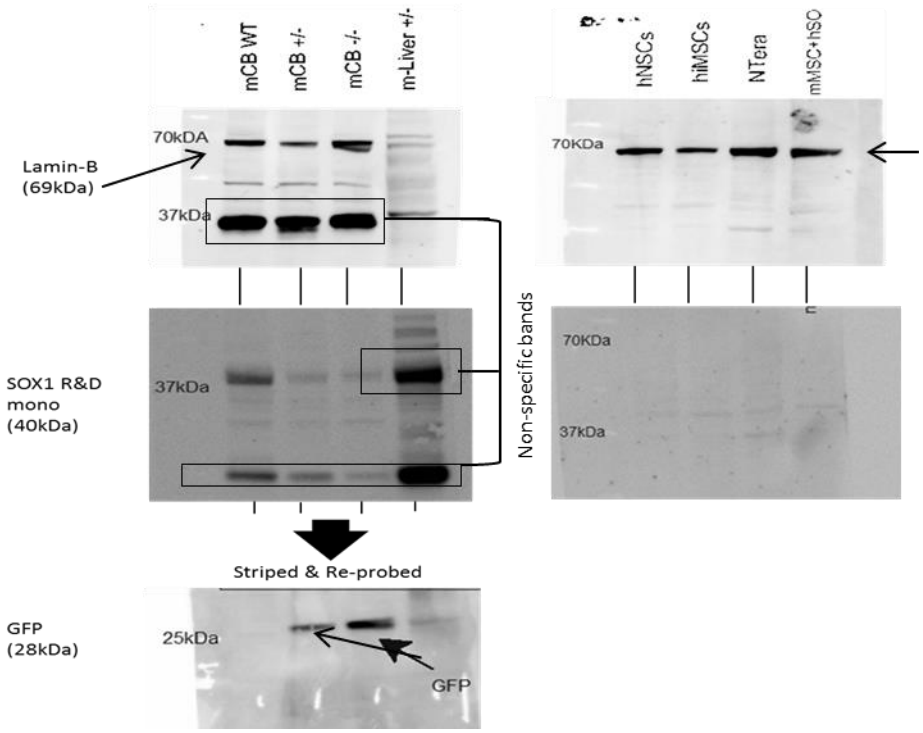


Figure 3-23: **Testing of different commercially available SOX1 antibodies:** Western blot membranes showing results for SOX1 antibodies, GFP and loading control Lamin-B.

Another SOX1 antibody (SOX1 rabbit monoclonal from Abcam) had been recently cited in the published literature for the detection of 39kDa SOX1 protein in human cell lines such as NTera and HeLa. Therefore, it was probed against different human cell lines and different genotypes of

mouse tissues (Figure 3-24). In the NTera and hNSCs cell line, a 39kDa SOX1 band was found to be very weak, which could be due to the fact that undifferentiated NTera might have low SOX1 protein expression or it might be down to the detection protocol needing optimisation. There were no bands for the other cell lines and in the mouse tissues (Figure 3-24). There were also strong non-specific bands in the human cell lines around 70kDa size, which could be due to cross reactivity in these cell lines.

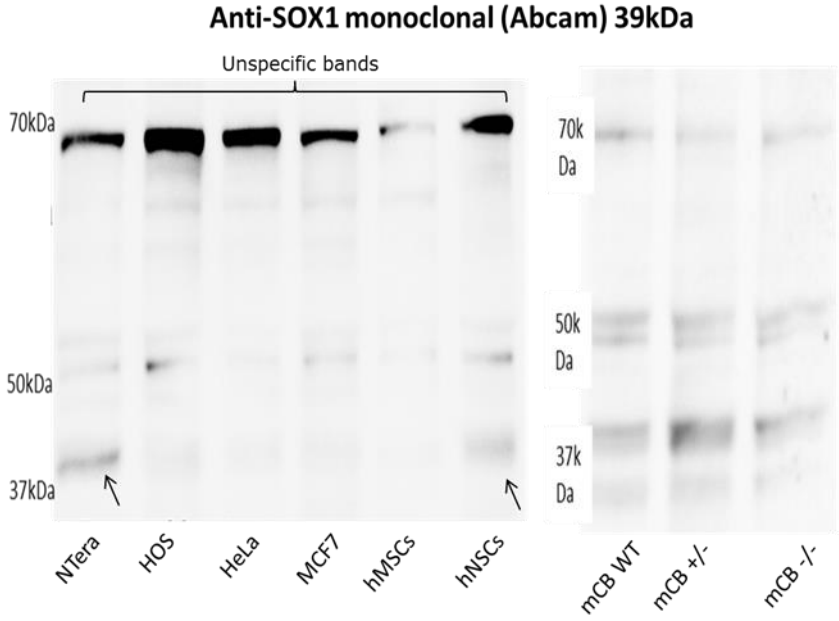


Figure 3-24: **Western blot optimization by using SOX1 monoclonal antibody**: SOX1 monoclonal (Abcam) antibody probed against different human cell lines and different mouse genotypes, Pointed arrows shows approximate size of SOX1 (39kDa) band in NTera and hNSCs cell lines.

3.3 Discussion

3.3.1 *SOX1* gene expression profile in different cancerous and normal cell lines

In this chapter, *SOX1* gene expression has been quantified in different cancerous and normal cell lines relative to its expression in ReN cells. The results showed differential gene expression of *SOX1* in the studied cell lines, NTera, HOS, MCF7 and T47D cell lines were found to express *SOX1* gene at different levels. Others cell lines - CACO2, Hs578T, HCT116, MRC5, MDA-MB-361, MDA-MB-231 and HeLa - were found to be negative for *SOX1* expression by RT-PCR (Figure 3-12, Figure 3-13). Results from qPCR analysis in these set of cell lines were consistent with RT-PCR, validating the reliability of qPCR data for *SOX1* expression. Interestingly, HeLa cell line has been found negative for *SOX1* expression while it has been reported in published literature that HeLa cell line express *SOX1* gene [192]; this could be due to the heterogeneity of HeLa cells that have diverged due to clonal selection in different laboratories [193]. MRC5 which is normal human lung fibroblast cell line appears to have no *SOX1* expression which is not surprising as *SOX1* expression in human is mainly confined to adult brain tissue.

CaCo2 and HCT116, colorectal carcinoma cell lines normally expressing stem cell markers like *SOX2, OCT4* and *NANOG* [194], do not express *SOX1*. Embryonal carcinoma cell line NTera which represent undifferentiated, pluripotent embryonic stem cell phenotype by expressing stem cell markers like *SOX2*, has been found to express *SOX1* gene at low level relative to ReN cells [195]. Very recently, in laryngeal squamous cell

carcinoma it has been found that *SOX1* act as a tumour suppressor and *SOX1* overexpression could downregulate *SOX2* expression while co-expression of *SOX1* and *SOX2* could reverse anti-tumour effect of *SOX1*[196]. SOX2 as a member of the SOXB1 subfamily of transcription factor has been reported as an oncogene in many cancer types [197-199], and its oncogenic role is suggested as important future prognostic factor and possible therapeutic interventions in cancer [200]. Therefore, it could be speculated that lacks of *SOX1* expression or co-expression of *SOX1* and *SOX2* might be linked with tumorigenesis. Further investigation is needed to analyse its relationship in these cell lines by performing functional assays such as knockdown of *SOX2* or ectopic expression of *SOX1* which might possibly explain the role of *SOX1* expression in these cancers.

Breast cancer is recognised as molecular heterogeneous disease, the cell lines for studying breast carcinoma are divided into different molecular subtypes based upon expression of markers ER (oestrogen receptor), PR (progesterone receptor) and HER2 (human epidermal growth factor receptor 2), each subtype has different prognosis and treatment responses [201]. Breast adenocarcinoma cell lines such as MCF7, T47D, and MDA-MB-361 are luminal type (ER^+, PR^+ and $HER2^{+/-}$) and Hs578T and MDA-MB-231 are basal type (ER^-, PR^- and $HER2^-$)[202]. In these cell lines, qPCR quantification has shown differential gene expression of *SOX1*, Luminal type cell lines such as MCF7 and T47D (ER^+, PR^+ and $HER2^-$) which shows a co-relation with good prognosis in breast cancer have been found to express *SOX1* gene [202] while basal type cell lines like Hs578T

and MDA-MB-231 (ER^-, PR^- and $HER2^-$) have shown no *SOX1* gene expression. These basal sub types are enriched for markers associated with the epithelial–mesenchymal transition and expression of features associated with cancer stem cells making them highly metastatic cell lines [203]. Interestingly, MDA-MB-361 which belongs to the luminal type with positive expression of all three markers (ER^+, PR^+ and $HER2^+$) [202] has shown no expression for *SOX1* contrary to the other luminal lines. *SOX1* expression in luminal type cell lines (breast cancer cell types with good prognosis) and lack of *SOX1* expression in basal cell lines (which are associated with aggressive metastatic tumour) backup the idea of possible anti-tumour effect of *SOX1* in cancer. *SOX1* gene expression needs further investigation in individual cancer types which might be helpful for future therapeutic and detection purposes. Nevertheless, *SOX1* differential gene expression in different cancerous cell lines and its expression pattern can possibly serve as discriminator between different cancers types which warrants further investigation at a clinical level.

3.3.2 Detection of SOX1 protein expression in different cell lines:

To study SOX1 protein function and its role in the regulation of gene transcription, it was necessary to identify a cell line which expresses SOX1 protein, and can be used as a positive control for studies in different cancerous cell lines. For example, It will be interesting to analyse *SOX1* gene expression in each cell lines and then to see whether the mRNA translate into the actual functional SOX1 protein in these cell lines. Western blot and Immunocytochemistry (ICC) techniques were used to

identify a cell line for SOX1 protein expression. Mouse brain tissue WT (mBrain WT) and mouse cerebellum Sox1+/- Gfp (mCB+/-) tissues were used as Sox1 positive controls. These were tested in parallel with human cell lines (Figure 3-24) by western blot and ICC, but no SOX1 signal was detected in these cell lines. Different anti-SOX1 antibodies were tested in case the problem was down to the detection of SOX1 protein but no specific signal for SOX1 has been observed. Only SOX1 SC antibody from Santa Cruz was able to give a band for SOX1 in mouse tissues samples but after the antibody validation test by different mouse tissue genotypes it was found that this antibody might be unspecific. High sequence similarities, between SOXB1 group of proteins suggests cross-reactivity with other SOX family of protein most probably with SOX2 or SOX3 (Section, 1.1.1). Similarly, SOX1 SC antibody was tried by ICC but no SOX1 signal was obtained for different human cell lines. The transfected mMSc+hSOX1 cell line, which has human *SOX1* cDNA expression, did not give detectable SOX1 protein expression, which might be due to the detection problem or might be due to the fact that the antibodies tried were not specific for SOX1 detection. Therefore, no commercially available SOX1 antibody was identified for specific SOX1 protein detection in human cell lines.

3.3.3 Epigenetic silencing of SOX1 gene expression through promoter hyper methylation

DNA methylation analysis in recent cancer studies have reported some developmental related genes that are differentially methylated and play important role in cancer progression and metastasis i-e, *SOX1, PAX1* and *LMX1A* etc. Among them *SOX1* has been proposed to be best discriminator between cancerous and normal cells showing wide variation of methylation pattern across different cancerous cells. So far, differential methylation pattern of *SOX1* gene has been reported in cervical, prostate and ovarian cancer cell lines [8, 65, 204-206]. Therefore, DNA methylation analysis of *SOX1* gene was carried out on wide range of both normal and cancerous cell lines (section 3.2.5) .

In this study, the observed *SOX1* promoter methylation pattern is consistent with previous published work as wide variation in *SOX1* promoter DNA methylation pattern across different cell lines was observed (Figure 3-15) showing lower level of DNA methylation in stem cell lines such as NTera, Human ES cells, hMSCs and ReN cells. Cancerous cell lines have shown differentiantial promoter methylation of *SOX1* such as MCF7, HOS, NTera, T47D and H9 were hypomethylated and Hs578T, HCT116, CaCo2, HeLa and MDA-MB-361/231 were found hypermethylated. *SOX1* gene expression in these cell lines were found to co-related with *SOX1* promoter methylation. Loss of *SOX1* expression through promoter hypermethylation in different cancer types has been already documented [205] The results showed that cancerous cell lines with promoter hypermethylation such as HCT116, CaCo2, HeLa and MDA-

MB-361/231 do not express *SOX1*. Epigenetic silencing of a gene, in particular DNA hyper methylation at its promoter region, has been already reported to contribute to carcinogenesis. Looking into the promoter methylation pattern of *SOX1* in these different cancerous cell lines data set, for example, Colorectal (HCT116, CaCo2), Breast (MDA-MB-361, Hs578T) and cervical carcinoma cell line (HeLa), which do not express *SOX1*, are completely methylated at *SOX1* promoter region (Figure 3-15, Figure 3-12), suggesting an epigenetic silencing of *SOX1* gene in these cell lines. On the other hand, MCF7 and T47D (breast adenocarcinoma), HOS (Osteosarcoma) and NTera (embryonal carcinoma) cell lines all express *SOX1* gene, expression of *SOX1* found in a variety of cancer types suggest that this could be an early event promoting cancerous transformation which is independent of the tissue of origin. However, it is theoretically possible that those cancers which have showed *SOX1* expression have arisen in multipotent progenitor with stem cell like attributes and thus that expression of *SOX1* is retained contributing and/or facilitating cancerous transformation rather than acquired in the cancer.

It has been known that changes in DNA methylation status of *SOX1* can significantly differentiate between pre-cancerous cervical cells and negative controls [204]. Therefore, It can be concluded that *SOX1* epigenetic silencing through promoter DNA hyper methylation is highly likely in cancer and its epigenetic profile can serve as discriminator between different cancer types. Role of *SOX1* in cancer development is still

emerging, *SOX1* is highly recommended as a detection or diagnostic candidate for clinical trials which might shed a light on whether *SOX1* can act as a detection (diagnostic) and/or a prognostic marker in cancer.

3.4 Conclusion

In conclusion, It was found that *SOX1* is differentially expressed in different cancer cell lines sugesting differential role of SOX1 in cancer and that epigenetic silencing of *SOX1* through promoter DNA methylation is likely dependent on the cancer type.

It was found that SOX1 gene expression co-relate with SOX1 promoter DNA methylation. Stem cell lines (NTera, Human ES cells, hMSCs and ReN cells) were found with lower level of DNA methylation. Different cancerous cell lines have shown differential promoter methylation of SOX1 such as MCF7, HOS, NTera, T47D and H9 were found hypomethylated and Hs578T, HCT116, CaCo2, HeLa and MDA-MB-361/231 were found hypermethylated. It has been suggested that SOX1 epigenetic profile and its expression pattern can serve as discriminator between different cancer types which needs further investigation.

In Breast carcinoma cell lines, *SOX1* was found to express in luminal cell lines (less aggressive) compared to lack of *SOX1* expression in basal cell lines (highly metastatic), which suggest possible anti-tumour effect of *SOX1.* It was also found that cancerous cell lines expressing stem cell marker like SOX2 (NTera, CaCo2 and HCT116) have differential SOX1 expression. It has been suggested that lacks of *SOX1* expression or co-

expression of *SOX1* and *SOX2* might be linked with tumourigenesis and their functional analysis might possibly explain the possible anti-tumour effect of *SOX1* in cancer.

4 Chapter: 04

SOX1 overlapping transcript (Linc403) and its relation to SOX1 expression

4.1 Introduction

SOX1 is a HMG-BOX transcription factor, involved in early embryogenesis and maintenance of neural stem cell [207]. SOX1, SOX2 and SOX3 belongs to SOXB1 subgroup of transcription factors; they have similar sequences, expression patterns and overexpression phenotypes [208]. SOX2 is a major transcriptional regulator in pluripotent stem cells [209]. *SOX2* gene maps to Chr3q26.3 locus, embedded into an intron of a long non coding RNA (lncRNA) called *SOX2* overlapping transcript (*SOX2*-OT) [210]. Human and mouse *SOX2* overlapping transcripts have multiple TSSs and are transcribed into several alternative transcript variants [211]. Recently, concomitant gene expression of *SOX2* and its *SOX2*-OT has been reported in breast, lung and oesophagus carcinoma [212-214]. Recent studies suggest a significant correlation between *SOX2*-OT and *SOX2* expression in cellular differentiation, pluripotency and carcinogenesis [211, 215-217]. *SOX2*-OT is differentially spliced into multiple transcript variants in stem and cancer cells [215, 218].

Similar to *SOX2*, *SOX1* is also embedded within an intron of a lncRNA called *SOX1-OT* (LINC00403; Figure 4-1). *SOX1-OT* is annotated in the NCBI RNA reference sequence collection (RefSeq) [28]. Presently, there

are two transcript variants annotated in RefSeq data, LINC00403 variant 1 and LINC00403 variant 2, but only transcript variant 1 overlaps the *SOX1* gene and is therefore known as *SOX1-OT*. *SOX1-OT* is found on chromosome 13 with genomic position chr13:111972310-112108015, leading to a genomic DNA size of 135.706kb and a 704bp RNA [28]. The *SOX1-OT* structure has a validated status in RefSeq and the reference sequences are derived from three different tissues which are Amygdala (GenBank: DA195709.1), foetal eye (GenBank: BQ184460.1.1) and Lung-carcinoid (GenBank: AI693652.1) [183].

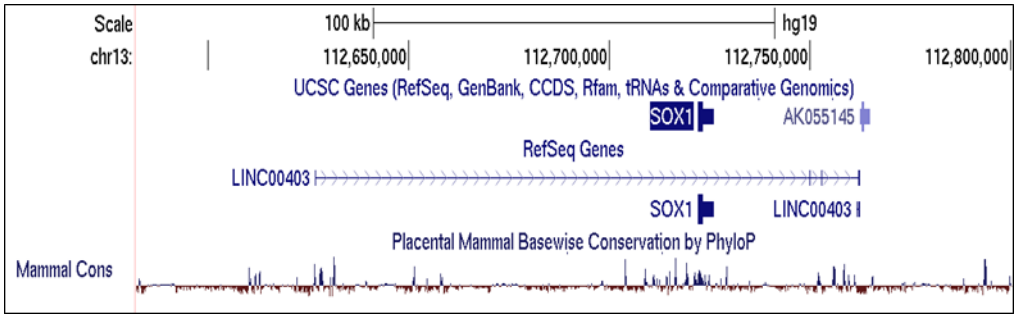


Figure 4-1: **Structure of LncRNA (LINC00403) overlapping SOX1 gene**: Images generated by UCSC browser (http://genome.ucsc.edu) human genome assembly (GRCh37/hg19) [28] showing the SOX1 overlapping transcript (LINC00403) annotated from RefSeq.

The structure of human *SOX1-OT* is not as well characterised as that of mouse *SOX1-OT* and nothing is known about its biological significance and function. This study describes the complex structure of *SOX1-OT*, its splicing variants and gene expression pattern in stem cell and cancer, and provides evidence of its potential role in transcriptional regulation of the *SOX1* gene.

4.2 Results

4.2.1 Structure architecture of SOX1-OT in ReN cells

In this study the structure of the *SOX1-OT* (LINC00403) was characterised in ReN cells by using RT-PCR primers in annotated exons of *SOX1-OT* (Figure 2-1). In addition, 5'RACE was also performed to identify the transcription Start Site (TSS) of the *SOX1-OT* in these cells. The results obtained have shown 5 novel exons (exon1a, 1b, 3a, 3b and 3c) and 9 novel transcript variants (V3-11) of *SOX1-OT* in ReN cells which were not previously annotated (Figure 4-2). Two main TSSs located in close genomic proximity to the *SOX1* gene has been identified for *SOX1-OT* (Figure 4-2A, bent arrows). It has been found that *SOX1-OT* spliced into several different transcript variants, including the two RefSeq annotated transcript variants 1-2 (Figure 4-2B) and 9 novel transcript variants 3-11 identified in this study (Figure 4-2C). Transcript variants 3-8 were identified by RT-PCR amplification and 9-11 by 5'RACE experiment. In this study, the 1st exon of the annotated transcript was not detected in the ReN cells.

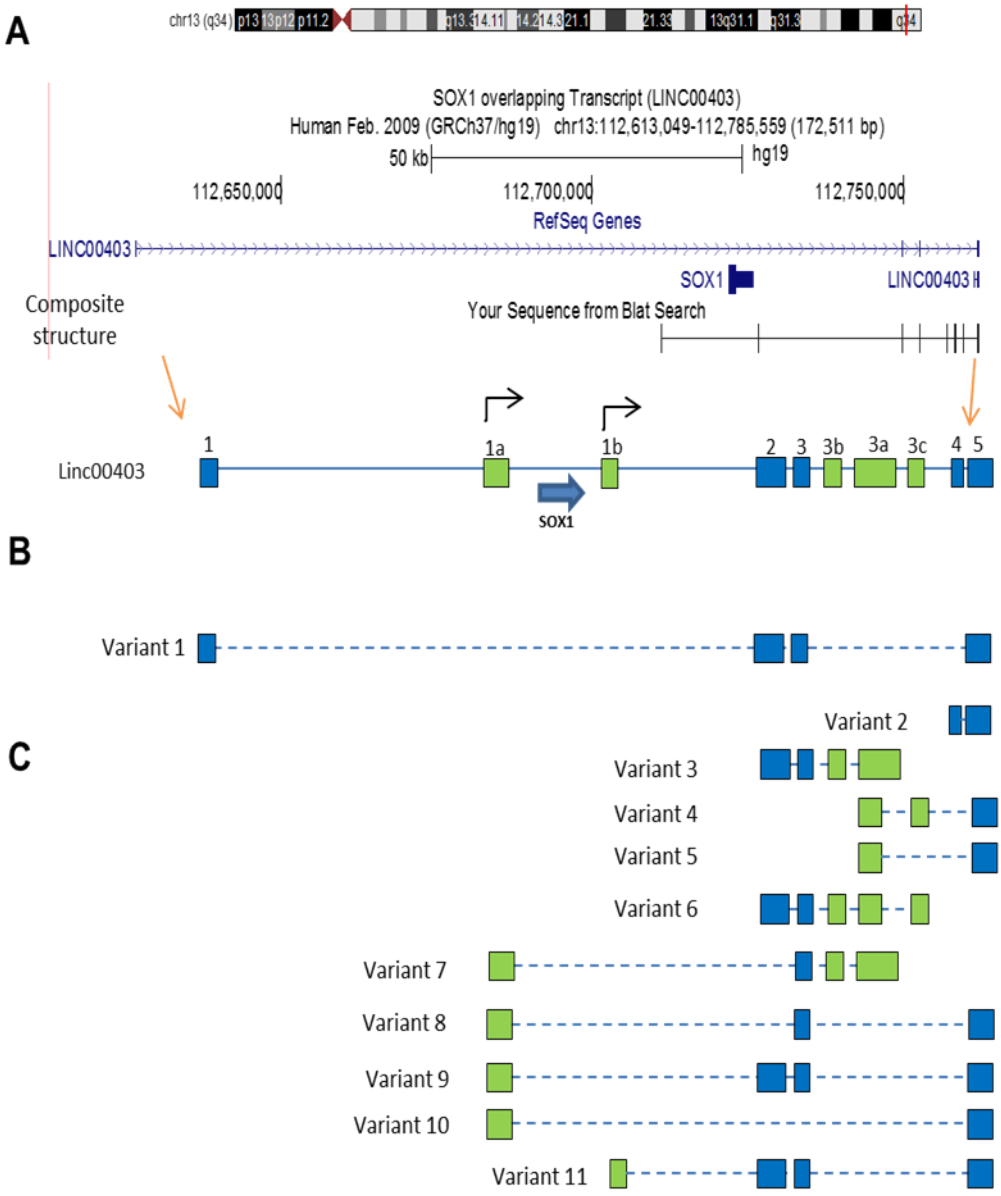


Figure 4-2: **SOX1-OT genomic structure and its transcript variants**. (A) Chromosome ideogram (top) with red line representing the region of interest, and UCSC genome generated images, http://genome.ucsc.edu (middle) showing the genomic locus of the annotated LINC00403 (*SOX1-OT*) and the newly detected annotated *SOX1-OT* structure (under Blat search) by RT-PCR and 5'RACE [178]. Composite *SOX1-OT* genomic structure (bottom) containing 10 exons with 5 RefSeq annotated exons (1, 2, 3, 4 and 5, blue box) and the newly identified exons 1a, 1b, 3a, 3b and 3c (green box). Primer binding sites are shown with blue arrows pointed in the direction of amplification. Potential transcription start sites (TSS) are shown with arrows pointing to the direction of transcription. (B) *SOX1-OT* Variants 1 and 2 as described in the UCSC human genome annotated transcripts provided by RefSeq. (C) New transcript variants 3-8 identified through RT-PCR amplification, and variants 9-11 identified by 5'RACE.

4.2.2 Comparison between the human and mouse SOX1 overlapping transcript

Similar to the human *SOX1* protein coding gene, the mouse *Sox1* gene has its own overlapping long non-coding RNA annotated as GM5607 in the RefSeq mouse genome dataset (NCBI accession: NR_027975.2) [219]. Mouse GM5607 transcript as a *Sox1* overlapping transcript (*Sox1-ot*), has a total of 8 annotated exons and is 50962bp long [178, 219]. Looking into the annotated human and mouse *SOX1* overlapping transcripts from RefSeq data in the UCSC genome browser, the mouse *Sox1-ot* is better characterised compared to the human *SOX1-OT*. Mouse *Sox1-ot* has a higher number of annotated exons (total of 8) compared to the human transcript (total of 5) [28]. The human annotated *SOX1-OT* is much large than mouse and its first 5' exon located well before upstream from *SOX1* gene as can be seen in the Figure 4-3, while the 1st exon of mouse *Sox1-ot* lies upstream of the *Sox1* gene in its close genomic proximity. The mouse *Sox1-ot* terminates further downstream compared to the human transcript and has more annotated exons in this region compared to human *SOX1-OT* (Figure 4-3, red circle) [28]. Peaks for species conservation show that this region is highly conserved both in human and mouse.

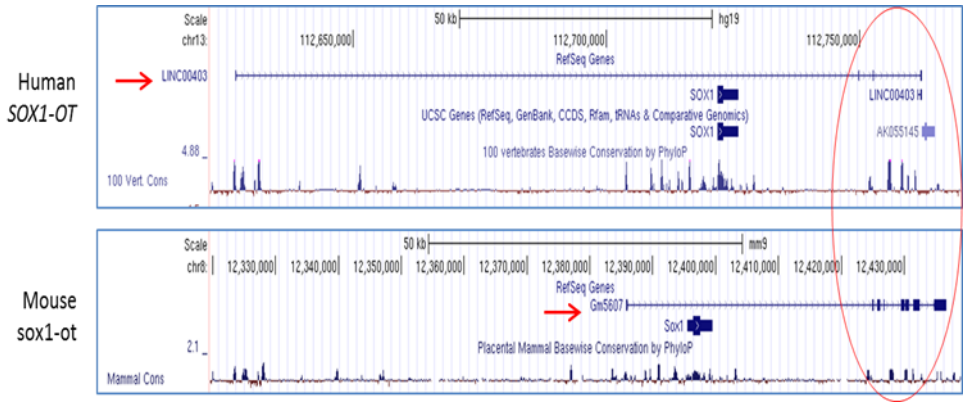


Figure 4-3: **Comparison for Human & Mouse SOX1 overlapping transcript**: overlapping lncRNA transcripts are pointed by bold red arrows and red circle shows the end region of both transcripts. Images generated through UCSC genome browser http://genome.ucsc.edu[28].

To identify evolutionary conserved regions in the human *SOX1-OT*, the ECR browser [220] was used for the human genome alignment across different species in order to generate evolutionary conservation heights for *SOX1-OT* (Figure 4-4). For *SOX1-OT*, three highly evolutionary conserved regions (Figure 7, a, b and c) were identified across different species. The regions 'b' and 'c' have a high percentage (75-85%) of sequence identity between human and mouse. The mouse *Sox1-ot* has annotated exons present at both regions compared to human *SOX1-OT* which has no annotated exons present in these evolutionary conserved regions (Figure 4-3). Therefore, to check if any unknown exons were present in the human *SOX1-OT* in the highly evolutionary conserved regions similar to the mouse *SOX1-OT* (Figure 4-4, b and c) RT-PCR primer were designed in these regions.

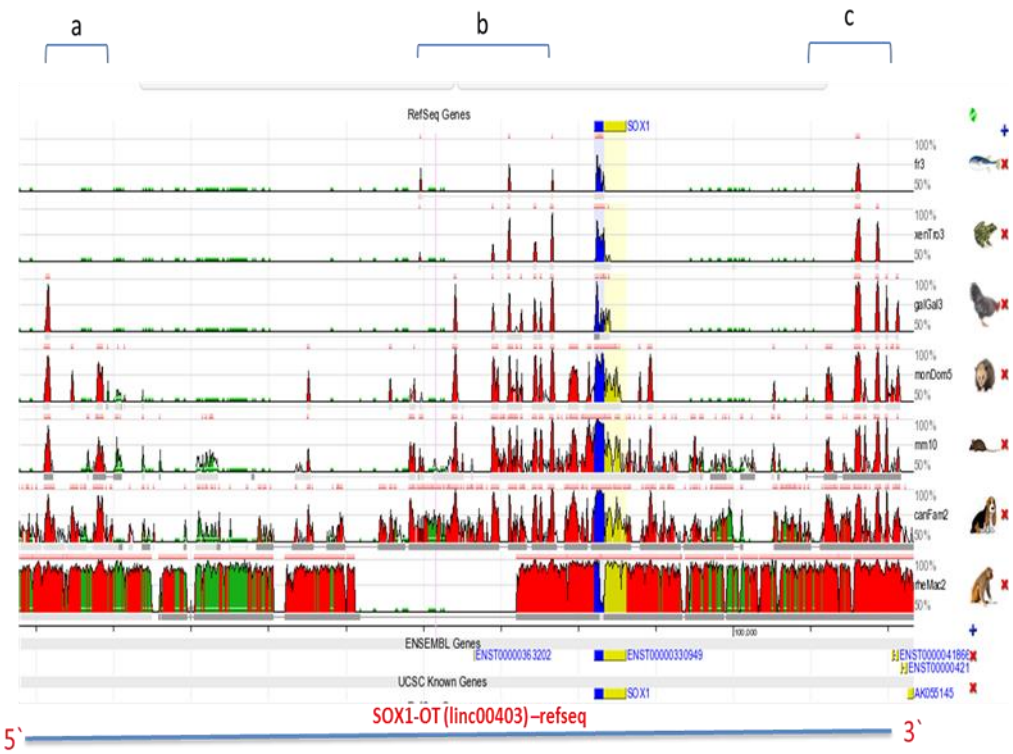


Figure 4-4: **Evolutionary conservation of the SOX1-OT genomic locus**: ECR browser (http://ecrbrowser.dcode.org) [220]. Pairwise genomic sequence alignment of human *SOX1-OT* genomic region with different species: Rhesus macaque (rheMac2), Dog (canFam2), Chicken (galGal3), Mouse (mm10), Frog (xenTro2), Opossum (monDom5) and Fugu (fr3). Regions with >50% sequence identity and min 200bp length have been identified as a, b and c.

Following are the results obtained by RT-PCR and 5'RACE which led to the discovery of unannotated exons and different transcript variants of *SOX1-OT* in ReN cells.

4.2.3 Investigation of SOX1-OT by RT-PCR:

RT-PCR was first used to detect expression of *SOX1-OT* in human ReN cells and to analyse the structure of the transcript compared to the annotated exons described in RefSeq [183]. ReN cells are a human neural progenitor cell line (neural stem cells) with the ability to readily differentiate into neurons and glial cells [158]. SOX1, as a neural marker, is express highly in neural stem cells (NSCs). Evidence in the current literature suggests that long non-coding overlapping transcripts regulate the nearby protein coding gene expression [221], therefore, in this case, to detect and analyse *SOX1-OT*, human ReN cells were the best available cell line to use.

4.2.3.1 Evidence of SOX1-OT expression in ReN cells:

To detect expression of *SOX1-OT* in human ReN cells, cDNA from ReN cells at different time point of neural differentiated at day 0, 2, 4 and 6 were used along with positive control gDNA. RT-PCR was performed using a gene specific primer pair (F4, R4) which binds within the last exon of the RefSeq annotated *SOX1-OT* (Figure 4-5A). The agarose gel images after RT-PCR are shown in the Figure 4-5, PCR products were obtained for all +RT samples from ReN cells differentiated at different time points from day 0-6, suggesting expression of *SOX1-OT* in ReN cells, -RT samples were negative, gDNA as a positive control had the expected specific size of band.

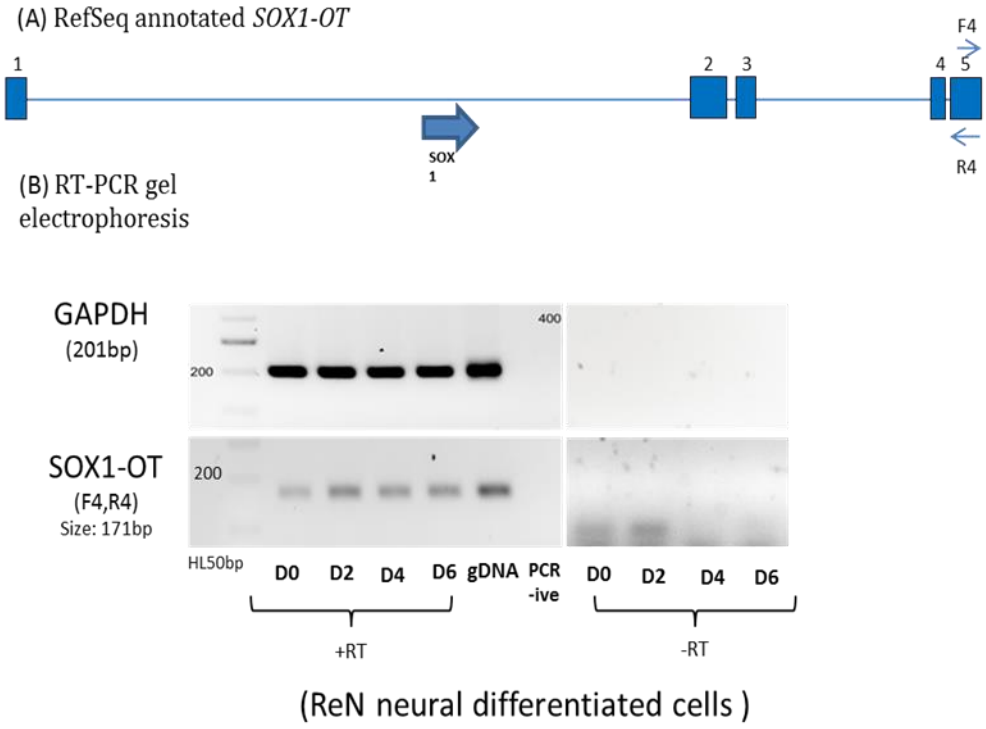


Figure 4-5: **Evidence of SOX1-OT detection in ReN cells:** RT-PCR results for primer pair (F4, R4) that binds within the last exon of the annotated *SOX1-OT*. *SOX1-OT* expression was detected in ReN cells samples collected from different time points of neural differentiation at day0-6. GAPDH was used as a loading control.

4.2.3.2 Identification of unannotated exons in the SOX1-OT:

The primer pair (F1, R1) was designed to amplify a region located between the annotated exon 3 and 4 of *SOX1-OT*, which was observed to be evolutionary conserved region across different species (Figure 4-3 and Figure 4-4), with mouse sharing approximately 75-85% sequence identity in this region [220]. Mouse *Sox1-ot* contains many annotated exons at this region compared to human *SOX1-OT* as discussed earlier. RT-PCR was performed at this specific genomic region of *SOX1-OT* (Figure 4-6), to see whether any unannotated exons might be present in human *SOX1-OT*. RT-PCR amplification by primer pair (F1, R1) showed the presence of a

detectable transcript signal which was then analysed by direct sequencing. (section, 2.5.4.6)

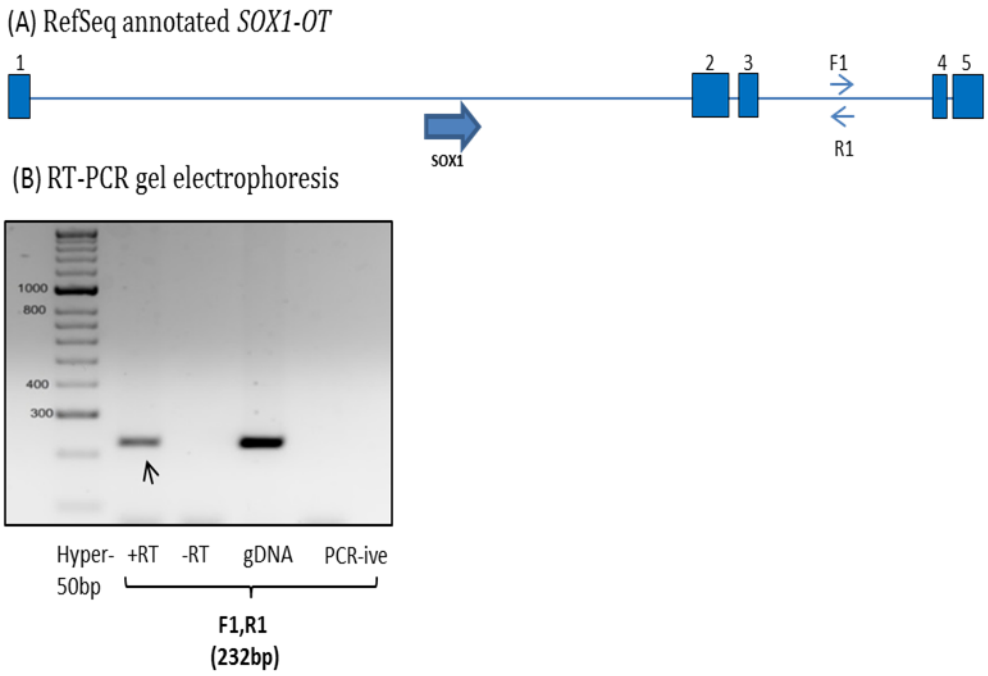


Figure 4-6: **Detection of unannotated exons in the SOX1-OT** (A) Primer pair (F1, R1) binding site for *SOX1-OT* and (B) Gel electrophoresis image showing RT-PCR amplified fragment in +RT sample (ReN cells) pointed by the arrow, gDNA was used as a positive control.

After direct sequencing, the amplified PCR sequence was aligned against the human genome (hg19) using BLAT and the alignment was visualised by UCSC genome browser. The alignment confirmed the amplified PCR sequences as part of an unannotated exon present in this region (Figure 4-7).

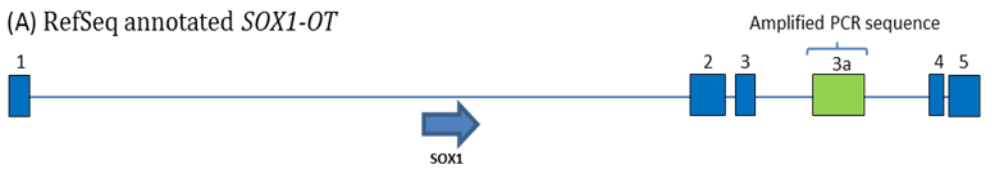


Figure 4-7: **Structure illustration of SOX1-OT after addition of newly detected exon**: blue boxes are the annotated exons, orange arrow represents SOX1 gene genomic location, green box is the newly identified unannotated exon-3a for *SOX1-OT*.

4.2.3.3 Structure determination of the SOX1-OT downstream of SOX1 gene:

After the detection of novel exon 3a at the genomic location downstream of *SOX1* gene in hNSCs (ReN) (Figure 4-7), further investigations were performed to characterise the sequence of the *SOX1-OT*. RT-PCR amplification was attempted using primers located between exon 1- 2, 1-3 and 1-5; however, it did not produce any detectable product (data not shown). In order to allow the identification of the whole length transcript and determine its structure, the transcript was divided into two parts upstream and downstream of the SOX1 gene, and different primer combinations were used in order to characterise each part. RT-PCR experiments performed in ReN samples are shown in the following subsections, and revealed that *SOX1-OT* (LINC00403) had different transcript variants and exons so far unannotated in the human genome (RefSeq). The RT-PCR products obtained were purified and sent for direct Sanger sequencing. Sequences received were aligned to human genome data by using UCSC genome browser. These results are discussed below in detail.

4.2.3.3.1 Transcript variant 1 and 3

The detection of transcript variant 1 & 3 by RT-PCR amplification are shown in Figure 4-8. The *SOX1-OT* variant-1 is the annotated transcript in the human genome (RefSeq) data. It has four (4) annotated exons. Transcript variant 1 was detected by RT-PCR in ReN cell differentiated at day-6 by using primer pair (F2,R3), see Figure 4-8B. After the identification of novel unannotated exon downstream of the transcript

(Figure 4-6B & Figure 4-7), primer pair (F2,R1) was designed to amplify the region starting from exon-2 in a sense direction and antisense primer pair binding in the recently identified exon-3a (Figure 4-8A). For primer pair (F2,R1) a PCR amplified product was obtained for the +RT sample around 800bp (Figure 4-8C) while the –RT sample and gDNA as a negative control were negative, as no genomic band was expected as this range because of amplicon would be > 10k in size. The amplified fragment was sequenced and aligned to UCSC human genome data set which has identified a new transcript variant-03 as it contained another novel unannotated exon between exon 3 and 3a. (Figure 4-8C).

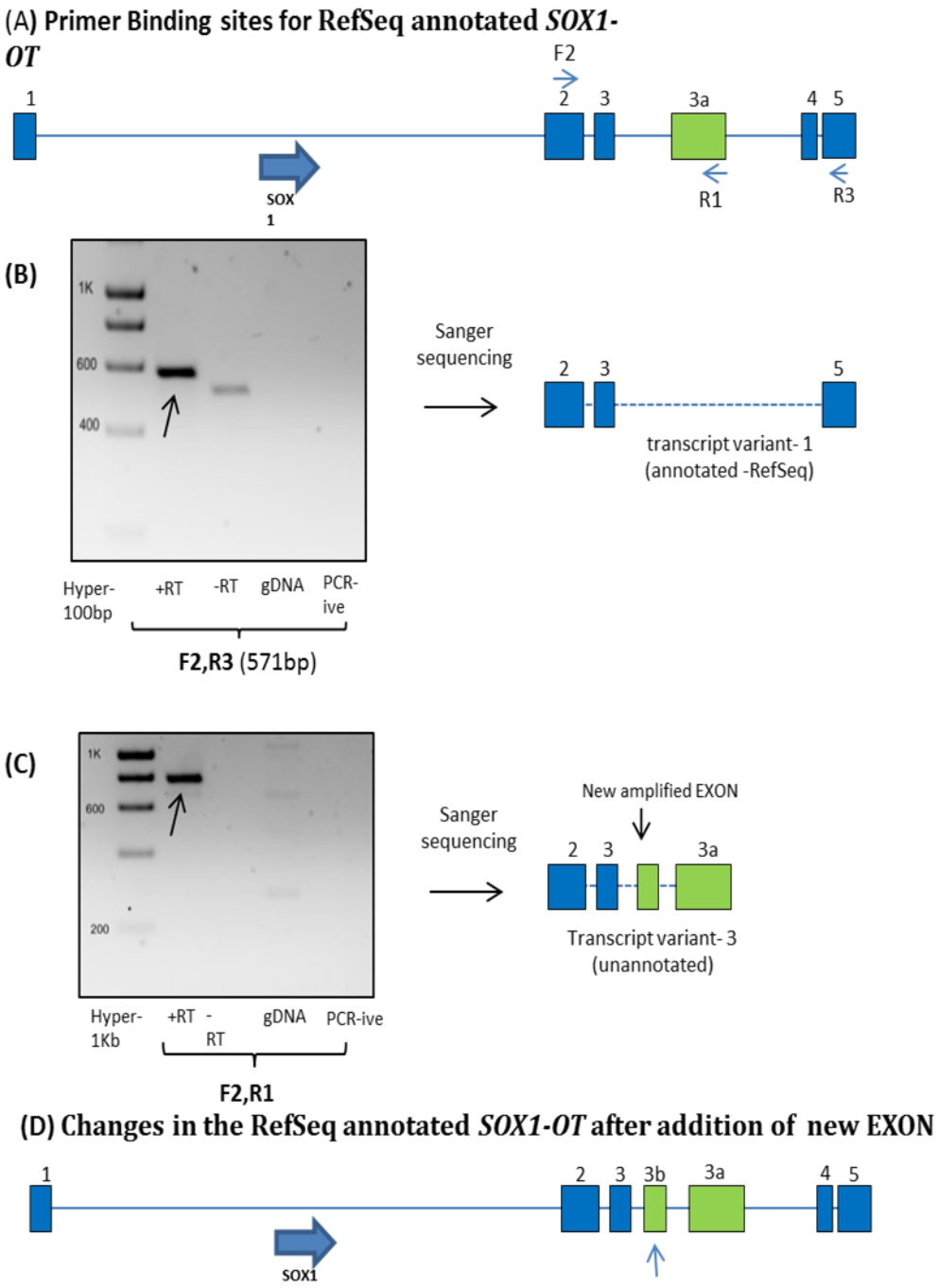


Figure 4-8: **Identifying part of the transcript variant 1 and 3**: (A) RT-PCR primers binding sites in the RefSeq annotated structure of *SOX1-OT*. (B) and (c) the Gel electrophoresis images; showing bands obtained for the +RT samples (ReN cells) pointed by the arrow, next to the images are the identified transcript variants through sanger sequencing. (D) The building structure of *SOX1-OT*, which has now two novels unannotated exons (3a and 3b) represented by a green box.

4.2.3.3.2 Transcript Variant-4 & 5:

Primer pair (F1,R3) binds to exon-3a in a sense direction and to the last exon-5 in antisense direction (Figure 4-9). Transcript variants 4 and 5 were identified during the analysis of *SOX1-OT* expression across different time points of neural differentiation of ReN cells (Day1-6). Transcript variant-4 has shown the presence of another novel exon (3c) in this region and expressed only at day2 of neural differentiation while transcript variant-5 lacking the new identified exon (3c) and only expressed at day-4. This analysis also showed an expression switch between transcript variant-4 and 5 at day-2 and day-4. There was no expression at day-0 (Figure 4-9).

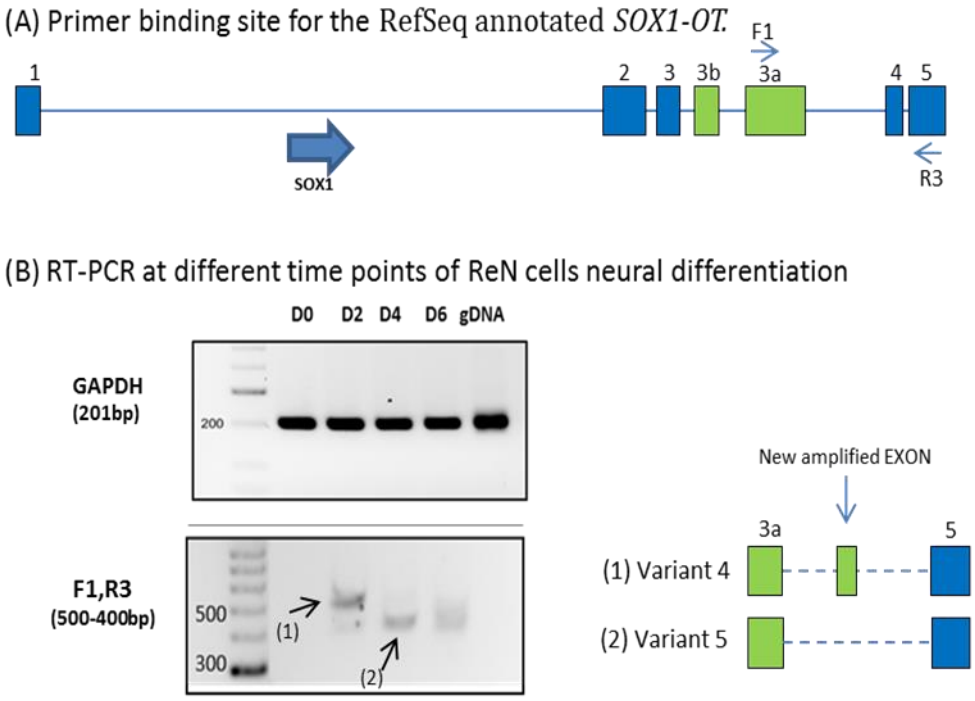


Figure 4-9: **Identification of Transcript variant 4 and 5:** (A) Illustration of primer pair (F1, R3) binding site in the *SOX1-OT*, (B) Gel electrophoresis image; RT-PCR for the ReN differentiated cells at different time points from day0, 2, 4 and 6. GAPDH used as a reference gene, PCR fragments 1 and 2 obtained for the primer pair (F1,R3) at day2 and 4 shows transcript variant 4 and 5 respectively. Structure of transcript variant-4 and 5; the exons expressed together for these transcripts are shown in a coloured

square box, blue boxes represent annotated exons and green box as novel unannotated exons.

There is another new entry of unannotated exon 3c for the *SOX1-OT* (Figure 4-10).

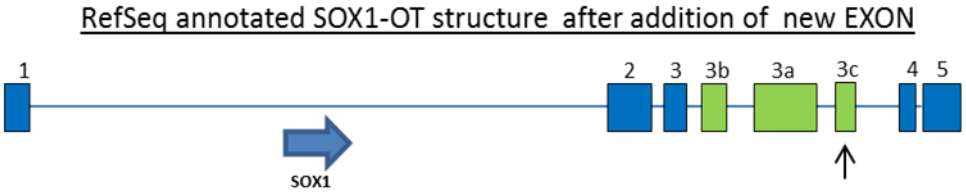


Figure 4-10: **Structure illustration of SOX1-OT after addition of new exon**: This diagram shows the building structure of *SOX1-OT*, which has now three novel unannotated exons (3a, 3b and 3c) represented by a green box.

4.2.3.3.3 Transcript Variant-6:

After the identification of exon 3c, primer pair (F2,R7) was designed to amplify the transcript from exon-2 in a sense direction and exon 3c in the antisense direction (Figure 4-11A). A PCR product was obtained for the +RT while the –RT was negative (Figure 4-11B), The gDNA sample was also negative as no band is expected at this range because the resultant PCR amplicon would be greater than 15kb in size. The two bands were obtained in +RT sample, one around 800bp and second just above 600bp, through Sanger sequencing the band just above 600bp was found to be an unspecific product. The result obtained from Sanger sequencing for the band in +RT (800bp, arrow pointed) has shown the presence of another transcript variant-6 (Figure 4-11B).

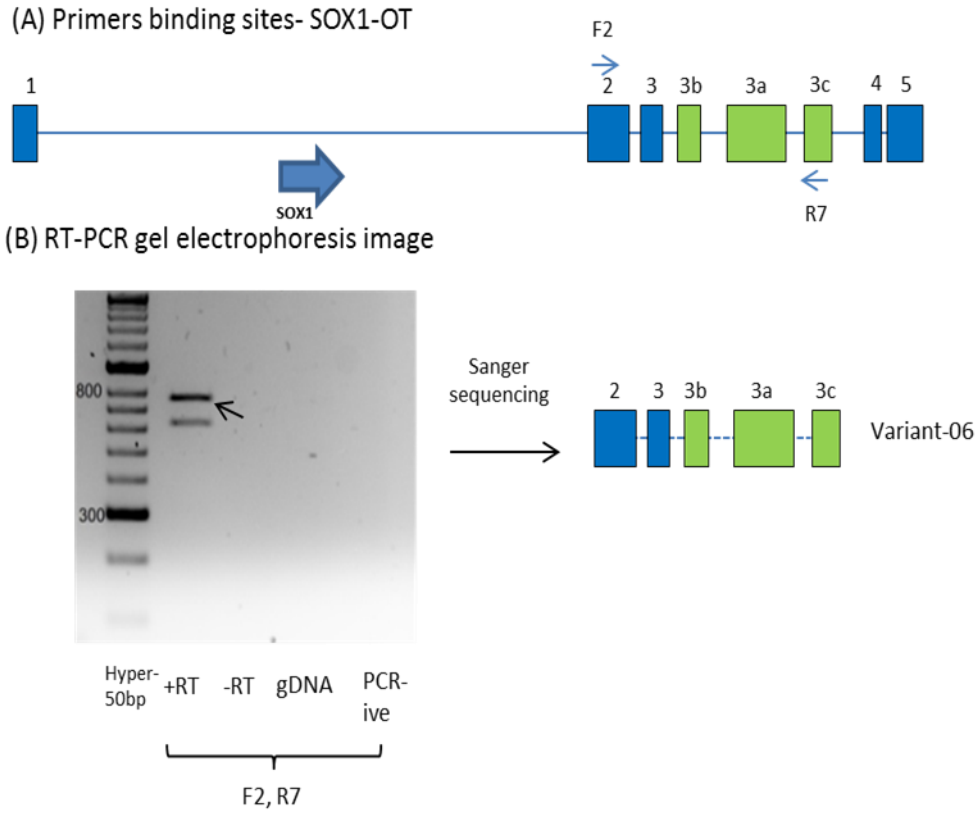


Figure 4-11: **Identification of Transcript variants 6:** (A) primer binding sites in the *SOX1-OT* for the primer pair (F2, R7). (B) Gel electrophoresis image; showing the bands obtained for the +RT sample (ReN cells). By Sanger sequencing the amplified PCR fragment (pointed arrow) was identified as transcript variant-6, illustrated next to the gel image having annotated exons (blue box) and unannotated exons (green box), the lines represent introns which spliced out during RNA processing.

Other transcript variants 7-8 and 9-11 that were identified in this study have been discussed in (section, 4.2.3.5) and (section, 4.2.4.4) respectively.

4.2.3.4 Structure of the overlapping transcript at location upstream of SOX1 gene:

After the identification of the *SOX1-OT* structure downstream of *SOX1*, the *SOX1-OT* structure upstream of *SOX1* was further investigated. The region upstream of *SOX1* gene is shown in Figure 4-12. Human *SOX1-OT* (LINC00403) and mouse *SOX1-OT* (GM5607) are aligned through UCSC

genome browser [28]. This sequencing alignment shows that the mouse gm5607 transcript starts upstream of the *SOX1* gene in a region with highly evolutionary conserved domains containing the 1st exon of the mouse gm5607 transcript. A primer was designed in this region (F6, R6) to see whether human *SOX1-OT* (LINC00403) has an exon present at same location to the 1st exon of mouse *Sox1-ot* (Figure 4-12).

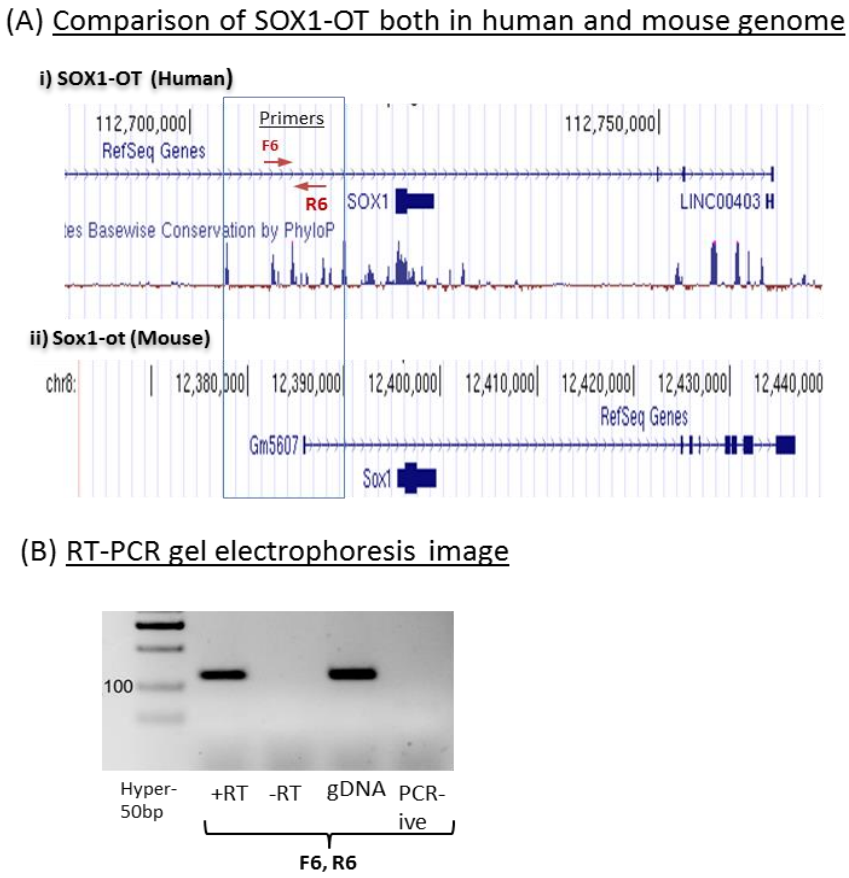


Figure 4-12: **Detection of new exon within SOX1-OT upstream of SOX1 gene:** (A) Human (LINC00403) and mouse (gm5607) overlapping transcript comparison, Arrows represents the Primers pair binding site for the human *SOX1-OT*, Image generated by USCS genome browser (http://genome.ucsc.edu) [28]. (B) RT-PCR (F6,R6) product run on 2% agarose gel, band obtained for the +RT sample (ReN cells D6), which was excised and send for sequencing suggested a presence of novel unannotated exon.

RT-PCR amplification using the prime pair (F6, R6) and sequencing of the PCR amplified fragment revealed that human overlapping transcript has

an exon present in this region corresponding to the equivalent genomic location of the mouse overlapping transcript (gm5607). PCR fragments obtained with the (F6, R6) primer pair are shown in Figure 4-12B.

After the detection of an unannotated exon at a genomic location upstream of the *SOX1* gene, this exon was number 1a in the structure of *SOX1-OT* identified so far by RT-PCR (Figure 4-13).

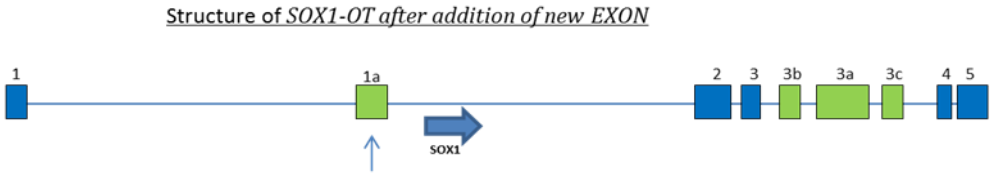


Figure 4-13: **Structure illustration of SOX1-OT after addition of new exon upstream of *SOX1***: The *SOX1-OT* diagram showing the on-going building structure of the transcript, which has now a new unannotated exon (1a) identified just upstream of SOX1-gene pointed by arrow. Blue box (annotated exons), green box (unannotated exons).

Human and mouse overlapping transcript comparison in the Figure 4-12A, suggested that ReN cells might have a *SOX1-OT* transcription start site similar to the mouse *SOX1-OT* and the annotated exon-1 of the transcript might not exist. Therefore, RT-PCR was performed to amplify 1st annotated exon of the *SOX1-OT* in the ReN cells differentiated at day-6, using the primer pair (F5a, R5a) which binds within exon-01 (Figure 4-14A). A fragment was amplified from the +RT sample in the ReN differentiated cell at Day6 (pointed arrow, Figure 4-14B) but the Sanger sequencing of the PCR fragment demonstrated that the amplified PCR fragment was a result from unspecific amplification of the primer pairs. The band obtained for the gDNA was found specific to the 1st exon determined by sequencing.

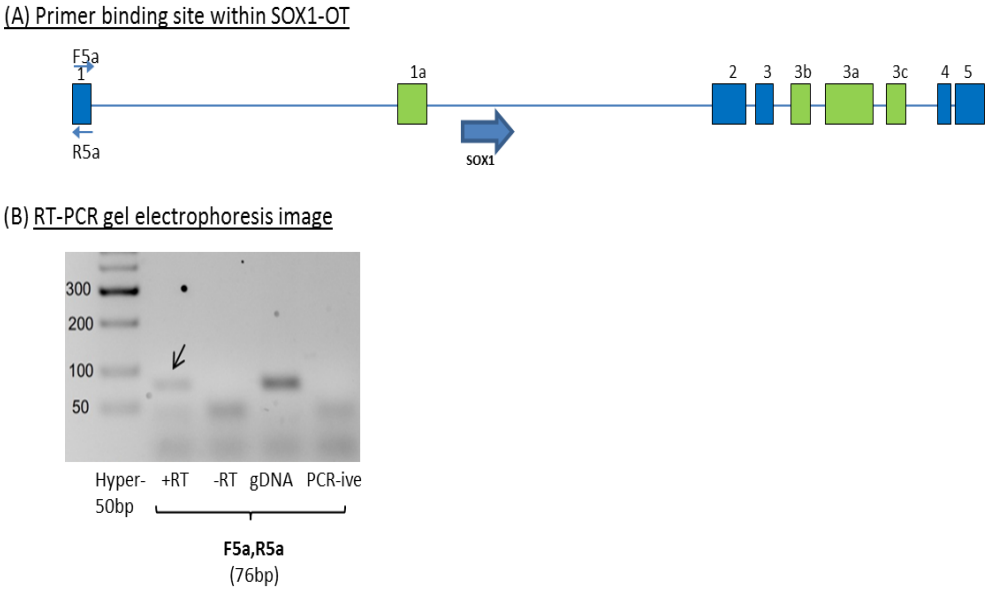


Figure 4-14: **Detection of annotated exon 1 of the SOX1-OT variant 1:** (A) Primer binding site for the primer pair (F5a, R5a) within the 1st exon of the *SOX1-OT*. (B) RT-PCR products obtained were run on 2% agarose gel, bands obtained were excised and sent for Sanger sequencing.

Next, the primer pair (F5a, R6a) was designed to amplify the region from 1st exon in the sense direction and the newly identified exon-1a in the anti-sense direction (Figure 4-15A). No band was detected for the prime pair in the +RT sample of the ReN differentiate cells at day-06, suggesting that if there is a 1st exon present in the transcript it is not co-expressed with exon-1a(Figure 4-15B).

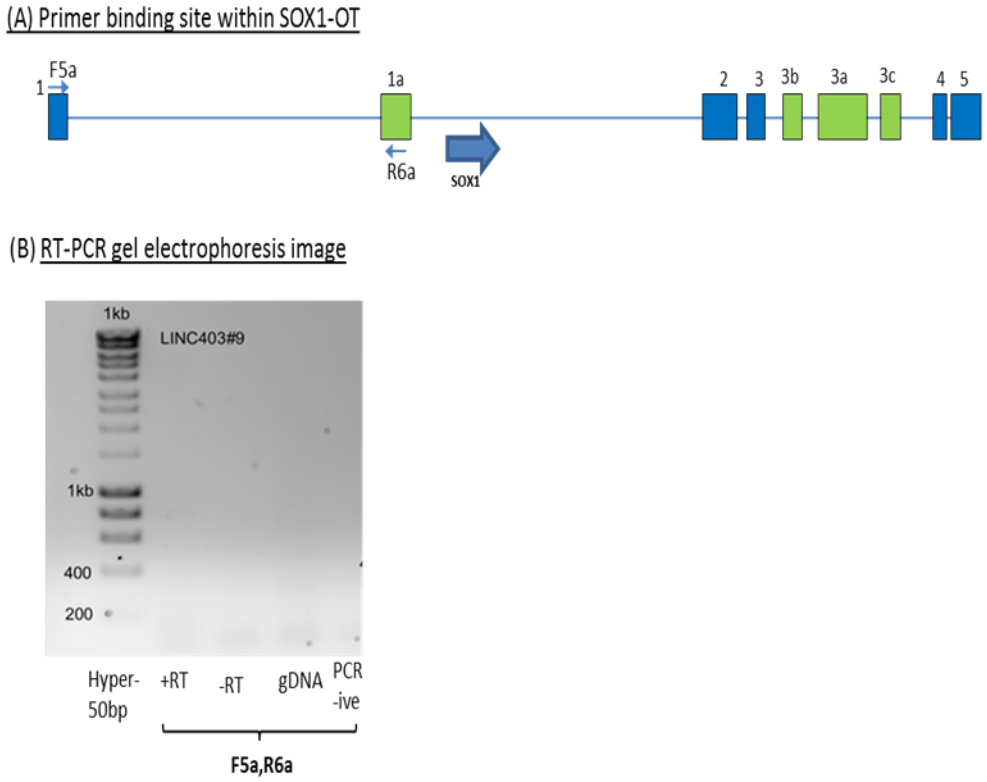


Figure 4-15: **RT-PCR detection of SOX1-OT upstream of SOX1:** (A) Primer pair amplified the region between 1st exon and newly identified exon 1a. (B) RT- PCR product was run on 2% agarose gel, no band was observed.

4.2.3.5 Detection of SOX1-OT having exons that overlap the SOX1 protein coding gene:

Primer pair (F6, R2) has primers in exon 1a and 2, allowing the amplification of the *SOX1-OT* with one primer upstream and one downstream of the SOX1 gene (Figure 4-16A). Fragments amplified from the +RT sample are shown in Figure 4-16B. Sequences obtained for the PCR amplicon were aligned to the human genome assembly through BLAT search in UCSC genome browser (Figure 4-16C). This was the first time that *SOX1-OT* transcript which spans the region including the SOX1 gene detected in the ReN differentiated cells at day 6.

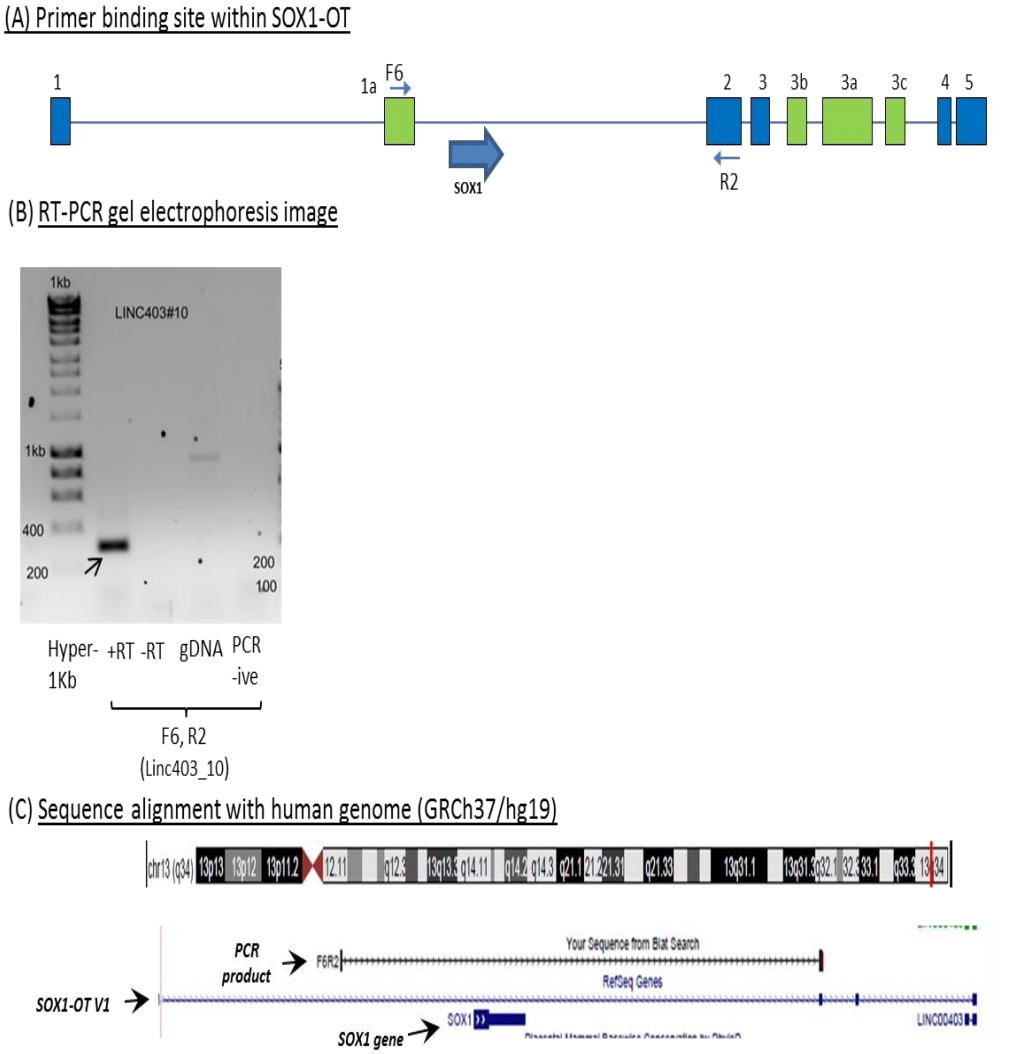


Figure 4-16: **Detection of SOX1-OT part that overlap the SOX1 gene**: (A) Primer binding sites for (F6, R2) are shown within *SOX1-OT*. (B) RT-PCR product was run on 2% agarose gel, PCR amplified fragment (arrow) was excised and send for Sanger sequencing (C)The sequence from the PCR product obtained was BLAT against the human genome (GRCh37/hg19) [178], image was generated by using the UCSC genome browser (http://genome.ucsc.edu) [28].

Detection of *SOX1-OT* was extended further to the downstream exons as illustrated by primers binding sites in Figure 4-17A. RT-PCR results obtained are shown in the Figure 4-17B. Results from the sequence alignment to the human genome by BLAT search are shown in the Figure 4-17C. It was found that primer pair (F6, R1) and (F6, R3) had amplified novel transcript variants of *SOX1-OT* which were number 7 and 8

respectively (Figure 4-17C), while primer pair (F6, R7) failed to detect the overlapping transcript.

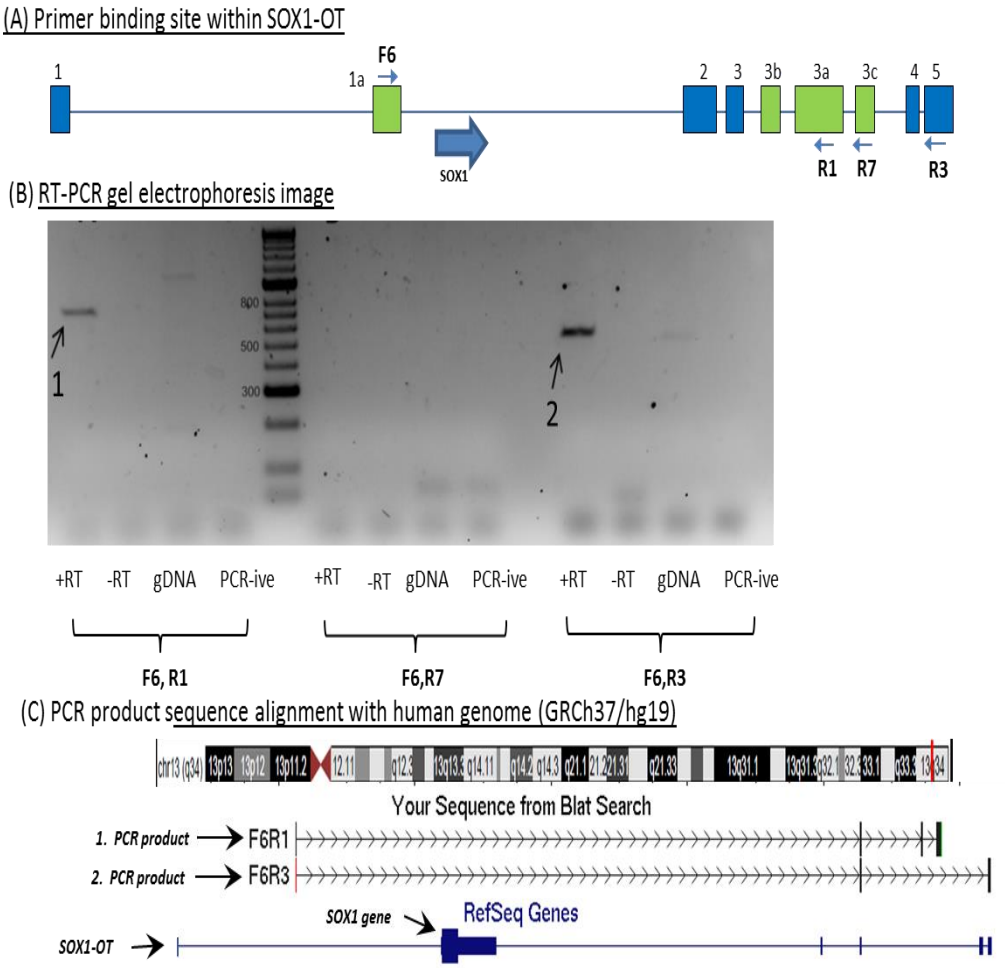


Figure 4-17: **Detection of SOX1-OT extended to downstream region of the transcript**: (A) Different RT-PCR primers binding sites are shown on the top. (B) RT-PCR product was run on 2% agarose gel, PCR amplified fragments (arrow) were excised and send for Sanger sequencing (C) The sequences from the RT PCR product obtained were BLAT against the human genome (GRCh37/hg19) [178], image was generated by using the UCSC genome browser (http://genome.ucsc.edu) [28]

SOX1-OT transcript variant 7 and 8 are transcripts which overlap the SOX1 gene. Transcript variant 7 and 8 are illustrated in the Figure 4-18.

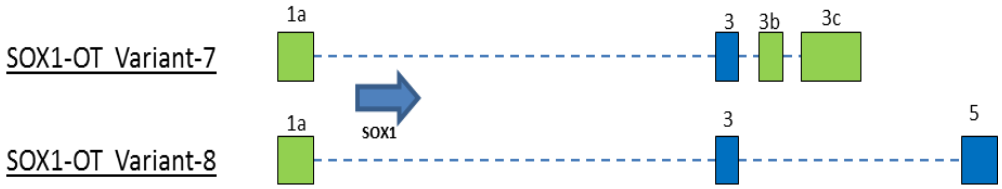


Figure 4-18 **Illustration of exons expression for the transcript variant 7 and 8:** Annotated exon (blue box) and unannotated exons (green box), dotted line represent introns which splice out during RNA processing.

In order to summarise the RT-PCR results, It was found that *SOX1-OT* has 6 novel transcript variants (Figure 4-19A) and 4 novel exons (Figure 4-19B) in the ReN cells. The 1st exon of the annotated transcript was not detected in the ReN cells. After the structure characterisation of *SOX1-OT* by RT-PCR, it was followed by 5'RACE experiment to identify a TSS for the *SOX1-OT*.

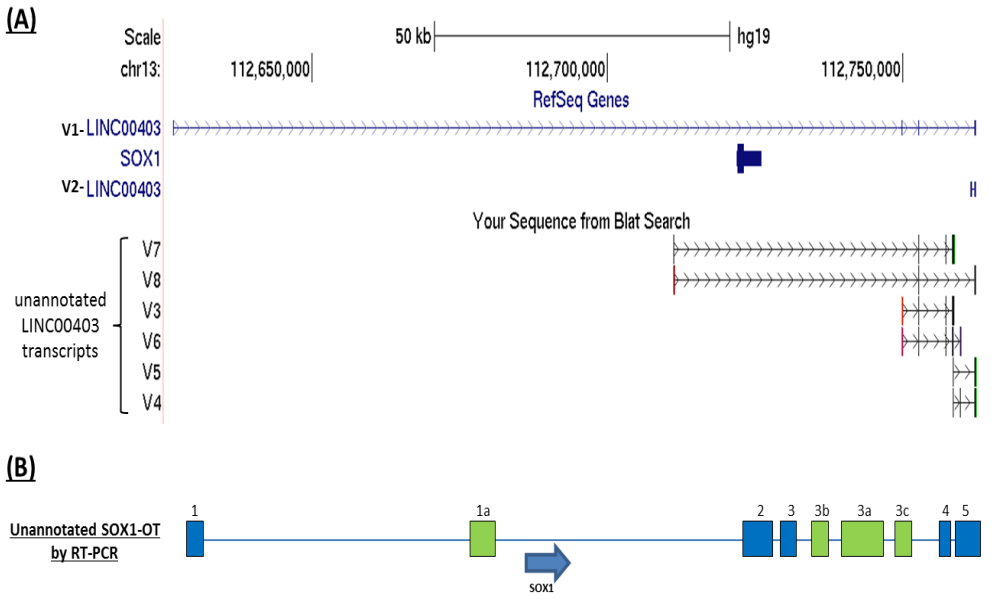


Figure 4-19: **Overview of the RT-PCR results:** (A) Schematic representation of RT-PCR identified different unannotated *SOX1-OT* variants, aligned to the human genome (GRCh37/hg19) through BLAT search [178]. Image generated by UCSC genome browser (http://genome.ucsc.edu) [28]. (B) Composite structure of the

unannotated *SOX1-OT* identified by RT-PCR. Annotated exon (blue box) and unannotated exons (green box) are shown.

4.2.4 5'RACE experiment to identify TSS of the SOX1-OT

5' RACE was designed to determine the transcription start sites of the isoforms of SOX1 overlapping transcripts containing exon 5. During the 5'RACE experiment, different validation steps were carried out to check for the presence of the desired transcript before proceeding to the next step.

4.2.4.1 GI-Primary PCR amplification:

After dC tailing of the cDNA, the GI primary PCR amplification was carried out with primer GI and a primer in exon5 (GSP2) partially overlapping at the 5' end with the primer used for reverse transcription (lane 2 and 3). The control RNA was amplified with GI primer and control primer cGSP1 (711bp) provided with the kit (lane 1). The control RNA is in-vitro transcribed RNA from the chloramphenicol acetyltransferase (CAT) gene that has been engineered to contain a 3' poly(A) tail. Gel image obtained from running half of the PCR reaction suggested that there are very faint bands for the primary PCR product in the +RT sample suggesting that RACE protocol has led to some degree of amplification. The control RNA has given band of the expected size which also proved that the 5'RACE protocol was working see Figure 4-20.

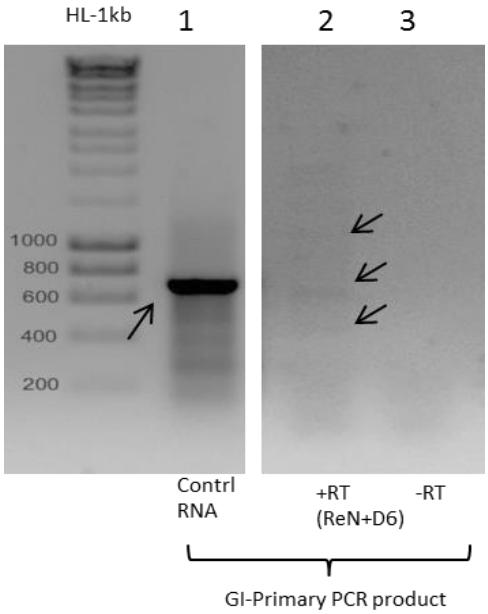


Figure 4-20: **GI-Primary PCR amplification**: ReN cells +day6; Primary PCR product from the 5'RACE has been run on 2% agarose gel. Band for the kit positive control has been obtained as pointed out by an arrow, while no clear band can be seen for the ReN cells+Day6 primary product.

4.2.4.2 AUAP secondary PCR amplification:

2µl of the primary GI amplification (+RT & -RT) was then used as template for the secondary amplification using the AUAP primer and a primer in exon 5 (GSP3) nested to the one used for the primary PCR. 10µL of secondary PCR reaction was loaded on 1% agarose gel. ReN differentiated cells at day6 have multiple amplified fragments (Figure 4-21, Lane 1) suggesting that different splice variants of *SOX1-OT* contain exon5, a finding consistent with the data obtained by RT-PCR. Lane-2 in Figure 4-21 is the No-RT control from primary PCR (-RT) for the ReN cells Day6 RNA; in this lane there is a fragment just below 600bp which could be due to presence of untailed RNA, to contamination or, as the template for this was the –RT product of the first round, could represent non-specific products.

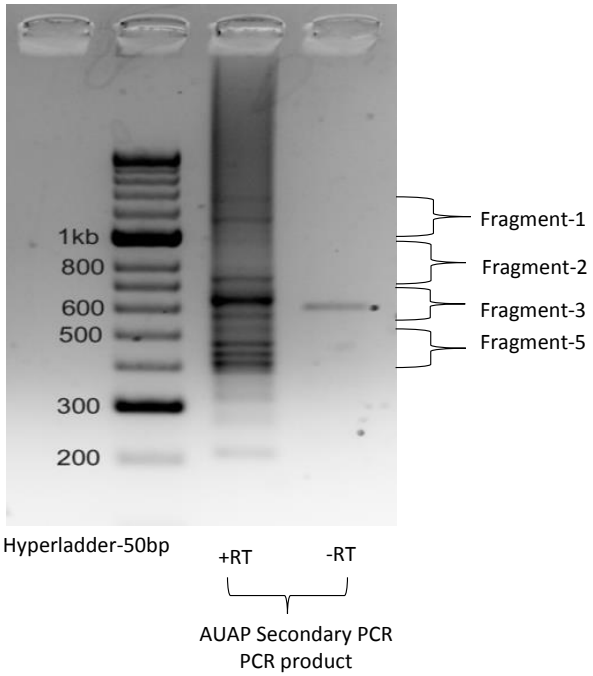


Figure 4-21: **AUAP secondary PCR amplification:** ReN cell+Day6 5'RACE- 10µL secondary PCR product was run on 2% agarose gel. Multiple bands (divided into different fragments 1-3 & 5) in the +RT sample suggest the presence of desired different transcript variants amplified by 5'RACE.

In order to verify the presence of the known target transcripts in the AUAP secondary PCR, 2µl of this was amplified by PCR using the *SOX1-OT* primer pairs to characterise the exon structure of the transcript (Figure 4-22). The primer pair (F3, R3) failed to amplify the transcript which is consistent with previous finding (data not shown). Primer pair (F6, R1) failed to amplify its target transcript variant that was identified earlier during RT-PCR analysis (Figure 4-17). Amplification products obtained with others different primer pairs have amplified different transcript variants of *SOX1-OT* suggesting that the desired transcripts are present in the secondary PCR products from the 5'RACE. The finding was compared with previously identified transcripts with same primer pairs, see Figure 4-17 for prime pair (F6, R3), see Figure 4-8 for (F2, R1) and (F2, R3), see

Figure 4-11 for (F2, R7) and see Figure 4-9 for (F1, R3). These finding indicated that the secondary PCR had amplified the target transcripts and the experiment was preceded further.

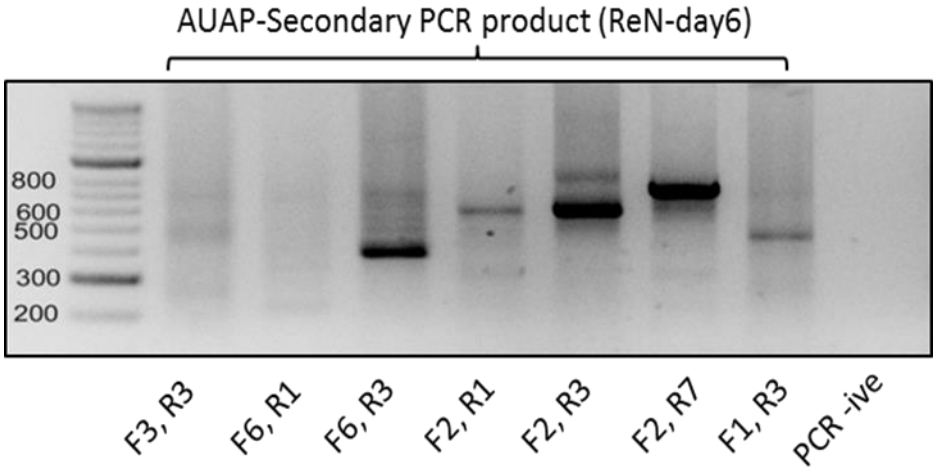


Figure 4-22: **RT-PCR performed on the 5'RACE AUAP-Secondary PCR product,** using different RT-PCR primer pairs that can detect different transcript variants of *SOX1-OT*. The PCR products obtained were run on the 2% agarose gel.

4.2.4.3 PCR detection of the desired insert in bacterial clones

5'RACE product or AUAP secondary PCR different sized fragments were isolated (1-3 & 5) from the gel, as shown in the Figure 23. The purified 5'RACE product were clone in a bacterial culture. A PCR was performed on overnight bacterial growth for individual clones (for fragment 1 and 3 only) to test which clones contained transcripts (an insert) for *SOX1-OT* (Figure 4-23). The negative clones were discarded.

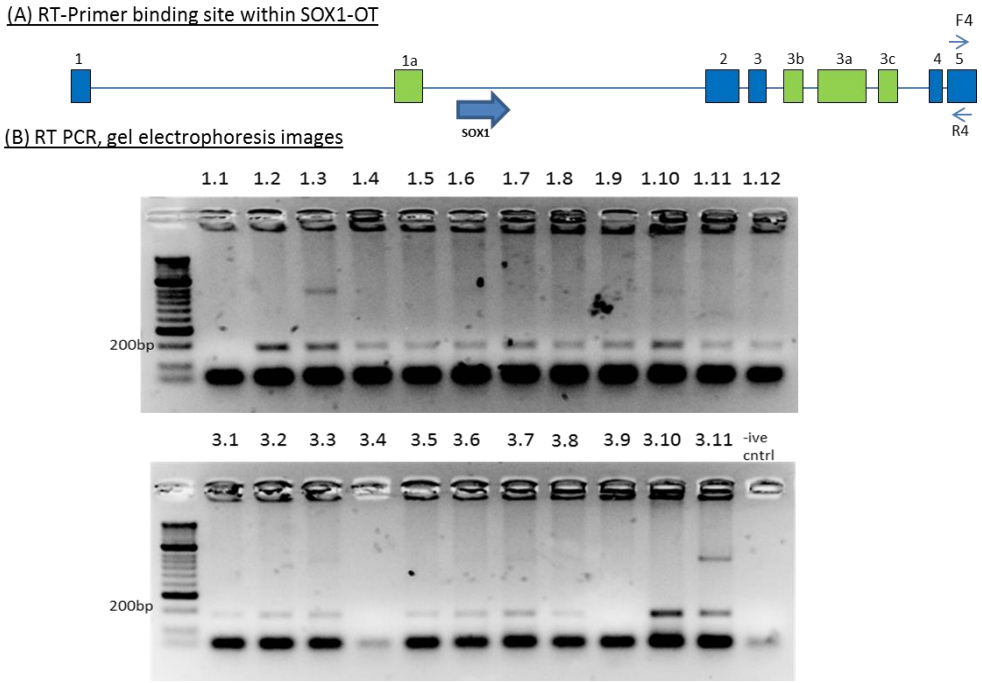


Figure 4-23: **PCR detection of the desired insert in bacterial clones:** (A) Schematic representation of the primer pair (F4, R4) binding site specific for the exon-5 of the *SOX1-OT*. (B) RT-PCR performed on the overnight bacterial culture from the fragment 1 and 3, by using prime pair (F4, R4) to check for the presence of desired insert in the plasmid.

Clones that were positive for *SOX1-OT* insert (5'RACE product) were further processed for plasmid DNA extraction. The plasmid DNA so obtained was then incubated with EcoRI to determine the range of insert sizes. Following are the gel images showing EcoRI enzyme digestion for fragment 1-3 & 5 (Figure 4-24). Different size fragments suggest different size insert into the plasmid as it was expected to be different transcript variants of *SOX1-OT*. All of the positive clones were sent for sequencing.

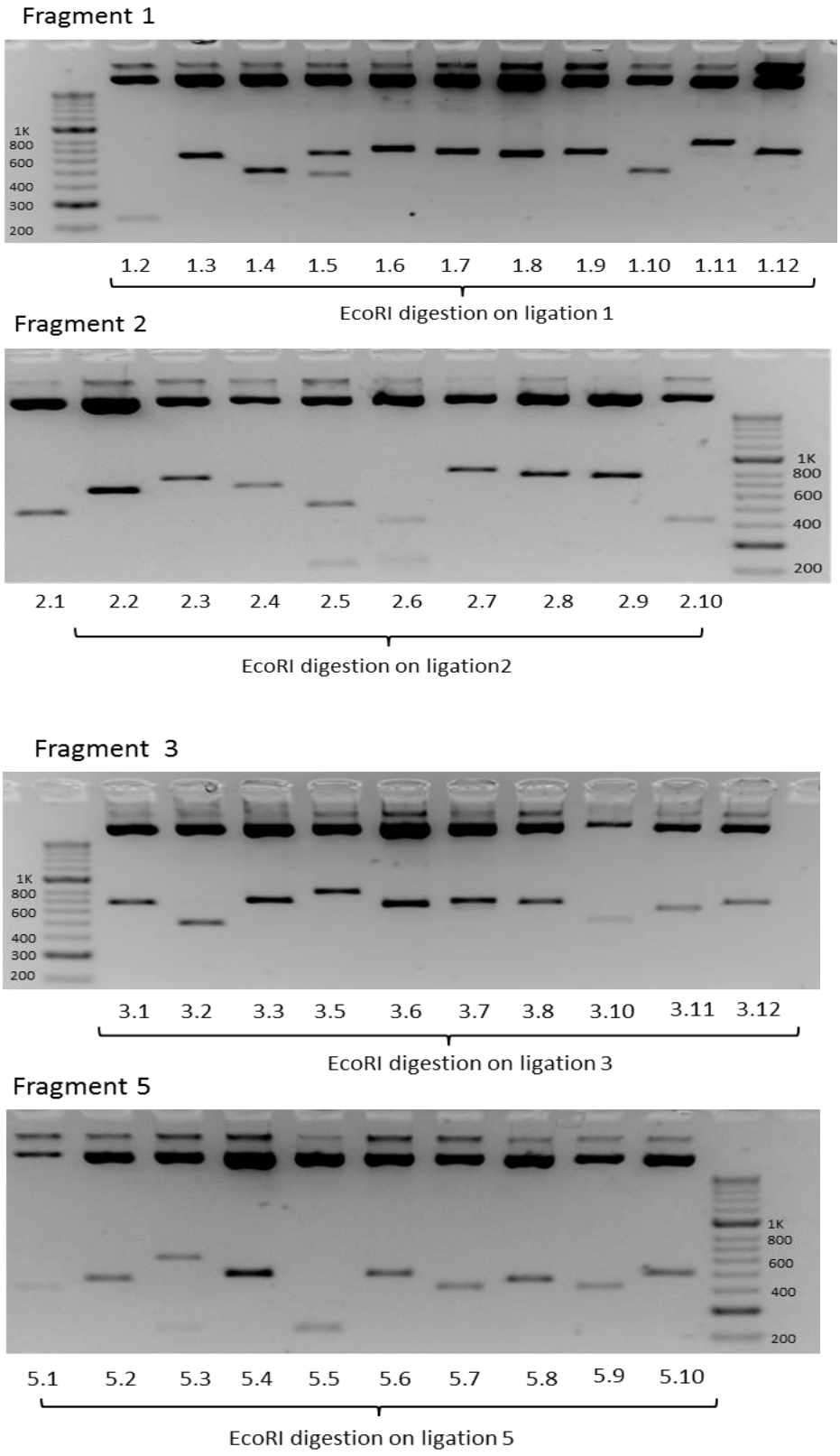


Figure 4-24: **Agarose gel images for EcoRI digestion product**: EcoRI digestion product from the fragment 1-3 & 5 were run on a 2% agarose gel; different sized bands suggest the presence of the desired different sized inserts in the plasmid.

4.2.4.4 Sanger sequences alignment to UCSC genome data

DNA plasmid was then sent to SourceBioscience for Sanger sequencing. The 5'RACE sequencing data were trimmed of the vector sequences and aligned to UCSC human genome browser assembly as shown in Figure 4-25.

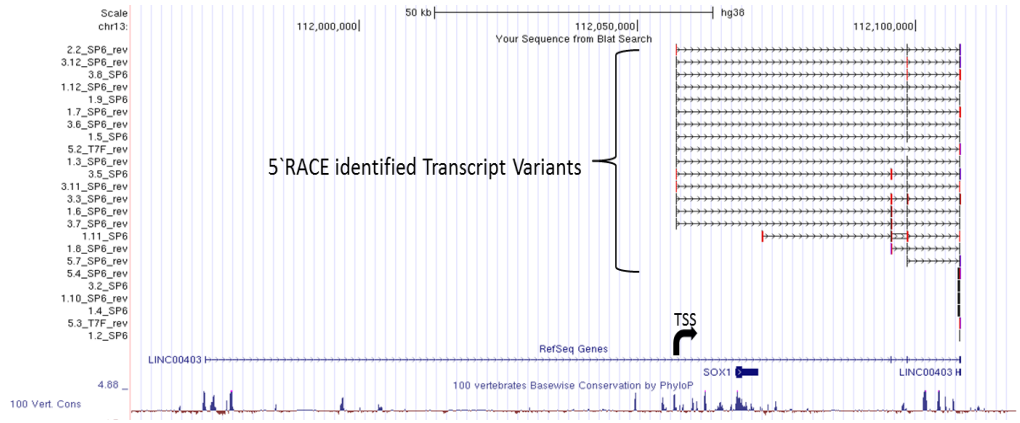


Figure 4-25: **Transcript variants amplified by 5'RACE**: Sequence alignment for all the *SOX1-OT* variants that were detected by 5'RACE. Sequences were aligned to the human genome data set (GRCh37/hg19) by BLAT search application [178], and image was generated by UCSC genome browser (http://genome.ucsc.edu) [28].

Results from the 5'RACE experiment has identified a novel unannotated exon (1b) at a position between exon 1a and exon 2 of the unannotated structure of *SOX1-OT* (Figure 4-26A). Three (3) novel transcript variants for the *SOX1-OT* have been identified by 5'RACE (Figure 4-26B) in addition to the 6 transcript variants that were identified by RT-PCR (Figure 4-20). At least one potential TSS can be recognised by 5'RACE (bend arrow, Figure 4-25). Interestingly the *SOX1-OT* transcript has TSS upstream of SOX1 gene within the same region identified by RT-PCR primers, see Figure 4-19. The 1st exon of the annotated *SOX1-OT* was found to be not amplified by 5'RACE which is consistent with RT-PCR findings. This suggest that ReN cells might not use the annotated 1st exon or that

transcript initiating at this exon might be expressed at very low levels which are not detectable by the used techniques.

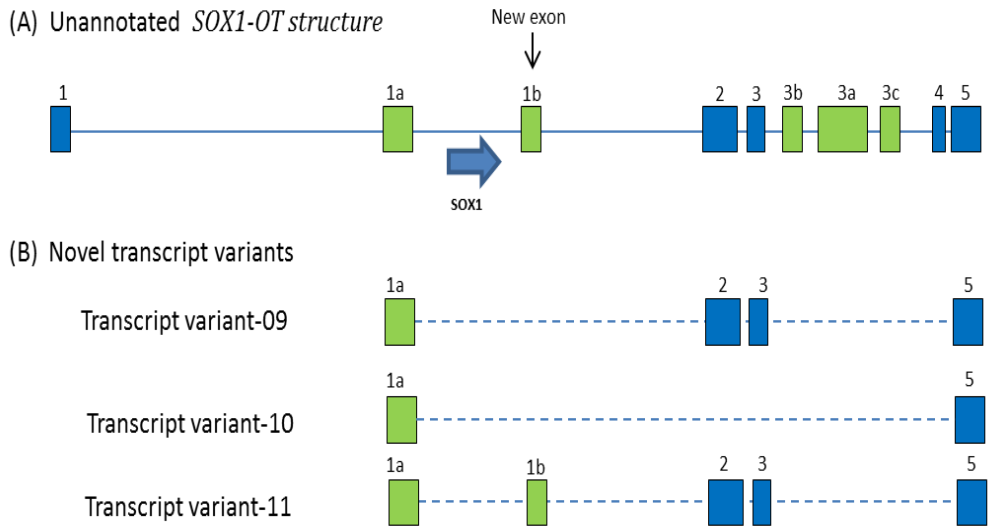


Figure 4-26: **Composite structure of SOX1 identified through RT-PCR and 5'RACE**: (A) Schematic illustration of unannotated structure of *SOX1-OT* in the ReN cells, showing position of the identified novel exon 1b by 5'RACE (B) Schematic illustration of novel transcript variants identified by 5'RACE experiment. Boxes represent exons; annotated exons (blue box) and unannotated exons (green box), dotted lines represent introns

4.2.5 PCR detection to test whether AK55143 gene is a part of SOX1-OT

Using an application (In other Genomes-convert) from the UCSC genome browser tool suite ([28], the genomic co-ordinates of the *SOX1-OT* identified in this study (chr13:112,613,049-112,785,559, GRCh37/hg19) were converted to mouse genome (GRCm38/mm10) coordinates chr8:12315941-12452701 (Figure 4-27). Interestingly, it suggested that the genomic location of the transcription termination end (TTE) of the mouse *SOX1-OT* was further downstream compared to that of the human *SOX1-OT*. The conversion of genomic coordinates of *SOX1-OT* shows that the mouse *SOX1-OT* transcription termination ends aligned with the human AK5145 gene which is located next to the termination end site for

human *SOX1-OT* (Figure 4-27). Therefore, we aimed to check whether human *SOX1-OT* has the same termination end site compared to mouse transcript; for this we designed RT-PCR primers to check if the AK55145 gene exon is part if the *SOX1-OT* transcript.

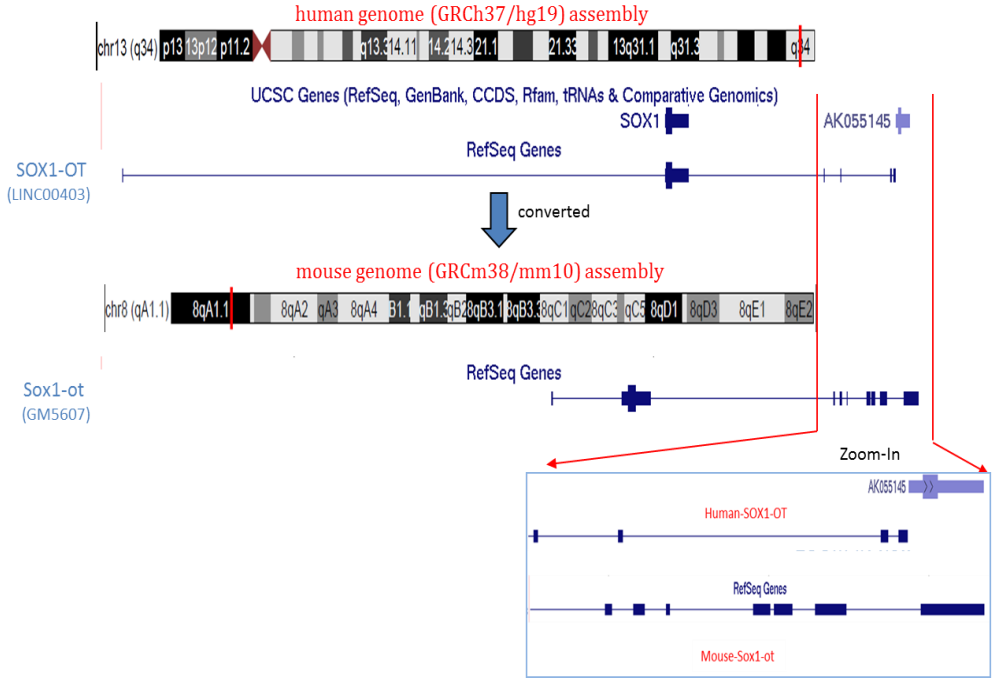


Figure 4-27: **Conversion of genome co-ordinates of SOX1-OT between human and mouse genome assembly**: *SOX1-OT* genomic region from human genome (GRCh37/hg19) assembly was converted into the mouse genome (GRCm38/mm10) assembly through UCSC genome browser tools [28]. The image was adopted from UCSCS genome browser (http://genome.ucsc.edu). The Zoom-in view shows the end region of SOX1 human and mouse transcript, the AK05145 gene aligning to the end part of mouse *SOX1-OT*.

First of all, expression of AK55145 was tested in ReN cell at day 0 and 6 using a primer pair (F12, R12) specific for this transcript. As shown in the Figure 4-28 this gene is only express at day6 of neural differentiation.

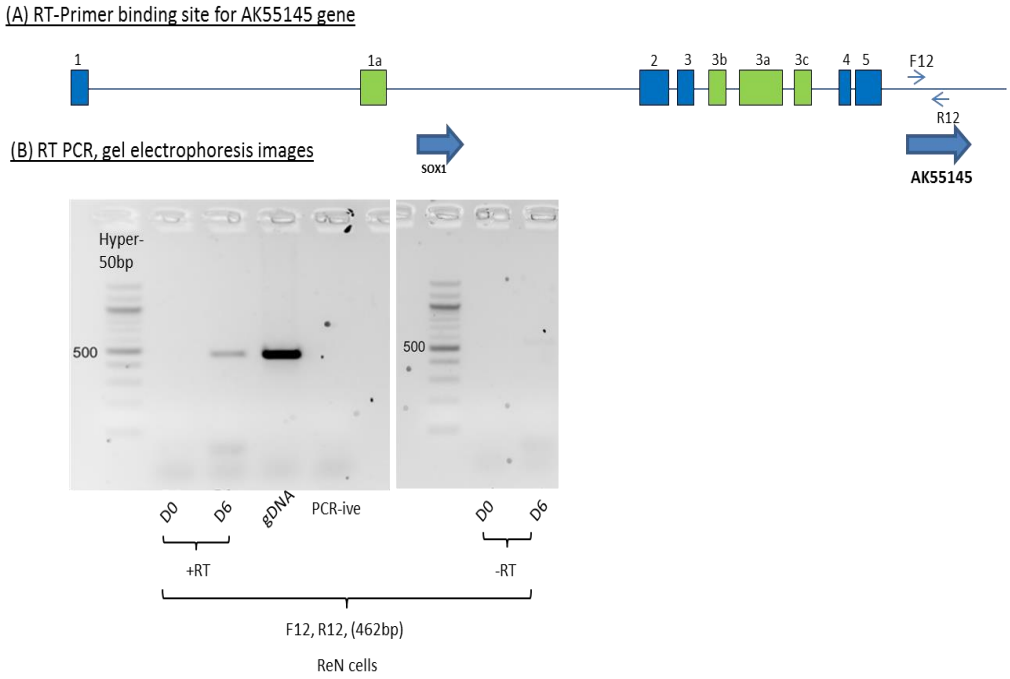


Figure 4-28: **PCR amplification of AK55145 gene:** (A) Illustration of primer binding sites for AK55145 has been shown (B) RT-PCR gel picture shows the bands obtained for AK55145 gene on day6 of differentiated ReN cells.

To test whether AK55143 is a part of *SOX1-OT*, A primer pair (F4, R12) was designed such that the sense (F4) primer binds to the last exon of *SOX1-OT* (exon 5) while the anti-sense (R12) binds to upstream of AK55143 gene (Figure 4-29A).

The RT-PCR result showed that the primer pair (F4, R12) amplified a fragment of 550bp in the +RT sample of day6 differentiated ReN cells (D6) and 762bp from gDNA which correspond to the distance between the two primers in genomic DNA. These data suggest that *SOX1-OT* last exon-5 and AK55143 are part of the same transcript. No fragment was amplified from Day-0 +RT sample which is consistent with the observation that no expression for *SOX1-OT* is observed in undifferentiated ReN cells (day 0), see Figure 4-29.

To confirm our findings, the fragments amplified from the Day6 +RT sample and from gDNA were excised, cleaned and sent for direct sequencing. Sequencing results have confirmed that AK55145 gene is actually a part of *SOX1-OT* transcript as shown in the UCSC genome browser generated images, see Figure 4-29.

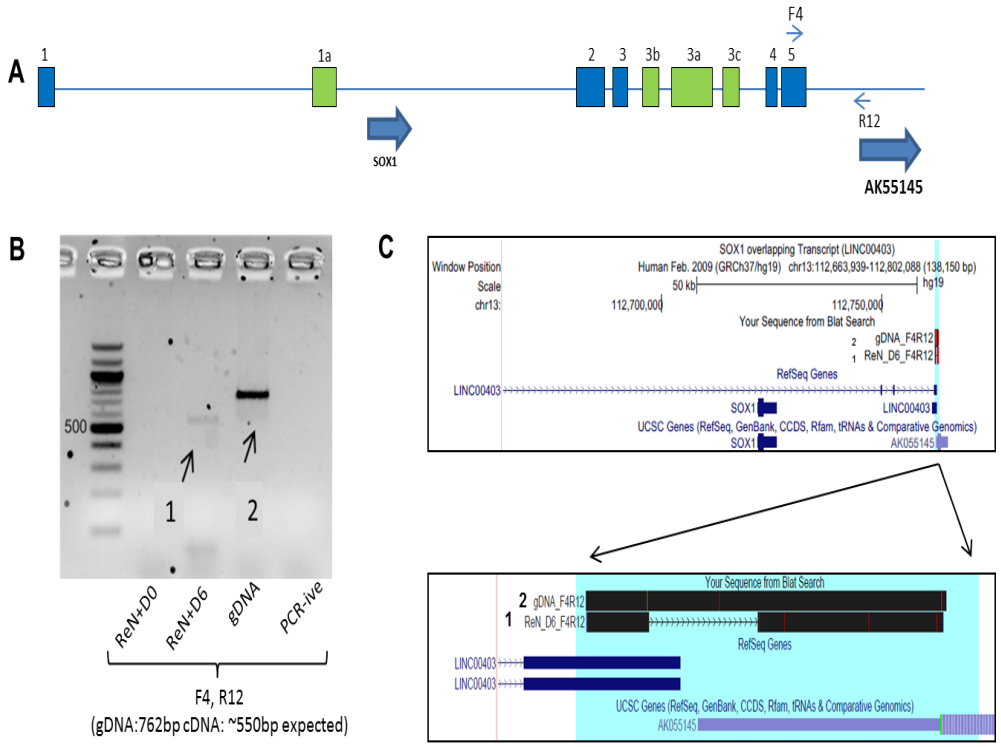


Figure 4-29: **RT-PCR amplification showing AK55145 is the part of the SOX1-OT**: (A) Primer binding sites for (F4, R12) is shown. (B) RT-PCR product was run on 2% agarose gel. (C) PCR fragment (1)+RT(ReN+D6) and (2)gDNA were sent for sanger sequencing, Sequences obtained were aligned to the human genome assembly(GRCh37/hg19) through BLAT search [178]. This image was generated by UCSC genome browser (http://genome.ucsc.edu) [28].

4.2.6 Comparison of SOX1-OT expression in different cancerous and normal cell lines:

Different cancerous and normal cell lines were used to test for the expression of *SOX1-OT* (Figure 4-30). GAPDH was used to assess template input. It was found that exon-5 of *SOX1-OT* is expressed highly in ReN cells differentiated at day6 (Figure 4.7, section 4.2.3.1). Other cell lines that

expressed exon-5 of *SOX1-OT* include NTera, MCF7, SH-SY5Y and T47D. Cell lines that were found negative for exon-5 of *SOX1-OT* are Hela, CaCo2, HOS, HCT116, MDA-MB231, Hs578T and MDA-MB361. This analysis was performed using primer pair (F4, R4) which amplifies exon-5, chosen because previous analysis found that exon-5 is normally expressed in most of the transcript variants therefore increasing chances of detection for *SOX1-OT* in the studied cell lines.

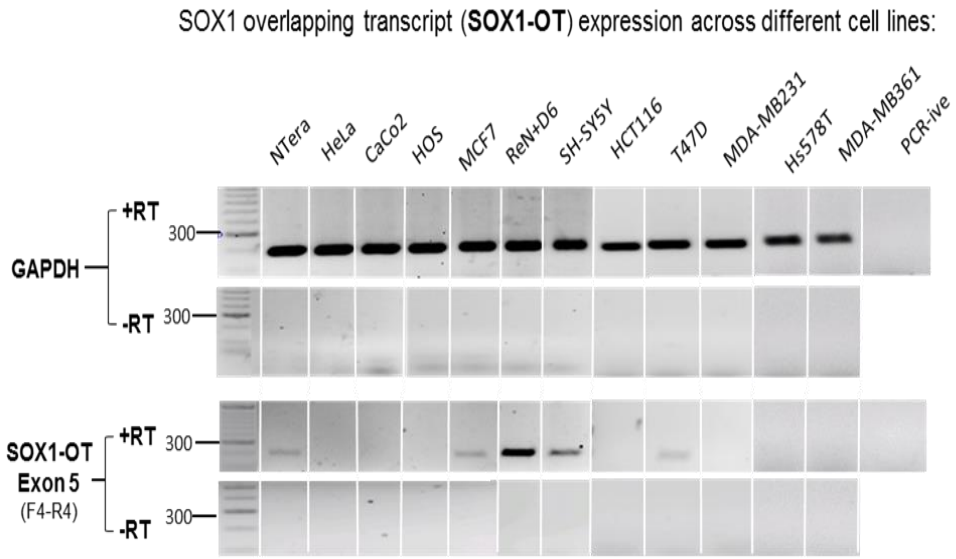


Figure 4-30: **SOX1-OT expression in different cancerous and normal cell lines**: RT-PCR gel images showing *SOX1-OT* expression across different cell lines, GAPDH was used as a reference gene.

Some of the cell lines were also analysed for the *SOX1-OT* variant which overlaps SOX1 gene (Figure 4-31). It was found that only ReN differentiated cells at Day-6 expressed the specific overlapping transcript variant and other cell lines were found negative.

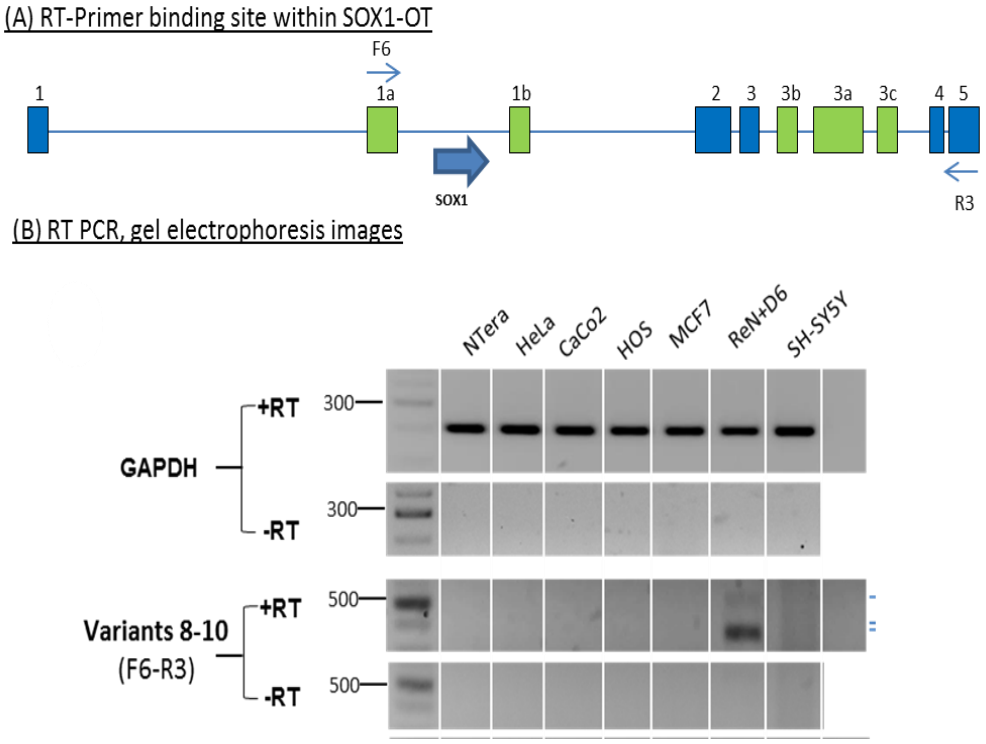


Figure 4-31: **Detection of SOX1-OT variants overlapping SOX1 gene across different cell lines:** (A) Schematic representation of the primer pair (F6, R3) binding site specific for the *SOX1-OT* region that overlaps SOX1 gene. (B) RT-PCR for the *SOX1-OT* across different cell lines, PCR product was run on 2% agarose gel.

4.2.7 Comparison of SOX1 overlapping transcript expression at different time points of ReN cells differentiation

Previous results have suggested that *SOX1-OT* is mostly expressed in ReN neural cells at day-6 of differentiation, therefore we were interested to analyse *SOX1-OT* expression at different time points (day0, 2, 4 and 6) during neural differentiation of ReN cells. Results obtained are shown in the Figure 4-32, *GAPDH* was used to assess template input. *SOX1-OT* containing exons 1a, 3a and 5 were initially analysed. *SOX1-OT* expression was found upregulated at day 6 as compared to day0. Overall, *SOX1-OT* expression increases along the neural differentiation of ReN cells as shown in the Figure 4-32.

Different RT-primer pair combinations were also used to detect some of the other variants at different time points of neural differentiation (Figure 4-32). It has been found that transcript V1 was expressed at low levels at day 0 and increased in expression at day2 which then remained constant at this higher level at day 4 and 6 of neural differentiation in ReN cells. While transcript V3 has no expression at day0, it appeared to be upregulated at day2 and then its expression decreases with neural differentiation of ReN cells. Transcript V4 and 5 were expressed at day 2, 4 and 6 while levels were below detection at day 0. Interestingly, expression of transcript V4 and 5 switched between day 2 and 4 and it is tempting to speculate that this switch may play a role in neural differentiation of ReN cells.

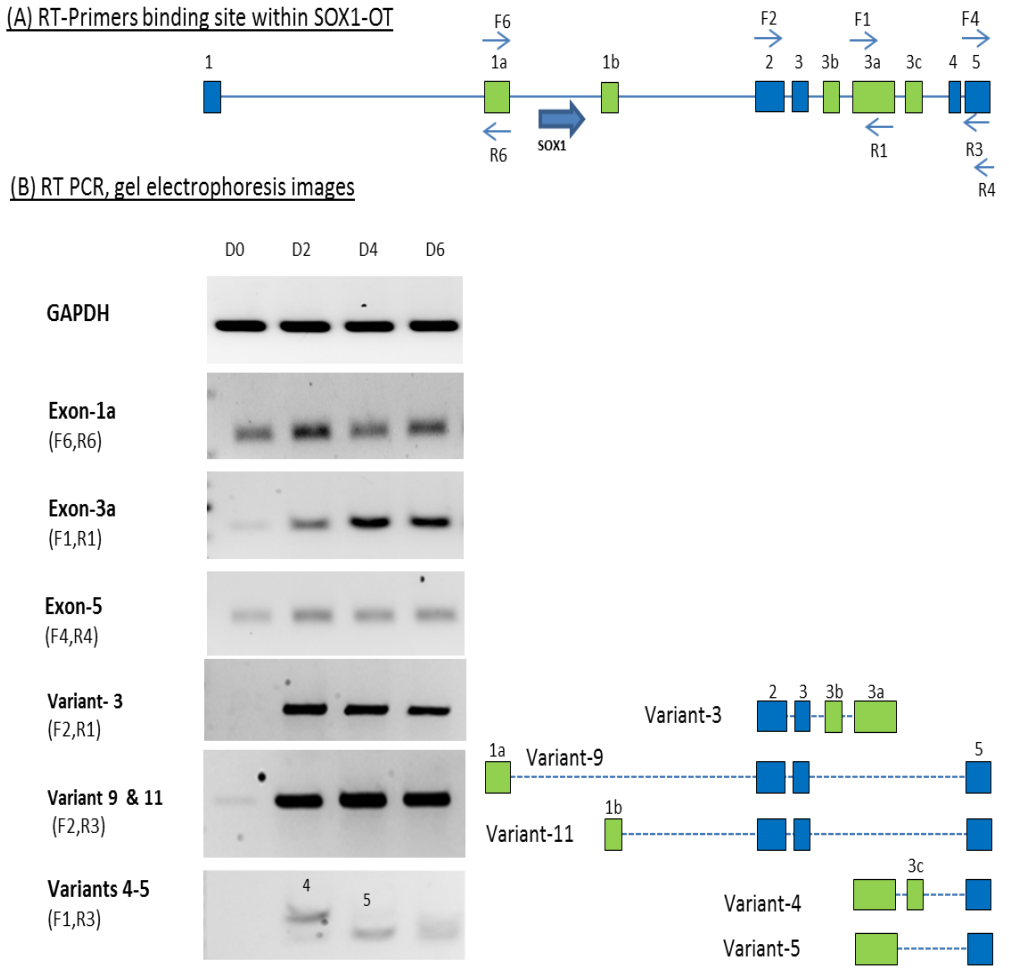


Figure 4-32: **RT-PCR detection of SOX1-OT variants and exons during neural differentiation**: (A) Schematic representation of different primer pairs binding site in the *SOX1-OT*, with blue arrows pointed in the direction of amplification (B) RT-PCR comparison of *SOX1-OT* expression at different time points of ReN cells differentiation at Day0-6, by using different primer pairs. Schematic illustration of different transcript variants structure has been shown next to the corresponding bands. GAPDH was used as a reference gene.

4.3 Discussion

Genome wide studies have reported large numbers of non-coding RNAs whose function and significance are not clear. To understand the complex transcriptome architecture, expression and regulation of genetic information it has become necessary to distinguish between mRNA and ncRNA transcripts [222]. Human *SOX1* overlapping transcript *(SOX1-OT)* is annotated as a long intergenic non-coding mRNA like transcript that has no significant coding potential, although proteomics studies have recently suggested that many predicted 'non-coding' RNAs actually code for very small peptides to mediate their cellular function [223]. In this study, the structure of the *SOX1-OT* was further characterised in ReN neural cells using two different techniques, RT-PCR and 5'RACE. Its possible role as a regulator of *SOX1* gene in neural stem cell differentiation and in cancer development has been discussed.

4.3.1 Characterization of the structure of *SOX1-OT*

In this study, it was found that in ReN cells human *SOX1-OT* has a complex structure which includes a few novel exons and different transcript variants which are unannotated in the human genome. The results showed that human *SOX1-OT* has a total of 10 exons, 5 of which (Exon1a, 1b, 3a, 3b, and 3c) are novel and previously unknown (Figure 4-2A). In addition to the two annotated transcript variants (V1-V2), we report 9 new transcript variants of *SOX1-OT* (V3-V11) which have not been previously reported in the literature (Figure 4-2C). Therefore, *SOX1-OT*

presents complex transcriptional features whose potential diverse biological significance and functions remain to be explored.

Identification of different transcript variants of *SOX1-OT* by RT-PCR and 5'RACE suggested more than one transcription start sites (TSS). FANTOM5 project tracks from the UCSC genome browser tool suite were used to identify the core regulatory sequences for *SOX1-OT* [28, 224]. FANTOM5 project data tracks were aligned to the unannotated *SOX1* mRNA sequence identified by 5'RACE and RT-PCR, see Figure 4-33 [178] [224]. The FANTOM5 project provides genome-wide mammalian gene expression data by mapping transcription start sites (TSSs), promoter region and enhancer in human and mouse primary cells, cell lines and tissues [225]. It has been found that two potential sites for the *SOX1-OT* have high peaks of total CAGE reads (Cap analysis for gene expression) as shown in the Figure 4-33. This observation is consistent with our finding by both RT-PCR and 5'RACE which also suggest a TSS for the *SOX1-OT* upstream of SOX1 gene, with genomic location in close proximity to the *SOX1* promoter region. Therefore, it is possible that the potential TSS upstream of *SOX1* gene might have potential regulatory activities for the *SOX1-OT* transcript and its close genomic location to the *SOX1* promoter region suggest it might also regulate *SOX1* gene expression. The likelihood of *SOX1-OT* acting as a regulator of *SOX1* is further supported by recent data indicating that *SOX2* transcription can be regulated by *SOX2*-OT within which *SOX2* lies [215].

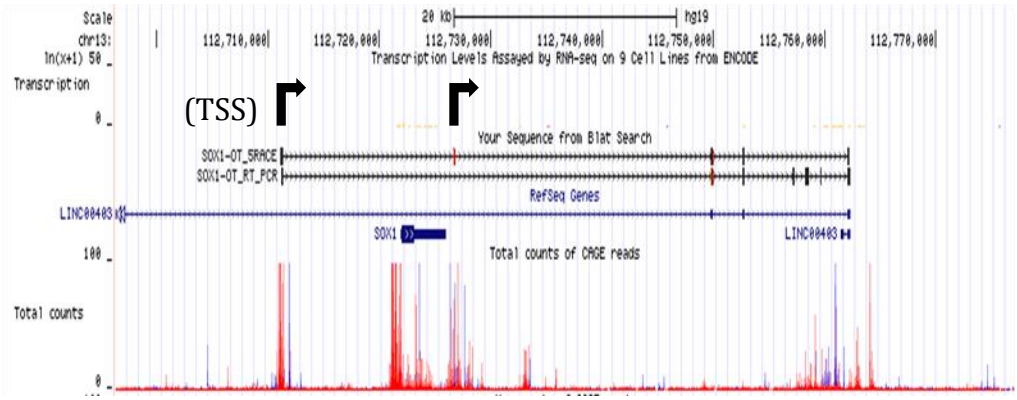


Figure 4-33: **Alignment of CAGE reads to the newly identified SOX1-OT mRNA sequence:** UCSC generated image (http://genome.ucsc.edu) [28] shows the FAMTON5 tracks [224, 226] providing peaks for total counts of CAGE reads i-e, 5'-end of the mapped CAGE reads represent TSS activities in the sample, this datahub is set up to provide the TSS activities in individual biological states and the identified regions. Two potential TSS sites for *SOX1-OT* has been show by black arrow. RED peaks:Total counts of CAGE reads forward, BLUE peaks:Total counts of CAGE reads reverse. CAGE analysis was performed across 975 human and 399 mouse samples, ncluding primary cells, tissues and cancer cell lines, using single-molecule sequencing.[226] *SOX1-OT* mRNA sequences were aligned to home genome assembly (GRCh37/hg19) by BLAT search [178].

The data presented in this chapter shows that the 1st exon of the RefSeq annotated transcript LINC00403 (*SOX1-OT*) is either absent or expressed at levels below the detection limit of our methods in ReN cells. However, it is important to note that the annotated structure of *SOX1-OT* has been obtained by combining information collected from three different tissues types which are amygdala (GenBank: DA195709.1), foetal eye (GenBank: BQ184460.1.1) and lung-carcinoid (GenBank: AI693652.1); this might explain the differences with this study which focused on characterizing the transcript in a well-defined cell type. The structure of *SOX1-OT* in ReN cells has the same sites for transcription start sites (exon-1a) identified by both RT-PCR and 5'RACE, see the complementary sequence alignment of *SOX1-OT* to the human genome in the Figure 4-34. In contrast, the differences between the information gathered by these two different

techniques are the downstream unannotated exons (3a, 3b and 3c) which were identified by RT-PCR but not picked up by 5'RACE (Figure 4-34). A possible explanation for this could be that 5'RACE anti-sense gene specific primer (GSP-1) was nested in the last exon-5 of the transcript, which means that only transcript variants containing the last exon-5 were detected by 5'RACE. Therefore, it is possible that these downstream exons are not co-expressed with exon-5 and are actually spliced out. In order to support this observation designing a 3'RACE with a sense primer in the exon-1a which will amplify the transcript in a sense direction (5'-3') till the last exon can possibly detect those transcript variants containing exons 3a, 3b and 3c.

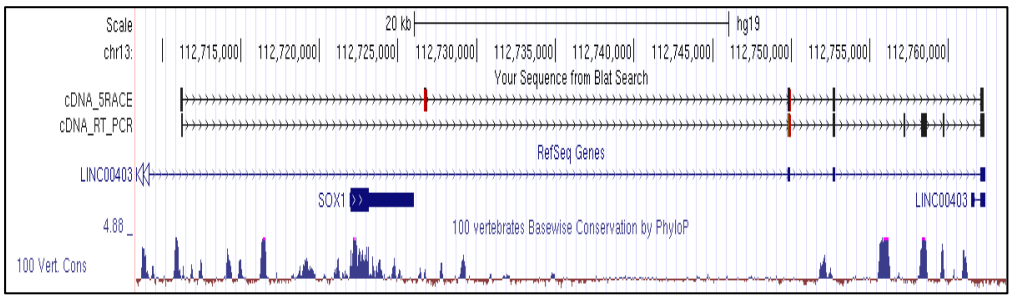


Figure 4-34: **SOX1-OT, cDNA sequences obtained through RT-PCR and 5'RACE**, were aligned to the human genome data (GRCh37/hg19) through BLAT search [178], This image was generated by UCSC genome browser (http://genome.ucsc.edu) [28].

Interestingly, the newly characterised structure of human *SOX1-OT* resembles that of the annotated mouse *Sox1-ot*. They both have TSS upstream of and near to the *SOX1* coding gene; moreover, though the 3` end of mouse overlapping transcript extend further downstream as compared to the current annotation of the human *SOX1-OT*. RT-PCR findings (Figure 4-29) extend human *SOX1-OT* to include the downstream AK55145 gene thus paralleling the mouse transcript 3'end. It was found

that this 3'end is only used in D6 ReN and not in earlier time points of differentiation; further work will be required to determine whether transcription termination end (TTE) usage is regulated in a cell type/ tissue/ differentiation stage specific manner.

4.3.2 Potential role of *SOX1-OT* in neural differentiation as a regulator of *SOX1*

To further analyse the relationship between *SOX1* and *SOX1-O*T expression in neural stem cells (ReN cells), their expression was evaluated over a time-course of differentiation (Figure 4-35). Relative quantification of *SOX1* expression at D0, D2, D4 and D6 of differentiation ReN cells showed that *SOX1* mRNA is significantly upregulated at day 2, 4 and 6 of neural differentiation compared to Day 0 (Figure 4-35B). Similarly, *SOX1-OT* expression was found up-regulated over the 6 day neural differentiation treatment, with a significant increase in signal within the first 2 days of treatment (Figure 4-35C). This indicated that *SOX1-OT* is detected alongside *SOX1* during neural differentiation of ReN cells. The observed co-regulation of *SOX1-OT* and *SOX1* transcription in neural differentiation is similar to what reported for *Sox2-ot* and *Sox2* during mouse neurospheres differentiation *in vitro* [215] Different transcript variants of *SOX1-OT* are differentially expressed during the course of neural differentiation. It is therefore feasible to speculate that co-expression of *SOX1-OT* and *SOX1* during neural differentiation might have co-regulatory role in pathways regulating neural differentiation. Furthermore, there has been a switch between transcript variant 4 and 5 from day 2 to 4, further supporting its possible regulatory role during neural differentiation.

Further experiments will be required to determine if *SOX1-OT* plays key functional roles in neural differentiation.

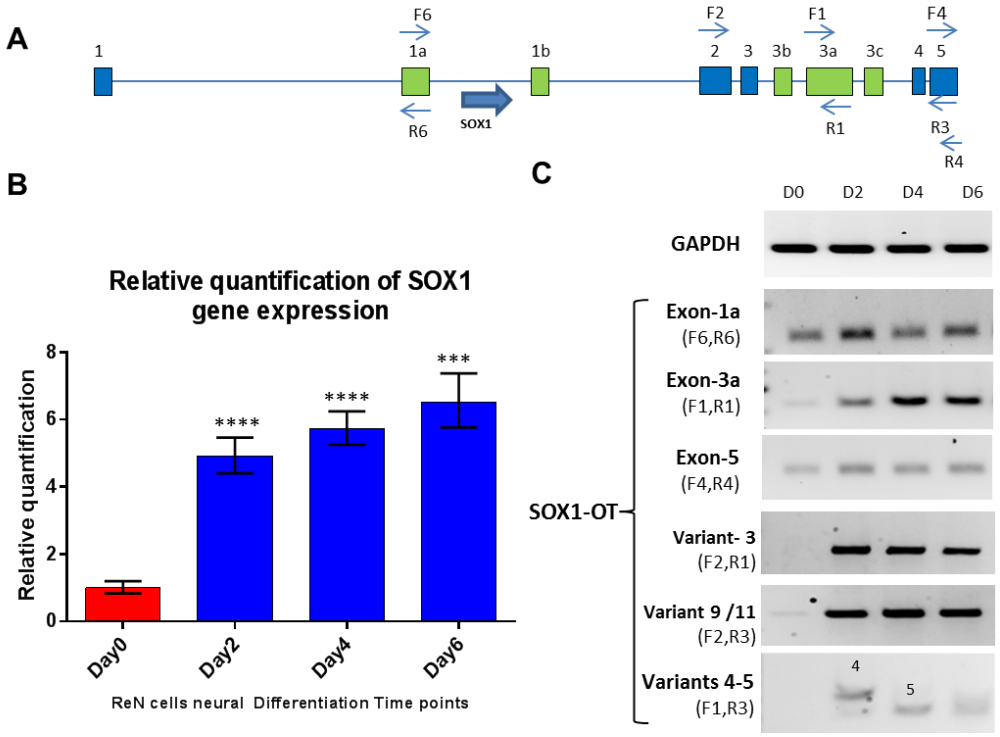


Figure 4-35: **Comparison of SOX1 and SOX1-OT expression across different time-points of ReN cells differentiation:** (A) Primer binding sites for the *SOX1-OT* RT PCR primer pairs. (B) Relative quantification of SOX1 gene expression analysed by qPCR across different time-points of ReN cells differentiation (day 0, 2, 4 and 6). SOX1 was significantly UP-regulated at day 2, 4 and 6 in comparison to D0. , n=3, ****P value <0.0001, 95% confidence interval, Error bars represent the standard error of the Δct's values.. (C) RT-PCR detection of *SOX1-OT* (exon1a, 3a and 5 and variants 3, 4, 5, 9, 11) in ReN cells undergoing differentiation (day 0, 2, 4 and 6).

4.3.3 *SOX1-OT* and *SOX1* are concomitantly expressed in different cancerous cell lines.

We detected co-expression of *SOX1-OT* and *SOX1* RNAs in NTera, T47D and MCF7 cancer cell lines. Concomitant expression of *SOX2* and its LncRNA *SOX2*-OT has been described in different cancer types and it was shown that *SOX2* gene expression is regulated by *SOX2*-OT. For example, *SOX2*-OT is upregulated together with *SOX2* and *OCT4* in esophageal

squamous cell carcinoma [212]. Moreover, co-expression of *SOX2*-OT and *SOX2* has been previously reported in NTera cell line that *SOX2*-OT is functionally associated with *SOX2* gene in pluripotency and tumorigenesis [215]. Also, concordant expression of *SOX2* and *SOX2*-OT has been reported in breast cancer and both are upregulated in cell suspension culture conditions that favours growth of stem cell phenotype [213]. Our finding of *SOX1-OT* expression in the NTera cell line which possesses stem cell like property indicates a potential role of *SOX1-OT* in pluripotency and cancer development. Similarly to what seen for *SOX2* and *SOX2*-OT, expression of *SOX1-OT* and *SOX1* in breast cancer cell lines (MCF7 and T47D) also shows a possible role of *SOX1-OT* in the regulation of *SOX1* expression in breast cancer. *SOX1* expression has been already reported in several cancer types such as hepatocellular carcinoma (HCC), prostate, cervical, ovarian and non-small cells lung cancer (NSCLC) [6-9, 12]. The data has shown that co-expression of *SOX1-OT* and *SOX1* in different cancer cell lines might hint to a functional role for *SOX1-OT* in tumorigenesis and stem cell pluripotency. Therefore, *SOX1-OT* might have a potential role in cancer by promoting *SOX1* expression; its expression in different cancer types in which *SOX1* has already been reported needs further investigation.

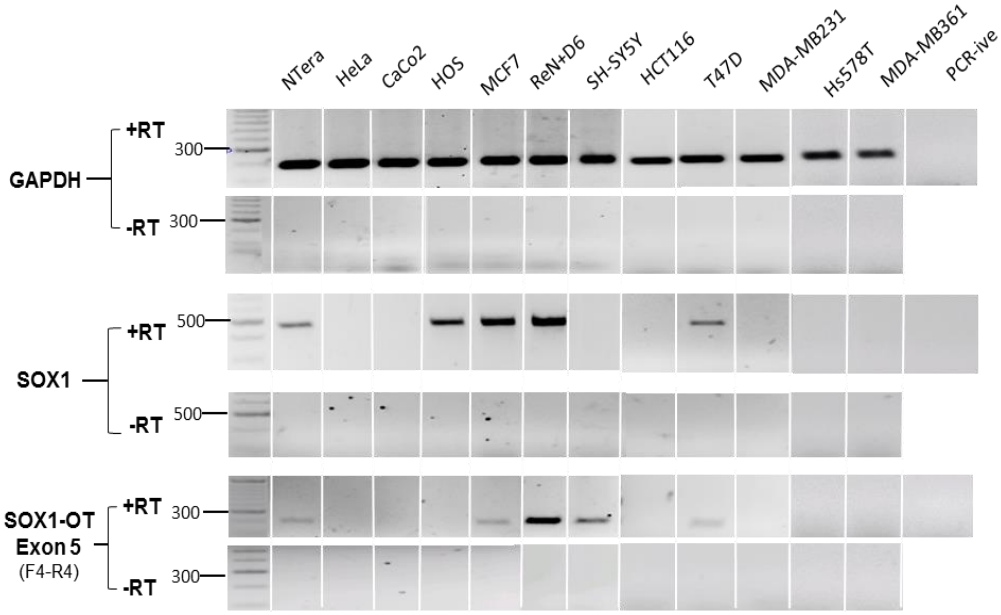


Figure 4-36 **Matching of SOX1 and SOX1-OT expression across different cell lines:** RT-PCR detection of *SOX1-OT* (Exon 10) and SOX1 expression across different human cell lines.

4.4 CONCLUDING REMARKS

In conclusion, it was found that *SOX1-OT* has a complex structure in ReN cell which includes two unannotated potential TSSs and at least ten (10) exons, Five (5) of which are unannotated in the human genome. *SOX1-OT* contains multiple transcript variants, among them 9 novel transcript variants (V3-V11) have been identified in this study. *SOX1-OT* is highly expressed in differentiated neural stem cell where their expression coincides with that of SOX1 indicating its potential co-regulatory role during neural differentiation. Furthermore, Co-expression of *SOX1-OT* and *SOX1* RNA was found in stem cell and different cancer cell lines, suggesting that *SOX1-OT* may play functional roles in stem cell pluripotency and carcinogenesis by regulating SOX1 expression. Future work needs to determine whether *SOX1-OT* has regulatory role in neural differentiation and potential as a novel marker or therapeutic target in cancer.

5 Chapter: 05

Prediction of SOX1 post-translational modifications and functional domains using bioinformatics approaches

5.1 Introduction

In addition to mechanisms regulating gene expression at a transcript level such as the epigenetic modifications and LncRNAs described in the previous chapters, a gene's function can also be regulated by modulating the properties of the protein it codes for through different types of post translational modifications (PTMs) [227]. There are several post-translational modifications that can occur and their nature will depend on the protein sequence context and on the identity of the particular amino acid. Post translational modifications have been previously discussed in details, (section, 1.5).

SOX1 protein function in stem cell maintenance and tumorigenesis has been well studied, but very little is known about its post translational modifications, which can significantly change a protein properties, in order to regulate its function. SOX1 in cancer is known to acts as a tumour suppressor gene and has been shown to inhibit tumour metastasis *in vitro* [6, 10, 11]. Contrary, SOX1 has been also found to act as an oncogene in prostate cancer where it promotes tumour invasion [12]. The work described in the previous chapters has shown that *SOX1* is differentially expressed in different cancer cell lines and that epigenetic silencing of *SOX1* through promoter DNA hyper-methylation is likely dependent on the

cancer type. *SOX1* expression was also found to coincide with the expression of the overlapping transcript *(SOX1-OT)*, hinting to a potential role *of SOX1-OT* in regulating *SOX1* expression in cancer. Besides these, SOX1 can be significantly regulated at protein level through different types of PTM. Information about the characteristics of the SOX1 protein and PTMs regulating its function has been scarce.

This chapter describes the datamining performed to predict potential post-translational modificatons and to identify functional domains of SOX1 by using online bioinformatics tools and publicly available databases [128]. Most of the PTMs databases that are available online are designed mostly based upon curated experimental evidences and/or wide range of algorithm-based prediction systems that can predict post-translationally modified residues within a known protein sequence [128]. The recent advances in PTM databases and prediction software facilitate the collection of experimentally verified PTM sites in a given protein sequence and, furthermore, the prediction of novel PTM sites. The lack of connection and integration of the many publicly available PTM databases leads to vastly heterogenic results making it difficult to identify potentially true PTMs in a given protein sequence. Therefore, Information from a variety of PTM databases was collected and compared to identify potentially key PTM sites within the SOX1 protein sequence. This will help to inform future experiments to study PTMs likely regulating human SOX1 protein functions.

5.2 Results

5.2.1 Collection of SOX1 PTM evidences from different databases.

5.2.1.1 PhosphoSitePlus® database query for SOX1 PTMs

Initially, the PhosphoSitePlus® database was searched to collect data about different types of PTM of SOX1 [228]. The result obtained is shown in Figure 5-1, the curated data from the PhosphoSitePlus® shows different types of predicted and already reported PTMs for the human SOX1 protein (Figure 5-1 A-B) [228]. The PhosphoSitePlus® database predictions were based on information curated from proteomic experimental data mainly performed through mass spectroscopy on different cell line and tissues [185, 229]. It has been shown that most of the predicted PTM residues in human SOX1 are experimentally verified in mouse SOX1 protein from different cells/tissues (Figure 5-1B) [230].

5.2.1.2 NetPhos3.1 server prediction of phosphorylation for SOX1

After the identification of predicted phosphorylated residues for human SOX1 through PhosphoSitePlus®, an additional search was performed on another phosphorylation specific prediction database called NetPhos3.1 server [186]. NetPhos3.1 gives phosphorylation prediction scores for each residue, helpful to indicate the confidence between a true phosphorylated site and those with very low chances to be phosphorylated. Results from the NetPhos3.1 can be used in addition to other available phosphorylation databases which can help to identify potential phosphorylated sites in a given amino acid sequence. In general, NetPhos3.1 predicts phosphorylated sites with a score of sensitivity in the range from 69% to

96% [15]. The result from NetPhos3.1 shows Serine (S), Threonine (Y) and Tyrosine (T) predicted phosphorylation sites within the SOX1 protein (Figure 5-2A). NetPhos3.1 found that residue S325 (Serine residue present at position 325) is highly likely to be modified with a prediction score of 0.905 (highlighted in yellow in Figure 5-2A). This prediction is supported by the result obtained with other analysis tools such as PhosphoSitePlus®, which has also predicted S325 to be phosphorylated in human SOX1 protein (Figure 5-1). Other residues which were found with high prediction score for phosphorylation are also shown in Figure 5-2B. Most of these phosphorylated residues are found within DNA binding domains of SOX1 such as HMG and SOXp domain (Figure 5-2B), Phosphorylation within these domains might potentially affect DNA binding properties of SOX1.

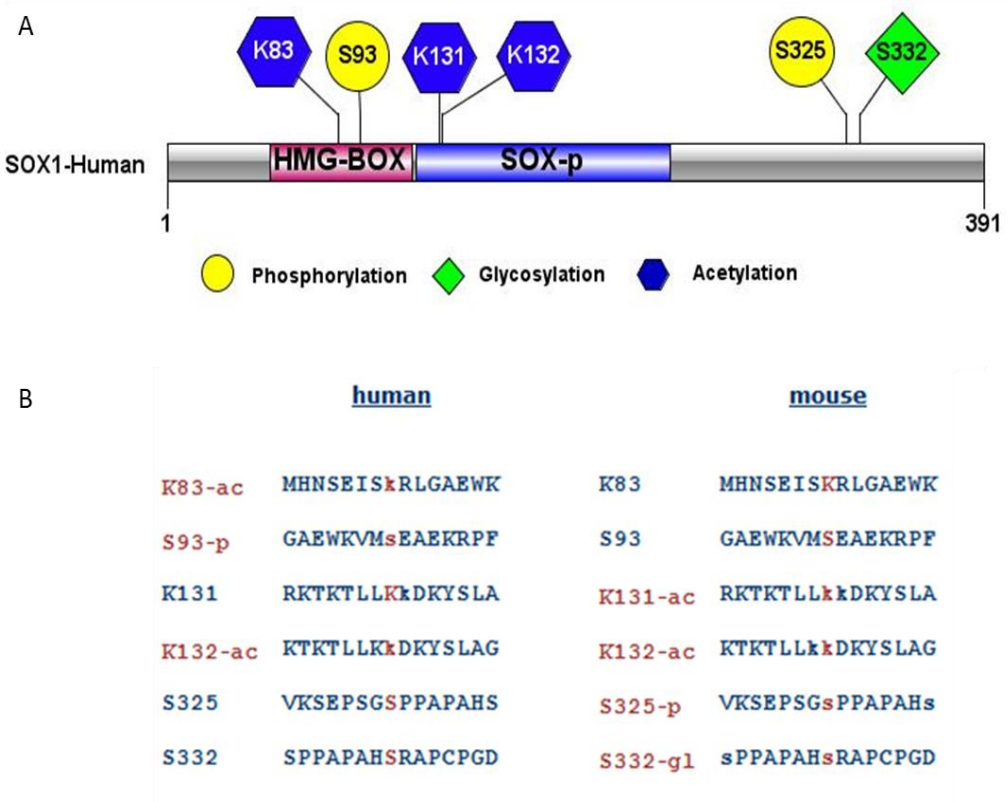


Figure 5-1 **Predicted PTMs residue of SOX1 by PhosphoSitePlus®:** (A) Schematic illustration of the different types of PTMs for the SOX1 protein collected from the PhosphoSitePlus® database; the IBS illustrator webserver was used for the diagram [180]. (B) Multiple sequence alignment for SOX1 protein in both human and mouse show different experimentally validated and predicted PTM residues, these are Phosphorylation (-p), Acetylation (-ac) and Glycosylation (-gl). Validated modified residues are shown in red colour and predicted modified residues are shown in blue, Screenshot of the image was adopted from the query result page of the PhosphoSitePlus® database [228].

A

```
>O00570_SOX1_HUMAN      391 amino acids
#
# netphos-3.1b prediction results
#
# Sequence                  # x   Context     Score   Kinase    Answer
# ------------------------------------------------------------------
# O00570_SOX1_HUMAN        12 S   TDLHSPGGA   0.963   unsp       YES
# O00570_SOX1_HUMAN       118 Y   EHPDYKYRP   0.942   unsp       YES
# O00570_SOX1_HUMAN       126 T   PRRKTKTLL   0.961   unsp       YES
# O00570_SOX1_HUMAN       135 Y   KKDKYSLAG   0.971   unsp       YES
# O00570_SOX1_HUMAN       172 S   QRLESPGGA   0.996   unsp       YES
# O00570_SOX1_HUMAN       271 S   SASPSGYGG   0.950   unsp       YES
# O00570_SOX1_HUMAN       325 S   EPSGSPPAP   0.904   unsp       YES
#
    MYSMMMETDLHSPGGAQAPTNLSGPAGAGGGGGGGGGGGGGGAKANQDR   #     50
    VKRPMNAFMVWSRGQRRKMAQENPKMHNSEISKRLGAEWKVMSEAEKRPF   #    100
    IDEAKRLRALHMKEHPDYKYRPRRKTKTLLKKDKYSLAGGLLAAGAGGGG   #    150
    AAVAMGVGVGVGAAAVGQRLESPGGAAGGGYAHVNGWANGAYPGSVAAAA   #    200
    AAAAMMQEAQLAYGQHPGAGGAHPHAHPAHPHPHHPHAHPHNPQPMHRYD   #    250
    MGALQYSPISNSQGYMSASPSGYGGLPYGAAAAAAAAAGGAHQNSAVAAA   #    300
    AAAAAASSGALGALGSLVKSEPSGSPPAPAHSRAPCPGDLREMISMYLPA   #    350
    GEGGDPAAAAAAAAQSRLHSLPQHYQGAGAGVNGTVPLTHI            #    400
%1  ...........S.....................................    #     50
%1  ................................................    #    100
%1  ................Y.......T.........Y.............    #    150
%1  ....................S............................    #    200
%1  ................................................    #    250
%1  ....................S............................    #    300
%1  ........................S........................    #    350
%1  ....................................
```

B

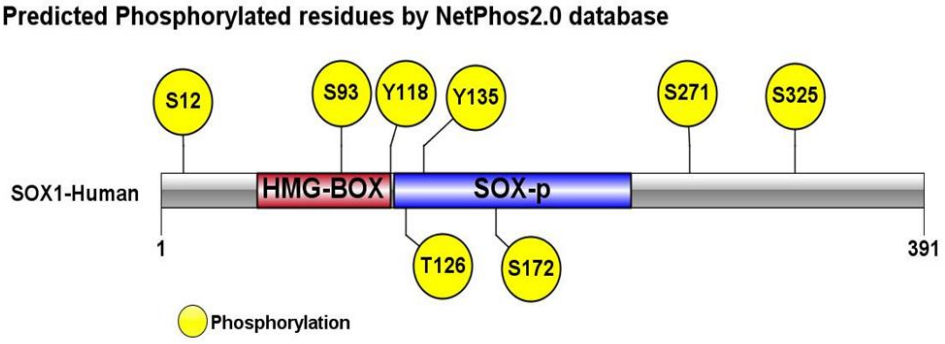


Figure 5-2 **Predicted Phosphorylation within SOX1 by NetPhos3.1:** (A) Result obtained from Human SOX1 protein query by NetPhos3.1. Each residue position is given. Column context has the modified residue motif and column Score has the predicted score for Serine, Threonine and tyrosine residues. Where the residue is marked by a dot means that the residue is not predicted to be phosphorylated, either because the score is below the threshold (0.9) or because the residue is not Serine, threonine or tyrosine (Screenshot of the query result page on NetPhos3.1 database [186]). (B) Illustration of predicted phosphorylation sites for SOX1 using NetPhos3.1; diagram was generated using the IBS illustrator webserver [180].

5.2.1.3 Yin-O-Yang server prediction of Yin-O-Yang effect within SOX1

The interplay between O-GlcNAc modification and phosphorylation which is termed Yin-Yang effect has very important roles in cell regulation and function. O-GlcNAc modification has been found to regulate function of tumour suppressor genes and oncogenes [139]. *SOX1* in cancer can act as a tumour suppressor gene or oncogene depending upon the different

cellular contexts. Therefore, in order to predict if any of the Serine or Threonine residues within SOX1 can have yin-yang effect, the SOX1 protein sequence was analysed using the YinOYang analysis tool [187, 231]. Results obtained are shown in Figure 5-3, showing all residues which are predicted with a potential to be O-GlcNAcylated (type of glycosylation) or phosphorylated or both. Sites which are reversibly and dynamically modified by O-GlcNAc or Phosphate groups at different times in the cell are shown in the YinOYang column of Figure 5-3A [187]. It was found that residue S325 is the only potential true residue which may be showing Yin-Yang effect (Figure 5-3A, line highlighted in red), while residues S308 and S332 have been shown with high chances of O-GlcNAc but very less likely to show Yin-Yang effect (Figure 5-3A, lines highlighted in yellow). It should be noted that S325 has been also predicted as phosphorylated residues by PhosphoSitePlus® and NetPhos3.1 database analysis tools (Figure 5-1, Figure 5-2).

A

SeqName	Residue	O-GlcNAc result	Potential (o-glcnac)	Thresh. (1)	Thresh. (2)	NetPhos potential (Thresh=0.9)	YinOYang?
O00570_SOX1	23	S	+	0.4694	0.3965	0.4848	
O00570_SOX1	195	S	++	0.6184	0.4411	0.5450	
O00570_SOX1	260	S	++	0.5722	0.4292	0.5289	
O00570_SOX1	267	S	+	0.4991	0.4495	0.5562	
O00570_SOX1	295	S	+	0.4802	0.4332	0.5343	
O00570_SOX1	307	S	+	0.4988	0.4255	0.5239	
O00570_SOX1	308	S	+++	0.6482	0.4314	0.5319	
O00570_SOX1	323	S	+	0.4006	0.3575	0.4323	
O00570_SOX1	325	S*	+	0.4019	0.3557	0.4298	0.905 *
O00570_SOX1	332	S	+++	0.5847	0.3889	0.4746	
O00570_SOX1	366	S	++	0.5571	0.4321	0.5328	
O00570_SOX1	389	T	+	0.5044	0.4271	0.5261	

B

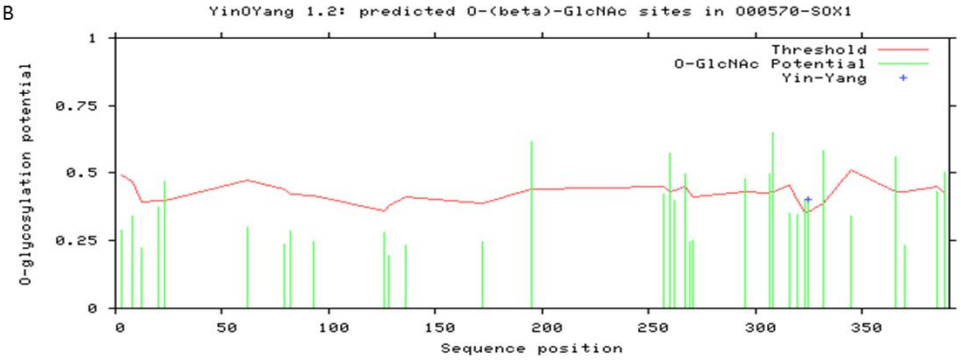


Figure 5-3 **Predicted YinOYang modified residues with SOX1**: Result obtained by YinOYang 1.2 Server showing Glycosylation (O-GlcNAc) and phosphorylation along with predicted residues with YinOYang effect. Asterisk mark represents YinOYang effect (B) Graphical representation of the result by YinOYang server. (Screenshots of image were taken from result page of the YinOYang server)[187, 231]

5.2.1.4 GSP-SUMO and JASSA databases query for SOX1

Sumoylation prediction query was performed for human SOX1 using GSP-SUMOv2.0 and JASSAv4 analysis tools. Both are comprehensive analysis tools for prediction of sumoylation sites and SIM (sumoylation interacting motif) within proteins (Refer to section 2.8.1.5). Results obtained are shown in the Table 11 and Table 12. It was found that human SOX1 has two potential sites for sumoylation at residues K131 and K319 as predicted by the two analysis tools.

Table 11 Result obtained by GSP-SUMOv.20 showing predicted sumoylated residues for SOX1

GSP-SUMOv2.0-Results for SUMO site

ID	Position	Peptide
O00570|SOX1_HUMAN	131	RKTKTLLKKDKYSLA
O00570|SOX1_HUMAN	319	GALGSLVKSEPSGSP

Table 12 **Result obtained by JASSAv4 showing predicted sumoylated residues for SOX1**, Best prediction score (PS) provided by the database has been shown as low or high for each prediction.

JASSAv4 database- Results for putative SUMO site

ID	Position	Peptide	Best PS
O00570|SOX1_HUMAN	131	RKTKTLLKKDKYSLA	Low
O00570|SOX1_HUMAN	319	GALGSLVKSEPSGSP	High

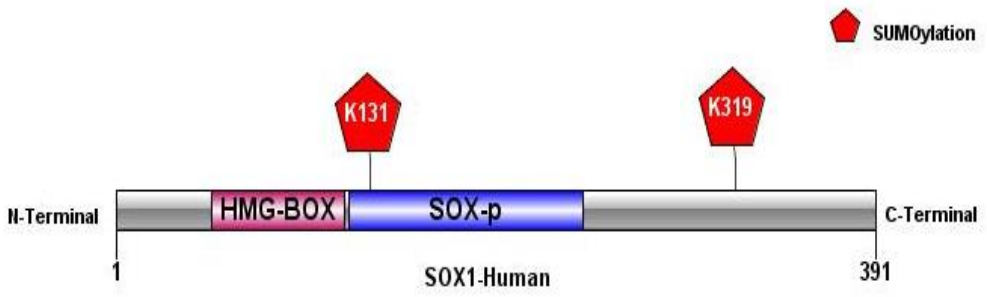


Figure 5-4 **Sumoylation predicted sites for SOX1**: Schematic representation of predicted sumoylated residues within SOX1 by GSP-SUMOv2.0 and JASSAv4 analysis tools, diagram was generated using the IBS illustrator webserver [180].

5.2.2 PTMs within highly conserved domains (HMG-BOX, SOXp) of SOX1

It has been found that the only experimentally verified phosphorylated site for human SOX1 protein is S93, found phosphorylated in a tumour tissue affected by Ischemia after 60min of dissection [232]. S93 is within the HMG-Box domain; which is the DNA binding portion of SOX proteins, the phosphorylation of S93 may alter the DNA binding abilities of SOX1 and therefore alter its function (Figure 5-1 and Figure 5-2).

Other types of PTM that have been found within DNA binding domain of SOX1 (HMG-BOX and SOXp) are acetylation of lysine residues at position K83, K131 and K132 (Figure 5-1). Residues K83 and K132 present within SOX1 HMG-Box and SOXp domains respectively, are experimentally verified acetylated residues curated by the PhosphoSitePlus® database (Figure 5-3B), [228]. While residue K131 has been experimentally verified as acetylated (K131-ac) in mouse SOX1 only, therefore it is highly likely that K131 might be post translationally modified on human SOX1 as well (Figure 5-1B). The above prediction arguments which are based upon sequence similarities between orthologous species can be supported by the observation that K132 is experimentally verified to be acetylated in both human and mouse SOX1 protein [233, 234]. Residue K131 has been also predicted as a potential site for sumoylation within SOXp domain of SOX1 (Table 11 and Table 12).

5.2.3 PTMs towards C-terminal region of SOX1 protein

The C-terminal region of SOX1 has been reported to play a role in transactivation of target genes [235]. Predicted PTM residues present in the C-terminal region of SOX1 are K319, S325 and S332 (Figure 5-1A, Figure 5-4). PTMs in this region may alter the transactivation ability of SOX1 therefore altering the transcriptional regulatory effect of SOX1.

Phosphorylated predicted residue S325 is located closer to the C terminal, outside of the HMG-box and SOXp domains (Figure 5-1A). Residue S325 has been predicted by two different databases (Figure 5-1 and Figure 5-2), S325 is the only potential true residue which may be showing Yin-Yang effect. This means that on residue S325 phosphorylation or O-GlcNAc modification can occur at different times (Figure 5-3A, line highlighted in red).

Residue S332 is predicted as glycosylated in human SOX1 protein as it has been found glycosylated in ortholog mouse SOX1 at position S332 (Figure 5-1B) [230].

The only sumoylation predicted site present towards C-terminal region is residue K319. It was found that sumoylation site K319 conform to the sumoylation consensus motif ($_{318}$V**K**SE$_{321}$) and has an adjacent S325 predicted to be a phosphorylation site ($_{318}$V**K**SEPSG**S**P$_{326}$). Studies have shown that phosphorylation may regulate sumoylation of a substrate [149]. Therefore, further analysis of this specific motif ($_{318}$V**K**SEPSG**S**P$_{326}$)

of SOX1 containing sumoylation and phosphorylation sites might provide useful information about regulation of SOX1 function.

5.2.4 Identification of putative conserved motif in a SOX1 protein

Evolutionary conserved regions usually point to key functional portions of a protein that need to be preserved between species. Multiple sequence alignment across different species was performed for SOX1 protein to determine if there was sequence conservation of the motif ($_{318}$VKSEPSGSP$_{326}$) identified through the sumoylation and phosphorylation profiling at C-terminal domain of SOX1. This analysis reveals that the target motif is highly conserved across different species (Figure 5-5A). SOX1, SOX2 and SOX3 are transcription factors having high sequence similarities, relatively similar expression pattern and belong to the same SOXB1 sub family. Therefore, the target motif within SOX1 $_{318}$VKSEPSGSP$_{326}$ was further analysed by multiple sequence alignment between these three proteins sequences. The target motif was found in a highly conserved region of the three proteins, and a consensus sequence xKSExSxxP at the target region was obtained, this consensus motif was termed SOXB1 consensus motif (Figure 5-5B).

5.2.4.1 SOXB1 consensus motif

In general, the sumoylation consensus motif usually starts with "ψKxE" in a protein sequence (ψ, can be any hydrophobic amino acid such as A, I, L, M, P, F, V or W while "x" can be any amino acid) [149]. The SOXB1 consensus motif obtained (xKSExSxxP) has both sumoylation site and adjacent serine residue for phosphorylation (Figure 5-6B). As discussed

before, sumoylation site usually start with any hydrophobic amino acid residue therefore “V” was intentionally kept “x” despite remaining conserved, this will help widen the detection of similar motifs in other un-related proteins, discussed in the next paragraph.

The SOXB1 consensus motif xKSExSxxP will be used to scan for matches against a collection of motifs in the ScanProsite database (human proteins only) in order to identify other proteins that share similar sequence patterns. The aim was to investigate whether the conserved target motif exists in other un-related proteins or if it is conserved only within the SOXB1 subgroup of transcription factors.

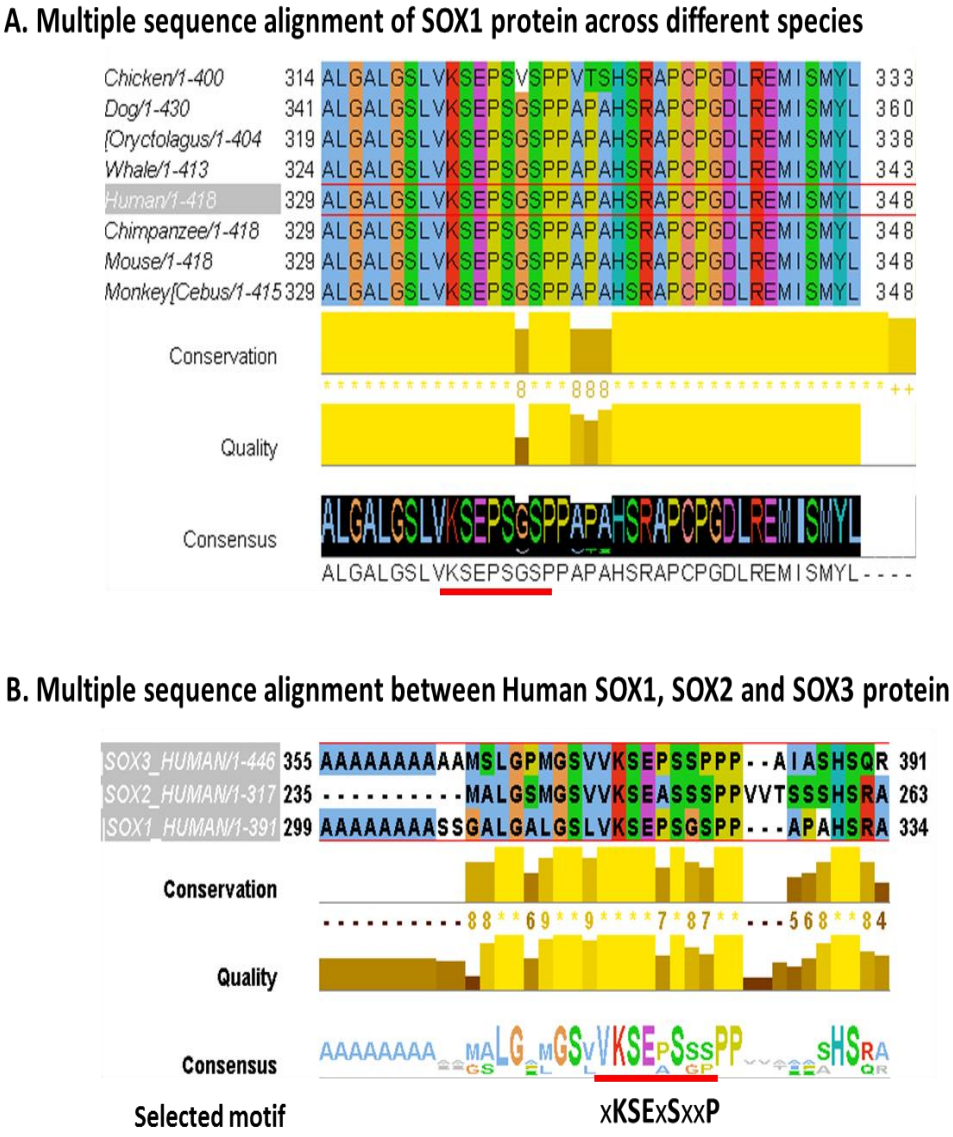


Figure 5-5 **Multiple Sequence alignment of SOX1 protein**: (A) Multiple Sequence alignment for SOX1 protein focusing on the target motif $_{318}$VKSEPSGSP$_{326}$, (underline red bar) (B) Multiple Sequence alignment between SOXB1 subgroup, focusing on the target motif region $_{318}$VKSEPSGSP$_{326}$ of SOX1 protein, consensus sequence obtained are underline red, selected motif also known as SOXB1 consensus motif has been also shown which was used to scan against protein database. Images were adopted from the EMBL-EBI resources, ClustalOmega alignment output page [184].

5.2.4.2 Identification of SOXB1 consensus motif in other un-related proteins

ScanProsite was used to generate a query for the SOXB1 consensus motif xKSExSxxP. In this way, the target motif was found in other 32 un-related human proteins (Figure 5-6). The proteins identified by ScanProsite were then retrieved from UniProt to compare their gene ontology. Interestingly, the majority of the proteins are transcription factors acting as transcription activators or repressors, sequence-specific DNA binding and chromatin binding [236]. A list of the transcription factors sharing the target SOXB1 consensus motif is presented in the Table 13.

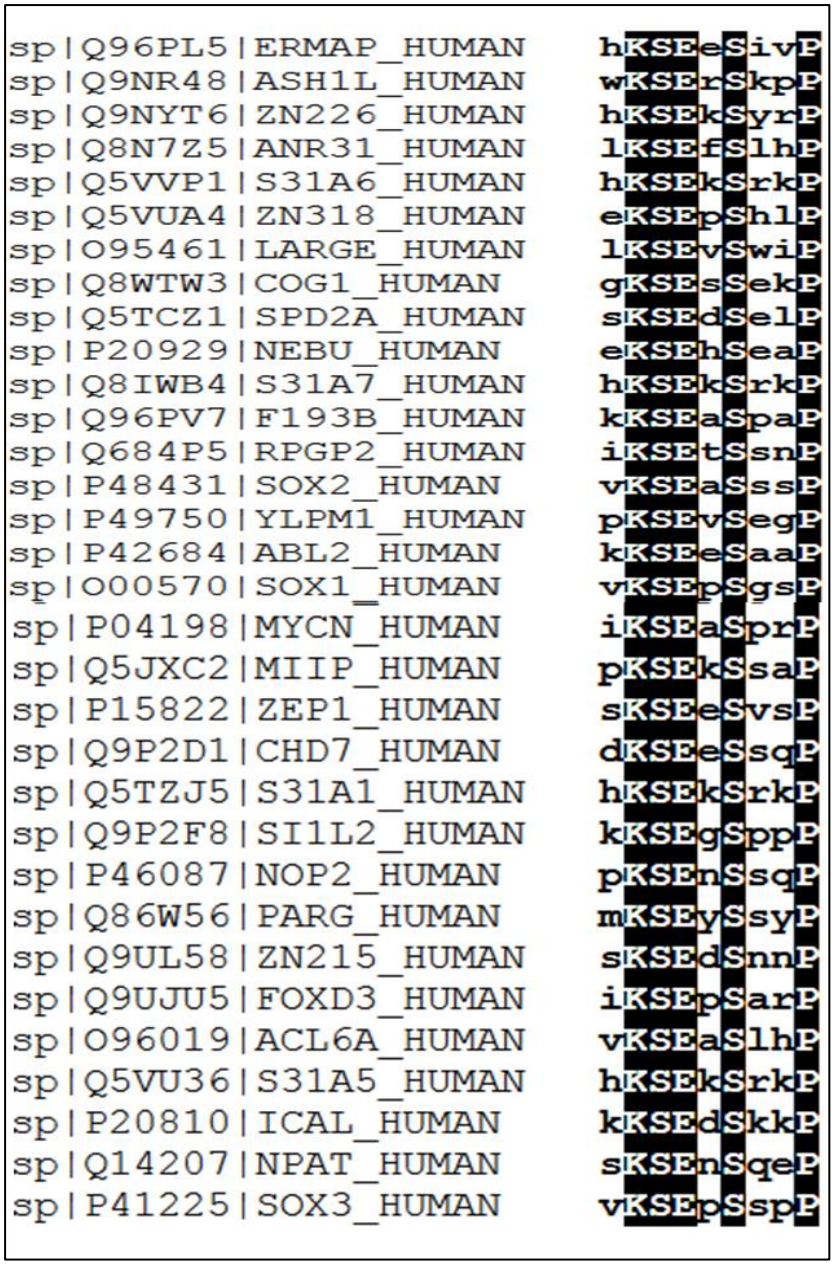


Figure 5-6 **ScanProsite query result**: showing a list of all identified 32 human proteins sharing the same consensus motif [181].

Table 13 shows the **ScanProsite database collection of all transcription factor proteins** sharing the SOXB1 consensus motif (xKSExSxxP). Each protein with culumn a, b, c [236] (obtained from Uniprot[237]) and d (provided by ScanProsite [181]) are shown.

(a)Uniprot No.	(b)Protein names	(c)Gene ontology (molecular function)	(d)Motif
P04198	N-myc proto-oncogene protein	Transcription factor activity, sequence-specific DNA binding	iKSEaSprP
Q5JXC2	Migration and invasion-inhibitory protein		pKSEkSsaP
Q9UJU5	Forkhead box protein D3	RNA polymerase II regulatory region sequence-specific DNA binding, transcriptional repressor activity, transcription factor activity	iKSEpSarP
Q5TCZ1	SH3 and PX domain-containing protein 2A	phosphatidylinositol binding ; superoxide-generating NADPH oxidase activator activity	sKSEdSelP
O00570	Transcription factor SOX-1	core promoter sequence-specific DNA binding; transcriptional activator activity, transcription factor activity.	vKSEpSgsP
P41225	Transcription factor SOX-3	RNA polymerase II transcription corepressor activity, transcription factor activity, sequence-specific DNA binding	vKSEpSspP
P48431	Transcription factor SOX-2	miRNA binding ; transcriptional activator/ factor/regulatory activities	vKSEaSssP
P15822	Zinc finger protein 40	sequence-specific DNA binding ; transcriptional repressor activity, transcription regulatory region DNA binding	sKSEeSvsP
Q9NYT6	Zinc finger protein 226	transcription factor activity, sequence-specific DNA binding	hKSEkSyrP
Q5VUA4	Zinc finger protein 318	nucleic acid binding zinc ion binding	eKSEpShlP
Q9UL58	Zinc finger protein 215	sequence-specific DNA binding ; transcription factor activity, sequence-specific DNA binding	sKSEdSnnP

Q9P2D1	Chromodomain-helicase-DNA-binding protein 7	chromatin binding; helicase activity, RNA polymerase II core promoter proximal region sequence-specific DNA binding	sKSEnSqeP
P20810	Calpastatin	Calcium-dependent cysteine-type endopeptidase inhibitor activity, endopeptidase inhibitor activity, poly(A) RNA binding	kKSEdSkkP
Q9NR48	Histone-lysine N-methyltransferase	chromatin binding ; DNA binding; histone methyltransferase activity	wKSErSkpP
O96019	Actin-like protein 6A	chromatin binding ; transcription coactivator activity	vKSEaSlhP
Q14207	Protein NPAT	protein N and C-terminus binding; transcription coactivator activity, transcription corepressor activity	sKSEnSqeP
P46087	Probable 28S rRNA	poly(A) RNA binding ; S-adenosylmethionine-dependent methyltransferase activity	pKSEnSsqP
Q9P2F8	Signal-induced proliferation-associated 1-like protein 2	GTPase activator activity	kKSEgSppP

5.2.4.3 Functional annotation clustering of different genes

DAVID software tools consists of an integrated biological knowledgebase and analytic tools that extract biological meaning from large gene or protein lists [189]. Table 13 containing list of proteins that shared the SOXB1 consensus motif were subjected to DAVID analysis. Result generated has shown that these proteins are mostly involved in biological processes like transcription regulation, DNA binding and transcription factor activity (Figure 5-7). DAVID analysis result snapshot has been shown in the Figure 5-7.

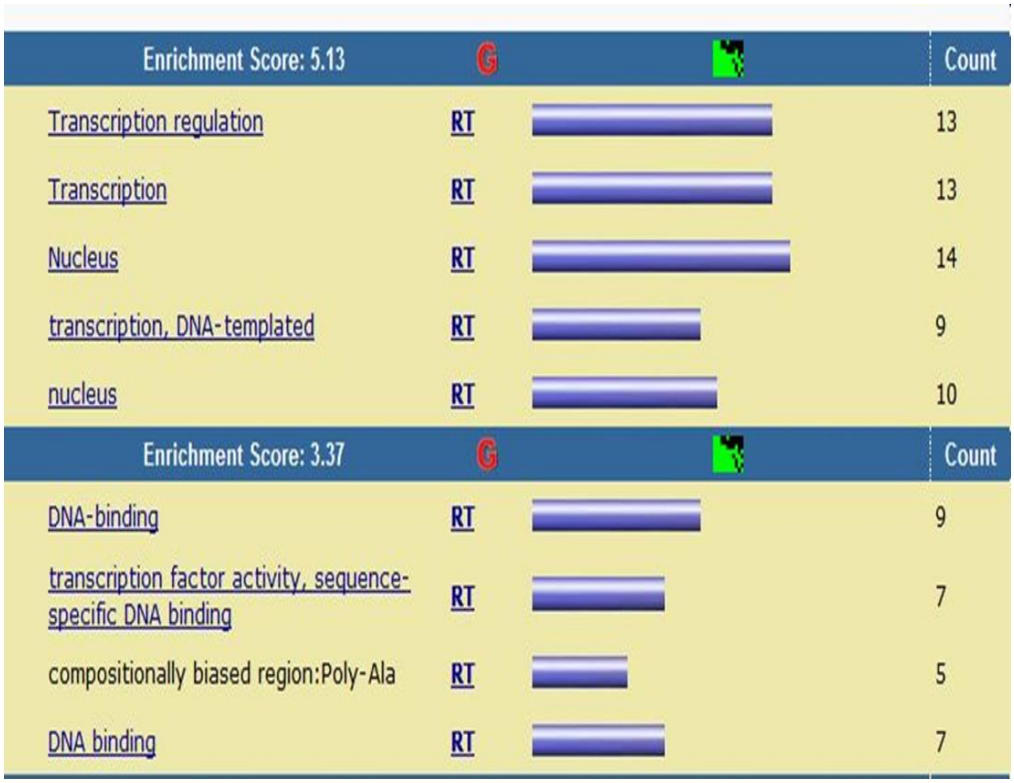


Figure 5-7: **DAVID functional annotation clustering** for the number of proteins that shared the same consensus motif. Given proteins have been clustered into two different groups, cluster 1 and 2. Column 'count' contains number of proteins involved in the respective term. The image was generated through DAVID functional annotation tools [190, 191].

5.2.4.4 Types of PTMs within the SOXB1 consensus motif

Comparison between the SOXB1 consensus motif across SOX1, SOX2 and SOX3 proteins has shown different types of PTM present at this region (Figure 5-8). SOX2 and SOX3 have phosphorylated serine residues in this conserved region. Sumoylation has been reported for human SOX2 at K245 which affects transcriptional activity of SOX2 [150], While SOX1 has only predicted sumoylation at K319 and phosphorylation at 325 (Figure 5-8).

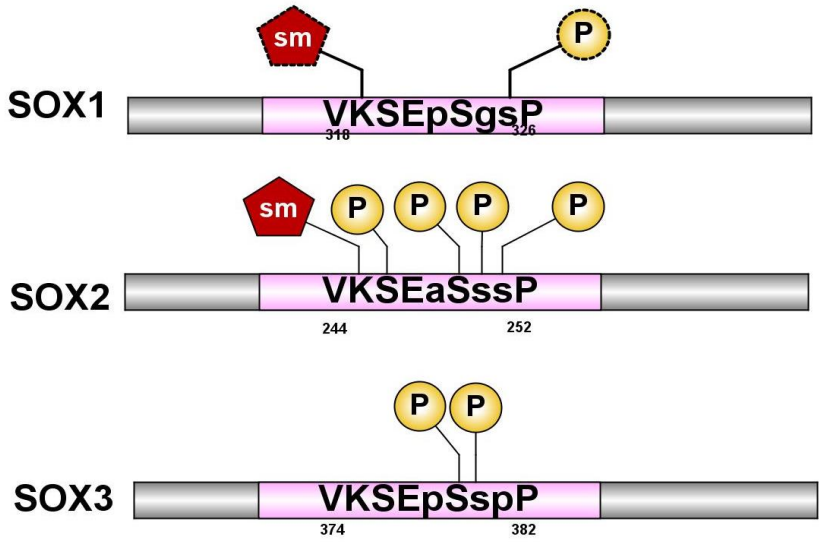


Figure 5-8 **Comparison of PTMs within SOXB1 conserved motif**: SOXB1 proteins multiple sequence alignment of the conserved motif showing Post translational modification, Conserved amino acids are shown in capital letters. P; Phosphorylation, sm; Sumoylation, Dotted lines shaped residue are predicted only. IBS illustrator was used for the diagram [180].

Altogether results collected from different databases have shown that the SOX1 protein contains a number of candidate sites for post translational modification as illustrated in Figure 5-9. It has been found that the C-terminal domain of SOX1 protein has a conserved motif (vKSEpSgsP) that contains sites for sumoylation and phosphorylation and is highly conserved within SOXB1 proteins (Figure 5-8). It was also found that SOXB1 consensus motif is shared among different un-related proteins, majority of them are transcription regulator proteins.

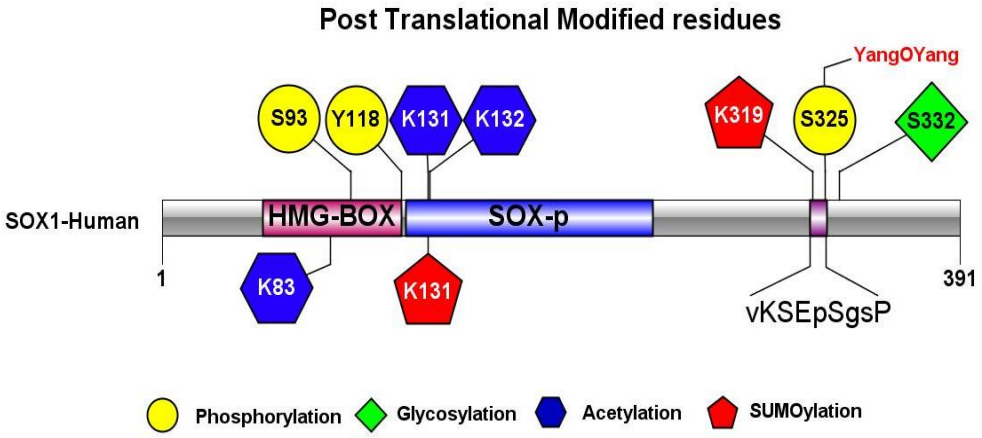


Figure 5-9 **Illustration of PTMs and functional domain of SOX1**: SOX1 protein structure with predicted post translational modifications site and SOXB1 consensus motif sequence at C-terminal region. IBS illustrator webserver was used to generate the diagram [180].

5.3 Discussion

5.3.1 Different types of PTMs might regulate SOX1 transcriptional activities

The HMG domain of SOX1 is an evolutionary conserved and functionally important region that binds to specific DNA sequences to bring conformational changes within chromatin structure [238]. It was found that the predicted post-translationally modified residues K83 (Acetylated), S93 (phosphorylated) and Y118 (phosphorylated) lie within the HMG domain of SOX1, making these modifications significant to SOX1 function. In mouse embryonic stem cell, SOX2 acetylated at residue K75 induces nuclear export of SOX2, while by blocking acetylation of this site retains SOX2 in nucleus and keeps regulating its target genes [239]. This acetylation site (K83 in SOX1) is highly conserved across different species in orthologues SOX1 proteins [239]. Therefore, it can be suggested that acetylation of SOX1 at K83 might also induce nuclear export machinery to retain SOX1 inside a nucleus in order to regulate its target genes.

Another domain of SOX1 is SOXp which is in close proximity to the HMG domain. This domain is found in the SOX family of proteins with two conserved sequence motif such as KKDK and LPG. The SOXp domain of SOXB1 proteins (SOX1, SOX2 and SOX3) has a binding site for the nestin neural enhancer and it has been documented that binding of SOX2 to the nestin enhancer upregulates the expression of nestin [240]. The acetylated residues K131 and K132 are located within the conserved sequence motif of the SOXp domain ($_{131}$**KK**DK), and acetylation at this conserved region might influence SOX1 binding to the nestin enhancer.

K131 has been also predicted as a sumoylation site (Figure 5-4), therefore, interplay between acetylation and sumoylation is likely to occur at this site. Studies have shown that cross talk between sumoylation and acetylation regulates transcription and DNA binding activity of the tumour suppressor p53 [241]. Therefore, it is likely that interplay between sumoylation and acetylation at K131 might be regulating transcription and DNA binding activity of SOX1.

Residue S325 that lies towards C-terminal domain of SOX1 was predicted as highly likely to be a true phosphorylation site that might show YingOYang effect (Interplay between O-GlcNAc and phosphorylation), Residue S325 has been experimentally verified phosphorylated in mouse SOX1 (Figure 5-1), It was found that phosphorylation site S325 is preceding proline residue (P326) and they lies in a close proximity to the sumoylation consensus motif ($_{318}$**VKSE**PSG**S**P$_{326}$) that has predicted sumoylated residue at K319. This finding is in line with the present literature that sumoylation which is dependent on proline directed phosphorylation is found adjacent to the sumoylation consensus motif [149]. These observations suggest that SOX1 motif $_{318}$**VKSE**PSG**S**P$_{326}$ might be a phosphorylation dependent sumoylation motif present at the C-terminal region of SOX1. The majority of transcription regulator proteins such as heat-shock factors (HSFs), GATA-1, and myocyte enhancer factor 2 contain a phosphorylation dependent sumoylation motif [149]. It has been known that phosphorylation-dependent sumoylation repressed transactivation capacities of HSF family proteins [149]. Therefore, SOX1

function as a transcription activator might be repressed through phosphorylation dependent sumoylation at this specific motif of SOX1 (sumo-phospho motif).

5.3.2 C-terminal of region SOX1 might act as putative functional domain

The SOX1 sumo-phospho motif ($_{318}$V**K**SEPSG**S**P$_{326}$) is present towards the C-terminal region containing different predicted PTMs. Yang et al. have reported occurrence of phospho-sumoyl switches at similar conserved region of SOX2 ($_{244}$V**K**SEASS**S**P$_{252}$) [242]. Tahmasebi et al. have shown that sumoylation at K245 within this region affects transcriptional activity of SOX2 [150]. Similarly, in mouse SOX2 sumoylation at K247 (equivalent to K245) inhibits SOX2 binding to Fgf4 enhancer and thus negatively regulates its transcriptional role through impairing DNA binding [151]. Furthermore, it was also found that phosphorylation of adjacent serine residues within the conserved motif do not affect transcriptional activity of SOX2 but rather are required for optimal sumoylation [150]. Similar to SOX2, SOX1 transcriptional activities might also be repressed by K319 sumoylation within the sumo-phospho motif, and phosphorylation at S325 might not affect the transcriptional role of SOX1 but is rather required for optimal sumoylation at K319.

5.3.3 Differential role of SOX1 in cancer

SOX1 has been found to supress tumour growth in many different cancer types [6, 10]. However, in prostate cancer progression SOX1 is expressed in more aggressive tumour and not in the less aggressive counterpart [12]. Autoantibodies to SOX1 are common in small-cell lung carcinoma (SCLC) and serve as a serological tumour marker for SCLC; SOX1 related auto immune response in SCLC is still elusive [111, 243]. This differential role of SOX1 in cancer is yet to be explored. In this study, among the proteins identified that shared similar pattern of motif along with SOX1 is the N-MYC proto-oncogene protein (N-MYC) and the Migration and Invasion-inhibitory protein (IIp45). N-MYC is a proto-oncogene expressed in a variety of human tumours and most frequently in neuroblastoma [244]. IIp456 protein is known to have tumour suppression effect by down-regulating adhesion-and-motility-associated genes in glioma cells invasion [245]. SOX1 protein function in cancer and the post translational modifications regulating it remain largely unexplored. The fact that SOX1 shares a short conserved motif with proteins like C-MYC and IIp456 (Table 13) having oncogenic and tumour suppressor properties respectively, might explain the differential role of SOX1 as a tumour suppressor or oncogene that depends on cancer types. Additional studies aimed to perform a functional analysis of SOX1 PTMs at the conserved motif ($_{318}$VKSEPSGSP$_{326}$), will be helpful in starting to unravel more about the role of SOX1 in cancer.

5.4 Conclusion

In conclusion, the data shown demonstrate that different types of PTMs might regulate SOX1 transcriptional activities at residue K83, K131, K319 and S325. It has been suggested that acetylation at K83 might retain SOX1 inside the nucleus to regulates transcription of its target genes. Interplay between sumoylation and acetylation at K131 might be regulating transcription and DNA binding activity of SOX1. It has been also suggested that the C-terminal region of SOX1 might act as a putative functional domain by repressing SOX1 transcriptional activities through phosphorylation dependent sumoylation at the motif $_{318}$V**K**SEPSG**S**P$_{326}$.

The SOXB1 consensus motif xKSExSxxP has been identified as a signature motif across different transcription regulator proteins. The SOX1 protein has been found to share a short conserved motif (xKSExSxxP) with proteins like C-MYC (proto-oncogene) and IIp456 (tumour suppressor effect), leading to the hypothesis that a functional analysis of SOX1 at the conserved motif ($_{318}$VKSEpSgsP$_{326}$), might provide useful information about the differential role of SOX1 in cancer.

6 Chapter 06

Final Discussion

SOX1 has been mainly reported to function as a tumour suppressor gene in different cancer types [6, 10, 246]. Contrary to this, *SOX1* has been reported as a oncogenes that promotes invasion of prostate cancer [12]. Moreover, in addition to its possible role as a tumour suppressor or oncogene, *Sox1* has been implicated as a marker of cancer stem cells (CSCs) in breast cancer [247]. To date, there has been no study addressing this differential role of *SOX1* in cancer and little is known about the role of *SOX1* gene regulation in cancer development.

This project aimed to highlight human *SOX1* gene regulation in the context of stem cells and cancer to identify several factors or mechanisms that could significantly regulate its function. In the first instant, Human *SOX1* gene expression and its promotor DNA methylation pattern was analysed in different stem cells and cancerous cell lines. It was found that *SOX1* is differentially expressed and its expression co-relates with its promoter DNA methylation in most of the cancer cell lines, except for the HOS cell line (Figure 6-1). *SOX1* lies within the transcriptional unit of an overlapping long non coding RNA gene, *SOX1-OT*, suggesting a potential regulatory relationship between the two. To address this, the structure of human *SOX1-OT* was characterised in ReN cells at different time points during a differentiation time course. *SOX1-OT* expression was further analysed in different stem cells and cancerous cell lines. It was found that *SOX1-OT* has a complex structure with many unannotated exons and

different transcript variants that are unannotated in the human genome. *SOX1-OT* was found highly expressed in differentiated neural stem cell and showed a switch between different transcript variants expression across different time points of neural differentiation. Co-expression of *SOX1-OT* and *SOX1* was also found in most of the stem cell and cancer cell lines (Figure 6-1).

Altogether, it was found that *SOX1* gene expression is co-related with its promoter DNA methylation, and co-expressed with *SOX1-OT* in different stem cell and cancer cell lines (Figure 6-1).

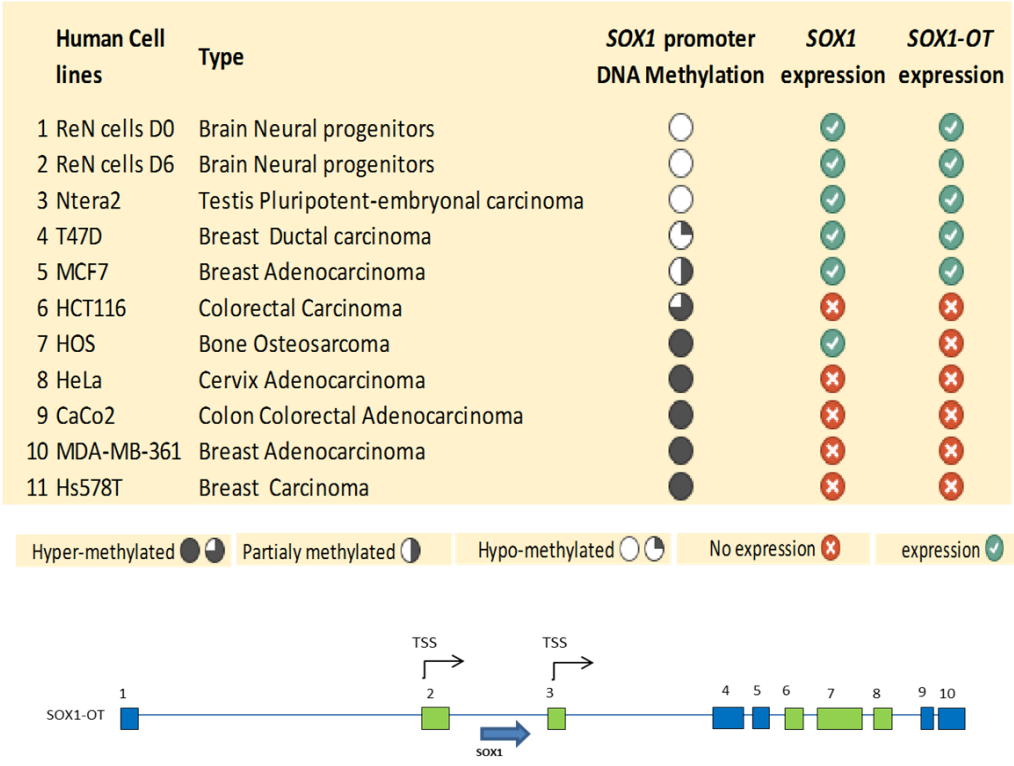


Figure 6-1: **Summary of SOX1 Methylation**, Gene expression and SOX1-OT results in all cell lines.

Additionally, different online bioinformatics databases were searched for SOX1 protein post translational modifications (PTMs). It was found that SOX1 protein contains many candidate sites for post translational

modification that might regulate SOX1 transcriptional activities (see section, 5.3). Data collected in this study strongly suggest that the human SOX1 protein is regulated at the post translational level, which could impact its function.

Detection of SOX1 protein expression in the studied cell lines was found challenging (see section, 3.2.7 and 3.2.8.2). Western blot and immunostaining experiments have shown that the anti-SOX1 antibodies were not specific for SOX1. It was also suggested that lack of SOX1 positive result might be down to its low expression. Therefore, further research will be required to test SOX1 protein expression on primary tissue/cells from the patients, as this might overcome the issue with low expression of SOX1. Although different commercially available antibodies were tested but problem with specificity was persistent, therefore future analysis will required to design SOX1 antibody at specific amino acid motif which does not share among other highly conserved proteins like SOX2 or SOX3.

6.1 Differential role of *SOX1* in cancer

Studies so far have shown that *SOX1* gene expression in cancer is suppressed by promoter methylation [7-11]. Epigenetic silencing of *SOX1* through promoter DNA methylation is likely dependent on the cancer type (section, 3.3.3). *SOX1* was found epigenetically silenced through promoter hypermethylation in the majority of cancer cell lines studied. Therefore, it is likely that loss of *SOX1* expression in these cancer types might be linked to tumourigenesis, which is in line with the published data about *SOX1* functioning as a tumour supressor gene [6, 10, 196].

In contrast to epigenetic silencing of *SOX1* in different cancer types, *SOX1* gene expression was also detected in a different set of cancer lines with a low level of promoter DNA methylation. Wright et al. have shown that *Brca1*-deficient mouse mammary tumours harbour heterogeneous cancer stem cell populations ($CD44^{+}/CD24^{-}$ and $CD133^{+}$), in which *Sox1* expression was significantly higher compared to the stem cell depleted population, indicating its possible role in cancer stem cell (CSC) regulation [247]. Recent research on CSC has identified *SOX2* involved in self-renewal of CSC in numerous cancer types [200], where it mediates self-renewal of CSC through EGFR signalling [200]. *SOX1* and *SOX2* belong to the same SOXB1 family and are functional redundant in diverse developmental events [23]. Therefore, *SOX1* may exert a similar function in the studied cancer lines. The *SOX1* expression observed in the cancer lines in this study may indicate a possible role of *SOX1* in maintaining pluripotency and/or self-renewal of CSC.

Altogether, it may be hypothesized that *SOX1* functions as tumour suppressor or oncogene depending on the regulatory network present in the particular cancer. Features regulating *SOX1*, such as promoter DNA methylation and PTM, are under the influence of wide variety of factors. These may, in turn, alter *SOX1* function. In addition, the availability of *SOX1* interacting partners, for example lncRNAs, provides an additional layer to further influence *SOX1* function.

6.2 Regulation of *SOX1* gene by long non-coding RNA (*SOX1-OT*)

Studies have shown that lncRNAs regulate transcriptional regulators of different developmental processes and play many different roles, at different points within the cell in processes like the regulation of pluripotency, stem cell differentiation and tumorigenesis [211, 215, 248]. For example, the lncRNA *ADINR* plays an important role in regulating the differentiation of human mesenchymal stem cells into adipocytes by modulating *C/EBPα*, which is a critical transcriptional regulator of adipogenesis [248].

This study has shown that *SOX1-OT* is highly expressed in differentiated neural stem cells. The multiple transcript variants of *SOX1-OT* appeared to be differentially expressed during neural differentiation, indicating a regulatory role in neural development. Furthermore, co-expression of *SOX1-OT* and *SOX1* in stem cells and different cancer cell lines (section, 4.3.3) suggest *SOX1-OT* may play a possible role regulating *SOX1* expression. These findings mirror what has already been described for the *SOX2* gene, another member of the SOXB1 family. *SOX2-OT* is an overlapping transcript of *SOX2* and mouse *Sox2-ot* is known to regulate transcription of the *Sox2* gene in different developmental processes such as embryonic and neural development [211]. Functional association between *SOX2-OT* and *SOX2* expression in pluripotency and tumorigenesis suggest a possible role of *SOX2-OT* regulating *SOX2* [215]. Therefore, evidences provided in the current study suggest that the lncRNA *SOX1-OT*

might play an important role in neural development and tumorigenesis by regulating *SOX1* expression.

The TSS of *SOX1* and *SOX1-OT* lie at close genomic location to each other (Figure 4-35) and there might be a possibility that these two transcripts are simply co-expressed by using the same promoter, rather than having a more complex regulatory interaction. The possibility of *SOX1-OT* and *SOX1* using same promoter is minimised by the RT-PCR results (Figure 4-36). It was found that HOS cells express *SOX1* but not *SOX1-OT* while SH-SY5Y cell line express *SOX1-OT* but not *SOX1*. Therefore, on the basis of this observation it can be suggested that transcription of these two transcripts is independent of each other. This can be further confirmed through future experiments by knockdown and overexpression of *SOX1-OT* in those cell lines that express both transcripts. This will determine its functional consequences on *SOX1* gene expression. Further experiments like Chromatin Immunoprecipitation (ChIP) analysis may also identify SOX1 binding sites within the *SOX1-OT* promoter highlighting the intriguing possibility of a regulatory feedback loop existing between them. These analyses will identify existence of a functional regulatory relationship between *SOX1* and *SOX1-OT.*

6.3 Regulation of SOX1 at the Post translational level

SOX1, despite being a key regulator during neurodevelopment and tumourigenesis, little is known about the regulation of SOX1 protein function at the PTM level. Data collected in this study strongly suggest that human SOX1 is regulated at the post translational level, which could impact its function.

Depending on the post translational modifications, a transcription factor might change its DNA binding activities and interactions with partner proteins within transcription regulatory complexes, consequently altering its function. For example, SOX2 sumoylation at K247 impairs its binding to the Fgf4 enhancer, which results in negative regulation of SOX2 transcriptional activities [151]. In another example, interaction of SOX2 with the nuclear export machinery by acetylation at K75 within the HMG domain retains SOX2 inside the nucleus to regulate its target genes [239]. Liu et al. have shown that SOX2 possesses variable functions depending on its PTMs, and as a result regulates pluripotency and differentiation of the stem cell [249]. Equivalent to the SOX2 findings at K247 and K75, SOX1 has been predicted to harbour a sumoylation site at K319 (section, 5.4.2) and acetylation at K83 (section, 5.4.1). The similarities between the predicted SOX1 and the known SOX2 PTMs highlight the importance of the *in silico* identified SOX1 PTMs as they may affect SOX1 function in a comparable way to the ones already described for SOX2.

6.4 Perspectives

In this study, potential regulatory mechanisms for *SOX1* gene regulation were analysed in the context of stem cells and tumourigenesis. These regulatory mechanisms studied were DNA methylation in the *SOX1* promoter region (section, 3.3.3), a *SOX1* long non-coding overlapping transcript (section, 4.3) and PTMs within the SOX1 protein (section, 5.3). Understanding these regulatory features will help to better model the

SOX1 transcription regulatory network in stem cell developmental processes and its role in cancer development.

Since these analyses were performed in well-established cell lines, further work is required to test these initial findings in primary tissue/cells from patients to determine physiological relevance of the cell line data.

Epigenetic silencing of *SOX1* through promoter DNA methylation is very common in cancer and has been proposed as a prognostic biomarker for the detection of different cancer types [7-11]. The DNA methylation pattern of the *SOX1* promoter and the lack of detectable gene expression in a panel of different cancerous cell lines analysed here may facilitate future research for the identification of *SOX1* as a detection marker in these cancer types. In addition, a possible *SOX1* role in CSC self-renewal or pluripotency requires further attention which might in the future identify *SOX1* as a regulator of CSC and possible therapeutic agent in cancer.

In addition to DNA methylation at the promoter region, other regulatory mechanisms like histone modifications and transcription factors binding can significantly influence transcriptional output. Further research is required to fully elucidate these transcriptional regulatory features influencing *SOX1* transcription. For example, ChIP-seq data available through ENCODE project can be used to gain information on *SOX1* promoter across different cell types and developmental stages, in relationship to *SOX1* expression. This will help to evaluate their functions in transcriptional regulation of *SOX1*.

The identified lncRNA, *SOX1-OT,* might have diverse biological significance and functions, which need to be explored. As suggested in this thesis, *SOX1-OT* might have a potential role in cancer by regulating *SOX1* expression. Analysis of *SOX1-OT* influencing *SOX1* enhancer or promoter activity during *SOX1* transcription could possibly identify the mechanism(s) through which *SOX1-OT* regulate *SOX1* expression. Cancer types that exhibit co-expression of *SOX1-OT* and *SOX1* need further attention. Gain or loss-of-function strategies would aid the determination of *SOX1-OT*'s mode of action and its relationship with *SOX1*. For example, *SOX1-OT* knock down in cancerous cell lines such as NTera, T47D and MCF7 might help determining its effects on *SOX1* expression in these cell lines. Cell migration assays need to be performed on these cell lines to see *SOX1-OT* knock down affects cell invasion ability. Knock down of *SOX1* in bone osteosarcoma cells (HOS) that are positive for *SOX1* expression but do not appear to express *SOX1-OT* (Figure 6-1), will be interesting to see its effect on *SOX1-OT* expression. Additionally, further work is required to fully explore epigenetic mechanisms such as DNA methylation, transcription factors binding and histone modifications at the promoter region of *SOX1-OT*, in conjunction with its gene expression. Identification of these features will help to understand their influence on the transcriptional regulation of *SOX1-OT*.

In vitro analysis of *SOX1-OT* expression alongside *SOX1* expression in neural stem cell differentiation (ReN) suggests its potential role in neural development. *SOX1* expression in conjunction with *SOX1-OT* expression

needs to be studied in vivo in tissues such as the brain in which *SOX1* is known to be expressed. One way to do that is transcriptome analysis of, for example, publicly available datasets. The expression analysis across different tissues *in vivo* will highlight whether *SOX1-OT* expression is independent of *SOX1* or not and additionally if *SOX1-OT* may potentially be involved in other developmental processes apart from neural development.

Better understanding of the *SOX1* and *SOX1-OT* transcriptional regulatory network may advance the understanding of its co-regulation in events like neural differentiation and tumorigenesis which may ultimately advance the development of cell therapies for neurodegenerative diseases and cancer.

The prediction of different types of PTMs within the SOX1 protein has identified several areas for further research that could help understand the function of SOX1 as a transcription factor. Data collected for SOX1 PTMs needs to be experimentally verified, in order to identify mechanisms through which SOX1 changes its function. For example, SOX1 plays a regulatory role during neuronal cell fate determination and differentiation. Kan et al. have shown that SOX1 binds to the Hes1 promoter attenuating Notch signalling that supresses neurogenesis [25], and the SOX1 HMG domain and C-terminus are both required for interacting with the Hes1 promoter [25]. Therefore, it could be speculated that PTMs predicted within these domains could have possible consequences on SOX1 binding to the Hes1 promoter. This study predicts

acetylation within the HMG domain at K83, which might retain SOX1 inside the nucleus. SOX1 is known to bind to β-catenin which attenuates Wnt signalling, further will be required to demonstrate whether blocking acetylation at K83 to retain SOX1 inside the nucleus influences the regulatory role of SOX1 in neural development and in cancer particularly.

Overall, this project presents new results on *SOX1* gene regulation at transcriptional and post transcriptional levels. Such information is required in order to elucidate *SOX1* gene regulation in neural development and cancer, which may have applications for therapeutic approaches for neurodegenerative diseases and cancer. Additionally, this study has now identified several putative post translational modifications which could regulate SOX1 if confirmed. The *in silico* identification of SOX1 PTMs and its putative functional domains will facilitate future work in understanding SOX1 function as a transcriptional regulator of neural stem cell fate and its differential role in cancer.

7 Chapter 07

References

1. Pevny, L.H., et al., *A role for SOX1 in neural determination.* Development, 1998. **125**(10): p. 1967-1978.
2. Kiefer, J.C., *Back to basics: Sox genes.* Developmental Dynamics, 2007. **236**(8): p. 2356-2366.
3. Kamachi, Y., M. Uchikawa, and H. Kondoh, *Pairing SOX off with partners in the regulation of embryonic development.* Trends in Genetics, 2000. **16**(4): p. 182-187.
4. Kan, L., et al., *Dual function of Sox1 in telencephalic progenitor cells.* Developmental Biology, 2007. **310**(1): p. 85-98.
5. Venere, M., et al., *Sox1 marks an activated neural stem/progenitor cell in the hippocampus.* Development, 2012. **139**(21): p. 3938-49.
6. Tsao, C.M., et al., *SOX1 Functions as a Tumor Suppressor by Antagonizing the WNT/beta-Catenin Signaling Pathway in Hepatocellular Carcinoma.* Hepatology, 2012. **56**(6): p. 2277-2287.
7. Apostolidou, S., et al., *DNA methylation analysis in liquid-based cytology for cervical cancer screening.* International Journal of Cancer, 2009. **125**(12): p. 2995-3002.
8. Su, H.Y., et al., *An epigenetic marker panel for screening and prognostic prediction of ovarian cancer.* International Journal of Cancer, 2009. **124**(2): p. 387-93.
9. Zhao, Y.X., et al., *Abnormal methylation of seven genes and their associations with clinical characteristics in early stage non-small cell lung cancer.* Oncology Letters, 2013. **5**(4): p. 1211-1218.
10. Lin, Y.W., et al., *SOX1 suppresses cell growth and invasion in cervical cancer.* Gynecol Oncol, 2013.
11. Guan, Z., et al., *SOX1 down-regulates beta-catenin and reverses malignant phenotype in nasopharyngeal carcinoma.* Mol Cancer, 2014. **13**: p. 257.
12. Mathews, L.A., et al., *Epigenetic regulation of CpG promoter methylation in invasive prostate cancer cells.* Molecular Cancer, 2010. **9**.
13. Pevny, L.H. and R. LovellBadge, *Sox genes find their feet.* Current Opinion in Genetics & Development, 1997. **7**(3): p. 338-344.
14. Sinclair, A.H., et al., *A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif.* Nature, 1990. **346**(6281): p. 240-4.
15. Wegner, M., *From head to toes: the multiple facets of Sox proteins.* Nucleic Acids Research, 1999. **27**(6): p. 1409-1420.
16. Denny, P., et al., *A Conserved Family of Genes Related to the Testis Determining Gene, Sry.* Nucleic Acids Research, 1992. **20**(11): p. 2887-2887.
17. Sarkar, A. and K. Hochedlinger, *The sox family of transcription factors: versatile regulators of stem and progenitor cell fate.* Cell Stem Cell, 2013. **12**(1): p. 15-30.
18. Schepers, G.E., R.D. Teasdale, and P. Koopman, *Twenty pairs of sox: extent, homology, and nomenclature of the mouse and human sox transcription factor gene families.* Dev Cell, 2002. **3**(2): p. 167-70.

19. Uchikawa, M., Y. Kamachi, and H. Kondoh, *Two distinct subgroups of Group B Sox genes for transcriptional activators and repressors: their expression during embryonic organogenesis of the chicken.* Mech Dev, 1999. **84**(1-2): p. 103-20.
20. Love, J.J., et al., *Structural Basis for DNA Bending by the Architectural Transcription Factor Lef-1.* Nature, 1995. **376**(6543): p. 791-795.
21. Bowles, J., G. Schepers, and P. Koopman, *Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators.* Developmental Biology, 2000. **227**(2): p. 239-255.
22. Bylund, M., et al., *Vertebrate neurogenesis is counteracted by Sox1-3 activity.* Nature Neuroscience, 2003. **6**(11): p. 1162-1168.
23. Miyagi, S., H. Kato, and A. Okuda, *Role of SoxB1 transcription factors in development.* Cell Mol Life Sci, 2009. **66**(23): p. 3675-84.
24. Graham, V., et al., *SOX2 functions to maintain neural progenitor identity.* Neuron, 2003. **39**(5): p. 749-65.
25. Kan, L.X., et al., *Sox1 acts through multiple independent pathways to promote neurogenesis.* Developmental Biology, 2004. **269**(2): p. 580-594.
26. Alcock, J., et al., *Expression of Sox1, Sox2 and Sox9 is maintained in adult human cerebellar cortex.* Neuroscience Letters, 2009. **450**(2): p. 114-6.
27. Nishiguchi, S., et al., *Sox1 directly regulates the gamma-crystallin genes and is essential for lens development in mice.* Genes & Development, 1998. **12**(6): p. 776-781.
28. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
29. Kan, L., et al., *Sox1 acts through multiple independent pathways to promote neurogenesis.* Developmental Biology, 2004. **269**(2): p. 580-94.
30. Ekonomou, A., et al., *Neuronal migration and ventral subtype identity in the telencephalon depend on SOX1.* Plos Biology, 2005. **3**(6): p. 1111-1122.
31. Elkouris, M., et al., *Sox1 Maintains the Undifferentiated State of Cortical Neural Progenitor Cells via the Suppression of Prox1-Mediated Cell Cycle Exit and Neurogenesis.* Stem Cells, 2011. **29**(1): p. 89-98.
32. Simons, B.D. and H. Clevers, *Strategies for homeostatic stem cell self-renewal in adult tissues.* Cell, 2011. **145**(6): p. 851-62.
33. Simona Casarosa, J.Z.a.L.C., *Systems for ex-vivo Isolation and Culturing of Neural Stem Cells, Neural Stem Cells* New Perspectives, Dr. Luca Bonfanti (Ed.) InTech, 2013.
34. Wen, S., H. Li, and J. Liu, *Dynamic signaling for neural stem cell fate determination.* Cell Adh Migr, 2009. **3**(1): p. 107-17.
35. Chapouton, P., R. Jagasia, and L. Bally-Cuif, *Adult neurogenesis in non-mammalian vertebrates.* Bioessays, 2007. **29**(8): p. 745-757.
36. Anastas, J.N. and R.T. Moon, *WNT signalling pathways as therapeutic targets in cancer.* Nature Reviews Cancer, 2013. **13**(1): p. 11-26.
37. Bray, S.J., *Notch signalling: a simple pathway becomes complex.* Nature Reviews Molecular Cell Biology, 2006. **7**(9): p. 678-689.
38. Bolos, V., J. Grego-Bessa, and J.L. de la Pompa, *Notch signaling in development and cancer.* Endocr Rev, 2007. **28**(3): p. 339-63.
39. Androutsellis-Theotokis, A., et al., *Notch signalling regulates stem cell numbers in vitro and in vivo.* Nature, 2006. **442**(7104): p. 823-6.
40. Moretti, J. and C. Brou, *Ubiquitinations in the notch signaling pathway.* Int J Mol Sci, 2013. **14**(3): p. 6359-81.

41. Allenspach, E.J., et al., *Notch signaling in cancer.* Cancer Biol Ther, 2002. **1**(5): p. 466-76.
42. Willert, K., et al., *Wnt proteins are lipid-modified and can act as stem cell growth factors.* Nature, 2003. **423**(6938): p. 448-52.
43. Clevers, H., *Wnt/beta-catenin signaling in development and disease.* Cell, 2006. **127**(3): p. 469-80.
44. Cheng, X.X., et al., *Correlation of Wnt-2 expression and beta-catenin intracellular accumulation in Chinese gastric cancers: relevance with tumour dissemination.* Cancer Lett, 2005. **223**(2): p. 339-47.
45. Kormish, J.D., D. Sinner, and A.M. Zorn, *Interactions between SOX factors and Wnt/beta-catenin signaling in development and disease.* Dev Dyn, 2010. **239**(1): p. 56-68.
46. Arce, L., N.N. Yokoyama, and M.L. Waterman, *Diversity of LEF/TCF action in development and disease.* Oncogene, 2006. **25**(57): p. 7492-504.
47. Levy, D.E. and J.E. Darnell, Jr., *Stats: transcriptional control and biological impact.* Nat Rev Mol Cell Biol, 2002. **3**(9): p. 651-62.
48. Shankaran, V., et al., *IFNgamma and lymphocytes prevent primary tumour development and shape tumour immunogenicity.* Nature, 2001. **410**(6832): p. 1107-11.
49. Bowman, T. and R. Jove, *STAT Proteins and Cancer.* Cancer Control, 1999. **6**(6): p. 615-619.
50. Mathews, L.A., et al., *Epigenetic regulation of CpG promoter methylation in invasive prostate cancer cells.* Mol Cancer, 2010. **9**.
51. Gong, L., et al., *Signal transducer and activator of transcription-3 is required in hypothalamic agouti-related protein/neuropeptide Y neurons for normal energy homeostasis.* Endocrinology, 2008. **149**(7): p. 3346-54.
52. Waddington, C.H., *The epigenotype.* Endeavour, 1942. **1**: p. 18-20.
53. Sharma, S., T.K. Kelly, and P.A. Jones, *Epigenetics in cancer.* Carcinogenesis, 2010. **31**(1): p. 27-36.
54. Lanctot, C., et al., *Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions.* Nat Rev Genet, 2007. **8**(2): p. 104-15.
55. Pollard T. D, E.W.C., *Cell Biology* 2002, Philadelphia: Saunders.
56. Cedar, H. and Y. Bergman, *Linking DNA methylation and histone modification: patterns and paradigms.* Nat Rev Genet, 2009. **10**(5): p. 295-304.
57. Delaval, K. and R. Feil, *Epigenetic regulation of mammalian genomic imprinting.* Curr Opin Genet Dev, 2004. **14**(2): p. 188-95.
58. Heard, E., et al., *Mammalian X-chromosome inactivation: an epigenetics paradigm.* Cold Spring Harb Symp Quant Biol, 2004. **69**: p. 89-102.
59. Shi, L. and J. Wu, *Epigenetic regulation in mammalian preimplantation embryo development.* Reprod Biol Endocrinol, 2009. **7**: p. 59.
60. Hirabayashi, Y. and Y. Gotoh, *Epigenetic control of neural precursor cell fate during development.* Nat Rev Neurosci, 2010. **11**(6): p. 377-88.
61. Jaenisch, R. and A. Bird, *Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.* Nature Genetics, 2003. **33**: p. 245-254.
62. Graff, J. and I.M. Mansuy, *Epigenetic dysregulation in cognitive disorders.* European Journal of Neuroscience, 2009. **30**(1): p. 1-8.
63. Goldberg, A.D., C.D. Allis, and E. Bernstein, *Epigenetics: A landscape takes shape.* Cell, 2007. **128**(4): p. 635-638.

64. Laird, P.W., *The power and the promise of DNA methylation markers.* Nature Reviews Cancer, 2003. **3**(4): p. 253-266.
65. Lai, H.C., et al., *Quantitative DNA methylation analysis detects cervical intraepithelial neoplasms type 3 and worse.* Cancer, 2010. **116**(18): p. 4266-74.
66. Takai, D. and P.A. Jones, *Comprehensive analysis of CpG islands in human chromosomes 21 and 22.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(6): p. 3740-3745.
67. Larsen, F., et al., *Cpg Islands as Gene Markers in the Human Genome.* Genomics, 1992. **13**(4): p. 1095-1107.
68. Kaneda, M., et al., *Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting.* Nature, 2004. **429**(6994): p. 900-3.
69. Jones, P.A. and D. Takai, *The role of DNA methylation in mammalian epigenetics.* Science, 2001. **293**(5532): p. 1068-70.
70. Geiman, T.M. and K. Muegge, *DNA methylation in early development.* Mol Reprod Dev, 2010. **77**(2): p. 105-13.
71. Chaligne, R. and E. Heard, *X-chromosome inactivation in development and cancer.* FEBS Lett, 2014. **588**(15): p. 2514-22.
72. Bannister, A.J. and T. Kouzarides, *Regulation of chromatin by histone modifications.* Cell Res, 2011. **21**(3): p. 381-95.
73. Phillips, T.S., K, *Chromatin remodeling in eukaryotes.* Nature Education, 2008.
74. Rodriguez-Paredes, M. and M. Esteller, *Cancer epigenetics reaches mainstream oncology.* Nature Medicine, 2011. **17**(3): p. 330-339.
75. Momparler, R.L., *Cancer epigenetics.* Oncogene, 2003. **22**(42): p. 6479-6483.
76. Ross, J.P., K.N. Rand, and P.L. Molloy, *Hypomethylation of repeated DNA sequences in cancer.* Epigenomics, 2010. **2**(2): p. 245-69.
77. Jones, P.A. and S.B. Baylin, *The fundamental role of epigenetic events in cancer.* Nat Rev Genet, 2002. **3**(6): p. 415-28.
78. Jones, P.A. and S.B. Baylin, *The fundamental role of epigenetic events in cancer.* Nature Reviews Genetics, 2002. **3**(6): p. 415-428.
79. Feinberg, A.P. and B. Vogelstein, *Hypomethylation of Ras Oncogenes in Primary Human Cancers.* Biochemical and Biophysical Research Communications, 1983. **111**(1): p. 47-54.
80. Sakai, T., et al., *Allele-Specific Hypermethylation of the Retinoblastoma Tumor-Suppressor Gene.* American Journal of Human Genetics, 1991. **48**(5): p. 880-888.
81. Muller-Tidow, C., et al., *DNA methylation of tumor suppressor genes in clinical remission predicts the relapse risk in acute myeloid leukemia.* Cancer Research, 2007. **67**(3): p. 1370-1377.
82. Herman, J.G. and S.B. Baylin, *Mechanisms of disease: Gene silencing in cancer in association with promoter hypermethylation.* New England Journal of Medicine, 2003. **349**(21): p. 2042-2054.
83. Das, P.M. and R. Singal, *DNA methylation and cancer.* Journal of Clinical Oncology, 2004. **22**(22): p. 4632-4642.
84. Feinberg, A.P. and B. Tycko, *Timeline - The history of cancer epigenetics.* Nature Reviews Cancer, 2004. **4**(2): p. 143-153.
85. Okano, M., et al., *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development.* Cell, 1999. **99**(3): p. 247-257.

86. Tost, J., *DNA Methylation: An Introduction to the Biology and the Disease-Associated Changes of a Promising Biomarker.* Molecular Biotechnology, 2010. **44**(1): p. 71-81.
87. Okano, M., et al., *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development.* Cell, 1999. **99**(3): p. 247-57.
88. Tahiliani, M., et al., *Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1.* Science, 2009. **324**(5929): p. 930-5.
89. Rasmussen, K.D. and K. Helin, *Role of TET enzymes in DNA methylation, development, and cancer.* Genes Dev, 2016. **30**(7): p. 733-50.
90. Seisenberger, S., et al., *Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1609): p. 20110330.
91. Wild, L. and J.M. Flanagan, *Genome-wide hypomethylation in cancer may be a passive consequence of transformation.* Biochim Biophys Acta, 2010. **1806**(1): p. 50-7.
92. Herman, J.G., et al., *Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma.* Proc Natl Acad Sci U S A, 1994. **91**(21): p. 9700-4.
93. Esteller, M., *Epigenetic gene silencing in cancer: the DNA hypermethylome.* Hum Mol Genet, 2007. **16 Spec No 1**: p. R50-9.
94. Shih, Y.L., et al., *Frequent concomitant epigenetic silencing of SOX1 and secreted frizzled-related proteins (SFRPs) in human hepatocellular carcinoma.* J Gastroenterol Hepatol, 2013. **28**(3): p. 551-9.
95. Roobol, M.J., A. Haese, and A. Bjartell, *Tumour markers in prostate cancer III: biomarkers in urine.* Acta Oncol, 2011. **50 Suppl 1**: p. 85-9.
96. Wolfgang, G., K. Michael, and A.S. Wolfgang, *Cancer Epigenetics : Methods and Protocols*, in *Cancer Epigenetics : Methods and Protocols*, M.V. Ramona G. Dumitrescu, Editor. 2012. p. 47-66.
97. Mathews, L.A., et al., *Epigenetic regulation of CpG promoter methylation in invasive prostate cancer cells.* Mol Cancer, 2010. **9**: p. 267.
98. Jemal, A., et al., *Global cancer statistics.* CA Cancer J Clin, 2011. **61**(2): p. 69-90.
99. Baylin, S.B. and J.G. Herman, *DNA hypermethylation in tumorigenesis: epigenetics joins genetics.* Trends in Genetics, 2000. **16**(4): p. 168-74.
100. Lai, H.C., et al., *Identification of novel DNA methylation markers in cervical cancer.* International Journal of Cancer, 2008. **123**(1): p. 161-7.
101. Shih Ie, M., et al., *Distinct DNA methylation profiles in ovarian serous neoplasms and their implications in ovarian carcinogenesis.* Am J Obstet Gynecol, 2010. **203**(6): p. 584 e1-22.
102. Nitta, K.R., et al., *Expression of Sox1 during Xenopus early embryogenesis.* Biochemical and Biophysical Research Communications, 2006. **351**(1): p. 287-293.
103. Rychel, A.L. and B.J. Swalla, *Development and evolution of chordate cartilage.* Journal of Experimental Zoology Part B-Molecular and Developmental Evolution, 2007. **308B**(3): p. 325-335.
104. Chizhikov, V.V. and K.J. Millen, *Control of roof plate formation by Lmx1a in the developing spinal cord.* Development, 2004. **131**(11): p. 2693-2705.

105. Yu, W., et al., *Global Analysis of DNA Methylation by Methyl-Capture Sequencing Reveals Epigenetic Control of Cisplatin Resistance in Ovarian Cancer Cell.* PLoS One, 2011. **6**(12).
106. Li, N., et al., *Cisplatin-induced downregulation of SOX1 increases drug resistance by activating autophagy in non-small cell lung cancer cell.* Biochem Biophys Res Commun, 2013. **439**(2): p. 187-90.
107. Baylin, S.B. and J.G. Herman, *DNA hypermethylation in tumorigenesis - epigenetics joins genetics.* Trends in Genetics, 2000. **16**(4): p. 168-174.
108. Holschneider, C.H. and J.S. Berek, *Ovarian cancer: Epidemiology, biology, and prognostic factors.* Seminars in Surgical Oncology, 2000. **19**(1): p. 3-10.
109. Tschernatsch, M., et al., *Anti-SOX1 antibodies in patients with paraneoplastic and non-paraneoplastic neuropathy.* Journal of Neuroimmunology, 2010. **226**(1-2): p. 177-180.
110. Sabater, L., et al., *SOX1 antibodies are markers of paraneoplastic Lambert-Eaton myasthenic syndrome.* Neurology, 2008. **70**(12): p. 924-928.
111. Lipka, A.F., J.J. Verschuuren, and M.J. Titulaer, *SOX1 antibodies in Lambert-Eaton myasthenic syndrome and screening for small cell lung carcinoma.* Ann N Y Acad Sci, 2012. **1275**: p. 70-7.
112. Graus, F., et al., *Anti-glial nuclear antibody: marker of lung cancer-related paraneoplastic neurological syndromes.* J Neuroimmunol, 2005. **165**(1-2): p. 166-71.
113. Sabater, L., et al., *SOX1 antibodies are markers of paraneoplastic Lambert-Eaton myasthenic syndrome.* Neurology, 2008. **70**(12): p. 924-8.
114. Sabater, L., et al., *Antibody repertoire in paraneoplastic cerebellar degeneration and small cell lung cancer.* PLoS One, 2013. **8**(3): p. e60438.
115. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
116. Fang, Y. and M.J. Fullwood, *Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer.* Genomics Proteomics Bioinformatics, 2016. **14**(1): p. 42-54.
117. Cheetham, S.W., et al., *Long noncoding RNAs and the genetics of cancer.* Br J Cancer, 2013. **108**(12): p. 2419-25.
118. Coetzee, S.G., et al., *FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs.* Nucleic Acids Res, 2012. **40**(18): p. e139.
119. Gibb, E.A., C.J. Brown, and W.L. Lam, *The functional role of long non-coding RNA in human carcinomas.* Mol Cancer, 2011. **10**: p. 38.
120. Sun, T., et al., *Emerging players in prostate cancer: long non-coding RNAs.* Am J Clin Exp Urol, 2014. **2**(4): p. 294-9.
121. Walsh, A.L., et al., *Long noncoding RNAs and prostate carcinogenesis: the missing 'linc'?* Trends Mol Med, 2014. **20**(8): p. 428-36.
122. Hauptman, N. and D. Glavac, *Long non-coding RNA in cancer.* Int J Mol Sci, 2013. **14**(3): p. 4655-69.
123. Hajjari, M. and A. Salavaty, *HOTAIR: an oncogenic long non-coding RNA in different cancers.* Cancer Biol Med, 2015. **12**(1): p. 1-9.
124. Wu, Y., et al., *Long Noncoding RNA MALAT1: Insights into its Biogenesis and Implications in Human Disease.* Curr Pharm Des, 2015. **21**(34): p. 5017-28.
125. Ji, P., et al., *MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer.* Oncogene, 2003. **22**(39): p. 8031-41.

126. Gutschner, T., M. Hammerle, and S. Diederichs, *MALAT1 -- a paradigm for long noncoding RNA function in cancer.* J Mol Med (Berl), 2013. **91**(7): p. 791-801.
127. Gutschner, T., et al., *The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells.* Cancer Res, 2013. **73**(3): p. 1180-9.
128. Blom, N., et al., *Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence.* Proteomics, 2004. **4**(6): p. 1633-1649.
129. Cohen, P., *The role of protein phosphorylation in human health and disease - Delivered on June 30th 2001 at the FEBS Meeting in Lisbon.* European Journal of Biochemistry, 2001. **268**(19): p. 5001-5010.
130. Latchman, D.S., *Transcription factors: an overview.* Int J Biochem Cell Biol, 1997. **29**(12): p. 1305-12.
131. Whitmarsh, A.J. and R.J. Davis, *Regulation of transcription factor function by phosphorylation.* Cell Mol Life Sci, 2000. **57**(8-9): p. 1172-83.
132. Jeong, C.H., et al., *Phosphorylation of Sox2 cooperates in reprogramming to pluripotent stem cells.* Stem Cells, 2010. **28**(12): p. 2141-50.
133. Takahashi, K. and S. Yamanaka, *Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.* Cell, 2006. **126**(4): p. 663-76.
134. Ouyang, J., et al., *Cyclin-dependent kinase-mediated Sox2 phosphorylation enhances the ability of Sox2 to establish the pluripotent state.* J Biol Chem, 2015. **290**(37): p. 22782-94.
135. Gong, C.X. and K. Iqbal, *Hyperphosphorylation of microtubule-associated protein tau: a promising therapeutic target for Alzheimer disease.* Curr Med Chem, 2008. **15**(23): p. 2321-8.
136. Benzeno, S., et al., *Identification of mutations that disrupt phosphorylation-dependent nuclear export of cyclin D1.* Oncogene, 2006. **25**(47): p. 6291-303.
137. Peter-Katalinic, J., *Methods in enzymology: O-glycosylation of proteins.* Methods Enzymol, 2005. **405**: p. 139-71.
138. Hart, G.W., *Dynamic O-linked glycosylation of nuclear and cytoskeletal proteins.* Annual Review of Biochemistry, 1997. **66**: p. 315-335.
139. Hart, G.W., et al., *Cross Talk Between O-GlcNAcylation and Phosphorylation: Roles in Signaling, Transcription, and Chronic Disease.* Annual Review of Biochemistry, Vol 80, 2011. **80**: p. 825-858.
140. Dias, W.B. and G.W. Hart, *O-GlcNAc modification in diabetes and Alzheimer's disease.* Mol Biosyst, 2007. **3**(11): p. 766-72.
141. Hart, G.W., et al., *Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease.* Annu Rev Biochem, 2011. **80**: p. 825-58.
142. Chaiyawat, P., et al., *Aberrant O-GlcNAcylated Proteins: New Perspectives in Breast and Colorectal Cancer.* Front Endocrinol (Lausanne), 2014. **5**: p. 193.
143. Mahajan, R., et al., *A small ubiquitin-related polypeptide involved in targeting RanGAP1 to nuclear pore complex protein RanBP2.* Cell, 1997. **88**(1): p. 97-107.
144. Zhang, D., et al., *A novel post-translational modification of nucleolin, SUMOylation at Lys-294, mediates arsenite-induced cell death by regulating gadd45alpha mRNA stability.* J Biol Chem, 2015. **290**(8): p. 4784-800.
145. Stielow, B., et al., *SUMO-modified Sp3 represses transcription by provoking local heterochromatic gene silencing.* EMBO Rep, 2008. **9**(9): p. 899-906.

146. Lin, X., et al., *SUMO-1/Ubc9 promotes nuclear accumulation and metabolic stability of tumor suppressor Smad4.* J Biol Chem, 2003. **278**(33): p. 31043-8.
147. Hilgarth, R.S. and K.D. Sarge, *Detection of sumoylated proteins.* Methods Mol Biol, 2005. **301**: p. 329-38.
148. Zhao, Q., et al., *GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W325-30.
149. Hietakangas, V., et al., *PDSM, a motif for phosphorylation-dependent SUMO modification.* Proc Natl Acad Sci U S A, 2006. **103**(1): p. 45-50.
150. Tahmasebi, S., et al., *Sumoylation of Kruppel-like factor 4 inhibits pluripotency induction but promotes adipocyte differentiation.* J Biol Chem, 2013. **288**(18): p. 12791-804.
151. Tsuruzoe, S., et al., *Inhibition of DNA binding of Sox2 by the SUMO conjugation.* Biochem Biophys Res Commun, 2006. **351**(4): p. 920-6.
152. Kim, K.I. and S.H. Baek, *SUMOylation code in cancer development and metastasis.* Mol Cells, 2006. **22**(3): p. 247-53.
153. Sarge, K.D. and O.K. Park-Sarge, *Sumoylation and human disease pathogenesis.* Trends Biochem Sci, 2009. **34**(4): p. 200-5.
154. Eifler, K. and A.C. Vertegaal, *SUMOylation-Mediated Regulation of Cell Cycle Progression and Cancer.* Trends Biochem Sci, 2015. **40**(12): p. 779-93.
155. Aubert, J., et al., *Screening for mammalian neural genes via fluorescence-activated cell sorter purification of neural precursors from Sox1-gfp knock-in mice.* Proc Natl Acad Sci U S A, 2003. **100 Suppl 1**: p. 11836-41.
156. Andrews, P.W., et al., *Pluripotent embryonal carcinoma clones derived from the human teratocarcinoma cell line Tera-2. Differentiation in vivo and in vitro.* Lab Invest, 1984. **50**(2): p. 147-62.
157. Pittenger, M.F., et al., *Multilineage potential of adult human mesenchymal stem cells.* Science, 1999. **284**(5411): p. 143-7.
158. Hoffrogge, R., et al., *2-DE proteome analysis of a proliferating and differentiating human neuronal stem cell line (ReNcell VM).* Proteomics, 2006. **6**(6): p. 1833-47.
159. Scherer, W.F., J.T. Syverton, and G.O. Gey, *Studies on the propagation in vitro of poliomyelitis viruses. IV. Viral multiplication in a stable strain of human malignant epithelial cells (strain HeLa) derived from an epidermoid carcinoma of the cervix.* J Exp Med, 1953. **97**(5): p. 695-710.
160. Biedler, J.L., et al., *Multiple neurotransmitter synthesis by human neuroblastoma cell lines and clones.* Cancer Res, 1978. **38**(11 Pt 1): p. 3751-7.
161. Rhim, J.S., et al., *Characterization of non-producer human cells induced by Kirsten sarcoma virus.* Int J Cancer, 1975. **16**(5): p. 840-9.
162. Sambuy, Y., et al., *The Caco-2 cell line as a model of the intestinal barrier: influence of cell and culture-related factors on Caco-2 cell functional characteristics.* Cell Biol Toxicol, 2005. **21**(1): p. 1-26.
163. Soule, H.D., et al., *A human cell line from a pleural effusion derived from a breast carcinoma.* J Natl Cancer Inst, 1973. **51**(5): p. 1409-16.
164. Laurent, L.C., et al., *Restricted ethnic diversity in human embryonic stem cell lines.* Nat Methods, 2010. **7**(1): p. 6-7.
165. Soule, H.D., et al., *Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10.* Cancer Res, 1990. **50**(18): p. 6075-86.
166. Jacobs, J.P., C.M. Jones, and J.P. Baille, *Characteristics of a human diploid cell designated MRC-5.* Nature, 1970. **227**(5254): p. 168-70.

167. Brattain, M.G., et al., *Heterogeneity of malignant cells from a human colonic carcinoma.* Cancer Res, 1981. **41**(5): p. 1751-6.
168. Cailleau, R., M. Olive, and Q.V. Cruciger, *Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization.* In Vitro, 1978. **14**(11): p. 911-5.
169. Hackett, A.J., et al., *Two syngeneic cell lines from human breast tissue: the aneuploid mammary epithelial (Hs578T) and the diploid myoepithelial (Hs578Bst) cell lines.* J Natl Cancer Inst, 1977. **58**(6): p. 1795-806.
170. Keydar, I., et al., *Establishment and characterization of a cell line of human breast carcinoma origin.* Eur J Cancer, 1979. **15**(5): p. 659-70.
171. Stangegaard, M., I.H. Dufva, and M. Dufva, *Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA.* Biotechniques, 2006. **40**(5): p. 649-57.
172. Nat, R., et al., *Neurogenic neuroepithelial and radial glial cells generated from six human embryonic stem cell lines in serum-free suspension and adherent cultures.* Glia, 2007. **55**(4): p. 385-99.
173. Archer, T.C., J. Jin, and E.S. Casey, *Interaction of Sox1, Sox2, Sox3 and Oct4 during primary neurogenesis.* Developmental Biology, 2011. **350**(2): p. 429-40.
174. Feng, N., et al., *Generation of highly purified neural stem cells from human adipose-derived mesenchymal stem cells by Sox1 activation.* Stem Cells Dev, 2014. **23**(5): p. 515-29.
175. Poloni, A., et al., *Glial-like differentiation potential of human mature adipocytes.* J Mol Neurosci, 2015. **55**(1): p. 91-8.
176. Livak, K.J. and T.D. Schmittgen, *Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.* Methods, 2001. **25**(4): p. 402-8.
177. Hall, T.A., *BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT*, in *Nucl. Acids*. 1999.
178. Kent, W.J., *BLAT - The BLAST-like alignment tool.* Genome Research, 2002. **12**(4): p. 656-664.
179. Apweiler, R., et al., *Update on activities at the Universal Protein Resource (UniProt) in 2013.* Nucleic Acids Research, 2013. **41**(D1): p. D43-D47.
180. Liu, W., et al., *IBS: an illustrator for the presentation and visualization of biological sequences.* Bioinformatics, 2015. **31**(20): p. 3359-61.
181. de Castro, E., et al., *ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W362-5.
182. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
183. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D733-45.
184. Sievers, F., et al., *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* Mol Syst Biol, 2011. **7**: p. 539.
185. Hornbeck, P.V., et al., *PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse.* Nucleic Acids Research, 2012. **40**(D1): p. D261-D270.

186. Blom, N., S. Gammeltoft, and S. Brunak, *Sequence and structure-based prediction of eukaryotic protein phosphorylation sites.* J Mol Biol, 1999. **294**(5): p. 1351-62.
187. Gupta, R. and S. Brunak, *Prediction of glycosylation across the human proteome and the correlation to protein function.* Pac Symp Biocomput, 2002: p. 310-22.
188. Beauclair, G., et al., *JASSA: a comprehensive tool for prediction of SUMOylation sites and SIMs.* Bioinformatics, 2015. **31**(21): p. 3483-91.
189. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.
190. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.
191. Bustin, S.A., et al., *The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments.* Clin Chem, 2009. **55**(4): p. 611-22.
192. Popovic, J., et al., *Expression analysis of SOX14 during retinoic acid induced neural differentiation of embryonal carcinoma cells and assessment of the effect of its ectopic expression on SOXB members in HeLa cells.* PLoS One, 2014. **9**(3): p. e91852.
193. Carson, S.D. and S.J. Pirruccello, *HeLa cell heterogeneity and coxsackievirus B3 cytopathic effect: implications for inter-laboratory reproducibility of results.* J Med Virol, 2013. **85**(4): p. 677-83.
194. Amini, S., et al., *The expressions of stem cell markers: Oct4, Nanog, Sox2, nucleostemin, Bmi, Zfx, Tcl1, Tbx3, Dppa4, and Esrrb in bladder, colon, and prostate cancer, and certain cancer cell lines.* Anat Cell Biol, 2014. **47**(1): p. 1-11.
195. Eini, R., et al., *Role of SOX2 in the etiology of embryonal carcinoma, based on analysis of the NCCIT and NT2 cell lines.* PLoS One, 2014. **9**(1): p. e83585.
196. Yang, N., et al., *SOX 1, contrary to SOX 2, suppresses proliferation, migration, and invasion in human laryngeal squamous cell carcinoma by inhibiting the Wnt/beta-catenin pathway.* Tumour Biol, 2015. **36**(11): p. 8625-35.
197. Basu-Roy, U., et al., *Sox2 maintains self renewal of tumor-initiating cells in osteosarcomas.* Oncogene, 2012. **31**(18): p. 2270-82.
198. Leis, O., et al., *Sox2 expression in breast tumours and activation in breast cancer stem cells.* Oncogene, 2012. **31**(11): p. 1354-65.
199. Saigusa, S., et al., *Correlation of CD133, OCT4, and SOX2 in rectal cancer and their association with distant recurrence after chemoradiotherapy.* Ann Surg Oncol, 2009. **16**(12): p. 3488-98.
200. Weina, K. and J. Utikal, *SOX2 and cancer: current research and its implications in the clinic.* Clin Transl Med, 2014. **3**: p. 19.
201. Bediaga, N.G., et al., *DNA methylation epigenotypes in breast cancer molecular subtypes.* Breast Cancer Res, 2010. **12**(5): p. R77.
202. Kao, J., et al., *Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery.* PLoS One, 2009. **4**(7): p. e6146.
203. Holliday, D.L. and V. Speirs, *Choosing the right cell line for breast cancer research.* Breast Cancer Res, 2011. **13**(4): p. 215.
204. Widschwendter, M., et al., *DNA methylation analysis in liquid-based cytology for cervical cancer screening.* International Journal of Cancer, 2009. **125**(12): p. 2995-3002.

205. Lai, H.C., et al., *Identification of novel DNA methylation markers in cervical cancer.* International Journal of Cancer, 2008. **123**(1): p. 161-167.
206. Mathews, L.A., et al., *Epigenetic regulation of CpG promoter methylation in invasive prostate cancer cells.* Molecular Cancer, 2010. **9**: p. -.
207. Kan, L., et al., *Sox1 acts through multiple independent pathways to promote neurogenesis.* Dev Biol, 2004. **269**(2): p. 580-94.
208. Archer, T.C., J. Jin, and E.S. Casey, *Interaction of Sox1, Sox2, Sox3 and Oct4 during primary neurogenesis.* Dev Biol, 2011. **350**(2): p. 429-40.
209. Zhang, S. and W. Cui, *Sox2, a key factor in the regulation of pluripotency and neural differentiation.* World J Stem Cells, 2014. **6**(3): p. 305-11.
210. Fantes, J., et al., *Mutations in SOX2 cause anophthalmia.* Nat Genet, 2003. **33**(4): p. 461-3.
211. Amaral, P.P., et al., *Complex architecture and regulated expression of the Sox2ot locus during vertebrate development.* RNA, 2009. **15**(11): p. 2013-27.
212. Shahryari, A., et al., *Two novel splice variants of SOX2OT, SOX2OT-S1, and SOX2OT-S2 are coupregulated with SOX2 and OCT4 in esophageal squamous cell carcinoma.* Stem Cells, 2014. **32**(1): p. 126-34.
213. Askarian-Amiri, M.E., et al., *Emerging role of long non-coding RNA SOX2OT in SOX2 regulation in breast cancer.* PLoS One, 2014. **9**(7): p. e102140.
214. Gure, A.O., et al., *Serological identification of embryonic neural proteins as highly immunogenic tumor antigens in small cell lung cancer.* Proc Natl Acad Sci U S A, 2000. **97**(8): p. 4198-203.
215. Shahryari, A., et al., *Long non-coding RNA SOX2OT: expression signature, splicing patterns, and emerging roles in pluripotency and tumorigenesis.* Front Genet, 2015. **6**: p. 196.
216. Ng, S.Y., R. Johnson, and L.W. Stanton, *Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors.* EMBO J, 2012. **31**(3): p. 522-33.
217. Gupta, R.A., et al., *Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis.* Nature, 2010. **464**(7291): p. 1071-6.
218. Saghaeian Jazi, M., et al., *Identification of new SOX2OT transcript variants highly expressed in human cancer cell lines and down regulated in stem cell differentiation.* Mol Biol Rep, 2016. **43**(2): p. 65-72.
219. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2007. **35**(Database issue): p. D61-5.
220. Ovcharenko, I., et al., *ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W280-6.
221. Wilusz, J.E., H. Sunwoo, and D.L. Spector, *Long noncoding RNAs: functional surprises from the RNA world.* Genes Dev, 2009. **23**(13): p. 1494-504.
222. Dinger, M.E., et al., *Differentiating protein-coding and noncoding RNA: challenges and ambiguities.* PLoS Comput Biol, 2008. **4**(11): p. e1000176.
223. Niazi, F. and S. Valadkhan, *Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs.* RNA, 2012. **18**(4): p. 825-43.
224. Consortium, F., et al., *A promoter-level mammalian expression atlas.* Nature, 2014. **507**(7493): p. 462-70.

225. Rosenbloom, K.R., et al., *ENCODE data in the UCSC Genome Browser: year 5 update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D56-63.
226. Kanamori-Katayama, M., et al., *Unamplified cap analysis of gene expression on a single-molecule sequencer.* Genome Res, 2011. **21**(7): p. 1150-9.
227. Wang, Y.C., S.E. Peterson, and J.F. Loring, *Protein post-translational modifications and regulation of pluripotency in human stem cells.* Cell Res, 2014. **24**(2): p. 143-60.
228. Hornbeck, P.V., et al., *PhosphoSitePlus, 2014: mutations, PTMs and recalibrations.* Nucleic Acids Res, 2015. **43**(Database issue): p. D512-20.
229. anti-AcK Antibody Used to Purify Peptides prior to MS2: Acetylated-Lysine (Ac-K2-100) Rabbit mAb Cat#: 9814, PTMScan(R) Acetyl-Lys Motif (Ac-K) Immunoaffinity Beads Cat#: 1989 , C.C.S.Y.B.H.D.c.c.T.S.S.o.A.U.t.P.P.p.t.M.
230. Trinidad, J.C., et al., *Global Identification and Characterization of Both O-GlcNAcylation and Phosphorylation at the Murine Synapse.* Molecular & Cellular Proteomics, 2012. **11**(8): p. 215-229.
231. Gupta, R., *Prediction of glycosylation sites in proteomes: from post-translational modifications to protein function*, in *CBS*. 2001.
232. Mertins, P., et al., *Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels.* Mol Cell Proteomics, 2014. **13**(7): p. 1690-704.
233. Li Y (2010) CST Curation Set: 10356; Year: 2010; Biosample/Treatment: tissue, h.u.D.v.t.S.-S.o.A.U.t.P.
234. Guo A (2007) CST Curation Set: 2844; Year: 2007; Biosample/Treatment: cell line, C.C.I.D.-S.-S.o.A.U.t.P.
235. Kamachi, Y., K.S. Cheah, and H. Kondoh, *Mechanism of regulatory target selection by the SOX high-mobility-group domain proteins as revealed by comparison of SOX1/2/3 and SOX9.* Mol Cell Biol, 1999. **19**(1): p. 107-20.
236. Huntley, R.P., et al., *The GOA database: gene Ontology annotation updates for 2015.* Nucleic Acids Res, 2015. **43**(Database issue): p. D1057-63.
237. UniProt, C., *UniProt: a hub for protein information.* Nucleic Acids Res, 2015. **43**(Database issue): p. D204-12.
238. Thomas, J.O., *HMG1 and 2: architectural DNA-binding proteins.* Biochem Soc Trans, 2001. **29**(Pt 4): p. 395-401.
239. Baltus, G.A., et al., *Acetylation of sox2 induces its nuclear export in embryonic stem cells.* Stem Cells, 2009. **27**(9): p. 2175-84.
240. Tanaka, S., et al., *Interplay of SOX and POU factors in regulation of the Nestin gene in neural primordial cells.* Mol Cell Biol, 2004. **24**(20): p. 8834-46.
241. Wu, S.Y. and C.M. Chiang, *Crosstalk between sumoylation and acetylation regulates p53-dependent chromatin transcription and DNA binding.* EMBO J, 2009. **28**(9): p. 1246-59.
242. Yang, X.J. and S. Gregoire, *A recurrent phospho-sumoyl switch in transcriptional repression and beyond.* Mol Cell, 2006. **23**(6): p. 779-86.
243. Kazarian, M. and I.A. Laird-Offringa, *Small-cell lung cancer-associated autoantibodies: potential applications to cancer diagnosis, early detection, and therapy.* Mol Cancer, 2011. **10**: p. 33.
244. Suenaga, Y., et al., *NCYM, a Cis-antisense gene of MYCN, encodes a de novo evolved protein that inhibits GSK3beta resulting in the stabilization of MYCN in human neuroblastomas.* PLoS Genet, 2014. **10**(1): p. e1003996.
245. Song, S.W., et al., *IIp45, an insulin-like growth factor binding protein 2 (IGFBP-2) binding protein, antagonizes IGFBP-2 stimulation of glioma cell invasion.* Proc Natl Acad Sci U S A, 2003. **100**(24): p. 13970-5.

246. Song, L., et al., *SOX1 inhibits breast cancer cell growth and invasion through suppressing the Wnt/beta-catenin signaling pathway.* APMIS, 2016. **124**(7): p. 547-55.
247. Wright, M.H., et al., *Brca1 breast tumors contain distinct CD44+/CD24- and CD133+ cells with cancer stem cell characteristics.* Breast Cancer Res, 2008. **10**(1): p. R10.
248. Xiao, T., et al., *Long Noncoding RNA ADINR Regulates Adipogenesis by Transcriptionally Activating C/EBPalpha.* Stem Cell Reports, 2015. **5**(5): p. 856-65.
249. Liu, K., et al., *The multiple roles for Sox2 in stem cell maintenance and tumorigenesis.* Cell Signal, 2013. **25**(5): p. 1264-71.

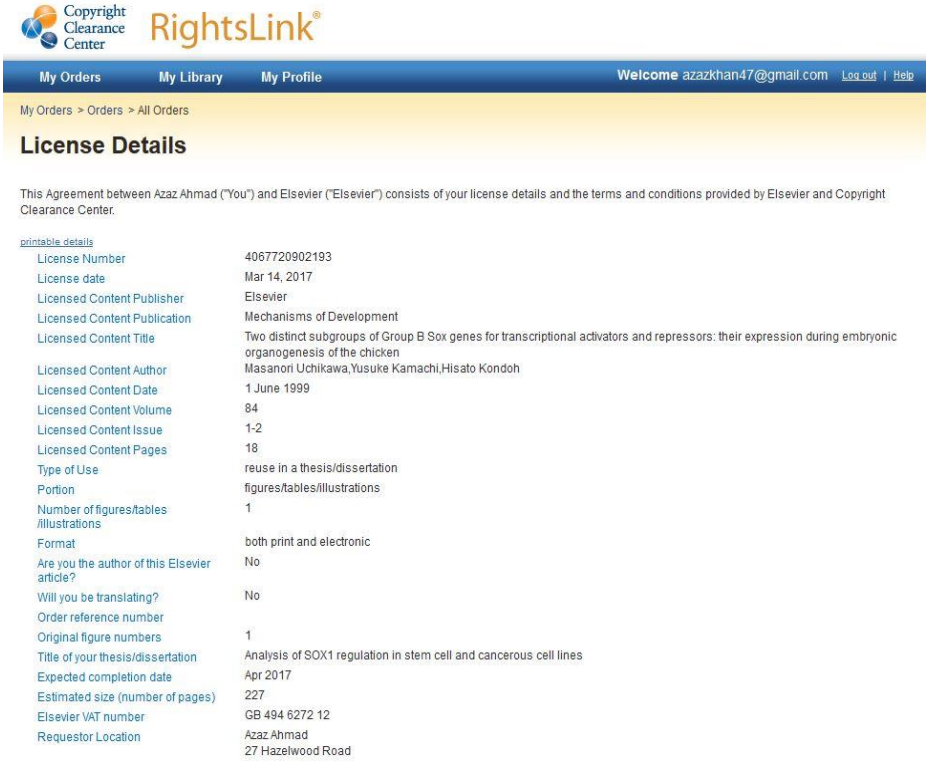
8 Appendix

Products	Catalogue no.
0% pre-cast Tris-Glycine gels	EC6075
2x Laemmli buffer	1610737
5'RACE System for Rapid Amplification of cDNA Ends	18374058
Animal Free Blocker	SP5030
Bovine gamma globulin	500 0001
Bradford assay dye reagent	500 0006
DAB peroxidase substrate kit	SK4100
Dapi-containing Vectashield	HP1200
DNase-I, Amplification grade kit	180868015
E.Coli, XL10 Gold® Ultracompetent cells	200314
Ethidium-bromide	E8751
EZ DNA Methylation-Gold™ Kit	D5005
HyperLadder™ 50bp	BIO-33053
Immun-Star™ Alkaline Phosphatase Substrate and enhancer kit	1705012
MinElute® PCR purification kit	28004
MiniElute gel extraction kit	28604
Nitrocellulose membrane	LC2001
Novex® Tris-Glycine SDS Running Buffer 10X	LC2675
NuPAGE® Transfer Buffer (20X	NP0006
One step PCR inhibitor removal kit	D6030
pGEM-T® Easy Vector	A1360
Platinum® Taq DNA Polymerase	10966026
Power SYBR® Green Master Mix	4368577

protein marker	1610375
QIAquick Gel Extraction Kit	28704
Quick-gDNA (Miniprep) kit	D3024
QuickLyse Mini Prep Kit	27405
RcoRI enzyme digestion	R0101S
RIPA lysis buffer	R0278
RNA Clean and Concentrator	R1015
SuperScript® III Reverse Transcriptase	1808044
T4 DNA ligase	M1801
Taqman Gene Expression Master Mix	4369016
TRI® Reagent	93289

Copyrights Licenses obtained for the published figures reproduced or edited in this study:

A) Copy right License obtained for the **Figure** 1-1, see below



Copyright Clearance Center RightsLink®

My Orders | My Library | My Profile | Welcome azazkhan47@gmail.com Log out | Help

My Orders > Orders > All Orders

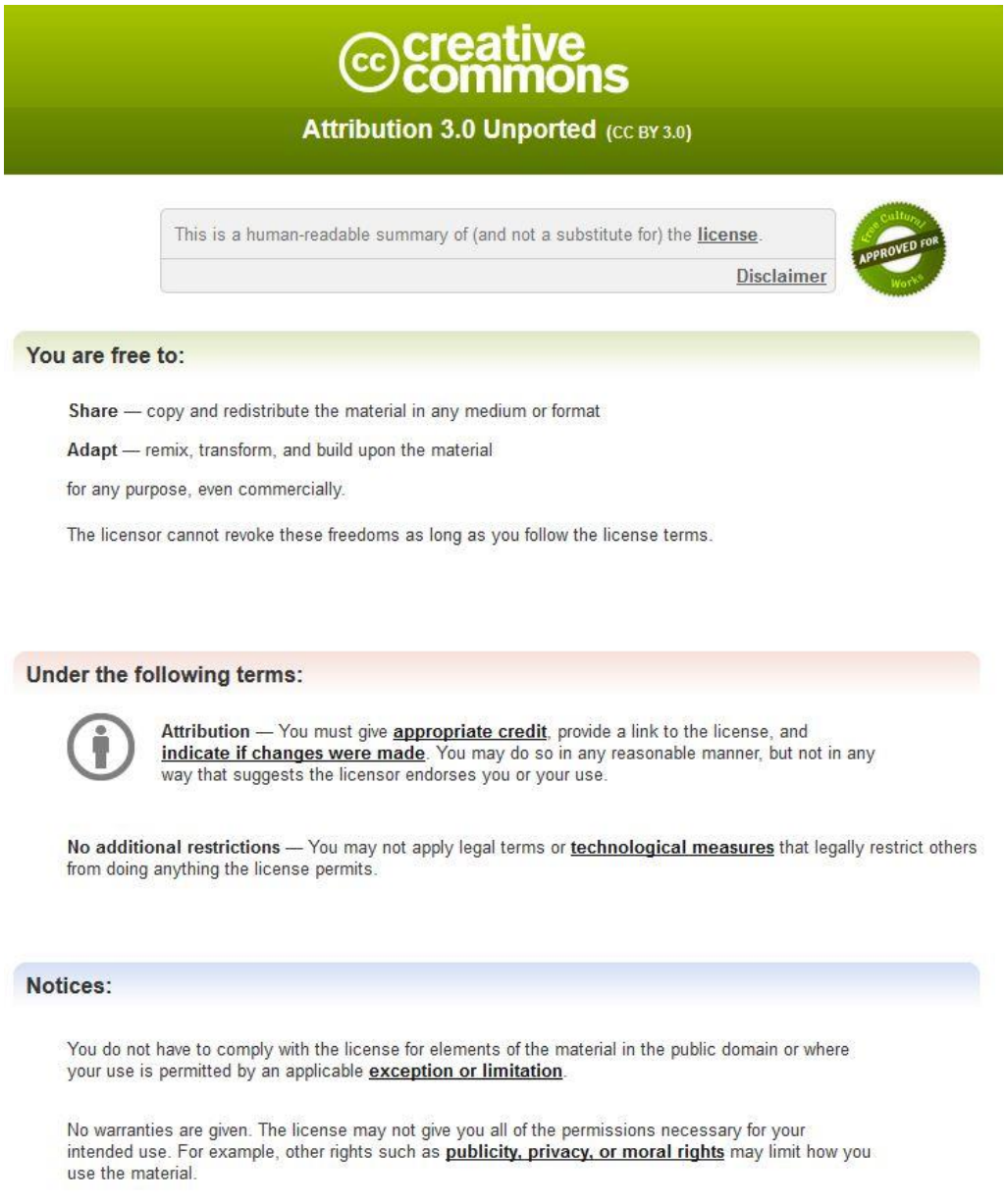
License Details

This Agreement between Azaz Ahmad ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

printable details

License Number	4067720902193
License date	Mar 14, 2017
Licensed Content Publisher	Elsevier
Licensed Content Publication	Mechanisms of Development
Licensed Content Title	Two distinct subgroups of Group B Sox genes for transcriptional activators and repressors: their expression during embryonic organogenesis of the chicken
Licensed Content Author	Masanori Uchikawa,Yusuke Kamachi,Hisato Kondoh
Licensed Content Date	1 June 1999
Licensed Content Volume	84
Licensed Content Issue	1-2
Licensed Content Pages	18
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Order reference number	
Original figure numbers	1
Title of your thesis/dissertation	Analysis of SOX1 regulation in stem cell and cancerous cell lines
Expected completion date	Apr 2017
Estimated size (number of pages)	227
Elsevier VAT number	GB 494 6272 12
Requestor Location	Azaz Ahmad 27 Hazelwood Road

B) Copy right License obtained for Figure 1-3, see below



creative commons

Attribution 3.0 Unported (CC BY 3.0)

This is a human-readable summary of (and not a substitute for) the license.

Disclaimer

APPROVED FOR Free Cultural Works

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material

for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give **appropriate credit**, provide a link to the license, and **indicate if changes were made**. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or **technological measures** that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable **exception or limitation**.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as **publicity, privacy, or moral rights** may limit how you use the material.