AN SSX4 KNOCK-IN CELL LINE MODEL AND *in silico* ANALYSIS OF GENE EXPRESSION DATA AS TWO APPROACHES FOR INVESTIGATING MECHANISMS OF CANCER/TESTIS GENE EXPRESSION

A THESIS SUBMITTED TO
THE DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS
AND THE INSTITUTE OF ENGINEERING AND SCIENCE OF
BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE

BY
DUYGU AKBAŞ AVCI
AUGUST 2009

Sevgili eşim Ender'e ve aileme…

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope, and in quality, as a thesis for the degree of Master of Science.

_____
Assist. Prof. Dr. Ali Güre

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope, and in quality, as a thesis for the degree of Master of Science.

_____
Assoc. Prof. Dr. Hilal Özdağ

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope, and in quality, as a thesis for the degree of Master of Science.

_____
Assist.Prof. Dr.Özlen Konu

Approved for the Institute of Engineering and Science

_____
Prof. Dr. Mehmet Baray

Director of the Institute of Engineering and Science

ABSTRACT

AN SSX4 KNOCK-IN CELL LINE MODEL AND *in silico* ANALYSIS OF GENE EXPRESSION DATA AS TWO APPROACHES FOR INVESTIGATING MECHANISMS OF CANCER/TESTIS GENE EXPRESSION

Duygu Akbaş Avcı
M.Sc. in Molecular Biology and Genetics
Supervisor: Assist. Prof. Ali O. Güre
August 2009, 104 pages

Cancer/testis (CT) genes mapping to the X chromosome (CT-X) are normally expressed in male germ cells but not in adult somatic tissues, with rare exception of oogonia and trophoblast cells; whereas they are aberrantly expressed in various types of cancer. CT-X genes are coordinately expressed and their expression is associated with poor prognosis in various types of cancer. The mechanisms responsible for the reactivation of CT-X genes during tumorigenesis are of great interest because of their prognostic and therapeutic value. In this study, we aimed to develop two approaches by which the mechanisms underlying the regulation of CT-X gene expression in cancer could be identified. Current evidence implicates promoter-specific demethylation as the key event inducing CT-X gene expression in cancer but the mechanisms of this epigenetic deregulation remain to be explored. We presume that coordinately expressed CT-X genes are regulated by common mechanisms. We, thus, decided that the study of a given CT-X gene could elucidate mechanisms pertinent to all.

Our first approach was to generate a a model whereby variations of the expression of an individual CT-X gene, namely SSX4, upon various manipulations could be easily monitored. For this pupose, we used the SSX4 targeting vector to generate an SSX4 knock-in (KI) lung cancer cell line (SK-LC-17) with a GFP reporter gene expressed from SSX4 promoter. SK-LC-17 is known to express SSX4 as well as other CT-X genes and its SSX4 promoter has been characterized in detail. We, thus, obtained one clone with homogenous GFP expression verified by sequencing for correct integration of SSX4 KI targeting vector. In the long-term, this cell line model will be used to identify transcriptional regulators of CT-X gene expression that function either in a direct manner as epigenetic controllers or indirectly as effectors upstream to epigenetic mechanisms.

Based on the fact that CT-X gene expression occurs coordinately in all tumor types, the second series of experiments described herein aimed to develop an approach whereby genes, which are differentially expressed between CT-X expressing (CT-X positive) and non-expressing (CT-X negative) tissues or cells could be identified. Towards this aim a meta-analysis of publicly available microarray datasets from different types of tumors and cancer cell lines was developed. Using this approach, the CT-X positive group was observed to contain gene expression signatures indicative of higher proliferative and metastatic capacity when compared to the CT-X negative group. Additional studies based on class prediction analysis in a lung cancer cell line dataset were performed to compensate for bias due to tissue specific differences between datasets obtained from the meta-analysis. Lastly, we selected a set of genes that behaved commonly in both meta-analysis and class prediction analysis to be validated in cancer cell lines with known CT-X expression profiles.

ÖZET

KANSER-TESTİS GEN İFADESİ MEKANİZMALARININ ARAŞTIRILMASI İÇİN İKİ YAKLAŞIM: SSX4 MODEL HÜCRE HATTININ OLUŞTURULMASI VE GEN İFADE VERİLERİNİN *in silico* ANALİZİ

Duygu Akbaş Avcı
Moleküler Biyoloji ve Genetik Yüksek Lisansı
Tez Yöneticisi: Yrd. Doç. Dr. Ali O. Güre
Ağustos 2009, 104 sayfa

X kromozomu üzerinde bulunan kanser-testis (CT-X) genleri normalde erkek eşey hücrelerinde ifade edilirken, oogonia ve trophoblast hücreleri dışında yetişkin vücut hücrelerinde ifade edilmezler; oysa ki birçok kanser türünde beklenmedik şekilde ifadeleri vardır. CT-X genleri eşgüdümlü olarak ifade edilir ve birçok kanserin kötü gidişatıyla ilişkilendirilmişlerdir. Tümör oluşumu sürecinde CT-X genlerinin yeniden etkinleşmesinden sorumlu mekanizmalar, prognoz ve terapiye yönelik değerlerinden dolayı önem taşımaktadırlar. Bu çalışmada, CT-X gen ifadesinin kontrolünde görevli mekanizmaları tanımlamak amacıyla iki yaklaşım geliştirmeyi hedefledik. Mevcut bulgular CT-X gen ifadesini tetikleyen kilit olay olarak promotora bağlı demetilasyonu işaret etmektedir; ancak bu epigenetik bozulmanın mekanizmaları açıklanmayı beklemektedir. CT-X genlerinin eşgüdümlü ifadelerinin ortak mekanizmalar tarafından düzenlendiğini öngördük. Bu nedenle bir CT-X genindeki mekanizmaların açığa çıkarılmasının tüm diğer CT-X genleriyle ilintili olacağına karar verdik.

İlk yaklaşımımız değişik dış etkenlerce oluşturulan, herhangi bir CT-X gen (bu çalışmada SSX4 geni) ifadesindeki değişimlerin izlenmesini sağlayan bir model oluşturmaktı. Bu amaçla bir SSX4 "knock-in (KI)" akciğer kanseri hücre hattı (SK-LC-17) oluşturmak için, SSX4 genini hedefleyen ve SSX4 promotorundan GFP (yeşil floresan protein) belirteç genini ifade eden bir vektör kullandık. SK-LC-17 akciğer kanseri hücre hattının SSX4 ve diğer CT-X genlerini ifade ettiği bilinmektedir ve bu hücredeki SSX4 promotor bölgesi ayrıntılı olarak tanımlanmıştır. Bu yüzden homojen olarak GFP ifade eden ve SSX4 KI vektörünün doğru olarak yerleştiğinin sekanslanarak doğrulandığı bir tektip hücre (klon) elde ettik. Uzun vadede, bu hücre hattı modeli CT-X gen ifadesinin – doğrudan epigenetik denetleyiciler olarak ya da dolaylı yoldan epigenetik mekanizmaları etkileyerek işleyen – transkripsiyona bağlı düzenleyicilerini bulmak için kullanılacak.

Bu çalışmada tanımladığımız ikinci deney serisi ile, CT-X gen ifadesinin tüm tümör tiplerinde eşgüdümlü olduğu gözönüne alınarak, CT-X ifadesi olan (CT-X pozitif) ve olmayan (CT-X negatif) doku ya da hücrelerde ayırt edici ifadeye sahip genleri bulmayı sağlayacak bir yaklaşım geliştirmeyi hedefledik. Bu amaca yönelik, değişik tümör ve kanser hücre hatlarına ait, ulaşılabilen veri gruplarını kullanarak bir "meta-analiz" yöntemi geliştirdik. Geliştirdiğimiz bu yaklaşımı kullanarak, CT-X pozitif grupların negatiflerle karşılaştırıldığında, yüksek bölünme ve metastaz kapasitesini işaret eden gen ifade imzalarını içerdiğini gözlemledik. Yaptığımız ek çalışmalarda, meta-analizde elde edilen veri grupları arasından dokuya özgü değişimlerin etkilerini gözardı etmek için, seçilen bir akciğer kanseri hücre hattı veri grubunda sınıf-tahmini (class-prediction) analizi yaptık. Son olarak, CT-X ifade profilleri bilinen hücre hatlarında onaylanmak üzere, hem meta-analizde hem de sınıf-tahmini analizinde ortak davranan bir gen seti belirledik.

ACKNOWLEDGEMENTS

I would like to thank the special people who had contributed this work in various ways.

I would like to express my gratitude to Assist. Prof. Ali O. Güre for his supervision, support and valuable suggestions throughout the course of my studies. He always shared his knowledge and experience with me and directed me toward new horizons. I am grateful for his patience, motivation, enthusiasm and understanding.

I am also grateful to Assist. Prof. Özlen Konu for supporting me at every stage of my graduate education and thesis work. Her un-ending energy has always inspired and motivated me.

I am grateful to Koray Doğan Kaya for his contribution to bioinformatics analyses and sharing his knowledge and ideas with us in this study. He was always supportive and helpful during this period.

I would like to thank Dr. Mayda Gürsel for her experimental support.

For their friendship and insights in seemingly troublesome challenges in the lab; I am grateful to Şükrü Atakan, Aydan Karslıoğlu, Derya Dönertaş, Sinem Yılmaz, Kerem Şenses, Esen Oktay, Rasim Barutçu, Şerif Şentürk, Pelin Gülay, Haluk Yüzügüllü, Özge Gürsoy Yüzügüllü, Ayça Arslan Ergül, Tülay Arayıcı, Onur Kaya, and Sinan Gültekin.

I would like to thank all the past and present members of the MBG laboratory.

It is impossible to express my endless love and thanks to my family and my husband. I will forever be grateful to them. I dedicate this thesis to them.

I was supported by project grants given to Dr. Ali O. Güre from TÜBİTAK and European Commission.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

5-azaDC	5-aza-2'-deoxycytidine
bp	Base pair
BRB	Biometric Research Branch
BORIS	Brother of the regulator of imprinted sites
cDNA	Complementary DNA
Ct	Cycle Threshold
CT	Cancer Testis
CTL	Cytotoxic T lymphocyte
C-terminus	Carboxyl terminus
ddH2O	Double distilled water
DMEM	Dulbecco's Modified Eagle's Medium
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic Acid
DNMT	DNA Methyltransferase
dNTP	Deoxyribonucleotide triphosphate
ds	Double strand
dsRNA	Double stranded RNA
E	Efficiency
EDTA	Ethylenediaminetetraacetic acid
EtBr	Ethidium Bromide
FBS	Fetal Bovine Serum
GAGE	G Antigen
GC-RMA	GeneChip Robust multichip average
GSE	Gene Expression Set
GEO	Gene Expression Omnibus
HAT	Histone acetyl transferase
HDAC	Histone deacetylase
HMT	Histone methyl transferase
IR	Inverted repeat
kb	Kilobase
LB	Luria-Bertani media
L1	LINE1 Repeat
MAGEA3	melanoma antigen family A, 3
miRNA	MicroRNA
mRNA	Messenger RNA
µg	Microgram
mg	Miligram
µl	Microliter
NaCl	Sodium chloride
NaOH	Sodium hydroxide
NEAA	Non-essential amino acid
ml	Mililiter
ncRNA	Noncoding RNA
nt	Nucleotide
N-terminus	Amino terminus
NY-ESO-1	cancer/testis antigen 1B
OATL	Ornithine Amino Transferase Like
PBS	Phosphate Buffered Saline

PCR	Polymerase Chain Reaction
Pc	Polycomb
piRNA	Piwi-interacting RNA
qPCR	Quantitative real-time PCR
RNA	Ribonucleic acid
RT-PCR	Reverse Transcription PCR
RP	RankProd
SEREX	Serological Screening of Expression Libraries
siRNA	Small Interfering RNA
SPANX	Sperm protein associated with the nucleus, X-linked
SSX	Synovial Sarcoma X-Translocation
TAE	Tris-Acetate-EDTA buffer
TF	Transcription Factor
Tm	Melting Temperature
TSA	Trichostatin A
TSS	Transcription Start Site
UV	Ultraviolet
v/v	volume/volume

1 INTRODUCTION

1.1 Cancer/Testis Genes

Cancer/testis (CT) genes are normally expressed in male germ cells but not in adult somatic tissues, with rare exception of ovary and trophoblast; whereas they are aberrantly expressed in various tumor types. They often encode tumor antigens that are immunogenic in cancer patients, as a result, they have the potential to be used as biomarkers and targets for immunotherapy. The first CT gene, termed melanoma antigen-1 or MAGE-1 (later renamed MAGEA1), was first isolated by genomic DNA expression cloning using melanoma-reactive cytotoxic T cells derived from a melanoma patient (van der Bruggen, Traversari et al. 1991). A range of other tumor antigen genes, including BAGE and GAGE1 were discovered using cytotoxic T cells isolated from the same patient in which MAGEA1 was discovered (Boel, Wildmann et al. 1995; De Backer, Arden et al. 1999). Since identification of tumor antigens utilizing T cell clones is a relatively difficult process, an easier approach, cDNA expression cloning using serum IgG antibody from cancer patients, called SEREX (serological analysis of recombinant cDNA expression libraries), was subsequently developed by Sahin et. al. (Sahin, Tureci et al. 1995). SSX2 and NY-ESO-1 were the first CT genes identified by SEREX (Sahin, Tureci et al. 1995; Chen, Scanlan et al. 1997). SEREX led to the identification of many additional CT genes. In addition to immunological approaches, many CT genes were found based on their specific mRNA expression profile utilizing high-throughput transcript techniques and analyses (like representational difference analysis (RDA), differential display, cDNA oligonucleotide array analysis, in silico expression analysis) in comparing transcriptomes of tumor versus normal or testis versus other tissues. (Gure, Stockert et al. 2000).

In recent years, the number of CT genes have rapidly increased.The Cancer/Testis gene database (CTdatabase) (http://www.cta.Incc.br) was newly created to gather and uniformly present the available information on CT genes (Almeida, Sakabe et al. 2009). The database provides basic gene, protein and expression information in normal and tumor tissues as well as immunogenicity in cancer patients. The CTdatabase now lists >130 RefSeq nucleotide identifiers as CT genes that belong to 83 gene families.

1.1.1 Genomic organization of CT genes

CT genes are divided between those that are encoded on the X-chromosome (CT-X genes) and those that are not (non-X CT genes). Most of CT-X genes are grouped in families that are embedded in tandem or inverted repeats (Warburton, Giordano et al. 2004). An analysis of the human X chromosome revealed that approximately 10% of the genes on the X-chromosome are CT genes (Ross, Grafham et al. 2005). The presence of CT-X genes as multi-gene families in large highly homologous inverted repeats suggests that CT-X genes mainly arose via segmental duplications.

The non-X CT genes, on the other hand, are distributed throughout the genome and do not generally form gene families or reside within genomic repeats (Simpson, Caballero et al. 2005).

1.1.2 Conservation

Comparison of human and chimpanzee genome showed that all human CT gene families are well conserved between the two species. The divergence rates were analyzed for human and chimpanzee CT gene orthologues and it was found that CT-X genes were evolving faster and undergoing stronger diversifying selection than non-X CT genes (Stevenson, Iseli et al. 2007).

On the other hand, CT genes are poorly conserved between human and mouse, with few exceptions (Stevenson, Iseli et al. 2007). All the MAGE genes identified until now are characterized by the presence of a large central region termed the MAGE homology domain (MHD). MAGE genes are classified into two subgroups, I and II, partly based on their expression profile. The type I MAGE genes have restricted expression pattern of CT genes wherase the type II MAGE genes are also expressed in normal tissues; in fact some of its members are not CT genes (Xiao and Chen 2004). Type I MAGE genes including MAGEA, MAGEB and MAGEC subfamilies are not conserved between human and mouse whereas the type II MAGE genes (mainly MAGED subfamily, necdin) have well-conserved mouse orthologues (Chomez, De Backer et al. 2001). Alignment of the MHD sequences between MAGE genes also revealed that type I and type II genes are phlogenetically distinct branches of the MAGE family. MAGE proteins from *Drosophila* and *Aspergillus* are most closely related to the type II MAGE proteins (Barker and Salehi 2002).

The mouse homologues of human SSX family are found as two subfamilies, Ssxa and Ssxb

(Chen, Alpen et al. 2003). In mouse, Ssxa has only one member whereas Ssxb contains at least 12 closely related members. In this regard, the Ssxb subfamily is more similar to the human SSX family. However, Ssxa and Ssxb sequences are about equally distant from the human SSX genes and there is no evidence that Ssxb is the evolutionarily ancestor of human SSX (Chen, Alpen et al. 2003). In contrast, all human and mouse SSX proteins share conserved KRAB (Kruppel-associated box) domain at the NH2 terminus and SSX-RD domain (SSX repression domain) at the COOH terminus, respectively. This implicates the functional importance of these protein domains (Chen, Alpen et al. 2003).

1.1.3 Expression

In the testis, CT-X genes are generally expressed in speramatogonia, which are proliferating germ cells whereas non-X CT genes are expressed during later stages of germ-cell differentiation, such as in spermatocytes and spermatids.

A recent study reported an in silico expression analysis of 153 CT genes in normal and cancer expression libraries. Based on the combined expression profiles from these libraries and RT-PCR analysis on a panel of 22 normal tissues, it was suggested that CT genes could be classified into 3 groups: (i) testis-restricted (expression in testis and placenta only), testis/brain-restricted (expression in testis, placenta and brain-regions only) and testis-selective (expression in other normal tissues as well) (Hofmann, Caballero et al. 2008). Of 153 genes, 7 CT genes were not identified in any library at all (2 CT-X and 5 non-X CT) and additional 8 CT-X genes were not present in any testis-annotated library. Testis-restricted and testis/brain-restricted CT genes are always expressed at lower intensities in placenta and brain than in testis, respetively. As shown in **Table 1.1**, most of the CT-X genes are testis-restricted or testis/brain-restricted compared to the non-X CT genes (Hofmann, Caballero et al. 2008).

Table 1.1: CT-X and non-X CT genes that show testis-restricted, testis/brain-restricted and testis-selective expression (Hofmann, Caballero et al. 2008)

	CT-X	Non-X CT
Testis-restricted	35	4
Testis-brain restricted	12	2
Testis-selective	26	59

The expression frequency of CT genes is variable in different tumor types. Melanoma, non-small cell lung cancer, hepatocellular carcinoma and bladder cancer have been identified as

high CT-gene expressors, with breast and prostate cancer being moderate and leukemia/lymphoma, renal and colon cancer low expressors (Hofmann, Caballero et al. 2008).

Expression analysis of CT-X genes in breast, melanoma and lung tumpors showed that CT-X genes are frequently co-expressed (Sahin, Tureci et al. 1998; Scanlan, Gure et al. 2002; Tajima, Obata et al. 2003; Gure, Chua et al. 2005). Besides co-expression of CT-X genes, it was shown that coordinately expressed CT-X genes are associated with poor prognosis in multiple myeloma, and non-small cell lung cancer. CT-X gene expression in these tumors is also significantly correlated with later stages of disease (Gure, Chua et al. 2005; Condomines, Hose et al. 2007). In addition, expressional analysis of individual CT-X genes showed that they are more frequently expressed in metastatic tumors than in primary tumors, indicative of a worse prognosis (Scanlan, Gure et al. 2002; Velazquez, Jungbluth et al. 2007).

1.1.4 Regulation of expression

Current evidence indicates that CT-X genes are activated by promoter-specific demethylation. So far, CT-X genes studied are induced by DNA methyltransferase (DNMT) inhibitor, 5-aza-2'-deoxycytidine (5-azaDC) treatment and, their promoter proximal regions are methylated in normal cells and tumor cells, which do not express CT-X genes (Weber, Salgaller et al. 1994; De Smet, Lurquin et al. 1999; Gure, Wei et al. 2002; Lim, Kim et al. 2005; Wischnewski, Pantel et al. 2006). On the other hand, it is interesting that SSX, MAGE and LAGE promoter-reporter constructs are active in both normal cells (fibroblasts) and cancer cell lines (AOG unpublished data) (Scanlan, Gure et al. 2002). This suggests that transcription factors required for the transcriptional activation of CT-X genes are present in both normal and tumor cells. Therefore, the mechanisms that normally lead to DNA methylation of CT-X promoters in normal cells are deregulated and the transcription factors are able to drive CT-X gene expression in cancer cells.

Genome-wide hypomethylation was firstly proposed as a mechanism to induce CT-X gene expression (De Smet, Lurquin et al. 1999). Hypomethylation of repeat sequences (LINE, SINE elements, etc.) cause genomic instability in cancer cells. Although there is an association between hypomethylation of L1 repeats and CT-X genes (Gure AO, unpublished data), genome-wide hypomethylation alone is not sufficient for the activation of CT-X genes as DNA is globally demethylated in colon cancer (Goelz, Vogelstein et al. 1985) , which is a low CT-expressor.

There are two studies investigating the role of DNMTs in epigenetic regulation of CT-X genes. Depletion of DNMT1, but not of DNMT3a and DNMT3b, in MZ2-MEL melanoma cells induced the activation of the MAGEA1 transgene, which was methylated in vitro and integrated into the genome (Loriot, De Plaen et al. 2006). In Hct116 colon cancer cells, the genetic knockout of both DNMT1 and DNMT3b could robustly induce MAGEA1, NY-ESO-1 and XAGE1 expression; whereas individual DNMT1 or DNMT3b knockout had a modest or negligible effect (James, Link et al. 2006).

Along with the DNA methylation, it was found that histone acetylation plays a secondary role as histone deacetyltransferase (HDAC) inhibitor, trichostatin A, by itself or in combination with 5DC could induce CT-X genes, including MAGE and SSX family members (Gure, Wei et al. 2002; Wischnewski, Pantel et al. 2006). It was shown that induction of GAGE gene expression in HEK293 cells by promoter-specific DNA demethylation is dependent on RNA transcription, following histone acetylation (D'Alessio, Weaver et al. 2007).

Moreover, it was suggested that BORIS (brother of the regulator of imprinted sites, a homologue of the abundant transcription factor CTCF) could induce CT-X gene expression. Unlike CTCF, BORIS is not expressed in normal cells whereas it is expressed in male germ cells. During spermatogenesis, its expression coincides with a marked decrease in CTCF expression (Hong, Kang et al. 2005). Both CTCF and BORIS were shown to bind MAGEA1 and NY-ESO-1 promoters. Ectopic expression of BORIS in normal fibroblasts induce demethylation of MAGEA1 and NY-ESO-1 promoters by displacing CTCF at these loci (Hong, Kang et al. 2005; Vatolin, Abdullaev et al. 2005).

Another insight for the regulation of CT-X genes comes from their organization into inverted repeats (IRs) on the X-chromosome (Warburton, Giordano et al. 2004). These inverted repeats containing CT-X genes could form different DNA structures, which may play a role in regulating CT-X gene expression. One of the large invereted repeats, MAGE/CSAG-IR, was proposed to extrude into a double cruciform DNA structure (Losch, Bredenbeck et al. 2007). Then, it was shown that in melanoma cell lines, MAGE and CSAG genes encoded in the MAGE/CSAG-IR are expressed coordinately and independent from the MAGEAs encoded outside the IR (Bredenbeck, Hollstein et al. 2008). It seems that the chromatin structre might be responsible for coordinate expression of CT-X genes in cancer, however, the difference in

this structure should be investigated for normal and cancer cells to understand CT-X gene activation.

1.1.5 Function

1.1.5.1 The function of CT-X genes

Most of CT-X genes do not have characterized biological functions in both the germ line and tumors. However, there is emerging data for MAGE genes mostly in terms of tumorigenesis (Simpson, Caballero et al. 2005). However, how they function in proliferating germ cells (spermatogonia) has remained to be elusive.

Using yeast two-hybrid screen, the transcriptional regulator SKI-interacting protein (SKIP) was identified as a binding partner for MAGEA1 (Laduron, Deplus et al. 2004). SKIP is a transcriptional regulator that connects DNA-binding proteins to coactivators or corepressors. MAGEA1 was found to inhibit the activitiy of SKIP-interacting transactivator, namely the intracellular part of Notch1, by binding to SKIP and recruiting histone deacetylase 1. This shows that MAGEA1 can act as trancriptional repressor (Laduron, Deplus et al. 2004). The function of MAGEA1 in the germ line has not been elucidated, but it is possible that pathways acting through SKIP are involved. It is highly probable that MAGEA1 represses the expression of genes required for differentiation in spermatogonia (Simpson, Caballero et al. 2005). MAGEA4 was similary identified in a yeast two-hybrid screen with the oncoprotein gankyrin; MAGE-A4 binds to gankyrin and suppresses its oncogenic activity (Nagao, Higashitsuji et al. 2003). Recently, MAGE-A3/6 was identified as a novel target of fibroblast growth factor 2-IIIb (FGFR2-IIIb) signaling in thyroid cancer cells, such that FGF7/FGFR2-IIIb activation resulted in H3 methylation and deacetylation of the MAGE-A3/6 promoter, to down-regulate gene expression (Kondo, Zhu et al. 2007).

Recent data indicate that expression of CT genes in cancer cells contributes directly to the malignant phenotype and response to therapy (Simpson, Caballero et al. 2005). It was found that cell lines, which express at least one of the three MAGE genes (MAGEA1, MAGEA2, and MAGEA3), were more resistant to TNF cytotoxicity than cell lines that expressed none of the MAGE genes (Park, Kong et al. 2002). Overexpression of MAGEA2 and MAGEA6 genes leads to acquisition of resistance to the chemotherapeutic drugs paclitaxel and doxorubicin in human cell lines (Glynn, Gammell et al. 2004). Besides MAGE gene family, expression of

GAGE family members, GAGE7C or GAGE7B, contributes directly to tumorigenesis by the inhibiton of apoptosis. Following their transfection into HeLa cells, GAGE7C or GAGEFB conferred resistance to apoptosis induced by either interferon-γ or by FAS (Cilensek, Yehiely et al. 2002).

1.1.5.2 The functions of non-X CT genes

Non-X CT gene products mostly have specific functions in spermatocytes during meiosis and in spermatids. The non-X CT genes SCP1 and SPO11 are components of the synaptonemal complex protein involved in chromosome reduction in meiosis (Keeney, Giroux et al. 1997; Pousette, Leijonhufvud et al. 1997). Their aberrant expression in cancer cells might cause abnormal chromosome segregation and aneuploidy (Simpson, Caballero et al. 2005). Another non-X CT gene, PLU-1, is a transcriptional co-repressor that interacts with the transcription factors BF-1 and PAX9 to regulate gene expression in the germ line (Tan, Shaw et al. 2003). It is most higly expressed in pre-meiotic spermatogonia, where it is proposed to repress the expression of genes required for the maintenace of germ cells in the testis, driving the germ cell differentiation (Madsen, Tarsounas et al. 2003). BRDT/CT-9 was found to mediate chromatin compaction folowing acetylation of histones and it is thougt to function in the elongating spermatids (Pivot-Pajot, Caron et al. 2003). And lastly, TPX1 and ADAM2 (a disintegrin and metalloproteinase domain 2) are expressed on the cell surface where TPX1 attaches spermatogenic cells to the surrounding Sertoli cells in the testis (Busso, Cohen et al. 2005) while the metalloproteinase ADAM2 participates in sperm–egg membrane binding (Evans 2001).

1.1.6 Immunogenicity of CT antigens

NY-ESO-1 is considered to be the most immunogenic CT-X antigen known to date as compared to other CT-X gene products, namely MAGEA1, MAGEA3 and SSX2 (Scanlan, Gure et al. 2002). Spontaneous immunity to NY-ESO-1 is common although immunological responses to NY-ESO-1 vary by individual, cancer type and grade of differentiation (Nicholaou, Ebert et al. 2006). Patients with advanced prostate cancer, neuroblastoma or melanoma are more likely to have detectable anti-NY-ESO-1 antibodies, with antibody responses are observed in up to 50% of patients whose tumors express NY-ESO-1. Simultaneous antibody and T-cell responses are commonly observed for NY-ESO-1 (Nicholaou, Ebert et al. 2006). More recent studies indicate that responses may be unmasked after depletion of regulatory T lymphocytes in vitro, suggesting that active suppression of

anti-NY-ESO-1 cellular immunity also occurs commonly (Nicholaou, Ebert et al. 2006).

Given the restricted expression pattern of CT-X genes and immunogenic properties of protein products of these genes, they present potential for use as therapeutic cancer vaccines. Vaccinations, with antigens specifically expressed by the tumor, are aimed at generating a specific anti-tumor response by triggering the immune system (Zendman, Ruiter et al. 2003). Initial clinical trials with NY-ESO-1 and MAGEA3 were disappointing. Following vaccination with NY-ESO-1 peptide, three of the five patients eventually developed disease progression (Jager, Gnjatic et al. 2000) In addition, injection of the MAGE-3.A1 peptide induced tumor regression in a significant number of the patients, even though no massive CTL (cytotoxic T lymphocyte) response was produced (Marchand, van Baren et al. 1999). Therefore, tumors escape from the attack by the immune system or a sustained immune response can not be developed by NY-ESO-1 and MAGEA3 peptides.

1.2 SSX gene family

Synovial sarcoma X-translocation (SSX) genes were first identified as fusion counterparts to SYT in in t(X;18)(p11.2;q11.2) chromosomal translocation that is present in 70% of synovial sarcomas (Clark, Rocques et al. 1994). The first member of the SSX gene family (HOM-MEL-40) identified by SEREX was SSX2 (Sahin, Tureci et al. 1995; Tureci, Sahin et al. 1996). By genome homology searches all 9 members of the SSX family together with 10 pseudogenes were subsequently identified (Gure, Tureci et al. 1997). SSX mapped to X chromosome within Xp11.2 (Clark, Rocques et al. 1994). SSX family members have high homology ranging from 89 to 95% at the nucleotide level and 77 to 91% at the amino acid level (Gure, Tureci et al. 1997). There are 2 SSX2 and 2 SSX4 genes oriented tail to tail and head to head, respectively. Normal testis expresses SSX1, 2, 3, 4, 5 and 7 but not 6, 8 and 9. Among tumor tissues, SSX1, 2 and 4 expression is found at substantial frequencies, whereas SSX3, 5 and 6 are rarely expressed and SSX7, 8 and 9 expression have not been detected (Gure, Wei et al. 2002). SSX proteins have two domains; one is Kruppel-associated box (KRAB) repression domain at the N terminus, the other is a repression domain (SSX-RD) at the C terminus (Lim, Soulez et al. 1998). They appear to be transcriptional regulators, whose actions are mediated primarily through association with or recruitment of Polycomb group repressors by the SSX-RD domain (Ladanyi 2001).

1.3 Epigenetic regulation of gene expression

Epigenetics is defined as heritable changes in gene expression that are not coded in the DNA sequence itself. Current literature demonstrates clearly the importance of epigenetic gene regulation in development, differentiation and proliferation. Epigenetic deregulation can result in human diseases such as cancer and neurodevelopmental disorders. In mammals, epigenetic processes mainly include DNA methylation, histone modifications, and noncoding RNA-mediated processes. They can not be thougt individually, they interact with each other and constitute a network to regulate gene expression. These epigenetic mechanisms, the crosstalk between them and how they are altered in cancer are summarized below.

1.3.1 DNA methylation

In mammals, methylation occurs almost exclusively at cytosines in the context of CpG dinucleotides (CpGs). Four DNA methyltransferases (DNMTs) sharing a conserved DNMT domain have been identified in mammals. DNMT1 maintains DNA methylation during replication by methylating the hemi-methylated sites (Bestor, Laudano et al. 1988). DNMT3a and DNMT3b are responsible for de novo methylation, as they are able to target unmethylated CpG sites (Okano, Xie et al. 1998). DNMT2 has only weak DNA methyltransferase activity in vitro and has recently been shown to efficiently methylate tRNA (Liang, Chan et al. 2002).

DNA methylation represses gene transcription either by directly preventing the binding of transcription factors to their promoters or through indirectly recruiting methyl-CpG binding proteins (MBDs). DNA methylation is essential for mammalian development, as DNMT3a$^{-/-}$ died at about 4 weeks after birth and DNMT3b$^{-/-}$ exhibited many developmental defects in mice (Okano, Xie et al. 1998). Mammalian DNA methylation has been implicated in a wide range of cellular functions, including tissue-specific gene expression, cell differentiation, cell fate determination, genomic imprinting, and X chromosome inactivation (Li and Zhao 2008).

The first genome-wide analysis of DNA methylation in the human genome showed that gene-rich domains including coding sequences contain high levels of DNA methylation. In colon cancer cells, gene-poor regions showed DNA hypomethylation supporting the hypothesis that global hypomethylation contributes to chromosomal instability and tumor progression (Weber, Davies et al. 2005). The bisulfite sequencing analysis of three human chromosomes confirmed that sequences outside of promoters have a high degree of DNA methylation (Eckhardt, Lewin et al. 2006). Thus in mammals DNA outside regulatory regions (intergenic

DNA, coding DNA and repeat elements) seems to be methylated (Weber and Schubeler 2007).

Genome-wide DNA methylation maps of human somatic (fibroblast) and germline cells showed that most CpG promoters having a high CpG content (HCPs) are unmethylated in both cell types, but a subset of CpG promoters having an intermediate CpG content (ICPs) are methylated only in primary cells wheras they are unmethylated in germ cells. These HCPs carry H3K4me2 which may protect them from DNA methylation. However, they are inactive and how activation of these accessible promoters is prevented is not known. The ICPs are mostly tissue-specific transcription factors , thereby they are repressed by DNA methylation in order to prevent alternative differentiation pathways. CT-X genes fall into HCP class and the mechanisms that are responsible for their methylation might be different than those that cause methylation of ICPs (Weber, Hellmann et al. 2007). How is DNA methylation targeted to the promoters including the promoters of CT-X genes? There are proposed models which are shown in **Figure 1.1**. There could be some protecting factors that prevents DNA methylation at promoters and loss of these factors may cause DNA methylation.Transcription of promoters could prevent DNA methylation but not always. For example; most HCPs do not have methylated CpG islands even though they are inactive (Weber, Hellmann et al. 2007). Some transcription factors such as Myc could interact with DNMTs and recruit them to the promoters (Brenner, Deplus et al. 2005). HMTs by direct interaction with DNMTs or the histone mark itself could recruit DNMTs to gene promoters which will discussed in the next section.

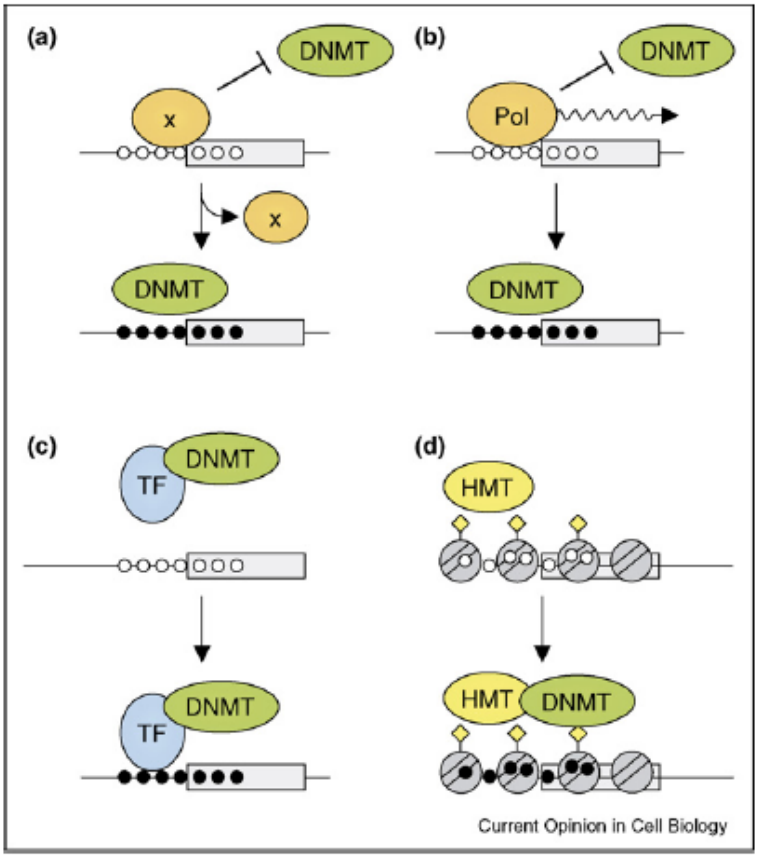


Figure 1.1: Models for targeting DNA methylation to the promoters in mammalian cells (Weber and Schubeler 2007). (a) The selective loss of an as-yet-unidentified protecting factor, X could target DNMTs to gene promoters . (b) Absence of transcription could initiate DNA methylation on some promoters. (c) Some transcription factors (TFs) have been proposed to interact with DNMTs and recruit them to their target sites. (d) HMTs DNMTs could be targeted by histone methylation through an interaction with the histone methyltransferase (HMT) or the histone mark itself. Box denotes first exon; circles denote methylated (black) or unmethylated (white) CpGs.

The DNA methylaton pattern in the human genome has functional importance in terms of gene expression and genome integrity. It was proposed that DNA methylation in the coding DNA inhibits cryptic transcriptional initiation outside gene promoters (Weber and Schubeler 2007). DNA methylation in gene-poor regions (repeat elements) serves to maintain genome integrity. In mammals, most repeats are found to be methylated (Rollins, Haghighi et al. 2006) and DNA methylation mediates their silencing (Walsh, Chaillet et al. 1998; Bourc'his and Bestor 2004; De La Fuente, Baumann et al. 2006). In cancer, genome-wide hypomethylation of repeat sequences lead to genome instability. Deletion of Dnmt1 and Dnmt3b induces chromosomal abnormalities in cancer cell lines (Karpf and Matsui 2005; Chen, Hevi et al. 2007). In addition, there is a strong association between LINE expression caused by DNA hypomethylation and overexpression of the c-MET oncogene in chronic myeloid leukemia. Then, it was found that transcription from the antisense promoter of a

LINE element within intron 2 of c-MET gene is driving its elevated expression (Roman-Gomez, Jimenez-Velasco et al. 2005). Besides genome-wide hypomethylation, there is gene-specific hypomethylation occurring in cancer cells, exemplified by CT-X genes. Which mechanisms are responsible for this pattern in cancer cells have not been known in detail. Despite the main roles of DNMTs in DNA methylation, current evidence does not implicate a reduction in their expression that contributes to cancer-related both gene-specific and genome-wide hypomethylation (Wilson, Power et al. 2007). One study showed an association between hypomethylation of BAGE loci (non-X CT gene) with hypomethylation of nearby juxtacentromic repeats and it was proposed that DNA hypomethylation may proceed into repeat sequences due to the mechanisms that cause hypomethylation of individual genes (Grunau, Sanchez et al. 2005). RNAs may be involved in DNA hypomethylation. One study reported that expression of an antisense RNA to the Sphk1 gene promotes region-specific hypomethylation (Imamura, Yamamoto et al. 2004).

1.3.2 Histone modifications

The nucleosome is the basic structural unit of chromatin, that consists of four core histones- H2A, H2B, H3 and H4- around which 146 bp DNA is wrapped. Histone proteins are subject to over 100 known post-translational modifications, including acetylation, methylation, ADP-ribosylation, ubiquitination, and phosphorylation. These modifications occur on the side chains of specific residues in the histone tails and cores and functionally impact transcription, replication, recombination, and repair (Mendenhall and Bernstein 2008).

All histone acetylations are associated with gene transcription whereas deacetylations are associated with gene repression. Active genes are characterized by high levels of H3K4me1, H3K4me2, H3K4me3, H3K9me1, and H2A.Z (a histone variant) surrounding transcription start sites (TSSs) and elevated levels of H2BK5me1, H3K36me3, H3K27me1, and H4K20me1 downstream of TSS and throughout the entire transcribed regions. In contrast, inactive genes are characterized by low or negligible levels of H3K4 methylation at promoter regions, high levels of H3K27me3 and H3K79me3 in promoter and gene-body regions; low or negligible levels of H3K36me3, H3K27me1, K3K9me1, and H4K20me1 in gene-body regions; and uniformly distributed and low levels of H2A.Z (Barski, Cuddapah et al. 2007).

Almost each of the histone modifications are exerted by different enzymes. In general, histone acetyltransferases (HATs) and deacetylases (HDACs) carry out (de-)acetylation; histone

methyltransferases (HMTs) - specifically lysine and arginine methyltransferases- and lysine demethylases carry out (de-)methylation; and serine/threonine kinases carry out phosphorylation. No arginine demethylases have been identified to date (Kouzarides 2007).

There are a number of indications that repressive histone modifications work hand-in-hand with DNA methylation to repress transcription (Fuks 2005). On the one hand, it is proposed that DNA methylation influences histone modification pattern. DNMTs and MBDs recruit repressor complexes containing HDACs (Bird 2002). On the other hand, it is proposed that histone modification is prerequisite for DNA methylation. In mammals, DNMTs interact with Suv39h H3K9 methyltransferases, and loss of H3K9 methylation in Suv39h-knockout embryonic stem cells showed impaired DNA methylation at major centromeric satellites (Lehnertz, Ueda et al. 2003). Moreover, DNA methylation comes after H3K9 methylation of $p16^{ink4a}$ tumor suppressor gene (Bachman, Park et al. 2003). However, how does crosstalk between DNA methylation and H3K9 methylation occur? There could be adaptor proteins such as HP1 that binds to methylated lysines or a direct interaction can occur between DNMT and H3K9 HMT (Fuks 2005).

The Polycomb group protein, EZH2 is an HMT that mediates H3K27 methylation (H2K27me3) and forms the Polycomb repressive complexes 2 and 3 (PRC2/3) with EED and SUZ12. PRC2/3 play a role Hox gene silencing, X-inactivaton and cancer metastasis. It was shown that EZH2 direct DNA methylation through direct binding with DNMTs (Vire, Brenner et al. 2006).

Lastly, H4K20me3 methylation by the Suv4–20h histone methyltransferases is a hallmark of pericentric heterochromatin (Schotta, Lachner et al. 2004; Martens, O'Sullivan et al. 2005). In cancer cells, a loss of H4K20me3 was observed (Fraga, Ballestar et al. 2005). Whether this loss involves the Suv4–20h enzymes remains to be proven. In the same cancer cells, the loss of H4K20me3 appeared to occur in the vicinity of pericentromeric repeats that show decreased DNA methylation (Fraga, Ballestar et al. 2005). A model was proposed how DNA methylation might be connected to H4K20me3 (Fuks 2005) **(Figure 1.2)**.

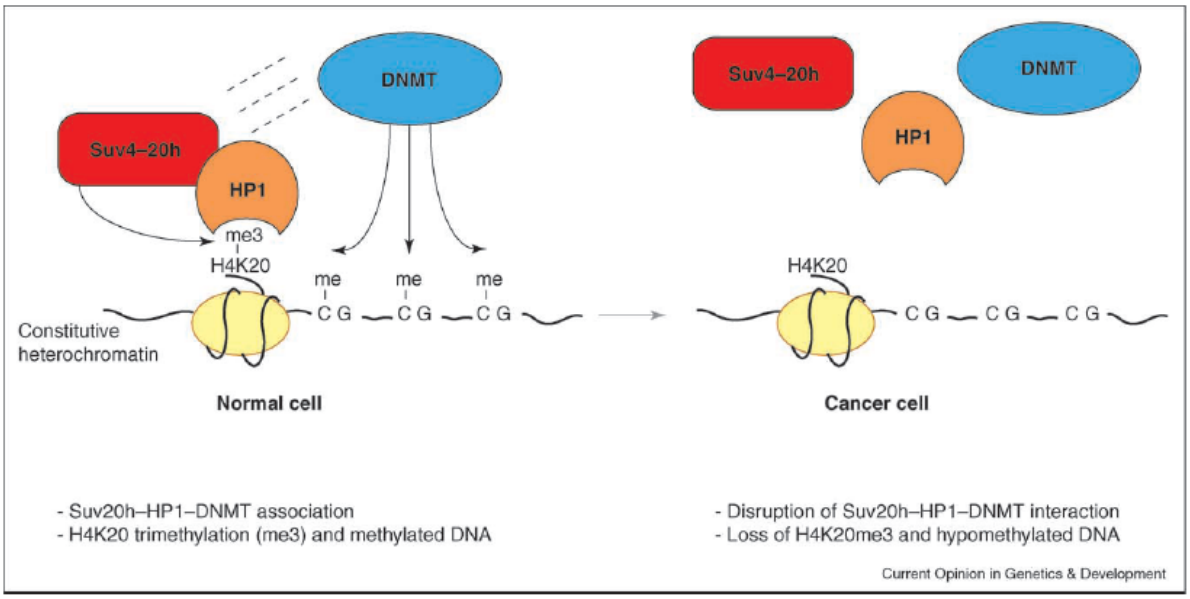


Figure 1.2: Model of how DNA methylation might be linked to H4K20me3 (Fuks 2005). Based on the data generated by Fraga et al., it was suggested that in normal cells, DNMT might interact with Suv4–20h. This interaction might be direct or through the HP1 protein and this would lead to H4K20me3 and methylation of DNA repeat sequences. Which comes first is not known. In cancer cells, the interaction of Suv4–20h, HP1 and DNMT would be disrupted by mutation, translocation, an inappropriate expression level, or defective post-translational modification of one of the partners. This would result in the observed DNA hypomethylation and decrease in H4K20 trimethylation.

1.3.3 Noncoding RNA mediated epigenetic gene regulation

Noncoding RNAs (ncRNAs) play a significant role in the control of epigenetic regulation, chromosomal dynamics, and long-range interactions. ncRNAs are either small or long.

1.3.3.1 Long noncoding RNAs

Long ncRNAs are generally longer than ~200 nucleotides and their expression is strictly regulated (Mercer, Dinger et al. 2009). Long ncRNAs can mediate epigenetic changes by recruiting chromatin remodelling complexes to specific genomic loci. One of the ncRNAs expressed from human homeobox (Hox) loci, silences transcription across 40 kb of the HOXD locus in trans by inducing a repressive chromatin state by recruitment of the Polycomb repressive complex PRC2 (Rinn, Kertesz et al. 2007). One of the long ncRNAs is Xist RNA, which play an essential role in X-chromosome inactivation. Xist RNA is expressed exclusively from the X chromosome to be inactivated. It is <17 kb (depending on species) and it is capped, spliced, and polyadenylated. After Xist RNA coating, the inactive X-chromosome is associated with repressive histone modifications such as H3K9me2 and H3K27 me3. Xist RNA has been shown to recruit EZH2 HMT that trimethylates H3K27 (Heard, Chaumeil et al. 2004). Long ncRNAs are also implicated in genomic imprinting. At

the Kcnq2 and Igf2 imprinted clusters, expression of ncRNAs from the unmethylated paternal alleles is required for silencing in cis. In Kcnq2 imprinted cluster, Kcnq1ot1 long ncRNA is expressed from the paternal allele. In this cluster, all paternally repressed genes were associated with repressive histone modifications such as H3K9me2 and H3K27me3, particularly in the trophectoderm-derived placenta. Then, it was demonstrated that EZH2 was required for imprinted gene repression *in vivo*. Kcnq1ot1 long ncRNA possibly recruits EZH2 to the repressed paternal allele and recruited EZH2 trimethylates H3K27 repressing gene expression on the paternal allele (Terranova, Yokobayashi et al. 2008).

1.3.3.2 Small noncoding RNAs

Small RNAs are characterized by their limited size (~20 -30 nucleotides) and their association with Argonaute (Ago) family proteins that have a role in all small RNA pathways. Ago proteins bind various <32 nt small RNAs which guide the Argonaute complexes to their regulatory targets. The Ago family proteins can be grouped into two classes: the Ago subfamily and the Piwi subfamily.At least three classes of small RNAs are encoded in human genome, based on their biogenesis mechanism and the type of Ago protein that they are associated with: microRNAs (miRNAs), endogenous small interfering RNAs (endosiRNAs or esiRNAs) and Piwi-interacting RNAs (piRNAs). Although these are the three main small RNAs known to date, numerous other small RNAs are still being discovered in the light of the recent developments (Kim, Han et al. 2009).

RNAi mediates heterochromatin formation in fission yeast. siRNAs generated from the heterochromatin regions were suggested to recruit H3K9 HMT to these loci. Because RNAi is central to heterochromatin formation, this study has challenged the intuitive belief that silent chromatin is not transcribed (and therefore, that RNA is not available or required to initiate silencing). RNAi-mediated chromatin effects have also been uncovered in organisms as diverse as *Tetrahymena*, *Drosophila*, and mammals, but the detailed mechanisms have yet to be revealed (Hall, Shankaranarayana et al. 2002; Volpe, Kidner et al. 2002; Bernstein and Allis 2005).

RNAi-like mechanisms are now known to play a critical role in mediating heterochromatic gene silencing and can prevent the mobilization of transposable elements (Bernstein and Allis 2005). RNAi-deficient *C. elegans* show high rates of transposition (Tabara, Sarkissian et al. 1999). In *Drosophila*, I elements (similar to mammalian LINE elements) can be silenced by previous introduction of transgenes expressing a small region of the transposon (Jensen,

Gassama et al. 1999; Bernstein and Allis 2005). In the mouse embryo, knock-down of Dicer results in an increase in the levels of retrotransposon transcripts (IAP and MuERVL) (Svoboda 2004). These results indicate that RNAi mechanism is important for maintainance of genomic stability and it may be a conserved mechanism across species (Bernstein and Allis 2005).

1.4 Combined analysis of microarray datasets: meta-analysis

With the implementation and wide-spread use of high-throughput microarray technology, there occurred a massive increase in publicly available datasets that can be used for subsequent analysis. However, direct comparison among heterogenous datasets was not possible due to the complicated experimental variables that are intrinsic to array experiments. For the elimination of these handicaps meta-analysis of microarray datasets appears to be a good and practical solution. Meta-analysis is a powerful tool for analyzing microarray experiments by combining data from multiple studies (Hong and Breitling 2008).Various papers have been published comparing data across labs generated by diffferent platforms (both Agilent and Affymetrix platforms) to determine whether they are comparable or not. Among different platforms, the Affymetrix platform provides by far the most consistent data across labs (Irizarry, Warren et al. 2005).

In recent years several meta-analysis methods have been proposed using different approaches. First, Fisher's inverse chi-square test computes a combined statistic form the *P*-values obtained from the analysis of the individual datasets. This method is easy to use and does not require additional analysis. However, by working with the *P*-values it is impossible to estimate the average magnitude of differential expression and one can obtain inconsistent fold changes (Hong and Breitling 2008). Secondly, Choi et al. used a t-like statistic (defined as an effect size) as the summary statistic for each gene from each individual datasets. They then proposed a hierarchical modeling approach to assess both intra- and inter-study variation in the summary statistic across multiple datasets and reports The approach has been implemented into a Bioconductor package *GeneMeta* facilitating its application (Choi, Yu et al. 2003). Lastly, the non-parametric rank product (RP) method has been introduced in another Bioconductor package (RankProd) (Hong, Breitling et al. 2006). It was initially proposed to identify differentially expressed genes between two conditions and based on calculating the rank products from replicate experiments (Breitling, Armengaud et al. 2004).

Then, it was developed to be used as a meta-analysis algorithm (Hong, Breitling et al. 2006). It is derived from biological reasoning about the fold-change criterion and identifies genes that are consistently found among the most up-regulated and down-regulated genes in a number of experiments (Hong and Breitling 2008) The rank product method offers several advantages over linear models or *t*-tests. It has increased power in low sample number and/or large noise settings. In addition, it has the ability to overcome the heterogeneity among multiple datasets and has been shown to be more consistent and reliable as compared to t-test based methods (Breitling and Herzyk 2005) Both the *t*-test based and RP method utilize permutation tests to assess the statistical significance, reporting the false discovery rate (FDR) of the identification based on combined statistics (Hong and Breitling 2008)

There are key points to be considered in conducting a meta-analysis. A recent review on the meta-anlysis presented a checklist for conducting meta-analysis of microarray datasets by dissecting the process to seven distinct issues (Ramasamy, Mondry et al. 2008). The first five issues were related to data acquisition and curation: identifying suitable microarray studies, extracting the data from studies, preparing the individual datasets, annotating the individual datasets, resolving the many-to-many relationship between probes and genes. Choosing the appropraite meta-analysis technique was presented as the sixth issue (Ramasamy, Mondry et al. 2008). The seventh issue of analyzing, presenting, and interpreting data was discussed briefly using an illustrative meta-analysis. Specifically, during the extraction of the data from the studies, in order to eliminate bias due to specific algorithms used in the original studies, it was recommended to obtain the feature-level extraction output (FLEO) files, such as CEL and GPR files, and converting them to gene expression data matrices (GEDMs) in a consistent manner. In addition when annotating the individual datasets, one can map probe-level identifiers such as I.M.A.G.E Clone ID, Affymetrix ID, or GenBank accession numbers to a gene-level identifier such as UniGene, RefSeq, or Entrez Gene ID (Ramasamy, Mondry et al. 2008).

2 OBJECTIVES AND RATIONALE

Cancer/testis (CT) genes are expressed at different frequencies in a wide range of cancer types. Previously, it was shown that coordinate expression of CT-X genes in non-small cell lung cancer (NSCLC) associates with poor prognosis (Gure, Chua et al. 2005). The mechanisms responsible for the reactivation of CT-X genes during tumorigenesis are of great interest because of their prognostic and therapeutic value. In this study, we aimed to develop two approaches by which the mechanisms underlying the regulation of CT-X gene expression in cancer could be identified. Current data suggests that CT-X gene expression is regulated by promoter specific methylation but the mechanisms of this regulation are not known. Our rationale is based on the hypothesis that coordinately expressed CT-X genes might be regulated by common mechanisms.

Our first approach was to generate a model by which the expression of an individual CT-X gene could be easily monitored. Such a model could then be used to test the effect of various manipulations on CT-X gene expression. For this approach we chose SSX4 since SSX4 promoter has been characterized in detail in this laboratory (Gure AO, unpublished data). We aimed to generate an SSX4 knock-in (KI) lung cancer cell line (SK-LC-17) with a GFP reporter gene expressed from SSX4 promoter. Such a cell would be visible by cytometry and manipulation of the genes' regulation would be easy to observe. We chose to generate this cell line using the SK-LC-17 lung cancer cell line since it is known to express SSX4 readily and its SSX4 promoter is known to be completely demethylated (Gure AO, unpublished data). A subsequent goal would be to transfect the knock-in cell line with a cDNA library prepared from a CT-X negative cell line and select the clones with repressed GFP expression by flow cytometry. We thus, would expect to obtain the clones with methylated SSX4, since SSX4 expression is repressed by promoter-specific methylation, and isolate the cDNA causing this modification. In this way, transcriptional repressors of CT-X gene expression that function either in a direct manner as epigenetic controllers or indirectly as effectors upstream to epigenetic mechanisms can be identified.

Our second approach was to utilize a meta-analysis of publicly available microarray data towards identifying genes whose expression (or the lack thereof) correlate with CT-X gene expression. We thus, wanted to simultaneously analyze datasets from tumor tissues

originating from various tissues. Since most public datasets are from a given tissue type, and we hypothesized that if all samples within a given dataset could be classified into CT-X positive and negative subgroups, that the comparison of these subgroups as a meta-analysis would reveal the CT-X-specific mechanisms instead of tissue specific differences. However, CT-X expression control might have tissue specific components as well, so we chose to include analyses that were limited to a given dataset, namely class prediction analyses.

3 MATERIALS AND METHODS

3.1 MATERIALS

3.1.1 Reagents

All laboratory chemicals were analytical grade from Carlo Erba (Milano, Italy), Merck (Schucdarf, Germany), Riedel-de Haën (Germany) and AppliChem (Darmstadt, Germany). Agar and yeast extract were supplied from BD Biosciences (USA). Tryptone was from Conda Laboratories (Spain). TRI Reagent (for RNA isolation) was purchased from Molecular Research Center, Inc (USA). Phenol:Chloroform:Isoamyl Alcohol 25:24:1 was purchased from Sigma-Aldrich (Belgium).

3.1.2 Kits

Qiagen Plasmid Mini-prep kit (for small-scale plasmid DNA isolation), Maxi-prep kit (for large-scale plasmid DNA isolation) and QiaQuick Gel Extraction kit (for recovery and extraction of DNA from agarose gel) were from Qiagen (Maryland, USA). PureLink Genomic DNA Mini kit (for small scale genomic DNA isolation) was obtained from Invitrogen (Germany).

3.1.3 Bacterial strains

The bacterial strain used in this work was: *E. coli* DH5α.

3.1.4 Enzymes

Restriction endonucleases were purchased from New England Biolabs (UK). T4 DNA Ligase was purchased from Fermentas (Germany).

3.1.5 PCR, Real-time PCR and cDNA synthesis reagents

For cDNA sythesis, DyNAmo cDNA Synthesis Kit was used (Finnzymes, Finland). DyNAzyme II Hot Start DNA Polymerase for the amplification of fragments up to 1 kb was purchased from Finnzymes (Finland). Elongase Enzyme Mix for the amplification of fragments up to 30 kb and the greater amplification of smaller fragments was purchased from Invitrogen (Germany). SYBR Green Master Mix and TaqMan Gene Expression Master Mix used in real-time PCR were obtained from Finnzymes (Finland) and Applied Biosystems (USA) respectively.

3.1.6 DNA Molecular Size Markers

GeneRuler DNA Ladder Mix (0.1-10 kbp) and GeneRuler 50 bp DNA ladder (50-1000 bp) were purchased from Fermentas (Germany).

3.1.7 Primers

The primers used in conventional PCR and quantitative real-time PCR analyses were synthesized by Iontek (Istanbul, Turkey). Pre-designed and synthesized FAM dye-labeled TaqMan MGB probes and unlabeled PCR primers for SSX4, NY-ESO-1, MAGEA3, GAGE and GAPDH used in real-time PCR were purchased from Applied Biosystems (USA).

3.1.8 Electrophoresis, photography and spectrophotometer

Horizontal electrophoresis apparatuses were from E-C Apparatus Corporation (USA). The power supply Power-PAC300 and Power-PAC200 was from BioRad Laboratories (USA). The Molecular Analyst software used in agarose gel profile visualizing was from Vilber Lourmat (France). Beckman Spectrophotometer Du640 was purchased from Beckman Instruments Inc. (USA) and Nanodrop ND-1000 Full-spectrum UV/Vis Spectrophotometer was purchased from Thermo Fisher Scientific (USA).

3.1.9 Tissue culture reagents

Dulbecco's modified Eagle's medium (DMEM), RPMI-1640 medium and Fetal Bovine Serum (FBS) were obtained from Biochrom (Germany). L-glutamine, non-essential amino acid, and penicillin/streptomycin mixture were from PAA (Austria). Trypsin was purchased from Sigma-Aldrich (USA). Hygromycin was purchased from BD Biosciences (USA).

3.1.10 Transfection reagents

Lipofectamine 2000 transfection reagent was obtained from Invitrogen (Germany). Opti-MEM I Reduced Serum Medium that was used to dilute Lipofectamine 2000 transfection reagent and DNA before combining them was obtained from Gibco (Invitrogen).

3.2 SOLUTIONS AND MEDIA

3.2.1 General solutions

50X Tris-acetic acid-EDTA (TAE): 121g Tris-base was first dissolved in 350 ml ddH2O. 18.6g EDTA and 28.6 ml glacial acetic acid were then added and the solution was brought to

500ml with ddH_20. Working solution (1X TAE) was prepared by diluting of 50X TAE to 1X with ddH_20.

Ethidium bromide: 10 mg/ml in water (stock solution), 30 ng/ml (working solution)

6X Gel loading dye solution: 10mM Tris-HCl (pH 7.6), 60mM EDTA (0.5M, pH8.0), 60% glycerol, 0.03% bromophenol blue or 0.03% xylene cyanol were mixed. Two different gel loading dyes were prepared; one with only bromophenol blue, the other with only xylene cyanol. Bromophenol blue co-migrates around ~300 bp DNA and it was used when analyzing larger DNA fragments. Xylene cyanol co-migrates around ~4000 bp and it was used when analyzing smaller DNA fragments.

3M Potassium-Acetate (KAc), pH 5.2: 29.4 g KAc was added to ~50 ml ddH2O. pH was adjusted to 5.2 by the addition of glacial acetic acid (30-35 ml). The solution was brought to 100 ml with ddH_20.

3.2.2 Microbiological media, solutions and media

Luria-Bertani medium (LB): Per liter; 10 g bacto-tryptone, 5 g bacto- yeast extract, and 10 g NaCl were dissolved in ddH_20 and autoclaved. LB agar plates contained additionally 15 g/L bacto agar.

Glycerol stock solution: Bacterial cultures were strored at -80°C in LB with a final concentration of 25% glycerol.

Carbenicillin: Stock solution was prepared as a 100 mg/ml solution byin ddH_20. It was sterilized by filtration and stored at -20°C. Working solution was100 μg/ml.

Transformation Buffer (TB): For TB, solutions of 0.5 M PIPES, 0.5 M $CaCl_2$, 1 M KCl, 1 M $MnCl_2$ and 1 M KOH were first prepared. TB contained 10 mM PIPES, 15 mM $CaCl_2$, 250 mM KCl and 55 mM $MnCl_2$. All components except $MnCl_2$ were added and pH was adjusted to 6.7 with 1 M KOH. The solution is was filter sterilized near fire and stored at 4^oC.

3.2.3 Cell culture solutions

Growth medium: 10% FBS, 1% penicillin/streptomycin, 1% L-glutamine (if the medium is

L-glutamine free), 1% non-essential amino acids were added to DMEM/RM1640 medium. Growth medium was stored at 4°C.

Freezing medium: 10% DMSO and 10% FBS were added to the growth medium to obtain a freezing medium with 10% DMSO and 20% FBS. The freezing medium was freshly prepared and kept on ice before use.

10X Phosphate-buffered saline (PBS): 40 g NaCl, 1 g KCl, 8.9 g Na_2HPO_4 and 1.2 g KH_2PO_4 were dissolved in ddH_20. The solution was brought to 500 ml with ddH_20. Working solution (1X PBS) was prepared by diluting 10X PBS to 1X in water (pH should be between 7.2 and 7.4) and autoclaved.

Hygromycin: Stock solution was 50 mg/ml solution in PBS and it was stored at 4°C. 50 μg/ml was used for stable cell line selection and maintenance.

3.3 METHODS

3.3.1 General Methods

3.3.1.1 Preparation of transformation-competent *E.coli* DH5α cells

50 ml LB was inoculated with a single colony from a freshly grown plate of *E. coli DH5*α strain and incubated overnight at 37°C, shaking at 200 rpm. With this fresh 50 ml of overnight culture, 200 ml LB in 1L flask was inoculated until an optical density 0.2 at 600 nm (OD_{600}) was reached. The culture was incubated at 20°C, shaking at 200 rpm. The culture should be in log phase growth that is the proper stage for making competent cells. Once the culture had an OD_{600} value between 0.5 and 0.6 (it will take 5-7 hours), it was poured into sterile tubes and incubated on ice for 10 minutes. The cells must be kept on ice throughout the rest of the procedure. The cells were harvested by centrifugation at 2500xg for 10 minutes at 4°C and the supernatant was discarded. The pellet was resuspended in 64 ml Transformation Buffer (TB) by gently pipetting up and down. Resuspended cells were left on ice for 10 minutes and centrifuged at 2500xg for 10 minutes. After the supernatant was removed away, the pellet was gently resuspended in 16 ml of ice-cold TB containing 7% DMSO and incubated on ice for 10 minutes. The competent bacteral cells were quickly aliquoted into cold 1.5 ml microcentrifuge

tubes (150 µl per tube) and snap-frozen with liquid nitrogen. Frozen competent cells were stored at -80°C.

3.3.1.2 *E.coli* DH5α transformation

150 µl competent *E. coli DH5α* cells were thawed on ice and polypropylene screw capped tubes were placed on ice as well. After thawing, 10 ng plasmid DNA or ligation mixture was added to competent cells. DNA-bacteria mixture was incubated on ice for 30 minutes after mixing gently without vortex. Then, the mixture was transferred to 42°C water bath for 45 seconds (heat-shock) and immediately placed on ice for 2-3 minutes. 850 µl of pre-warmed LB was added onto cells, which were then cultured for 1 hour at 37°C with shaking at 200 rpm. After 1-hour incubation, samples were centrifuged at 13000 rpm for 20 seconds, and the supernatant was removed away but leaving approximately 100 µl LB. The pellet was resuspended in the remaining LB. The bacteria cells were plated out on LB-agars with selection agents such as ampicillin. The plates were incubated overnight at 37°C without shaking to allow the growth of the transformants.

3.3.1.3 Long term storage of bacterial strains

To keep bacterial cells including plasmid in it or as empty for future experiments and to have a stock of strain in a laboratory is necessary. The most frequently used method is "Glycerol-Stock" method. A single colony picked from either an agar plate or a loop-full of bacterial stock was inoculated into 5 ml LB (with a selective agent if necessary) in 15 ml screw capped tubes. Tubes were incubated overnight at 37 °C, shaking at 200 rpm. For glycerol stock, 500 µl of bacteria cell culture was added into 500 µl of sterile 50% (v/v) glycerol in LB. This mix was snap-frozen with liquid nitrogen and stored at -80 °C.

3.3.1.4 Plasmid DNA purification

3.3.1.4.1 Small-scale plasmid DNA purification (mini-prep)

The plasmid that is prepared in small-scale was used for sequencing or cloning procedures. 1.5 ml of bacterial cell culture grown overnight was used for isolation of plasmid DNA with Qiagen plasmid mini-prep kit according to the manufacturer's instructions. Plasmid DNA was eluted in a total volume of 50 µl ddH_2O.The quality of miniprep was checked by loading 250 ng of final yields on agarose gel and visualizing under U.V.

3.3.1.4.2 Large-scale plasmid DNA purification (midi-prep)

The plasmid that is prepared in large-scale was used for sequencing or mammalian cell transfection procedures. 100 ml of bacterial cell culture grown overnight was used for isolation of plasmid DNA with Qiagen plasmid midi-prep kit according to the manufacturer's instructions. Plasmid DNA was eluted in a total volume of 100 µl ddH_2O. The quality of midiprep was checked by loading 250 ng of final yields on agarose gel and visualizing under U.V.

3.3.1.5 Phenol/chloroform DNA extraction and ethanol precipitation

This method was preferred to remove proteins from a DNA sample (plasmid DNA or genomic DNA). The procedure given here was for isolation of plasmid DNA from large-scale restriction enzyme reactions. If the sample volume is small (<250 µl), the sample was diluted to 350-500 µl. Before use, phenol:chloroform:isoamylalcohol 25:24:1 (v/v) was shaked vigorously. An equal volume of phenol: chloroform: isoamylalcohol was added to the sample. After the sample was mixed by vortex for 1 minute, it was centrifuged at 13000 rpm for 10 minutes at room temperature (RT) in a bench-top centrifuge to separate the aqueous phase, which contains DNA from the organic phase that contains proteins. The aqueous phase was removed to a new tube. To remove traces of phenol, an equal volume of chloroform was added to the tube and the sample was mixed by vortex for 1 minute before being centrifuged at 13000 rpm for 2 minutes at RT. In order to precipitate DNA, 0.1 volume of KAc pH 5.2 was first added and then 100% ethanol (EtOH) was added until 70% EtOH was reached in DNA solution. The DNA solution was incubated at -80^oC for 2 hours or overnight at -20^oC. After incubation, the precipitated DNA was recovered by centrifugation at 13000 rpm for 10 minutes at 4^oC. The supernatant was discarded and the pellet was washed with 70% EtOH. EtOH was removed away after centrifugation at 13000 rpm for 5 minutes at 4^oC. The DNA pellet was air-dried for 15 minutes and resuspended in sterile ddH_2O and stored at -20^oC.

3.3.1.6 Genomic DNA purification from cultured cells

Cultured cells that were grown in 10 mm tissue culture dishes to 80-90% confluency were washed with 1X PBS, trypsinized, and centrifuged at 1000 rpm for 5 minutes at 4^oC. The pellet was resuspended in 5 ml 1X PBS and centrifuged at 1000 rpm for 5 minutes at 4^oC. Then, it was stored at -80^oC until used for genomic DNA isolation. Genomic DNA was isolated by use of "PureLink genomic DNA mini kit" following manufacturer's instructions. Genomic DNA was eluted in a total volume of 100 µl. The quality of genomic DNA was

checked by loading 250 ng of final yields on agarose gel and visualizing under U.V.

3.3.1.7 Total RNA Extraction from cultured cells

Exponentially growing monolayer cultures were washed with PBS, trypsinized, and centrifuged at 1000 rpm for 5 minutes at 4^{o}C. The pellet was resuspended in 5 ml 1X PBS and centrifuged at 1000 rpm for 5 minutes at 4^{o}C. Then, it was stored at -80^{o}C until used for RNA isolation The total RNA isolation from cell line pellets was performed using TRI reagent (Trizol). Per 5-10 x 10^{6} cells, 1 ml Trizol was added to the pellet. The pellet was dissolved in Trizol by pipetting up and down until a homogenous solution was obtained. The mixture was tranferred to a 1.5 ml eppendorf tube and incubated for 5 minutes at RT. Per 1 ml Trizol, 200 µl chloroform was added to the tube. After shaking vigorously for 15 seconds, the mixture was centrifuged at 13.000 rpm for 15 minutes at 4^{o}C. The aqueous phase that contains RNA was transferred to a new tube and 500 µl isopropanol was added to it. After mixing by inverting gently, the mixture was incubated for 10 minutes at RT and then centrifuged at 13.000 rpm for 10 minutes at 4^{o}C. The supernatant was discarded and the RNA pellet was washed with 1 ml of %75 EtOH (the pellet should move in this step). After centrifugation at 13.000 rpm for 5 minutes at 4^{o}C, the supernatant was removed away and the pellet was air-dried for 3-5 minutes. The RNA pellet was dissolved in 250 RNase-free ddH_2O and RNA was incubated for 15 minutes at 55 ^{o}C to increase its solubility. The concentration of the isolated RNA and OD 260/280 ratio (between 1.8-2.1) were measured with the NanoDrop ND-1000 spectrophotometer. Isolated RNAs were snap-frozen with liquid nitrogen and stored at -80^{o}C.

3.3.1.8 Quantification and qualification of nucleic acids

Concentration and purity of the double stranded (ds) nucleic acids (plasmid and genomic DNA) and total RNA were determined by using the RNA and ds DNA options on Nanodrop ND-1000 Full-spectrum UV/Vis Spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA). OD 260/280 ratio indicated the purity of nucleic acids and should be generally between 1.8-2.1 for them.

3.3.1.9 Restriction enzyme digestion of DNA

For cloning, 5-10 µg plasmid DNA was digested in a 50 µl reaction volume that was incubated overnight. Restriction enzyme reaction mixtures of PCR products were incubated for 4-6 hours. Diagnostic restriction enzyme digestions in order to verify the correctly ligated construct were carried out in a 20 µl reaction volume that was incubated for 2-4 hours and

0.5-1 µg plasmid DNA was used. Reactions were carried out with the appropriate reaction buffer and conditions according to the manufacturer's recommendations. Digestion of DNA with two different restriction enzymes was also performed in the appropriate common reaction buffer and conditions recommended by the manufacturer.

3.3.1.10 DNA extraction from agarose gel

DNA extraction from agarose gel was performed using QiaQuick Gel Extraction kit (Qiagen) according to the manufacturer's instructions. Recovered DNA was used for either DNA ligation or DNA sequencing in this study.

3.3.1.11 DNA ligation

T4 DNA ligase was used for DNA insert ligation into vector DNA. Ligation was performed in 10 µl or 20 µl reaction volumes depending on the concentration of vector (backbone) and insert DNA. Usually, 100 ng of vector DNA was used. According to the molar ratio of vector:insert DNA, the amount of insert DNA was calculated using the formula below:

$$\frac{\text{ng of vector x kb size of insert}}{\text{kb size of vector}} \text{ x molar ratio of } \frac{\text{insert}}{\text{vector}} = \text{ng of insert DNA}$$

Reactions were carried out with 10X or 2X T4 DNA ligase buffer and were incubated at RT or 16^{o}C overnight.

3.3.1.12 Agarose gel electrophoresis of DNA

DNA fragments were fractionated by horizontal electrophoresis by using
standard buffers and solutions. DNA fragments less than 1 kb were generally separated on 1.5% or 2.0 % agarose gel, those greater than 1 kb (up to 11 kb) were separated on 1 % agarose gels. Agarose gels were prepared by completely dissolving agarose in 1x TAE electrophoresis buffer to required percentage in microwave and ethidium bromide was added to final concentration of 30 µg/ml. The DNA samples were mixed with 6X DNA loading buffer and loaded onto gels. The gel was run in 1x TAE at different voltage and time depending on the size of the fragments at room temperature. Nucleic acids were visualized under U.V. and GeneRuler DNA molecular size markers (Fermentas) were used to estimate the fragment sizes. DNA Ladder Mix (0.1-10 kbp) was loaded for products sizes of over 1kb and 50 bp DNA ladder (50-1000 bp) for product sizes of below 1kb.

3.3.2 Computational Analyses

The sequences of genes that were chosen for validation of class-prediction analysis in GSE4824 lung cancer cell lines dataset were obtained from NCBI (National Center for Biotechnology Information). The exon-intron information of these genes were derived using Ensembl Genome Browser (http://www.ensembl.org). Restriction endonuclease maps of the plasmid DNAs were analyzed using the online NEBcutter2 (http://tools.neb.com/NEBcutter2/) tool. Primers were designed using Primer3 online primer design tool tool (http://frodo.wi.mit.edu/) (Rozen and Skaletsky 2000). The results of the DNA sequencing of engineered constructs were visualized using Chromas-v1.45 available for download at http://www.technelysium.com.au/chromas14x.html. The alignments of nucleic acid were performed by using the NCBI BLAST (nucleotide blast and blast2Sequences) algorithm available at the web page (http://www.ncbi.nlm.nih.gov/BLAST) and ClustalW algorithm provided by EMBL-EBI at http://www.ebi.ac.uk/Tools/clustalw2/index.html (Thompson, Gibson et al. 1997).

3.3.3 Vector construction

SSX4 knock-in (KI) vector was generated and sequence-verified by Dr. Ali O. Güre in Cornell University, USA. Additional restriction enzyme digestions of the KI vector were performed in Bilkent University. In addition, enhanced green flurescent protein (EGFP) of the KI vector was sequenced with the primers given **Table 3.1**. The steps in the generation of SSX4 KI vector were as follows:

1- EGFP was amplified using XbaI-HindIII containing PCR primers from pHygEGFP plasmid and cloned into the XbaI-HindIII digested SSX4 A3-B pGL3 luciferase reporter vector including containing the corresponding SSX4 5' genomic sequence (see **Figure 4.1** for A3 and B primer locations).

2- 3' SSX4 homology sequence was amplified with BamHI-SalI containing PCR primers and was cloned downstream of EGFP.

3- Since BamHI and BglII generated compatible cohesive ends, hygromycin resistance gene (PGK/HYG) with its own 5'regulatory region and PolyA signal obtained by BglII digestion was cloned into BamHI digested construct between GFP and SSX4 3' homology sequence.

4- ß-Actin promoter driven diphteria toxin (β-actin/DTA) was cloned into NotI-KpnI digested construct, upstream of SSX4 5' homology sequence.

The final SSX4 KI vector (12 kb in length) in the linear form is shown in **Figure 3.1**.

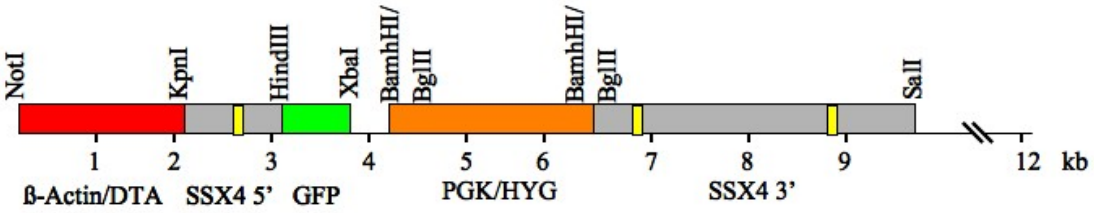


Figure 3.1: SSX4 KI vector. The components of SSX4 KI construct were indicated below the boxes. Yellow boxes showed first, forth and fifth exons of SSX4.

Table 3.1: The sequencing primers that were used to sequence EGFP

Primer ID	5'-3' sequence
P16 (Fwd)	GTCCTGAGGCTGGAAAGACTCA
M4 (Rev)	ATTCATCGATCGCAGATCCTTATCG
P50 (Fwd)	CAGGCTGTTTCTCTCGCAGGTG
M45 (Rev)	TAATAGCGAAGAGGCCCGCAC

*Fwd: Forward Rev:Reverse

The construct lacking ß-Actin promoter driven diphteria toxin (DTA) was used as a control to observe if GFP was expressed from the SSX4 promoter, following transient transfection into SK-LC-17 lung cancer cells. This plasmid is referred to as "Step 6". Another control plasmid was generated in order to test correct amplification of primers used in nested PCR by cloning EGFP obtained by XbaI-HindIII digestion from SSX4 KI plasmid into XbaI-HindIII digested SSX4 A1-B pGL3 luciferase reporter vector. The reaction setup for this cloning experiment is given in **Table 3.2**. First, SSX4 KI and SSX4 A1-B pGL3 vectors were double digested to obtain a 0.75 kb insert (EGFP) and a 5.6 kb vector DNA, respectively. The reaction was incubated at 37^{o}C overnight and digestion products were purified on an agarose gel. After DNA extraction from the gel, ligation reaction was set using 3:1 molar ratio of insert/vector and incubated at 16 ^{o}C overnight. As a control, only vector DNA was used for ligation reaction. The ligated DNA was transformed into competent *E.coli* DH5α cells

Table 3.2: The reaction setup for cloning EGFP into SSX4 A1-B pGL3 luciferase vector

XbaI&HindIII digestion	Ligation of insert into vector
4 µg plasmid DNA	100 ng vector
1 µl XbaI	40.2 ng insert
1 µl HindIII	1 µl 10X LigationBuffer
5 µl NEB Buffer 2	1 µl T4 DNA Ligase
5 µl 10X BSA	completed to 10 µl with ddH2O.
completed to 50 µl with ddH2O.	

3.3.4 Testing for Correctly Integrated Vector by Nested PCR

Two pairs of primers were used in two successive runs of PCR in a nested PCR reaction; the second pair amplifies a secondary target within the initial amplification product. Nested PCR was performed to screen SSX4 KI clones. Forward primers were 640 to 500 bp upstream from SSX4 5' homology sequences and thus corresponded to genomic DNA of the cell, whereas reverse primers aligned to the EGFP sequence that would be expected to be part of the construct. The sequences of primers that were used in the first (1^{st}) and the second (2^{nd}) runs of nested PCR are listed in **Table 3.3**.

Table 3.3: Primers used in nested PCR to test for correct vector insertion*

Primer ID	5'-3' sequence	Tm (^oC)	Amplicon length (bp)
P4 (Fwd)	AGAATGAGATGGGAGGATTGACCAAG	63^oC	1848
M4 (Rev)	GTGCAGATGAACTTCAGGGTCAGC		
A2.1 (Fwd)	ATATTTCTCGAGCACTATTCACAATAGCAAAGAC	63^oC	1580
M26 (Rev)	GGTGAGACTGCTCCCAGTGC		
P4 (Fwd)	AGAATGAGATGGGAGGATTGACCAAG	65^oC	1914
M7 (Rev)	AAGCACTGCACGCCGTAGGTC		
A2.1 (Fwd)	ATATTTCTCGAGCACTATTCACAATAGCAAAGAC	65^oC	1702
M4 (Rev)	GTGCAGATGAACTTCAGGGTCAGC		

*P4&M4 and A2.1&M26; P4&M7 and P4&M7 and A2.1&M4 primer pairs were used in two successive runs of nested PCR.

The reaction setup for nested PCR and PCR conditions optimized for primer pairs were given in **Table 3.4** and **Table 3.5**, respectively.

Table 3.4: The reaction setup for nested PCR

Reagents	1^{st} round of PCR reaction	2^{nd} round of PCR reaction
5X Buffer B	5 µl	5 µl
10 mM dNTP mix	0.5 µl	0.5 µl
10 µM forward primer	0.5 µl	0.5 µl
10 µM reverse primer	0.5 µl	0.5 µl
Elongase enzyme mix	0.5 µl	0.5 µl
DNA	100 ng genomic DNA /12 pg plasmid DNA	2 µl of 5:125 diluted 1^{st} PCR product
ddH_2O	completed to 25 µl	completed to 25 µl

Table 3.5: PCR conditions for primer pairs used in nested PCR

Initial denaturation	94°C 30 seconds
40 cycles	94°C 30 seconds
	63°C-65°C 30 seconds
	68°C 30 seconds
Final extension	68°C 10 minutes

3.3.5 Tissue culture

3.3.5.1 Cell lines

Lung cancer cell line SK-LC-17 was cultured in RPMI-1640. 15 HCC derived cell lines (Huh7, FOCUS, Mahlavu, Hep40, Hep3B, Hep3B-TR HepG2, PLC/PRF/5, SK-Hep1, Snu182, Snu387, Snu398, Snu423, Snu449 and Snu475) were used in this study. Snu182, Snu387, Snu398, Snu423, Snu449 and Snu475 and Hep40 were cultured in RPMI-1640 and the rest were cultured in DMEM. 14 colon cancer cell lines SW620, SW480, SW837, SW48, Hct8, Hct15, Hct116, Colo205, KM12, LoVo, Co115, HT29, WiDr, LS513 were cultured in RPMI-1640. 7 breast cancer cell lines MCF-7, MDA-MB-231, MDA-MB-157, CAMA1, SK-BR-3, BT20 and BT-474 were cultivated in DMEM. Melanoma cell lines SK-MEL-28 and SK-MEL-30 were cultured in RPMI-1640.

3.3.5.2 Growth conditions of cell lines

Dulbecco's modified Eagle's medium (DMEM) or RPMI 1640 supplemented with 10% FCS, 1% penicillin/streptomycin mixture, 1% NEAA, 1% L-glutaime (if the medium is L-glutamine free) was used to culture the cell lines used in this study. The cells were incubated in a humidified incubator at 37°C supplied with 5% CO_2. SSX4 KI transfected stable clones were cultured in parental cell line's culture medium + 50 μg/ml hygromycin. The cells were passaged before reaching confluence. The growth medium was aspirated and the cells were washed once with autoclaved sterile 1X PBS. Trypsin was added to the flask to remove the monolayer cells from the surface (0.3 ml per 25-cm2 flask, 1 ml per 25-cm2 flask) and incubated at 37°C for 2-10 minutes depending on the cell line. After adding fresh medium onto cells to neutralize trypsin, cell suspension was transferred to a 15 ml sterile falcon tube and centrifuged at 900 rpm for 3 minutes. Supernatant was discarded and the pellet was resuspended with growth medium to be transferred to either fresh petri dishes or fresh flasks at required dilutions (from 1:2 to 1:10). All media and solutions used for culture were kept at 4°C (except stock solutions) and warmed to 37°C before use.

3.3.5.3 Thawing cyropreserved cell lines

One vial of the frozen cell line from the liquid nitrogen tank was taken and immediately put into ice. The vial was immersed in 37^{o}C water bath until the external part of the cell solution was thawed (takes approximately 1-2 minutes). The cell solution was quickly poured into a 15 ml sterile falcon tube containing 5 ml warm culture medium. The cells were centrifuged at 900 rpm for 3 minutes at RT. Supernatant was discarded and the pellet was resuspended in 5 ml culture medium to be plated into 25-cm^2 flask. After overnight incubation in a humidified incubator at 37^{o}C supplied with 5% CO2, culture mediums were refreshed.

3.3.5.4 Cyropreservation of cell lines

Exponentially growing cells were harvested by trypsinization and neutralized with growth medium. The cells were precipitated at 900 rpm for 3 minutes. The pellet was suspended in a cold freezing medium containing 10%DMSO and 20%FCS. 1ml of cell suspension was placed into 1ml screw capped cryotubes. The tubes were first frozen at -20^{o}C for 1hour and then left at -80°C overnight. The next day, the tubes were transferred into the liquid nitrogen storage tank.

3.3.5.5 Transfection of SK-LC-17 lung cancer cells

Transfections were performed using Lipofectamine 2000 transfection according to the instructions of the supplier (Invitrogen). For transient transfection with Step 6 construct or the pHygEGFP plasmid, 6-well plates were used. Exponentially growing SK-LC-17 cells were plated in 6 well-plates at a concentration of 6.5 x 10^5 cells/well, a day before transfection. The cells were incubated overnight and reached 90-95% confluency. Transfection was performed in cell culture medium with a 1:2, DNA (μg) to Lipofectamine 2000 (μl) ratio. Briefly for one well, 4 μg of plasmid and 8 μl of Lipofectamine 2000 were diluted in 100 μl OptiMem I Reduced-Serum medium or cell culture medium lacking FBS. The solutions were mixed gently and the one containing Lipofectamine 2000 was incubated for 5 minutes at RT. The diluted DNA was combined with diluted Lipofectamine 2000 (total volume > 200 μl) and after a brief vortex, the mixture was left to RT for the formation of transfection complex for 30-45 minutes. During incubation, the growth medium of cells was changed with OptiMem I Reduced-Serum medium lacking antibiotics. Then, transfection complex was pipetted dropwise onto the cells and the plate was mixed gently by rocking it back and forth. The medium was replaced with fresh growth medium after 5 hours. After 48-72 hours incubation, cells were harvested and used for subsequent experiments.

Stable transfection of SK-LC-17 cells with SSX4 KI construct was performed in 150 mm petri dishes with 7.5 x 10^6 cells plated one day before transfection. 56 µg of NotI linearized SSX4 KI plasmid and 112 µl of Lipofectamine 2000 were diluted in 3.5 ml OptiMem I Reduced-Serum medium or cell culture medium lacking FBS. 48 hours following transfection, cells were trypsinized and re-plated at lower density into 75 cm^2 flasks. Cells were cultivated in the presence of 50 µg/ml hygromycin for 1-week. After 1-week, cells were harvested and diluted to 0.5 cell/100 µl of cell suspension before they were seeded in 96-well plates in the presence of hygromycin (100 µl per well).

3.3.5.6 Flow cytometry analysis

Flow cytometry analysis was performed in a BD FACSCalibur system. Transiently transfected SK-LC-17 cells with pHygEGFP plasmid and Step6 construct, as well as stable SSX4 KI clones and parental SK-LC-17 cells were harvested by trypsinization and resuspended in 1X PBS. Resuspended cells were kept on ice and in the dark until analysis in flow cytometry. Upon adjustment of instrument settings, 20.000 cells were counted for the analysis.

3.3.6 cDNA synthesis

cDNA was synthesized by using DyNAmo^{TM} cDNA Synthesis Kit according to the manufacturer's instructions. Briefly; for 1X reaction 1µg of total RNA, 1µl of random hexamers and required amount of RNase-free ddH_2O were mixed in a total 8µl volume. Prepared mixes were incubated at 65°C for 5 minutes and chilled on ice. Consequently, 10µl of 2X RT reaction buffer and 2µl of M-MuLV RNase H^+ reverse transcriptase (including RNase inhibitor) were added to complete the reaction volume to 20µl. Then, the reaction mixes were incubated in a ABS thermal cycler 9700 programmed as follows: 25^oC for 10 minutes, 37°C for 60 minutes, 85^oC for 5 minutes and 4^oC hold. Each cDNA sample was diluted at a ratio of 1:2 with ddH_2O and stored at -20^oC to be used as a PCR template for further experiments.

3.3.7 Primer design for expression analysis by real-time quantitative RT-PCR

Primer pairs used in the validation of microarray analysis results of lung cancer cell lines were designed using Primer3. They targeted exon-exon junctions or different exons in order to prevent amplification of possible contaminating genomic DNA (Primer 3 reference). The primer pair was designed such that it would only be able to produce a longer amplicon from

genomic DNA or would not be able to amplify the corresponding genomic DNA region in a given PCR condition (critical parameter was extension time). The primer pair used for amplification of the housekeeping gene, GAPDH was described before (Sayan, Sayan et al. 2001). Primers used for expression analysis have been designed strictly considering these criteria, and are listed in **Table 3.6**.

Table 3.6:Sequences of the primers used for validation analysis*

Gene symbol	5'-3' Sequence (upper: forward; lower:reverse)	Tm (^{o}C)	Amplicon length (bp)
SCD	ACACTTGGGAGCCCTGTATG	60^{o}C	152
	GCAGCCGAGCTTTGTAAGAG		
RAD21	AACCAATGCCAACCATGACT	60^{o}C	152
	CCTCTTCCTCTTGGCTTTTG		
HSPH1	CACAGCCCCAGGTACAAACT	60^{o}C	144
	TGGGCTTTTTAGCTTCTGGA		
HSP90B1	CTGTCTGGGACTGGGAACTT	60^{o}C	197
	TGGAGCAGATGTGGGTACAA		
LAPTM4B	ATACGGCAACTGCCTCCTAA	60^{o}C	146
	CGGTAGCAGTTCCAAACACA		
NFYC	CGCCATGGCAATTACAAAAT	60^{o}C	119
	GCTCGGCAGGAGTTACAGAC		
SSRP1	GAAGGAGGAAGGCAAGATCC	60^{o}C	178
	CTTCTCATCCCGGTCACTGT		
TWIST	CCGGAGACCTAGATGTCATTG	60^{o}C	148
	CACGCCCTGTTTCTTTGAAT		
LIMK2	TCAGGTTTGCCAAAGGAATC	60^{o}C	91
	ATGAGGCAGTTGTGCGAGTT		
GAPDH	GGCTGAGAACGGGAAGCTTGTCAT	60^{o}C	151
	CAGCCTTCTCCATGGTGGTGAAGA		

*GAPDH was used for normalization of qPCR data.

3.3.8 Real-time quantitative PCR (qPCR)

3.3.8.1 Taqman probe-based qPCR of lung, colon, breast and hepatocellular carcinoma (HCC) cell lines

Taqman probe-based qPCR was performed using Taqman Gene Expression Assays (Applied Biosystems) for four CT-X genes; NY-ESO-1, MAGEA3, SSX4 and GAGE. Each assay contained a pre-designed FAM-dye labeled Taqman MGB probe and two unlabeled oligonucleotides. Taqman MGB probes were sequence-specific dual-labeled probes with a fluorophore (e.g. FAM) at 5' end, a non-fluorescent quencher (NFQ) and minor groove binder (MGB) at 3'end. Sequences of probes used in qPCR are given in **Table 3.7**. Each probe had a unique assay ID except GAPDH since it was purchased as an endogenous control that can be

used for singleplex PCR reactions. "_m" in assay ID indicates an assay whose probe spannes an exon junction and will not detect genomic DNA. The probes for CTAG1A; CTAG1B, multiple GAGE genes and MAGEA3 were designed to be complementary to the 1st-2nd exon junction whereas SSX4 and SSX4B probes were designed to span the 3rd-4th exon junction. The probe for GAPDH was located in 3rd exon. GAPDH was used for normalization of qPCR data. Neither probe sequence nor primer sequences were provided by Applied Biosystems. Only context sequence was provided as the nucleotide sequence surrounding the probe. Multiple GAGE genes represented GAGE7, GAGE12I, GAGE2A, GAGE8, GAGE4, GAGE5, GAGE1, GAGE13,
GAGE12H, LOC729408, GAGE12D, GAGE12J, GAGE6, GAGE12E, GAGE12C, GAGE12G.

Table 3.7: The probes used in qPCR.

Assay ID /Product ID	Gene symbol	Context sequence	Tm (^{o}C)	Amplicon length (bp)
Hs00265824_m1	CTAG1B, CTAG1A	GCTTGAGTTCTACCTCGCCATGCCT	60^{o}C	103
Hs00275620_m1	Multiple GAGE genes	ACTGAGATTCATCTGTGTGAAATAT	60^{o}C	69
Hs00366532_m1	MAGEA3	GGTGAGGAGGCAAGGTTCTGAGGGG	60^{o}C	145
Hs02341531_m1	SSX4, SSX4B	AACCACAGGAATCAGGTTGAACGTC	60^{o}C	101
4333764F	GAPDH	-	60^{o}C	122

qPCR was performed on a Stratagene MX3005P Real-time PCR System (USA). The PCR reaction was set according to the manufacturer's recommendations. Briefly for 1X reaction; 10 µl of 2X Taqman Gene Expression Master Mix, 1 µl of 20X Taqman Gene Expression Assay (probe-primer mix), and 2 µl of 1:1 diluted cDNA template were mixed in a total volume of 20 µl. 12 µl mineral oil was added to cover top of the mixture to prevent evaporation. After an initial 2 minutes of incubation at 50^{o}C [uracil-DNA-glycosylase (UDG) activation] and 10 minutes of initial denaturation at 95°C, 2-step thermal cycling was performed at 94°C for 15 seconds, 60°C for 1 minute for a total of 40 cycles. If the master mix contains dNTP mix with dUTP, UDG treatment can prevent the reamplification of carryover-PCR products by digesting any dU-containing DNA(Longo, Berninger et al. 1990). Ten-minute incubation at the 95°C substantially reduces UDG activity. Because UDG is not completely deactivated during the 95°C incubation, it is important to keep the annealing temperatures greater than 55°C and to refrigerate PCR products at 2 to 8°C in order to prevent

amplicon degradation.

The master mix also contained ROX passive reference dye which was used to normalize fluorescent fluctuations due to well-to-well changes in concentration or volume. ROX does not take part in the PCR reaction and its fluorescence remains constant during the PCR reaction. Since the emission wavelenghts of FAM (517 nm) and ROX (607 nm) are different, they could be used together in the same reaction. Stratagene MX3005P sofware performed normalization by dividing the fluorescence intensity of FAM by the fluorescence intensity of ROX and obtained a ratio defined as the Rn (normalized reporter) for a given reaction tube.

3.3.8.2 qPCR of lung cancer cell lines using SYBR Green I

qPCR using SYBR Green I was performed for validation analysis in lung cancer cell lines using Stratagene MX3005P Real-time PCR System (USA). 1X reaction contained 10 µl of 2X SYBR Green PCR Master Mix, 0.4 µl 25 µM forward and reverse primers, and 2 µl of 1:1 diluted template cDNA were mixed in a total volume of 20 µl. After an initial 2 minutes of incubation at 50^{o}C (UDG activation) and 15 minutes of denaturation at 95°C, thermal cycling was performed at 94°C for 30 seconds, 60-62°C for 30sec (optimized for each primer pair), 72°C for 30 seconds for a total of 50 cycles and a final extension step at 72°C for 10 minutes. In order to validate the production of a single target-specific PCR product, the amplification was followed by a melt curve protocol with an initial step at 55°C for 30 seconds and 80 repeats of 0.5°C increments with 15 seconds dwell time, from 55°C to 95°C.

3.3.8.3 Calculation of relative expression using ΔΔCt formula

In cell lines and tissues, the relative expression ratio (R) of transcripts (target gene) was measured based on a modified ΔΔCt formula (Pfaffl 2001) and normalized to *GAPDH* (reference gene). In $R= (E_{target})^{\Delta Ct_{target}(\text{control-sample})} / (E_{ref})^{\Delta Ct_{ref}(\text{control-sample})}$ formula, E_{target} and E_{ref} reflect PCR efficiencies of the primers for target genes and reference genes, respectively. We assumed the PCR efficiencies of the primers used in Taqman probe-based qPCR as 2.0.
In cell lines, *GAPDH* was the reference gene. ΔCt values were obtained by subtracting Ct values of individual genes (sample) from the average Ct value of all cell lines for that gene (control). All reactions were performed in duplicates. A no-template control of nuclease-free water was included in each run. Relative expression tables were established by representing ΔΔCt values in log2 base, and in all subsequent analyses these values were used.

3.3.9 Bioinformatic analyses

3.3.9.1 Data retrieval for meta-analysis

Tumor and cell line microarray datasets were obtained from the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) and Array Express (http://www.ebi.ac.uk/microarray-as/ae/) repositories, and in one case from http://data.cgt.duke.edu/oncogene.php. Inclusion criteria were as follows:

1. Only datasets from solid tumors and tumor cell lines without any drug treatment were used.
2. Datasets were analyzed only if they were available as raw data within CEL files.
3. Datasets were restricted to either the Affymetrix HG-U133A or HG-U133Plus2 platforms.
4. Datasets were used if they showed differential CT-X gene expression in the cluster analysis; as in a dataset, some tumor samples express CT-X genes coordinately whereas other tumor samples do not express CT-X genes (**section 3.3.9.5**).

Table 3.8 shows datasets that met these inclusion criteria. Although other tumor datasets (colon, prostate etc.) that fitted into the first three criteria (1-3) were obtained, in the cluster analysis, they did not show differential CT-X gene expression pattern and therefore they were not included in meta-analysis. Samples corresponding to normal tissues were excluded from these datasets.

Table 3.8: Tumor and cell line microarray datasets used in meta-analysis

GSE number/Array Express ID	Tumor type / Cancer cell lines	Sample size	Array platform	Reference
E-GEOD-GSE7390[#]	Node-negative breast tumors	198	HG-U133A	(Desmedt, Piette et al. 2007)
GSE6008*	Ovarian tumors	99	HG-U133A	(Hendrix, Wu et al. 2006)
GSE10072*	Lung adenocarcinoma tumors	58	HG-U133A	(Landi, Dracheva et al. 2008)
E-TABM-36[#]	Hepatocellular carcinoma tumors	57	HG-U133A	(Chiang, Villanueva et al. 2008)
GSE4824*	Lung cancer cell lines	70	HG-U133A	(Zhou, Peyton et al. 2006)
GSE5720*	NCI-60 Cancer cell panel	60	HG-U133A	(Shankavaram, Reinhold et al. 2007)
GSE9843*	Hepatocellular carcinoma	91	HG-U133Plus2	(Chiang, Villanueva et al. 2008)
GSE3141[$]	Primary lung tumors	111	HG-U133Plus2	(Bild, Yao et al. 2006)
GSE9891*	Ovarian tumors	385	HG-U133Plus2	GEO Datasets, Bowtell D, 2008
GSE5460*	Breast tumors	127	HG-U133Plus2	GEO Datasets, Richardson AL, 2007

Dataset sources: [#]Array Express, *GEO, [$]http://data.cgt.duke.edu/oncogene.php.

3.3.9.2 Normalization of raw data within CEL files

After raw data were downloaded for each dataset, they were preprocessed individually using GeneSpring GX 9.0.6 software (Agilent Technologies). Data were normalized with the GC-RMA algorithm. This algorithm performed three tasks in the following order: background correction, quantile normalization and probe summarization. Baseline transformation was not performed subsequent to probeset summarization. After GC-RMA normalization, expression values of probesets were obtained in log2 scale.

3.3.9.3 Quality control on samples of individual datasets

Quality control (QC) analysis was performed for each dataset using GeneSpring GX 9.0.6 software. The following analyses were carried out during QC analysis:

- Principal Component Analysis (PCA) calculated the PCA scores for each sample and depicted them in a 3D scatter plot. This analysis was useful when viewing of separations between groups of replicates. For the datasets used in this study, although some samples were grouped separately, most samples were clustered together in a 3D PCA plot.

- The correlation analysis calculated the correlation coefficients using Pearson correlation coefficient for each pair of arrays and displayed them as a correlation table. Correlation coefficients ranged around 0.7-1.0 which were acceptable.
- 3'/5' ratios for actin and GAPDH probesets were calculated for each array, reflecting RNA sample quality.The hybridization quality was analyzed by hybridization controls for each array. All the samples in terms of 3'/5' ratios and the hybridization quality were included in further analyses according to the QC results.

3.3.9.4 Hierarchical clustering analysis of tumor and cell line datasets

Seven CT-X gene families were selected to be used in cluster analysis and grouping of datasets according to their expression values. The selected CT-X genes and their corresponding probeset IDs on Affymetrix HG-U133A and HG-U133Plus2 arrays are shown in **Tables 3.9** and **3.10**, respectively. Since there is considerable homology among members of a given CT gene family, these probesets were analyzed to see whether they were specific for the target transcripts using "Probe Match" tool that was available in NetAffx Analysis Center on the web page of Affymetrix (http://www.affymetrix.com/analysis/index.affx). Two probesets (215885_at for SSX2 and 215932_at for MAGEC2) present on both HG-U133A and HG-U133Plus2 were found not to be unique and were excluded from further analysis.

Table 3.9: Probesets used on Affymetrix HG-U133A array

CT-X Gene Families	Probe Set ID	Gene Symbol
NY-ESO-1	210546_x_at	NY-ESO-1
NY-ESO-1	211674_x_at	NY-ESO-1
NY-ESO-1	217339_x_at	NY-ESO-1
GAGE	208283_at	GAGE1
GAGE	207086_x_at	GAGE1, GAGE12, GAGE13, GAGE2, GAGE4, GAGE5, GAGE6, GAGE7,GAGE8
GAGE	207739_s_at	GAGE1, GAGE12, GAGE13, GAGE2, GAGE4, GAGE5, GAGE6, GAGE7,GAGE8
GAGE	208155_x_at	GAGE1, GAGE12, GAGE4, GAGE5, GAGE6, GAGE7
GAGE	206640_x_at	GAGE12, GAGE13, GAGE2, GAGE4, GAGE5, GAGE6, GAGE7
GAGE	208235_x_at	GAGE12, GAGE7
MAGEA	207325_x_at	MAGEA1
MAGEA	210295_at	MAGEA10
MAGEA	210503_at	MAGEA11
MAGEA	210467_x_at	MAGEA12
MAGEA	214603_at	MAGEA2
MAGEA	209942_x_at	MAGEA3
MAGEA	214254_at	MAGEA4
MAGEA	214642_x_at	MAGEA5
MAGEA	214612_x_at	MAGEA6

MAGEA	210274_at	MAGEA8
MAGEA	210437_at	MAGEA9
MAGEB	207534_at	MAGEB1
MAGEB	206218_at	MAGEB2
MAGEB	207579_at	MAGEB3
MAGEB	207580_at	MAGEB4
MAGEB	207581_s_at	MAGEB4
MAGEC	206609_at	MAGEC1
MAGEC	220062_s_at	MAGEC2
MAGEC	216592_at	MAGEC3
SPANX	220922_s_at	SPANXA1,SPANXA2, SPANXB1, SPANXB2, SPANXC, SPANXE,SPANXF1
SPANX	220921_at	SPANXB1, SPANXB2, SPANXF1
SPANX	220217_x_at	SPANXC
SSX	206626_x_at	SSX1
SSX	206627_s_at	SSX1
SSX	215881_x_at	SSX10, SSX2, SSX3, SSX5, SSX7, SSX9
SSX	207493_x_at	SSX2
SSX	210497_x_at	SSX2
SSX	216471_x_at	SSX2
SSX	207666_x_at	SSX3
SSX	211670_x_at	SSX3
SSX	211731_x_at	SSX3
SSX	208586_s_at	SSX4
SSX	210394_x_at	SSX4
SSX	211425_x_at	SSX4
SSX	208528_x_at	SSX5

Table 3.10: Probesets used on Affymetrix HG-U133Plus2 array.

CT-X Gene Families	Probe Set ID	Gene Symbol
NY-ESO-1	210546_x_at	NY-ESO-1
NY-ESO-1	211674_x_at	NY-ESO-1
NY-ESO-1	217339_x_at	NY-ESO-1
GAGE	208283_at	GAGE1
GAGE	207086_x_at	GAGE1, GAGE12, GAGE13, , GAGE2, GAGE4, GAGE5, GAGE6, GAGE7,GAGE8
GAGE	207739_s_at	GAGE1, GAGE12, GAGE13, GAGE2, GAGE4, GAGE5, GAGE6, GAGE7,GAGE8
GAGE	208155_x_at	GAGE1, GAGE12, GAGE4, GAGE5, GAGE6, GAGE7
GAGE	206640_x_at	GAGE12, GAGE13, GAGE2, GAGE4, GAGE5, GAGE6, GAGE7
GAGE	208235_x_at	GAGE12, GAGE7
MAGEA	207325_x_at	MAGEA1
MAGEA	210295_at	MAGEA10
MAGEA	210503_at	MAGEA11
MAGEA	210467_x_at	MAGEA12
MAGEA	214603_at	MAGEA2
MAGEA	1553830_s_at	MAGEA2
MAGEA	209942_x_at	MAGEA3
MAGEA	214254_at	MAGEA4
MAGEA	214642_x_at	MAGEA5

MAGEA	1553585_a_at	MAGEA5
MAGEA	214612_x_at	MAGEA6
MAGEA	210274_at	MAGEA8
MAGEA	210437_at	MAGEA9
MAGEB	207534_at	MAGEB1
MAGEB	1552913_at	MAGEB18
MAGEB	206218_at	MAGEB2
MAGEB	207579_at	MAGEB3
MAGEB	207580_at	MAGEB4
MAGEB	207581_s_at	MAGEB4
MAGEB	1552858_at	MAGEB6
MAGEC	206609_at	MAGEC1
MAGEC	220062_s_at	MAGEC2
MAGEC	216592_at	MAGEC3
SPANX	220922_s_at	SPANXA1,SPANXA2, SPANXB1, SPANXB2, SPANXC, SPANXE,SPANXF1
SPANX	224032_x_at	SPANXA1, SPANXA2, SPANXC, SPANXE
SPANX	220921_at	SPANXB1, SPANXB2, SPANXF1
SPANX	220217_x_at	SPANXC
SSX	206626_x_at	SSX1
SSX	206627_s_at	SSX1
SSX	215881_x_at	SSX10, SSX2, SSX3, SSX5, SSX7, SSX9
SSX	207493_x_at	SSX2
SSX	210497_x_at	SSX2
SSX	216471_x_at	SSX2
SSX	207666_x_at	SSX3
SSX	211670_x_at	SSX3
SSX	211731_x_at	SSX3
SSX	208586_s_at	SSX4
SSX	210394_x_at	SSX4
SSX	211425_x_at	SSX4
SSX	208528_x_at	SSX5

Following normalization, only probeset IDs corresponding to the selected CT-X genes (**Table 9** and **Table 3.10**) were imported to GeneSpring GX 9.0.6 software as a .txt file to acquire their normalized expression values from individual datasets that were used for cluster analysis. Hierarchical cluster analysis was performed using average linkage method with Euclidean distance matrix measure. As a second method of analysis we carried out class prediction analysis for the GSE4824 dataset (lung cancer cell lines) using BRB Array Tools (http://linus.nci.nih.gov/BRBArrayTools).

3.3.9.5 CT-X grouping of tumor and cell line datasets

To group samples into CT-X expressors and non-expressors normalized data for individual datasets were first exported from GeneSpring into R and the“grouping” algorithm defined in

Appendix A was utilized. The manipulations performed by this script are listed below. Briefly, the script ranks probests for each sample from the highest to the lowest and uses this information to determine a cut-off value above which CT-X gene expression is considered positive. Two separate grouping analyses were performed, one for tumor samples and one for cell lines. Each grouping analysis was performed seperately for HG-U133A and HG-U133Plus2 based data. The grouping script carried out the following commands:

1. The expression value for each CT-X gene family (e.g. SSX, MAGE, GAGE etc.) was defined:
    a. For each sample, the "CT-X gene value" corresponding to the average value of several probesets of the same CT-X gene was determined (for those genes for which several probesets existed; e.g. 208586_s_at, 210394_x_at and 211425_x_at are three probesets each hybridizing with SSX4).
    b. Then the "CT-X gene family value", corresponding to the average value of individual CT-X genes within the same family was detemined.
2. Normalized data of individual datasets were combined into a single expression matrix and the average rank value corresponding to the expression value of 3.5 was determined for each matrix. This corresponded to the 10,856th probeset (rank) for the combined cell line expression matrix, and to 10,001st rank for combined tumor expression matrix (**Table 4.3**), among 22,283 probests ranked from the highest to the lowest expression value (for HG-U133A). For the combined tumor expression matrix formed using HG-U133Plus2 platform, this corresponded to the 28, 581st rank among 54,675 probesets.
3. To determine if a given sample could be considered positive for a given CT-X gene expression, its "CT-X gene family value" had to be higher than that of the average rank calculated in step 2 within each sample data.
4. For each of the seven CT-X gene families, each sample was assigned a number based on whether it was positive (1) or negative (0) for a given CT-X family value.
5. The script then calculated the sum of all values (0 or 1) for each sample for each CT-X gene family and grouped the samples into 3 categories: Samples with expression for at least 2 CT-X gene families (CT-X positive group, "1"), those with expression of only one CT-X gene family (CT-intermediate group: "-1"), and without CT-X expression (CT-X negative "0").

3.3.9.6 Meta-analysis

All the scripts written in R that performed data pre-processing, and meta-analysis are listed in **Appendix A**.

3.3.9.6.1 Data pre-processing

1. Using GeneSpring, each sample within all datasets that were categorized as CT-X positive "1" and - negative "0" were selected and intermediate "-1" samples were discarded.
2. Then, for each sample, those probesets with normalized expression values below the cut-off rank, as defined by the grouping script explained in **section 3.3.9.5**, were filtered out.
3. Among the remaining probesets only those that were common to all samples were then combined in an expression data matrix using the R based "pre-processing script". Thus, 19,421 probesets for tumor samples, and 18, 411 probests for the cell line data remained.

3.3.9.6.2 Meta-analysis using Bioconductor RankProd package

The resulting combined-filtered expression data matrices were used by the RankProd package (Hong, Breitling et al. 2006) in the R environment to identify genes differentially expressed between CT-X positive and CT-X negative groups.

The RankProd algorithm applies a series of calculations:

1. Once two experimental conditions (the two experimental groups: A & B) are defined (in our case the CT-X negative and CT-X positive groups), for a given dataset, and a fold change value (FC) is calculated for each probeset that corresponds to the ratio of its value in a sample in A (#a1), compared to that of a sample in B (#b1). The algorithm then calculates a second FC value for #a1 by measuring the ratio of its value to that of another sample in B (#b2). The algorithm thus calculates all FC values for every possible pairwise comparison for a given probeset in each dataset.
2. After a FC is found for all probesets, it is ranked within each comparison such that the probeset with the greatest fold change is assigned a rank: 1/n, where n equals the total number of probesets.
3. Then all rank values obtained for one probeset are multiplied, generating the rank product (RP). The rank products of a probeset are combined from different datasets by taking their geometric mean.

4. To evaluate the significance of RP values, random permutations of expression values are performed for each array and then RP values are calculated as described above. This determines how unlikely it is to observe the same RP for a probeset by chance, thus converting from the RP value to an E-value, similar to that of BLAST. For 100 random permutations, the average expected value E (RP) ≈ x(RP)/100 is calculated which refers to how many simulated RP values smaller than or equal to a real RP value are found.
5. Then, for each gene g, the percentage of false positives (PFP) which corresponds to a false discovery rate is calculated if this gene would have a significant fold change ($p<0.05$).

A total of 100 permutations were performed to obtain significant genes at 5% FDR rate. RankProd generated up-regulated and down-regulated gene lists and also two respective plots showing estimated FDR (PFP) versus the number of identified genes. To obtain biologically relevant genes, probesets with less than 1.2 fold change were excluded from further analysis.

3.3.9.6.3 Validation of the rank-product method using HG-U133Plus2 tumor datasets

HG-U133Plus2 contains all probesets (with a few exceptions) contained within HG-133A. Normalized data of individual datasets were first combined into a single expression matrix. To validate the initial RankProd data, we therefore, filtered in those probesets used for the initial meta-analysis with HG-U133A based data (corresponding to ~19,415 probesets) from HG-U133Plus2 tumor datasets by an R based script. We thus generated a second data matrix. Using these probesets that were thus identical between the two platforms, a second meta-analysis was performed by RankProd.

3.3.9.7 Class prediction analysis of GSE4824 lung cancer cell line dataset

The raw data was normalized by GC-RMA algorithm using BRB Array Tools. The class prediction analysis of BRB Array Tools creates a multivariate predictor in order to determine to which of the two classes (in our case; CT-X positive and negative) a given sample belongs (BRB Array Tools User's Manual). Different predictors (nearest centroid, compound covariate, support vector machine etc.) were utilized simultaneously to obtain an output probeset list that disciminates CT-X positive and negative lung cancer cell lines at 0.001 significance.

3.3.9.8 Finding common probesets between different analyses by CROPPER

CROPPER is a free web-based software that can perform cross-platform/cross-species combinations of genomic data derived from heterogenous sources (http://katiska.uku.fi/~jmpaanan/cropper/). In this study, it was used to find common probesets from the same platform between different analyses.

3.3.9.9 DAVID functional annotation clustering

DAVID (the database for annotation, visualization, and intergrated discovery) bioinformatic resources was used in order to analyze probeset lists derived from both meta-analysis and class prediction analysis (http://david.abcc.ncifcrf.gov/). Among four useful DAVID's major analytic modules, functional annotation clustering was employed in this study. First, the probeset list was submitted to DAVID and HG-U133A platform was set as the background for subsequent analysis. GOTERM_BP_ALL from gene ontology and BBID, BIOCARTA, KEGG_PATHWAY from pathways were selected for DAVID functional annotation clustering. Enrichment score 1.3 is equivalent to non-log scale 0.05 (Huang da, Sherman et al. 2009).

4 RESULTS

4.1 Generation of SSX4 knock-in SK-LC-17 cell line

4.1.1 SSX4 knock-in vector

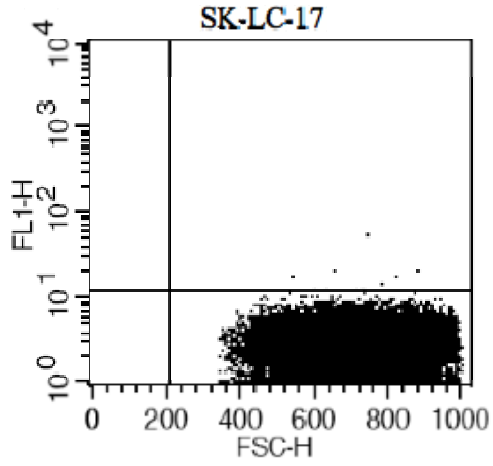
The promoter architecture of SSX4 gene was taken into account while designing the SSX4 knock-in (KI) vector. The SSX4 promoter was previously characterized (Gure AO, unpublished data) by luciferase reporter constructs that were generated by PCR amplification of genomic DNA with forward (A1, A2, etc.) and reverse (B) primers shown in **Figure 4.1**. Luciferase reporter experiments showed that the minimal promoter activity of SSX4 lay between primers A3.3 and A4. Interestingly, this promoter showed low but significant promoter activity in the reverse direction demonstrating bidirectional activity.

Figure 4.1 Sequence of the SSX4 promoter-proximal region: Exons 1 and 2 (containing translation initiation code in red) are shown in yellow. Ornithine Aminotransferase Like (OATL) pseudogene which is disrupted by an L1 repeat is shown in light blue. Alu repeats at the 5' end of the promoter are indicated. The primers that were used to generate promoter-reporter constructs are indicated above their respective sequences (boxed).

Since L1 repeats were localized upstream of A3 forward primer, the 5' homology sequence of the KI vector started from the A3 primer and ended with B primer sequences in order to prevent integration of the KI vector into L1 repeat containing regions in the genome. Although the SSX 3' homology sequences was amplified using SSX4 specific sequences, subsequent sequencing of this region after it was cloned into the KI vector revealed that it was derived from SSX7 which is highly homologous to SSX4 but not expressed in either testis or cancer cells. The KI vector contained two selection markers both of which were driven by their own promoters; phosphoglycerate kinase (PGK) promoter- driven hygromycin antibiotic gene provided positive selection whereas β-actin-driven diphteria toxin-A gene (DTA) provided negative selection. Following homologous recombination between the KI vector and the SSX4 gene mediated by SSX4 3' and 5' homology sequences, DTA toxin is expected to be cleaved off. Non-homologous recombination events, on the other hand, would not require DTA elimination and thus die due to the toxicity induced by DTA. The sequence of the KI vector showing its components is given in **Appendix B**.

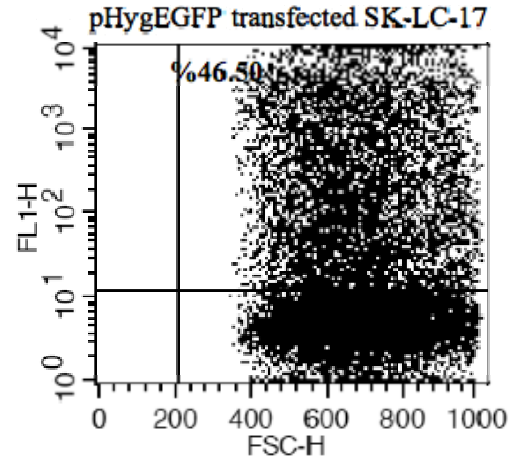
4.1.2 Screening of SSX4 KI clones for GFP expression by flow cytometry

Integration of KI vector into SSX4 gene would lead to expression of EGFP from SSX4 promoter. In order to check whether GFP could indeed be expressed from the SSX4 promoter, we transfected SK-LC-17 cells transiently with the KI vector lacking diphteria toxin (Step6) and performed flow cytometry analysis. As a positive control, we used transfected SK-LC-17 cells with pHygEGFP plasmid which has the EGFP under the control of the strong CMV promoter (**Figure 4.2**).



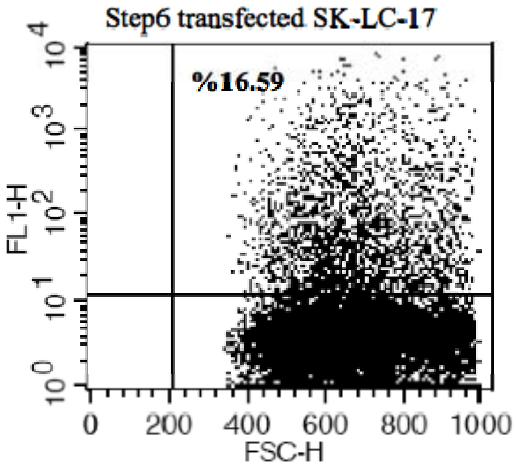
Gated Events: 17875
Total Events: 20000

Quad	Events	% Gated	% Total
UL	0	0.00	0.00
UR	9	0.05	0.04
LL	0	0.00	0.00
LR	17866	99.95	89.33



Gated Events: 16889
Total Events: 20000

Quad	Events	% Gated	% Total
UL	0	0.00	0.00
UR	7853	46.50	39.27
LL	0	0.00	0.00
LR	9036	53.50	45.18



Gated Events: 15501
Total Events: 20000

Quad	Events	% Gated	% Total
UL	0	0.00	0.00
UR	2572	16.59	12.86
LL	0	0.00	0.00
LR	12929	83.41	64.64

Figure 4.2: Dot plot analysis of untransfected SK-LC-17 cells and the same cells transiently transfected with pHygEGFP and the Step6 construct. 20.000 cells were counted and gated for the analysis. The cutoff for GFP expression intensity was in reference to that obtained for the untransfected SK-LC-17 cells. The number and percentage of cells within each quadrant are shown in tables next to each dot plot. The upper right quadrant (UR) indicates GFP positive cells.

As shown in Figure 4.3, the percentage of cells within the UR quadrant for untransfected cells were 0.05%, compared to 46.5% for pHygEGFP, and 16.59% for Step 6 KI vector transfected cells. These results showed that the SSX4 promoter as contained within the KI vector was functional upon transient transfection into SK-LC-17 cells. The Step 6 vector is 9840 bp compared to pHygEGFP (5793 bp), which possibly accounts for its lower transfection efficiency. In addition, GFP expression intensity of the pHygEGFP transfected cells was higher as relative to Step6 transfected cells since GFP was expressed from the strong CMV promoter.

.

Following verification of GFP expression from SSX4 promoter, 7.5 x 10^6 SK-LC-17 cells were stably transfected with the KI construct and clone selection was performed by growing the cells in hygromycin at 50 mg/ml, corresponding to the concentration at which >90% of the parental cells are killed in a week. Although initially more than a hundred clones survived hygromycin selection, about 30 of them died subsequently, possibly reflecting clones that had DTA integrated into their genome which showed its effect later on in the drug selection process. Thus the insertion efficiency was about 1:10^5, which is very high. We screened 70 stable clones for GFP expression by flow cytometry and chose 42 of them with GFP expression. We observed that only one clone (clone #70) could be classified as 100% GFP positive, while all others had partial positivity (**Table 4.1 and Figure 4.6**).

4.1.3 Determination of KI insertion site of GFP expressing SSX4 KI vector transfected clones by nested PCR

To determine correct insertion of the KI vector, we used forward primers homologous to genomic DNA sequences that would be expected to locate 640 to 500 bp 5' to the forward end of the KI vector, combined with reverse primers homologous to EGFP sequence, not normally found in genomic DNA of the cells. A control plasmid was generated in order to test the amplification efficiency of these primers in nested PCR. This control plasmid contained the SSX4 promoter-proximal region (A1-B) that thus contained sites for the forward primers (shown in **Figure 4.1**) and the EGFP sequence. Therefore, it represented the final DNA sequence upon the correct insertion of KI vector into the SSX4 gene. Two independent nested PCR reactions were performed using two sets of primer pairs for each reaction. **Figure 4.3** shows the location of these primers within the control plasmid sequences. In the first run of reaction (40 cycles), we did not observe any amplification products either specific or non-specific (data not shown). The 2nd run of reaction using the primary products yielded specific amplicons for 7 of the 42 stable clones (**Figures 4.4 and 4.5**). The products amplified by P4&M7 and A2.1&M4 primer pairs were gel-purified and sequenced with A2.1 and M4 primers.

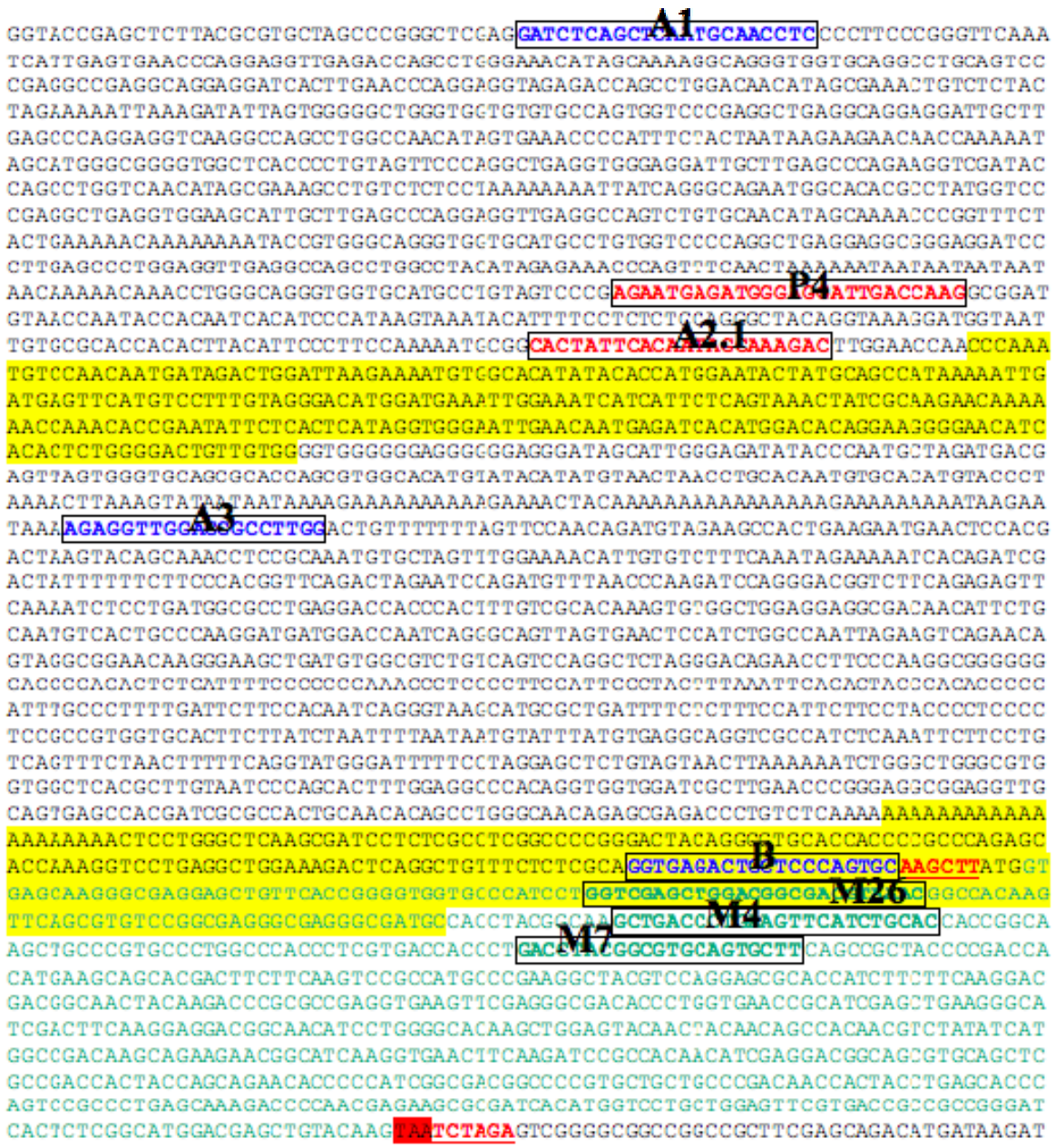


Figure 4.3.: Primers used in nested PCR in context of the SSX4 5' region after correct KI vector insertion. The region between primers A3 and B is contained within the KI vector as SSX4 5' homology sequence. Sequences shown in green are of EGFP and would not be expected to occur in the genomic DNA of untransfected cells. P4 and A2.1 are forward primers whereas M26, M4, and M7 are reverse primers. The primer pairs that were used in each nested PCR reaction was as follows: P4&M4 → A2.1&M26; P4&M7 → A2.1&M4. Sequences that were typically confirmed by sequencing analyses are highlighted in yellow.

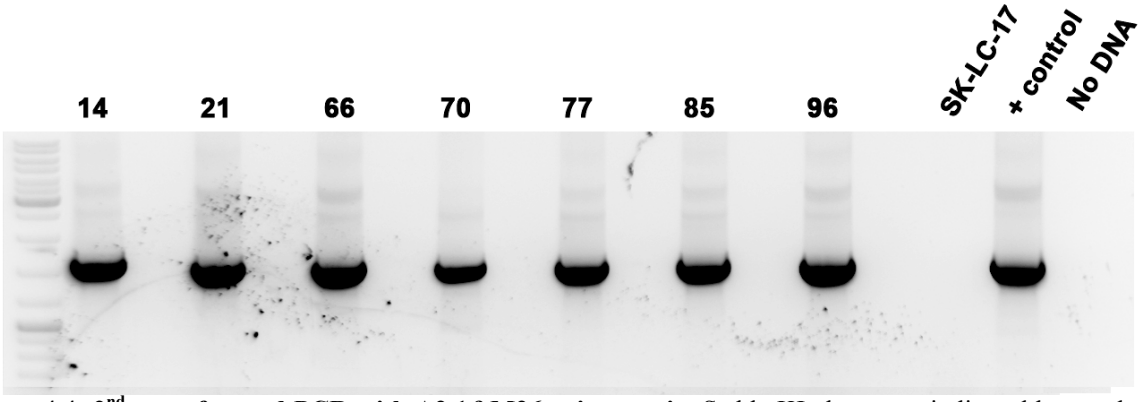


Figure 4.4: 2nd run of nested PCR with A2.1&M26 primer pair: Stable KI clones are indicated by numbers. The fist run was carried out using the P4&M4 primer pair in **Figure 4.3.** SK-LC-17 was used as a negative control while the control plasmid was used as a positive control. The expected amplicon is 1580 base pairs and was obtained from all clones shown. These products were gel-purified for sequencing.

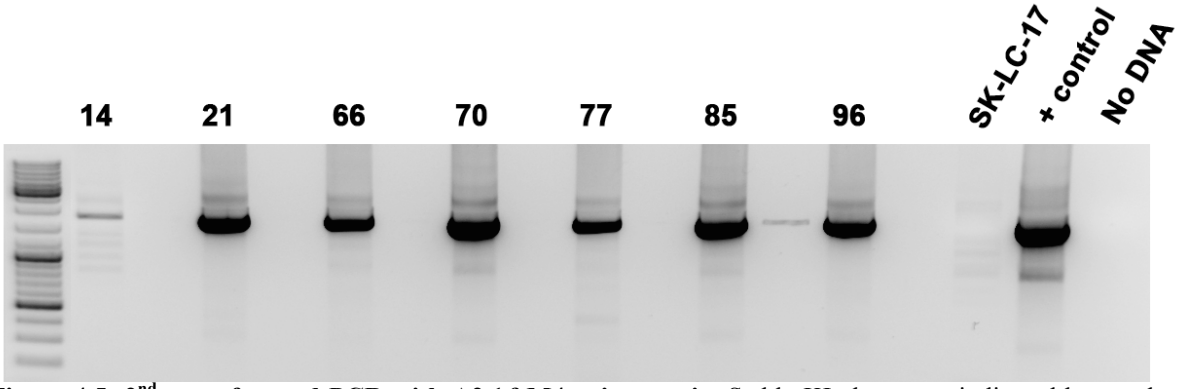


Figure 4.5: 2nd run of nested PCR with A2.1&M4 primer pair: Stable KI clones are indicated by numbers. The first run was carried out using P4&M7 primer pair in **Figure 4.3.** The amplicon was 1702 base pairs. The PCR for clone #14 was subsequently repeated and a band of similar intensity to the others was obtained. These products were gel purified and sequenced with the same PCR primer pair.

4.1.4 Sequencing of the amplified products for individual stable clones

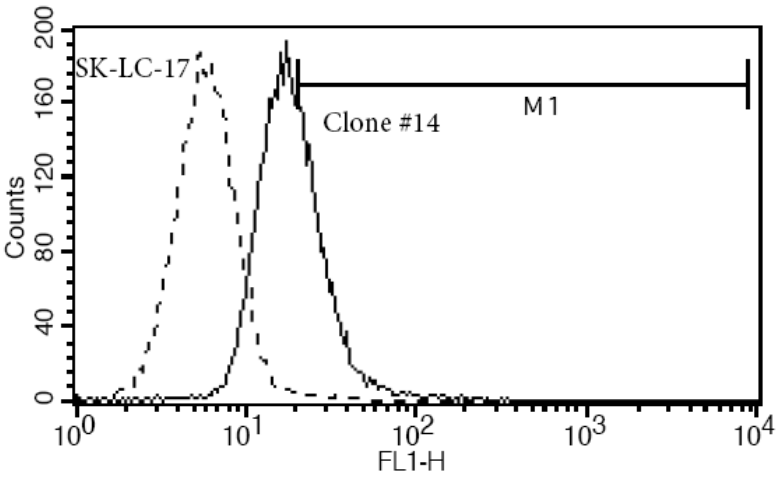
Figure 4.3 shows the sequenced regions of the amplified products with A2.1 forward and M4 reverse primers in yellow. According to the BLAST results given in **Appendix C**, the sequenced region with A2.1 forward primer aligned to the expected sequence in the SSX4 promoter-proximal region at %99 identity, where the first exon of Ornithine Aminotransferase Like (OATL) pseudogene lies in the genomic DNA whereas it aligned to the genomic clones that contained other family members of SSX gene family at < %95 identity (**Appendix C**). Moreover, the sequenced region with M4 reverse primer aligned perfectly to the SSX4 minimal promoter (**Appendix C**). This verified correct insertion of the KI vector into SSX4 gene for clones #14, 21, 70, 77, 85 and 96 except clone #66.

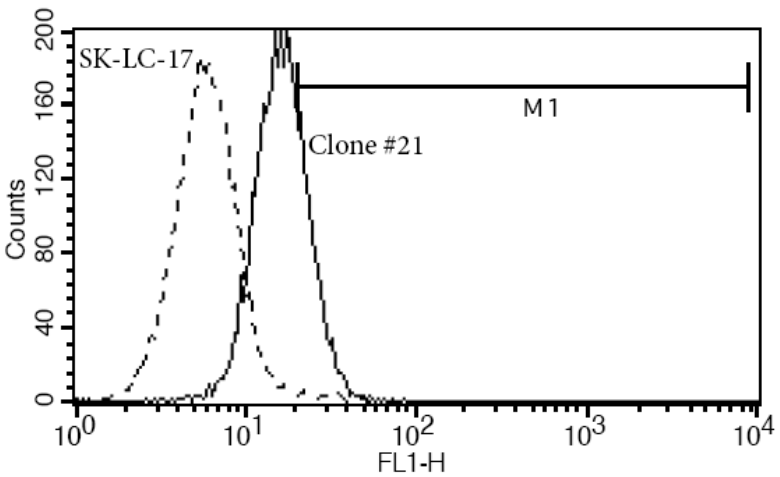
4.1.5 Flow cytometry analysis of SSX4 KI clones that were verified by nested PCR

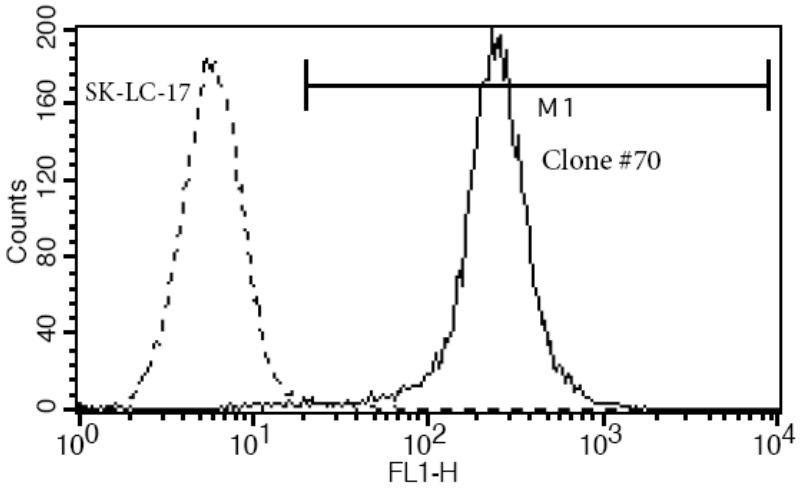
We observed a heterogeneous GFP expression of SSX4 KI clones except Clone #70 in the initial screening by flow cytometry (**Figure 4.6**). The percentage of GFP expressing cells and their GFP expression intensity and the corresponding histogram plots are given in **Table 4.1** and **Figure 4.6**, respectively.

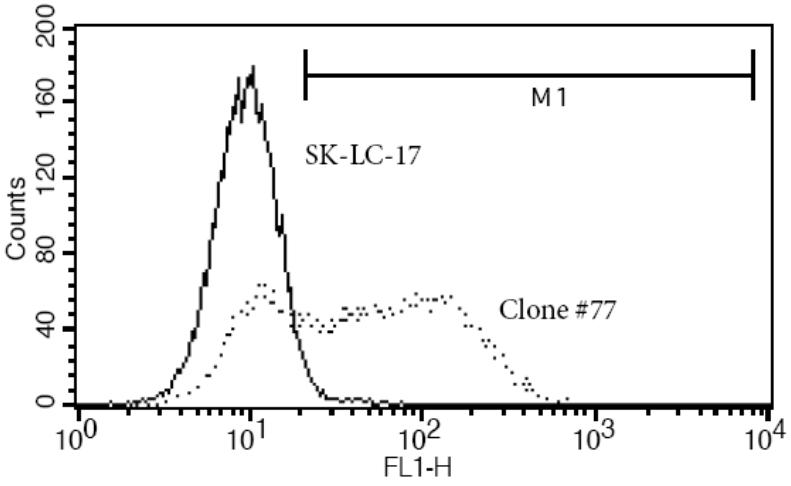
Table 4.1: The percentage of GFP expressing cells and their GFP expression intensity

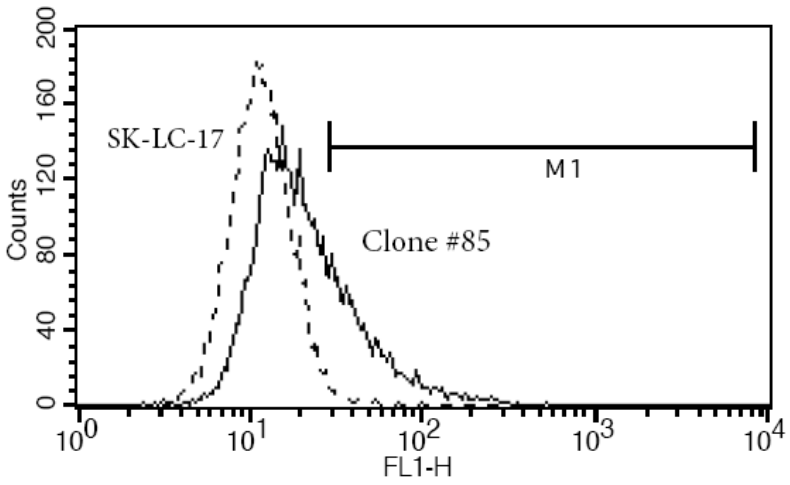
	% GFP positive cells (M1)	GFP expression intensity
SK-LC-17	0.50	32.39
Clone #14	36.74	28.53
Clone #21	24.65	24.18
Clone #70	99.51	251.38
Clone #77	68.05	105.25
Clone #85	23.82	53.36
Clone #96	41.87	67.99






SK-LC-17
M1
Clone #70
Counts
0
40
80
120
160
200
10^0
10^1
10^2
10^3
10^4
FL1-H


M1
SK-LC-17
Clone #77
Counts
0
40
80
120
160
200
10^0
10^1
10^2
10^3
10^4
FL1-H


SK-LC-17
M1
Clone #85
Counts
0
40
80
120
160
200
10^0
10^1
10^2
10^3
10^4
FL1-H

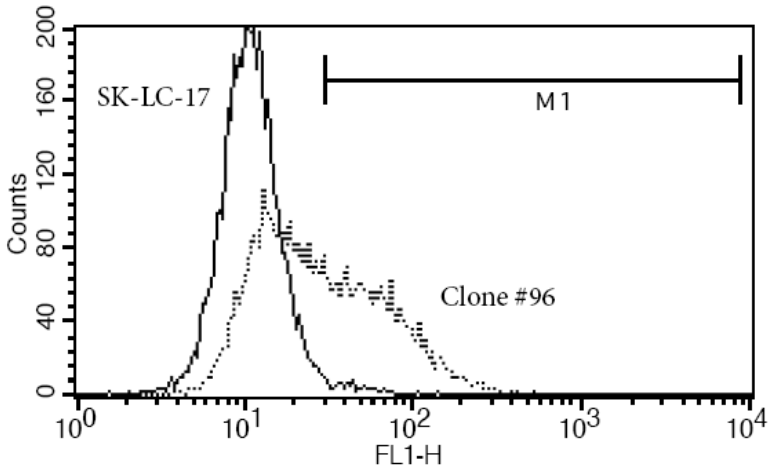
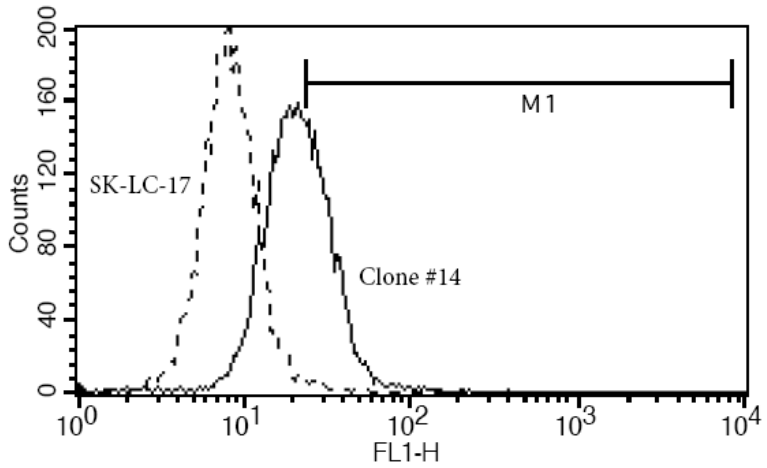


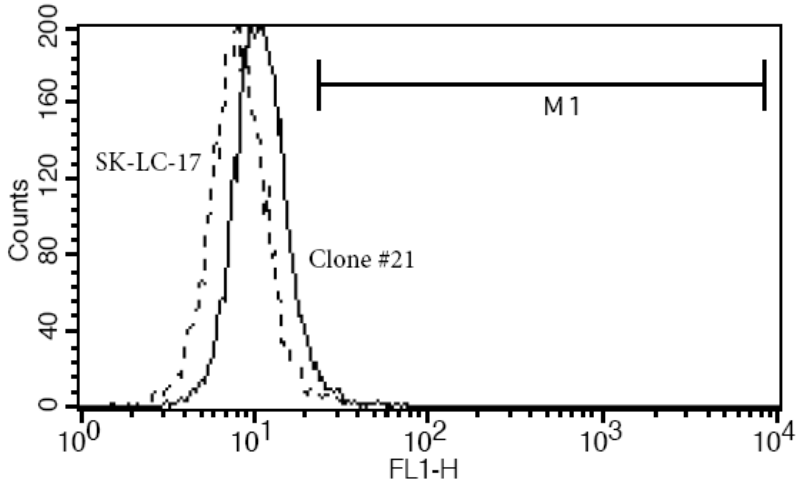
Figure 4.6: Histogram plot analysis of SSX4 KI clones: The percentage of GFP expressing cells (M1) and their GFP expression intensity are given in **Table 4.1.** 20.000 cells were counted and gated for the analysis. Cutoff for GFP expression intensity was based on that observed for untransfected SK-LC-17 cells.

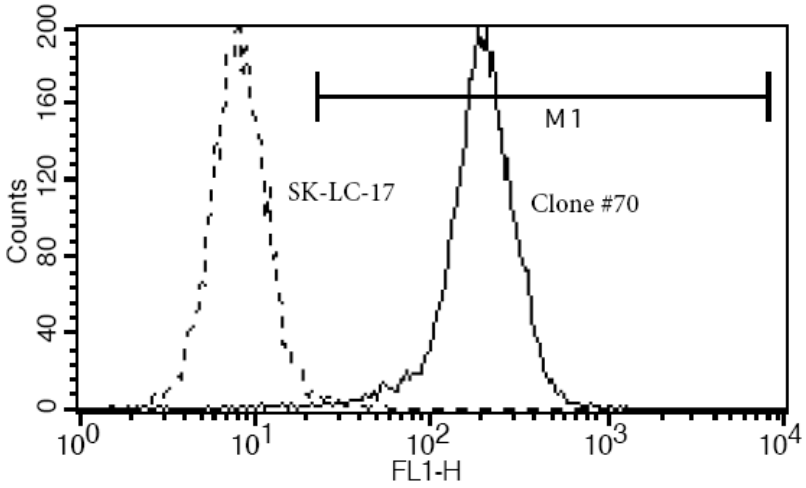
We performed a second flow cytometry analysis of SSX4 KI clones in order to see whether clones were stably expressing GFP (**Table 4.2 and Figure 4.7**).

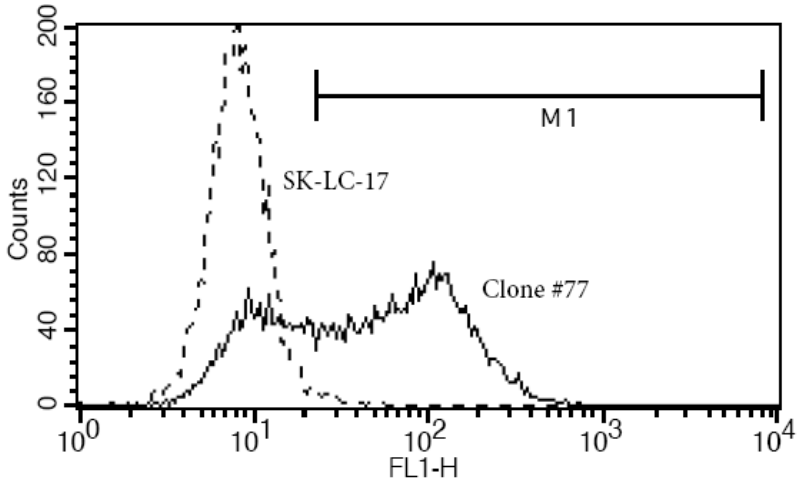
Table 4.2: The percentage of GFP expressing cells and their GFP expression intenstiy

	% GFP positive cells (M1)	GFP expression intensity
SK-LC-17	0.62	28.07
Clone #14	36.12	32.11
Clone #21	1.30	29.52
Clone #70	99.72	199.39
Clone #77	66.33	98.21
Clone #85	2.94	38.05
Clone #96	52.43	61.24




M1
SK-LC-17
Clone #21
Counts
FL1-H


M1
SK-LC-17
Clone #70
Counts
FL1-H


M1
SK-LC-17
Clone #77
Counts
FL1-H

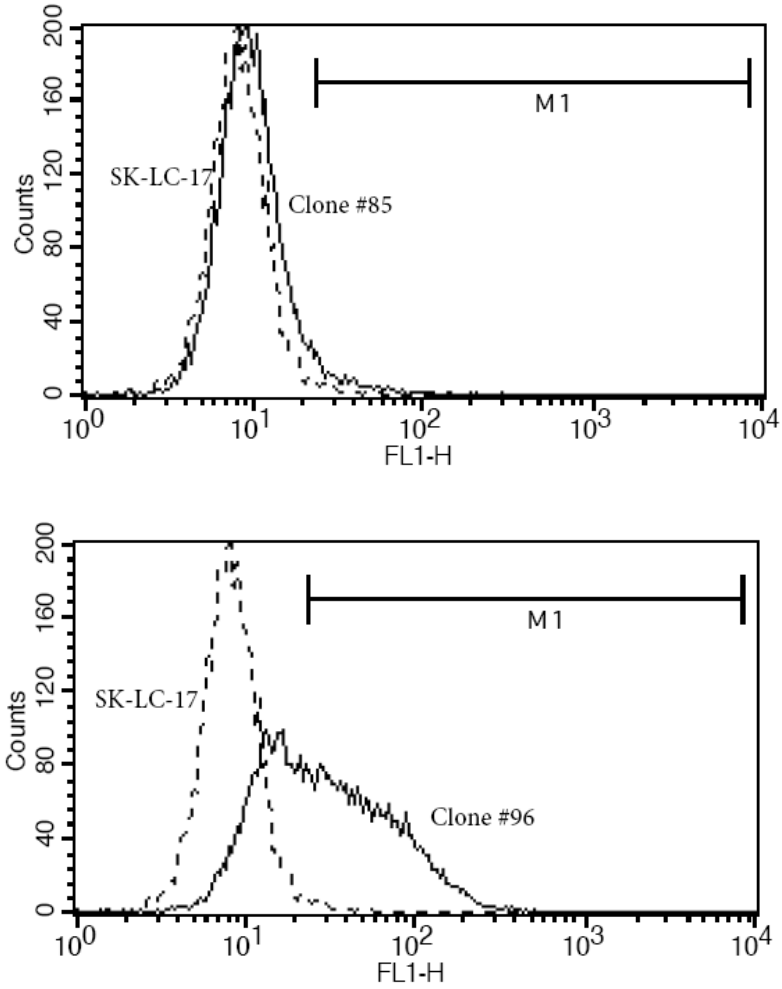


Figure 4.7: Histogram plot analysis of SSX4 KI clones. The percentage of GFP expressing cells (M1) and their GFP expression intensity are given in **Table 4.2.** 20.000 cells were counted and gated for the analysis. Cutoff for GFP expression intensity was based on that observed for untransfected SK-LC-17 cells.

Comparison of these flow cytometry analyses showed that clone #85 and #21 lost their GFP expression. We observed only minor differences in terms of percentage of GFP expressing cells and GFP expression intensity for clones #14, #70, #77 and #96.

4.1.6 Quantitative real-time PCR data for SSX4 gene in KI clones

There are 2 SSX4 genes (SSX4; SSX4B) located head to head orientation in the Xp11.23. Since they are nearly identical copies, we could not differentiate whether SSX4 KI construct integrated into one of them or both by sequencing results. Quantification of SSX4 expression in these clones by qPCR would possibly help us understand this. If KI construct integrated into both SSX4 and SSX4B, the clones would not express SSX4. If it integrated into one of them, the clones would have SSX4 expression reduced by %50. We compared SSX4 transcript levels of clones with that of Clone #1 which gave no band in nested PCR (negative

control), and SK-LC-17. According to qPCR results, Clone 21 and 85 had reduced SSX4 expression whereas others expressed SSX4 even at higher levels than SK-CL-17 (**Figure 4.8**).

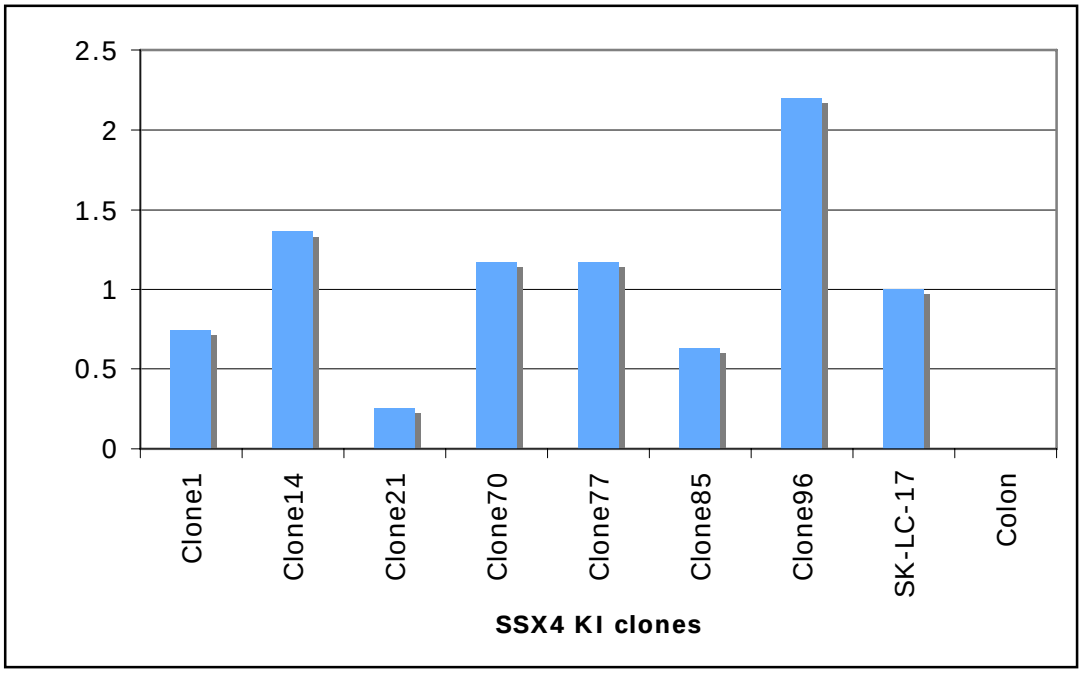


Figure 4.8: Relative SSX4 expression in SSX4 KI clones. SSX4 expression was checked by qPCR using SSX4 Taqman MGB probe in SSX4 KI clones. Relative expression was calculated using ΔΔCt method and SSX4 expression in SK-LC-17 was taken as the reference. Colon cDNA was used as a negative control to show that SSX4 is not expressed in normal tissues.

4.2 Meta-analysis of tumor and cell line microarray datasets

4.2.1 Hierarchical clustering analysis of tumor and cell line microarray datasets showed coordinate CT-X gene expression

We obtained raw data for 2 cell line datasets and 8 tumor datasets from GEO and Array Express (**Table 3.8**). 4 of tumor datasets were generated by using Affymetrix HG-U133A platforms and the rest were generated by using HG-U133Plus2 platforms. The raw data were normalized with GC-RMA algorithm using GeneSpring GX 9.0.6 software (Agilent Technologies). All the arrays within datasets were included in further analyses according to the quality control (QC) analysis of GeneSpring GX 9.0.6 (**section 3.3.9.6.2**, data not shown). Subsequent to normalization, we used seven CT-X gene families, NY-ESO-1, SSX, GAGE, MAGEA, MAGEB, MAGEC and SPANX, for the hierarchical clustering analysis of tumor and cell line datasets, individually (**Table 3.9** and **Table 3.10**). The clustering analysis

demonstrated that the selected CT-X genes were coordinately expressed in all tumor and cell line datasets, and that the samples could be categorized into three distinct groups; CT-X positive group expressing more than one CT-X gene family, CT-X intermediate group expressing only one CT-X gene family and CT-X negative group showing no CT-X gene expression. Two representative cluster images, one for the GSE4824 dataset containing lung cancer cell lines and GSE10072 composed of lung tumor samples, are shown in **Figures 4.9** and **4.10,** respectively.

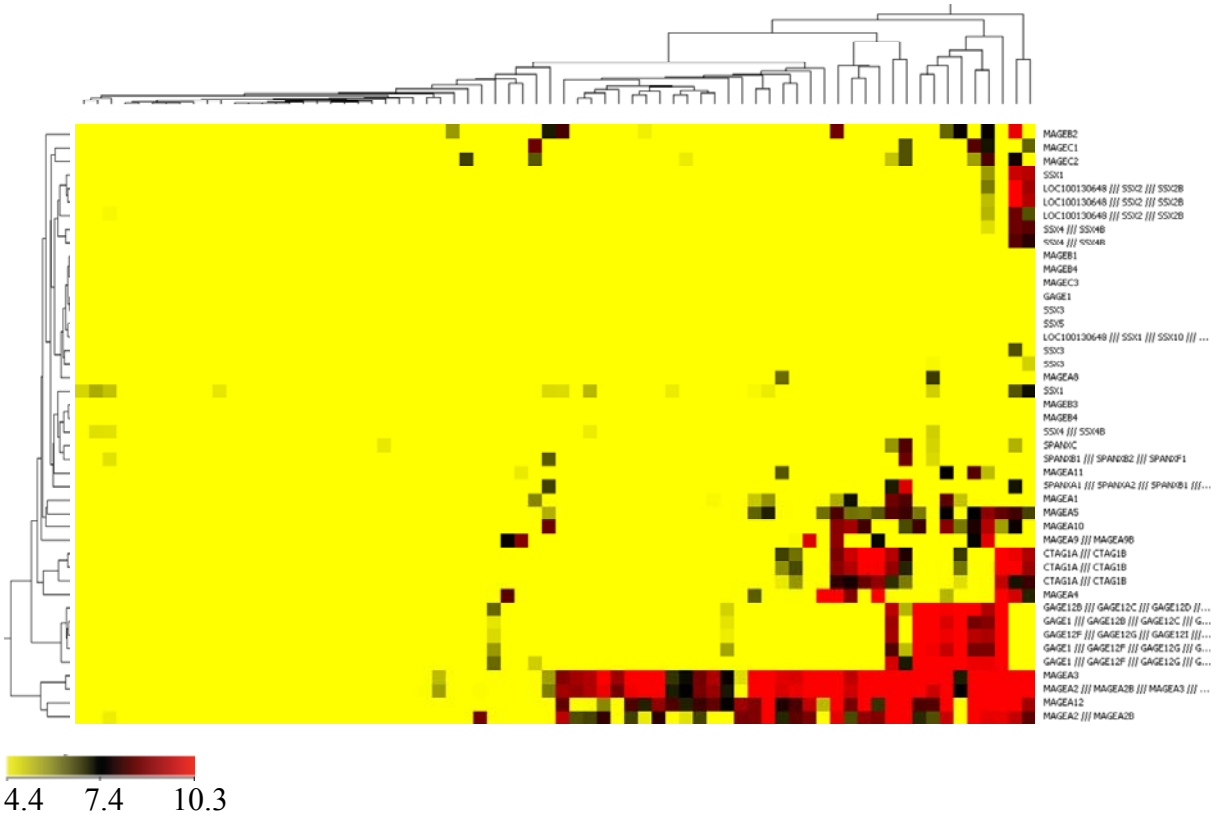


Figure 4.9: Hierarchical clustering analysis of lung cancer cell lines (GSE4824 dataset). The normalized expression values of 44 probesets corresponding to the selected CT-X genes were used to establish the hierarchical clustering using GeneSpring GX 9.06. Average linkage method and Euclidean distance as a matrix measure were used. Color code below the heatmap shows the range of normalized expression values. The tree at the top of the heatmap represents samplewise clustering. The tree on the left of the heatmap represents genewise clustering.

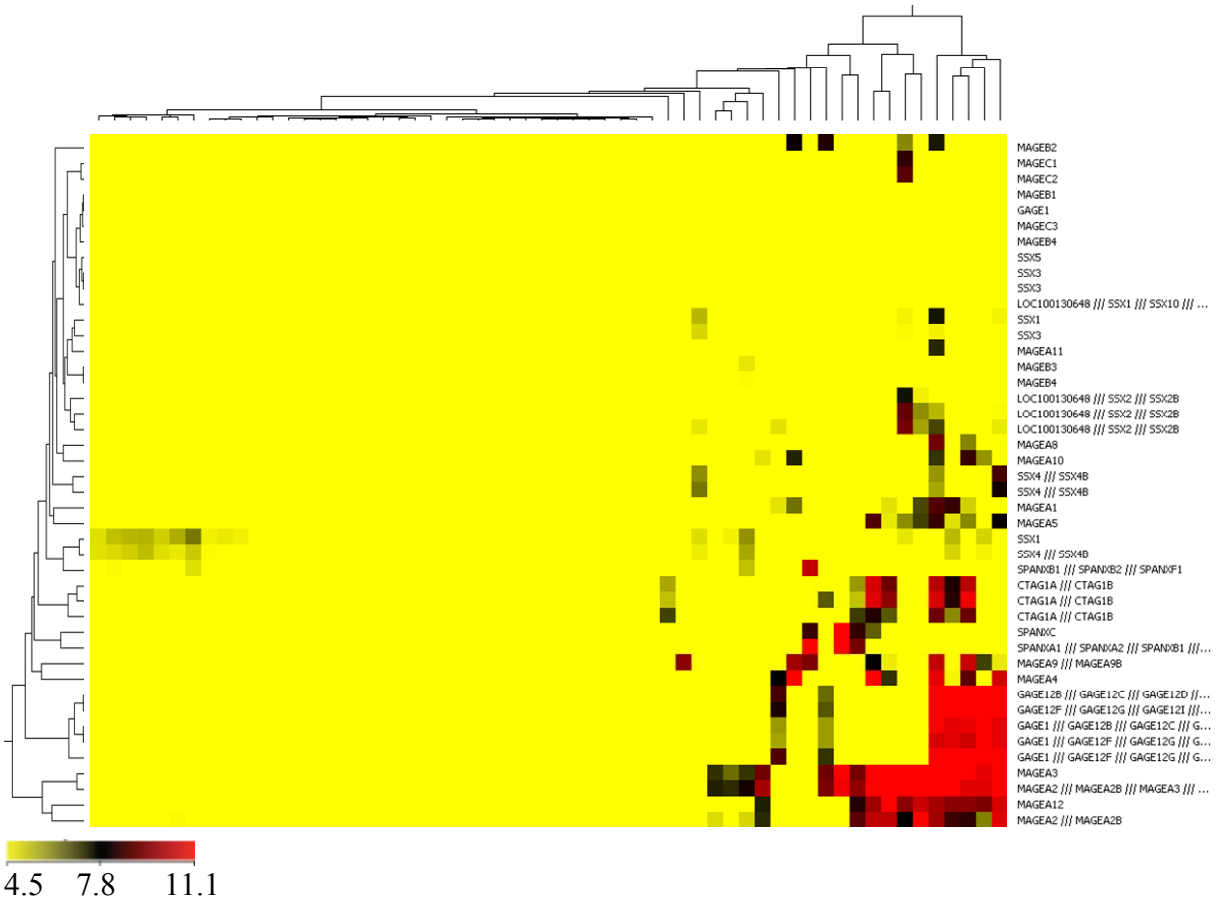


Figure 4.10: Hierarchical clustering analysis of lung adenocarcinoma tumors (GSE10072 dataset). The normalized expression values of 44 probesets corresponding to the selected CT-X genes were used to establish the hierarchical clustering using GeneSpring GX 9.06. Average linkage method and Euclidean distance as a matrix measure were used. Color code below the heatmap shows the range of normalized expression values. The tree at the top of the heatmap represents samplewise clustering.The tree on the left of the heatmap represents genewise clustering.

Lung and HCC tumors expressed CT-X genes at a higher frequency than ovarian and breast tumors, verifying previous reports on CT-X gene expression patterns of different tumors (Hoffman O 2008; Simpson AJ, 2005). In addition, the expression frequency of CT gene families was variable in tumor datasets: SSX genes were less frequently expressed in breast, lung, and ovarian tumors with the exception of HCC, whereas MAGE-A and GAGE genes were more frequently expressed in all tumors. All the members of one gene family were not expressed simultaneously in a dataset. This was expected for SSX gene family. It was previously shown that SSX1, 2 and 4 are expressed at substantial levels whereas SSX3, 5 and 6 are rarely expressed and SSX7, 8 and 9 expression are not detected among tumor tissues (Gure, Wei et al. 2002).

4.2.2 CT-X grouping of tumor and cell line datasets

To compare CT-X positive and negative samples by RankProd analysis, described in the following section, we first identified these samples as described in detail in **section 3.3.9.5**. Grouping was performed separately for cell line and tumor datasets. Using the combined data for tumor and cell line datasets, we determined an average rank value as the cut-off above which a sample would be considered positive for expression. The average rank value was slightly different for the combined data of cell line datasets and tumor datasets that were generated either by HG-U133A or HG-U133Plus2 arrays as shown in **Table 4.3**. In this way, we determined CT-X expression intensity for each array and generated three CT-X groups based on the clustering analysis for each dataset. The number of samples in CT-X positive, intermediate and negative groups are listed for cell line and tumor datasets in **Tables 4.4**, **4.5 and 4.6,** respectively.

Table 4.3: The average rank value of 3.5 in the combined data for cell line datasets and tumor datasets (HG-U133A and HG-U133Plus2)

	The average rank value
Cell line datasets (HG-U133A)	10856th rank (%48.7)
Tumor datasets (HG-U133A)	10001st rank (%44.9)
Tumor datasets (HG-U133Plus2)	28581st rank (%52.3

Table 4.4: The number of samples in CT-X positive, negative and intermediate groups for cancer cell line datasets

GSE number	**# Samples in CT-X positive group**	**# Samples in CT-X negative group**	**# Samples in CT-X intermediate group**
GSE4824/LCCL*	20	28	22
GSE5720/NCI-60 CLP**	24	25	11

*LCCL: lung cancer cell lines; **CLP: cell line panel

Table 4.5: The number of samples in CT-X positve, negative and intermediate groups for tumor datasets generated by HG-U133A arrays

GSE number/Array Express ID	**# Samples in CT positive group**	**# Samples in CT negative group**	**# Samples in CT intermediate group**
E-TABM-36/HCC*	16	37	4
GSE10072/LADC**	11	41	6
E-GEOD-GSE7390/BT[&]	14	164	20
GSE6008/OT[&&]	8	78	13

*HCC: Hepatocellular carcinoma, **LADC: Lung adenocarcinoma, [&]Node-negative breast tumors, [&&]Ovarian tumors

Table 4.6: Number of samples in CT-X positive, negative and intermediate groups for tumor datasets generated by HG-U133Plus2 arrays

GSE number	# Samples in CT positive group	# Samples in CT negative group	# Samples in CT intermediate group
GSE9843/HCC*	28	44	19
GSE9891/OT&&	38	209(42***)	38
GSE3141/LT**	33	52	26
GSE5460/ BT&	5	97(48***)	25

*HCC: Hepatocellular carcinoma, &&Ovarian tumors, **LT: Lung tumors, &Breast tumors,
*** The numbers of samples had to be reduced for meta-analysis due to memory limitations of the hardware utilized.

4.2.3 Meta-analysis of tumor datasets

The probesets with normalized expression values below the corresponding values of the mean rank value (**Table 4.3**) were filtered out from each array using an R script. Then common probesets were combined in a data matrix with an R script to be used in meta-analysis (**Appendix A**). After data pre-processing on HG-U133A platform, 19,421 probesets out of 22,283 probesets remained for the combined tumor data.

RankProd was utilized to identify differentially expressed genes between CT-X positive and negative groups. Meta-analysis of tumor datasets resulted in the identification of 1875 probesets that were up-regulated in the CT-X positive group compared to CT-X negative group, with a FC≥1.2; a P-value of <0.05 and 1881 down-regulated probesets satisfying the same criteria (**Table 4.7**).

Table 4.7: The number of probesets that were identified in the meta-analysis of tumor datasets (HG-U133A) with a FC≥1.2 and FC≥1.5 at 0.05 significance

	FC≥1.2, P<0.05	FC≥1.5, P<0.05
#probe sets down-regulated in CT-X positive group	1881	309
#probe sets up-regulated in CT-X positive group	1875	336

4.2.3.1 Validation of the rank-product method using tumor datasets generated using HG-U133Plus2 arrays

In order to test the reliability of the rank-product method we performed a second meta-analysis with RankProd using data from tumor samples generated using the HG-U133Plus2 platform. The datasets selected for validation analysis were from the same tumor types as

those used for generating the HG-U133A based data, namely, HCC, primary lung, breast and ovarian tumors, shown in **Table 3.8**.

To generate data from the HG-133Plus2 based datasets that could be comparable to that obtained from HG-U133A based datasets, we filtered in those probesets used for the initial meta-analysis with HG-U133A based data (corresponding to ~19,415 probesets) from HG-U133Plus2 based tumor data by an R based script. After we combined normalized data from individual datasets, we ran the RankProd package. We compared the meta-analysis of the HG-U133A and the filtered HG-U133Plus2 based data and identified 681 common probesets that were up-regulated in CT-X positive group compared to CT negative group, with a fold change (FC) ≥1.2 and an FDR (False Discovery Rate) or P-value of ≤0.05; and 824 common probesets that were down-regulated and satisfying the same criteria. **Table 4.8** lists the number of probesets identified in each meta-analysis and the number of probesets common to both.

Table 4.8: The number of probesets that were identified in the meta-analysis of HG-U133A and HG-U133Plus2 based data and the number of probesets that were common between them (FC≥1.2, P≤0.05).

	#down-regulated in CT-X positive	#up-regulated in CT-positive
Meta-analysis of the HG-U133A based data	1881	1875
Meta-analysis of the HG-U133Plus2 based data	1841	1807
Common number of probesets	824	681

4.2.3.2 DAVID functional annotation clustering analysis of common probesets between meta-analysis of HG-U133A and HG-U133Plus2 based data

In order to understand the biological meaning of the output probeset lists, we performed DAVID functional annotation clustering analysis of common probesets that were down-regulated as well as for those that were up-regulated, more than 1.2 fold (FC≥1.2, P≤0.05) in the CT-X positive group compared to the CT-X negative group. This analytic module of DAVID clusters functionally similar terms associated with the gene or probeset list into functional annotation groups and then ranks these groups by giving them enrichment scores (Huang da, Sherman et al. 2009). The enrichment score given to each functional annotation group is the geometric mean of all the P-values of each annotation term (GO term) in the

group. Since enrichment score 1.3 is equivalent to a P-value of 0.05, we focused on the annotation groups with scores ≥1.3.

DAVID functional annotation clustering analysis identified 10 annotation clusters with a fold enrichment of ≥1.3 using 688 DAVID IDs among 824 down-regulated common probesets with a FC ≥1.2 and a P-value of P≤0.05. The first three functional annotation groups with the highest enrichment score values are listed in **Table 4.9**. Percentage (%) indicates the ratio of DAVID IDs that belong to the annotation term (e.g. "~response to wounding") to total DAVID IDs.

Table 4.9: The functional annotation groups for down-regulated common probesets (FC≥1.2, p≤0.05) in the CT-X positive group compared to the CT-X negative group

Functional Group 1 / Enrichment score 4.91			
Term	**%**	**P-value**	**FDR***
GO:0009611~response to wounding	7.27%	6.50E-09	1.24E-05
GO:0009605~response to external stimulus	9.01%	5.42E-08	1.03E-04
GO:0006954~inflammatory response	4.80%	1.49E-05	0.028333967
GO:0006950~response to stress	11.34%	3.99E-05	0.075976292
GO:0006952~defense response	6.40%	5.31E-04	1.006668669
GO:0050896~response to stimulus	17.88%	0.031113311	45.25675988
Functional Group 2 /Enrichment score 4.69			
Term			
GO:0050817~coagulation	2.62%	4.18E-06	0.007970879
GO:0007596~blood coagulation	2.47%	1.54E-05	0.029440702
GO:0050878~regulation of body fluid levels	2.76%	2.51E-05	0.047791569
GO:0007599~hemostasis	2.47%	3.11E-05	0.059306511
GO:0042060~wound healing	2.62%	7.01E-05	0.133501166
Functional Group 3 / Enrichment score 3.23			
Term			
GO:0050727~regulation of inflammatory response	1.16%	1.68E-04	0.319095341
GO:0031347~regulation of defense response	1.16%	1.68E-04	0.319095341
GO:0050729~positive regulation of inflammatory response	0.73%	0.001225652	2.310702665
GO:0031349~positive regulation of defense response	0.73%	0.001225652	2.310702665
GO:0048583~regulation of response to stimulus	1.16%	0.001650018	3.098893738

*FDR: False Discovery Rate

DAVID functional annotation clustering analysis identified 21 annotation clusters with a fold enrichment of ≥1.3 using 570 DAVID IDs among 681 up-regulated common probesets with a FC≥1.2 and a P-value of P≤0.05. The first three functional annotation groups with the highest enrichment score values are listed in **Table 4.10.**

Table 4.10: The functional annotation groups for up-regulated common probesets (FC≥1.2, p≤0.05) in the CT-X positive group compared to the CT-X negative group

Functional Group 1 / Enrichment score 25.94			
Term	**%**	**P-Value**	**FDR**
GO:0000278~mitotic cell cycle	12.28%	5.04E-33	9.60E-30
GO:0022403~cell cycle phase	12.63%	5.56E-32	1.06E-28
GO:0000279~M phase	11.23%	1.54E-31	2.94E-28
GO:0007067~mitosis	9.82%	9.54E-30	1.82E-26
GO:0022402~cell cycle process	17.89%	1.02E-29	1.94E-26
GO:0007049~cell cycle	19.47%	1.27E-29	2.42E-26
GO:0000087~M phase of mitotic cell cycle	9.82%	1.70E-29	3.24E-26
GO:0051301~cell division	9.30%	9.42E-27	1.80E-23
GO:0000074~regulation of progression through cell cycle	10.35%	2.00E-12	3.82E-09
GO:0051726~regulation of cell cycle	10.35%	2.18E-12	4.15E-09
Functional Group 2 / Enrichment score 8.73			
Term			
GO:0000070~mitotic sister chromatid segregation	2.63%	4.46E-12	8.49E-09
GO:0000819~sister chromatid segregation	2.63%	8.02E-12	1.53E-08
GO:0007059~chromosome segregation	3.33%	8.14E-12	1.55E-08
GO:0030261~chromosome condensation	1.40%	4.52E-06	0.008616287
GO:0007076~mitotic chromosome condensation	1.23%	1.72E-05	0.032750855
Functional Group 3 / Enrichment score 7.04			
Term			
GO:0007017~microtubule-based process	7.19%	2.93E-19	5.58E-16
GO:0007051~spindle organization and biogenesis	2.63%	9.32E-15	1.78E-11
GO:0000226~microtubule cytoskeleton organization and biogenesis	3.86%	4.07E-13	7.76E-10
GO:0007010~cytoskeleton organization and biogenesis	9.47%	1.16E-11	2.20E-08
GO:0007018~microtubule-based movement	3.51%	1.18E-09	2.24E-06
GO:0030705~cytoskeleton-dependent intracellular transport	3.68%	9.23E-09	1.76E-05
GO:0046907~intracellular transport	8.25%	1.46E-04	0.277609046
GO:0051649~establishment of cellular localization	9.65%	1.57E-04	0.299349739
GO:0051641~cellular localization	9.65%	2.82E-04	0.536302863
GO:0051234~establishment of localization	16.32%	0.558041728	99.99998262
GO:0051179~localization	18.07%	0.751988594	100
GO:0006810~transport	14.74%	0.784476042	100

*FDR: False Discovery Rate

We also performed DAVID functional annotation clustering analysis with the probeset lists generated either by the meta-analysis of HG-U133A or HG-U133Plus2 based data, and found identical functional annotation groups with the highest enrichment score values between these two analyses (data not shown). Similarly, the common down- and up-regulated probesets gave us the same functional annotation groups that are listed in **Tables 4.9** and **4.10.** This indicated the consistency of RankProd analysis which could generate almost identical biological output even when different set of tumor samples were utilized.

According to the DAVID functional annotation clustering analysis results, the probesets related to immune response and coagulation were downregulated whereas the probesets related to mitosis, cell cycle and cytoskeleton organization and biogenesis were upregulated in the CT-X positive group as compared to the CT-X negative group. It seems that tumors expressing CT-X genes have a higher proliferation rate and metastatic capacity (in regard to cytoskeleton organization and biogenesis) than tumors that are not expressing CT-X genes.

4.2.4 Meta-analysis of cancer cell line datasets

After data pre-processing on HG-U133A platform (explained in **section 3.3.9.6.1**), 18,411 probe sets remained for the combined cell line data. RankProd was utilized to identify differentially expressed genes between CT-X positive and negative groups.

Meta-analysis of cell line datasets resulted in the identification of 1211 probesets that were up-regulated in CT-X positive group compared to CT-X negative group, with a FC≥1.2; a P-value of ≤ 0.05 and 1539 down-regulated probesets satisfying the same criteria (**Table 4.11**).

Table 4.11: The number of probesets that were identified in the meta-analysis of cell line datasets with a FC≥1.2 and FC≥1.5 at 0.05 significance.

	FC≥1.2, P≤0.05	FC≥1.5, P≤0.05
#probe sets down-regulated in CT-X positive group	1539	816
#probe sets up-regulated in CT-X positive group	1211	615

We compared the meta-analysis of cancer cell line datasets and tumor datasets both of which were generated by using HG-133A arrays and identified 227 common probesets that were up-regulated in CT-X positive group compared to CT-X negative group, with a fold change (FC) ≥1.2 and a P-value of ≤0.05; and 444 common probesets that were down-regulated and satisfying the same criteria.

We utilized DAVID functional annotation clustering annalysis of probesets that were down-regulated and up-regulated more than 1.5 fold change (FC≥1.5, P≤0.05) in CT-X positive group compared to CT-X negative group, respectively. This analysis identified 7 annotation clusters with a fold enrichment ≥1.3 using 501 DAVID IDs among 615 up-regulated probesets with a FC≥1.5 and a P-value of p≤0.05. The first four functional annotation groups with enrichment scores ≥1.3 are listed in **Table 4.12**. Percentage (%) indicates the ratio of

DAVID IDs that belong to the annotation term (e.g. cellular localization) to total DAVID IDs.

Table 4.12: The functional annotation groups for up-regulated probesets (FC ≥ 1.5, p≤0.05) in the CT-X positive group compared to the CT-X negative group

Functional Group 1 / Fold enrichment 3.26			
Term	**%**	**P-Value**	**FDR**
GO:0051641~cellular localization	11.60%	2.08E-06	0.003973984
GO:0051649~establishment of cellular localization	11.20%	4.68E-06	0.008915578
GO:0016043~cellular component organization and biogenesis	25.60%	1.91E-05	0.036356043
GO:0046907~intracellular transport	9.20%	2.63E-05	0.050123855
GO:0032940~secretion by cell	4.60%	3.26E-04	0.61937338
GO:0051179~localization	25.40%	5.84E-04	1.107972291
GO:0008104~protein localization	8.80%	7.29E-04	1.379625832
GO:0045184~establishment of protein localization	8.20%	0.001475057	2.774665446
GO:0033036~macromolecule localization	8.80%	0.00227037	4.240262173
GO:0051234~establishment of localization	21.40%	0.004359533	7.991088354
GO:0015031~protein transport	7.20%	0.007548161	13.44859852
GO:0006886~intracellular protein transport	4.80%	0.011383711	19.6071726
GO:0046903~secretion	4.60%	0.01263466	21.52429841
GO:0006810~transport	20.00%	0.01295533	22.00871527
Functional Group 2 / Fold enrichment 3.19			
Term			
GO:0006888~ER to Golgi vesicle-mediated transport	2.20%	5.13E-05	0.097786048
GO:0048193~Golgi vesicle transport	2.60%	3.18E-04	0.605192487
GO:0032940~secretion by cell	4.60%	3.26E-04	0.61937338
GO:0045045~secretory pathway	3.80%	9.92E-04	1.873650133
GO:0016192~vesicle-mediated transport	6.40%	0.001059073	1.999651131
GO:0046903~secretion	4.60%	0.01263466	21.52429841
Functional Group 3 / Fold enrichment 2.26			
Term			
GO:0046148~pigment biosynthetic process	1.40%	0.001101871	2.07965719
GO:0006583~melanin biosynthetic process from tyrosine	0.80%	0.001202233	2.267028356
GO:0006582~melanin metabolic process	0.80%	0.001202233	2.267028356
GO:0042438~melanin biosynthetic process	0.80%	0.001202233	2.267028356
GO:0042440~pigment metabolic process	1.40%	0.002176929	4.069161211
GO:0019748~secondary metabolic process	1.40%	0.008135582	14.41993726
GO:0009072~aromatic amino acid family metabolic process	1.00%	0.009161013	16.09085073
GO:0006570~tyrosine metabolic process	0.80%	0.011028345	19.05451945
GO:0006725~aromatic compound metabolic process	2.20%	0.01353118	22.87151411
KEGG_PATHWAY_hsa00350:Tyrosine metabolism	0.60%	0.590723139	99.99873526
Functional Group 4 / Fold enrichment 1.95			
Term			
GO:0007049~cell cycle	9.80%	0.001762793	3.307340064
GO:0000279~M phase	4.20%	0.002046006	3.828940052
GO:0007067~mitosis	3.60%	0.002251613	4.205938394
GO:0000087~M phase of mitotic cell cycle	3.60%	0.00250899	4.675890718
GO:0022403~cell cycle phase	4.80%	0.003433177	6.345445001
GO:0022402~cell cycle process	8.00%	0.010295896	17.90406135
GO:0000278~mitotic cell cycle	4.00%	0.019215595	30.91689662

GO:0051301~cell division	2.60%	0.097904121	85.9715428
GO:0000074~regulation of progression through cell cycle	5.00%	0.159942087	96.39293745
GO:0051726~regulation of cell cycle	5.00%	0.160815004	96.46371961

*FDR: False Discovery Rate

DAVID functional annotation clustering analysis found 4 annotation clusters with a fold enrichment ≥1.3 using 710 DAVID IDs among 816 down-regulated probesets with a FC ≥1.5 and a P-value of p≤0.05. The first three functional annotation groups were listed in **Table 4.13.**

Table 4.13: The functional annotation groups for down-regulated probesets (FC ≥ 2.0, p≤0.05) in the CT-X positive group compared to the CT-X negative group

Functional Group 1 / Fold enrichment 6.53			
Term	**%**	**P-Value**	**FDR**
GO:0002376~immune system process	11.83%	3.19E-09	6.08E-06
GO:0006955~immune response	8.87%	2.83E-07	5.39E-04
GO:0050896~response to stimulus	21.69%	2.79E-05	0.053169586
Functional Group 2 / Fold enrichment 4.93			
Term			
GO:0048519~negative regulation of biological process	13.94%	2.38E-07	4.55E-04
GO:0048523~negative regulation of cellular process	13.10%	1.49E-06	0.002837092
GO:0009892~negative regulation of metabolic process	5.07%	0.004607553	8.427016603
Functional Group 3 / Fold enrichment 4.52			
Term			
GO:0012501~programmed cell death	10.28%	7.42E-07	0.00141389
GO:0006915~apoptosis	10.14%	1.08E-06	0.002050884
GO:0043067~regulation of programmed cell death	7.75%	2.13E-06	0.004065203
GO:0030154~cell differentiation	18.31%	2.35E-06	0.004485707
GO:0048869~cellular developmental process	18.31%	2.35E-06	0.004485707
GO:0008219~cell death	10.42%	2.75E-06	0.005247704
GO:0016265~death	10.42%	2.75E-06	0.005247704
GO:0042981~regulation of apoptosis	7.61%	3.43E-06	0.006535566
GO:0048468~cell development	13.52%	7.85E-06	0.014967071
GO:0043065~positive regulation of apoptosis	4.51%	1.24E-05	0.023565591
GO:0043068~positive regulation of programmed cell death	4.51%	1.47E-05	0.027949661
GO:0006917~induction of apoptosis	3.66%	1.39E-04	0.264801062
GO:0012502~induction of programmed cell death	3.66%	1.51E-04	0.287157479
GO:0043069~negative regulation of programmed cell death	2.96%	0.023217395	36.09659327
GO:0043066~negative regulation of apoptosis	2.82%	0.038411808	52.6050396
GO:0006916~anti-apoptosis	2.11%	0.067378006	73.54439926

*FDR: False Discovery Rate

We observed nearly the same pattern for cancer cell lines as we observed for tumors with a few exceptions. Interestingly, the probesets that function to induce apoptosis were downregulated in the CT-X positive group as compared to the CT-X negative group.

4.2.5 Clustering analysis of probesets that were identified by meta-analysis of cell line datasets in GSE4824 lung cancer cell line dataset

In meta-analysis, we simultaneously analyzed datasets from tumors and tumor cell lines originating from various tissues as reported in previous sections. We showed coordinate expression of CT-X genes in our datasets (**Figures 4.9 and 4.10**). By meta-analysis, we primarily aimed to find CT-X regulatory mechanisms that are common and drive CT-X gene expression in cancer, regardless of tissue origin. However, we also consider different cancer types might have unique mechanisms underlying CT-X gene expression. We, therefore, wanted to test if the probesets identified by meta-analysis could discriminate CT-X positive and negative samples when datasets were analysed individually. Therefore, we selected 25 probesets ($p<0.0001$) with the highest fold change values from the up-regulated and down-regulated probeset lists that were generated by the RankProd analysis of cell line datasets and then we performed a sample based clustering analysis with these probesets on the same samples used for the meta-analysis from one of the datasets (GSE4824). Indeed, hierarchical clustering analysis showed that these 25 selected probesets could not successfully classify lung cancer cell lines into the CT-X positive and negative groups although there was accumulation of samples of a given class in a particular cluster; this inability of discrimination might be due to variable expression pattern of probesets between samples (**Figures 4.11** and **4.12**). Some of these probesets also might be unique in their expression to other dataset used (GSE5720) in meta-analysis.

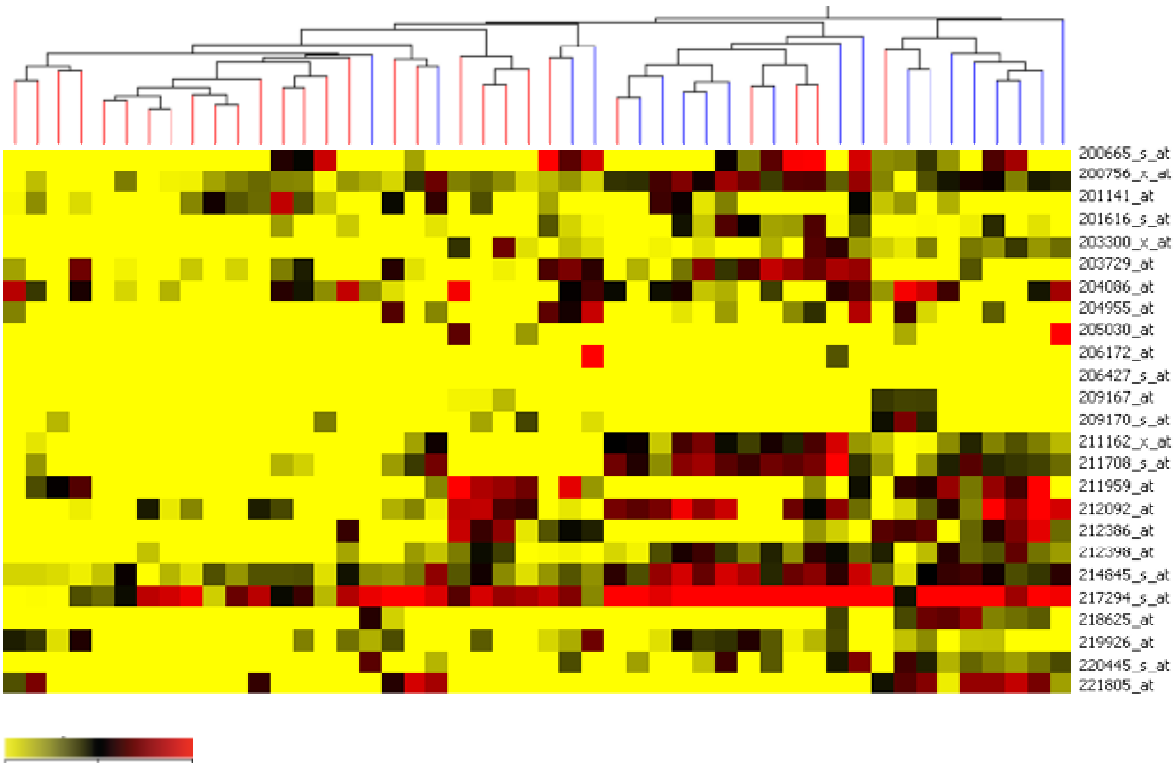


Figure 4.11: Hierarchical clustering analysis of CT-X positive and CT-X negative lung cancer cell lines (GSE4824) The normalized expression values of 25 probe sets ($p<0.0001$) up-regulated in CT-X positive lung cancer cell lines as relative to CT-X negative lung cancer cell lines were used to establish a sample based hierarchical clustering using GeneSpring GX 9.06. Average linkage method and Euclidean distance as a matrix measure were used. Color code below the heatmap shows the range of normalized expression values.The tree at the top of the heatmap represents samplewise clustering. Blue indicates CT-X positive cell lines wheras red indicates CT-X negative cell lines.

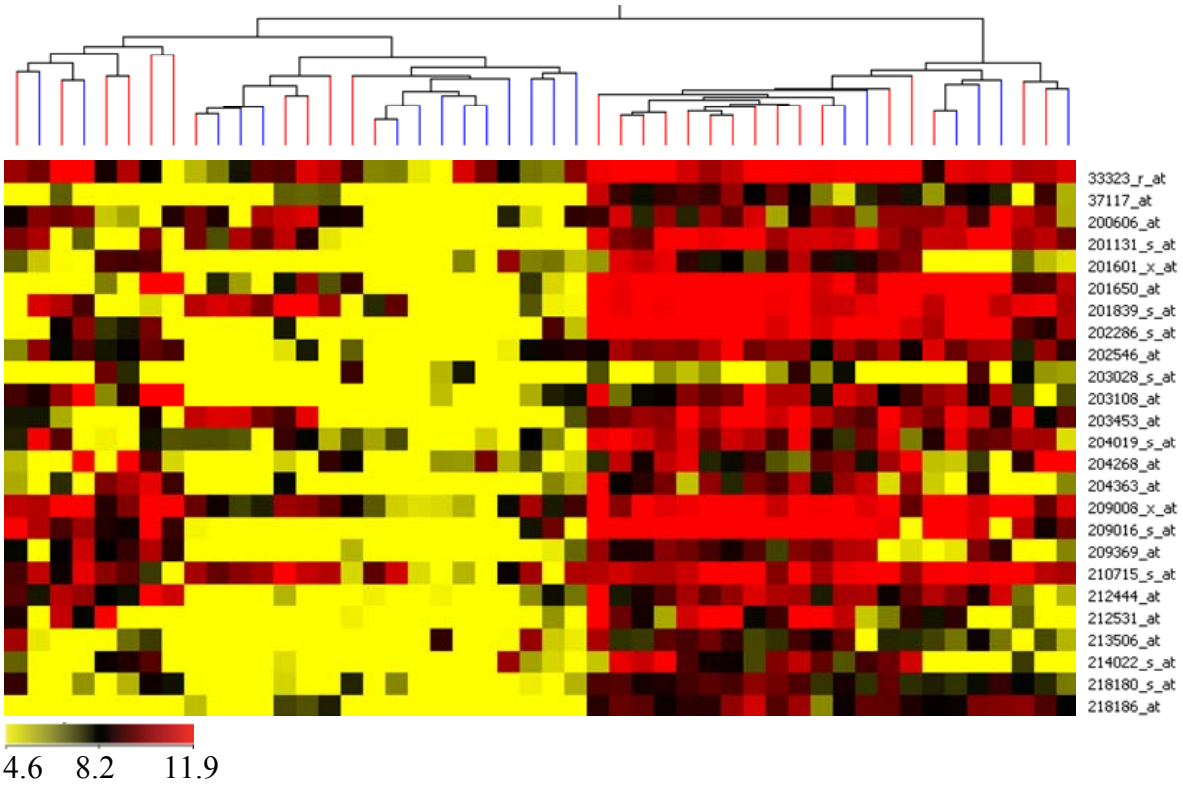


Figure 4.12: Hierarchical clustering of CT-X positive and CT-X negative lung cancer cell lines (GSE4824). The normalized expression values of 25 probe sets ($p<0.0001$) down-regulated in CT-X positive group compared to CT-X negative group were used to establish a sample based hierarchical clustering using GeneSpring GX 9.06. Average linkage method and Euclidean distance as a matrix measure were used. Color code below the heatmap shows the range of normalized expression values. The tree at the top of the heatmap represents samplewise clustering. Blue indicates CT-X positive cell lines wheras red indicates CT-X negative cell lines.

4.2.6 Class prediction analysis of GSE4824 lung cancer cell line dataset via BRB Array Tools

The fact that the probesets that were most significant in meta-analysis are not perfect identifiers of CT-X positive and negative samples could likely be due to dataset-specific differences of the fold change (FC) value of those probesets identified by the meta-analysis. This could be because of the underlying algorithm of the RankProd analysis (**section 3.3.9.6.2**); as it first determines the expression difference (FC) of a given probeset by pair-wise comparison within arrays in each dataset and calculates the rank product according to the rank of each FC value. Then, it combines the rank products from different datasets and determines the signifance of this rank product. Therefore, it is likely that an FC value of a probeset given in the end might be one that belongs to the most significant rank product and it might not reflect the actual fold change value which is expected to be more substantial. Moreover, the RankProd analysis relies on determination of an FC value without considering the correlation structure among probesets thus it is not sensitive to the variability around the FC values associated with each dataset. On the other hand, classification methods (class prediction) such as support vector machines take advantage of the relationships among the samples used in the analysis to extract the best discriminating combination of probesets maximizing the separation between groups, i.e., herein CT-X positive and negative samples. We, therefore, decided to perform class prediction analysis in the lung cancer cell line dataset, to reveal differences of gene expression that could clearly classify lung cancer cell line samples based on CT-X expression. We hoped that this type of analysis would reveal additional differences that could be validated by RT-PCR subsequently.

The raw data of the GSE4824 dataset were normalized with GC-RMA algorithm using BRB Array Tools. Class prediction analysis was carried out at 0.001 significance using CT-X positive and CT-X negative lung cancer cell lines according to our previous classification (**Table 4.4**). All the predictors of the class prediction analysis including nearest centroid, compound covariate, support vector machines correctly discriminated CT-X positive and CT-

X negative lung cancer cell lines giving %96 percent of classification on average (**section 3.3.0.7**). Class prediction analysis resulted in the identification of 88 up-regulated and 48 down-regulated probesets at 0.001 significance. Although excluding CT-X genes from the class prediction analysis reduced the percent of classification on average of CT-X positive and negative lung cancer cell lines, almost the same probeset list with the same number of probesets was generated (data not shown). In the heatmap generated by clustering analysis with these probesets, CT-X negative (as shown by **0**) and positive (as shown by **1**) cell lines could be clearly differentiated. Interestingly, the probesets classified a subgroup of CT-X negative samples together with CT-X positive samples, but not the other way arround, possibly suggesting that mechanisms underlying CT-X gene expression might effect the cells in a two-tiered fashion, where some samples induce CT-X gene expression upon this mechanistic alteration, while others do not.

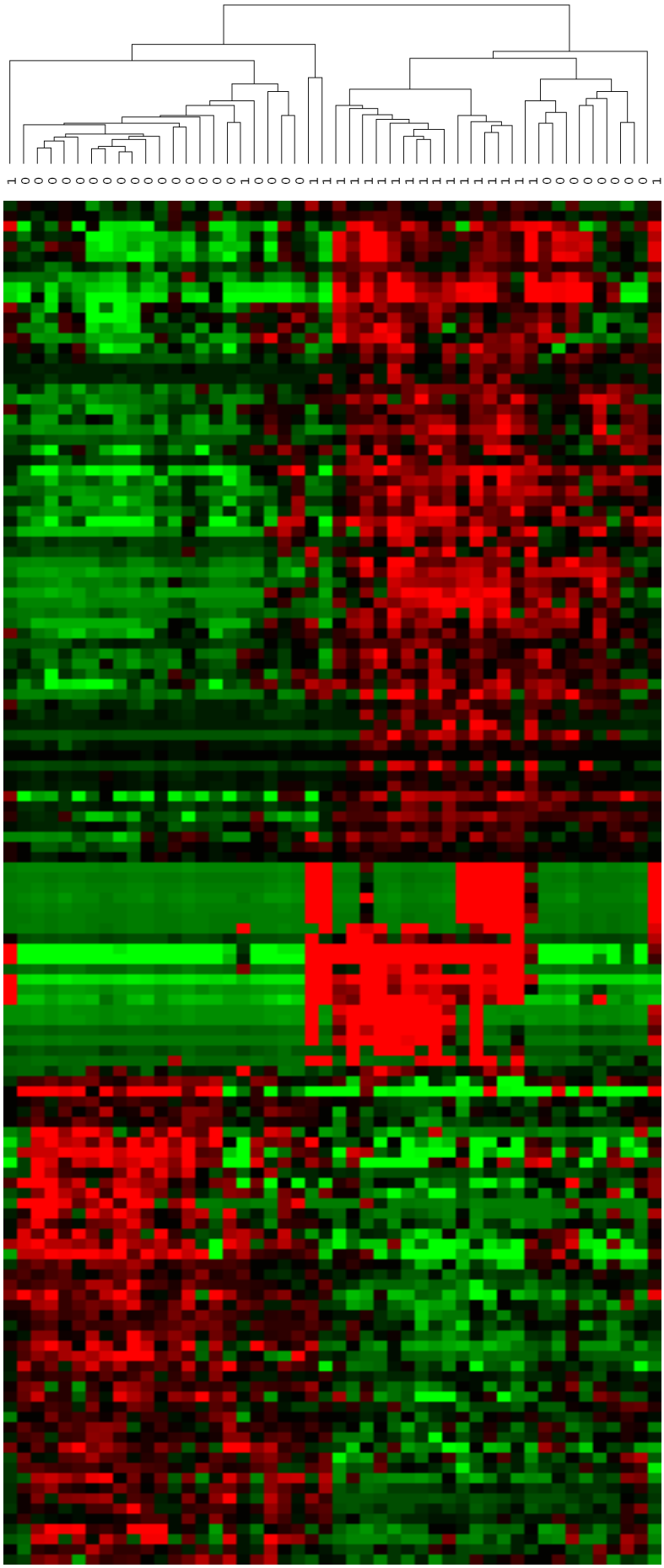


Figure 4.13: Hierarchical clustering of CT-X positive and CT-X negative lung cancer cell lines using the probesets generated by the class prediction analysis The mean-centered normalized expression values of 88 up-regulated and 48 down-regulated probesets were used to establish a hierarchical clustering using BRB Array Tools. Average linkage method and spearman rank correlation as a matrix measure were used. Red indicates upregulation wheras green indicates downregulation. At the top of the heatmap, CT-X positive cell lines are shown by **1** wheras CT-X negative cell lines are shown by **0**.

4.2.6.1 DAVID Functional annotation clustering analysis of probesets found by the class prediction analysis and selection of the probesets for validaton in lung cancer cell lines

By using DAVID tools, we investigated the functions of individual probesets that were generated at 0.001 significance by class prediction analysis of lung cancer cell lines. We found that a subset of genes that were upregulated in the CT-X positive group as relative to the CT-X negative group, which are expressed in different cancer cells and associate with increased proliferation rate, metastasis and angiogenesis (Yang, Mani et al. 2004; Kasper, Vogel et al. 2005; Scaglia and Igal 2005; Muchemwa, Nakatsura et al. 2008; Scaglia and Igal 2008; Shea, Wells et al. 2008). In addition, some of them confer drug resistance or sensitivity to cancer cells (Atienza, Roth et al. 2005; Kasper, Vogel et al. 2005; Zhang, Wang et al. 2009). However, these genes were not enriched in the CT-X negative group.

Of the 88 up-regulated and 48 down-regulated probe sets identified by the class-prediction analysis, 67 up-regulated and 30 down-regulated probesets were identical to that obtained by our meta-analysis We take this to further support the validity of the meta-analysis. We chose 8 up-regulated probesets and 1 down-regulated probeset that behaved commonly in both meta-analysis and class prediction analysis to be validated in lung cancer cell lines. **Table 4.14** lists these probesets with fold changes and p-values given by meta-analysis and class pediction analysis and their associated biological process in normal tissues.

Table 4.14: The probesets selected for validation in lung cancer cell lines

		Class prediction analysis**		Meta-analysis*		
Gene symbol	**Gene name**	**Fold change**	**P -value**	**Fold change**	**P- value**	**Biological process**
SCD	stearoyl-CoA desaturase (delta-9-desaturase)	8.0472729	0.0002313	4.1718815	0	Fatty acid biosynthetic process
RAD21	RAD21 homolog (S. pombe)	5.2286351	0.0005711	1.9766752	0	DNA repair
HSPH1	heat shock 105kDa/110kDa protein 1	3.9502552	0.0004841	2.0466639	0	Protein metabolic process&response to unfolded protein
HSP90B1	heat shock protein 90kDa beta (Grp94), member 1	3.4590574	0.0001664	1.805706	0	RNA processing&RNA binding
LAPTM4B	lysosomal protein transmembrane 4 beta	2.5006464	0.0008625	1.9290123	0	Transport / Localization
NFYC	nuclear transcription factor Y, gamma	2.2005275	0.0009699	1.601794	0.0001	Regulation of transcription
SSRP1	structure specific recognition protein 1	2.1555008	0.0005962	1.4148274	0.0041	Transcription&DNA replication&DNA repair
TWIST1	twist homolog 1 (Drosophila)	2.1365362	0.0009206	2.1834061	0	Transcription
LIMK2	LIM domain kinase 2	-3.0896689	0.0005682	-2.0033	0	Protein modification process (protein kinase activity

*Meta-analysis of 2 cancer cell line datasets (**section 4.2.4**)
**Class prediction analysis of GSE4824 lung cancer cell line dataset

4.3 Expression analysis of four CT-X genes in lung, colon, breast and HCC cancer cell lines to determine CT-X positive and negative cell lines

Validation of the expression of the above described genes identified commonly by the meta-analyis and class prediction analysis requires that we test tissues or cell lines that have been thoroughly caracterized with regard to their CT-X expression patterns. We, therefore, quantified the relative expression of four CT-X genes; namely SSX4A/SSX4B, CTAG1A/CTAG1B (NY-ESO-1), MAGEA3 and multiple GAGE genes using specific Taqman MGB (FAM labelled) probes by qPCR. We used ΔΔCt values of SK-LC-17 samples as the reference. cDNA from normal colon tissue that is known not to express any of the CT-X genes was used as a negative control.

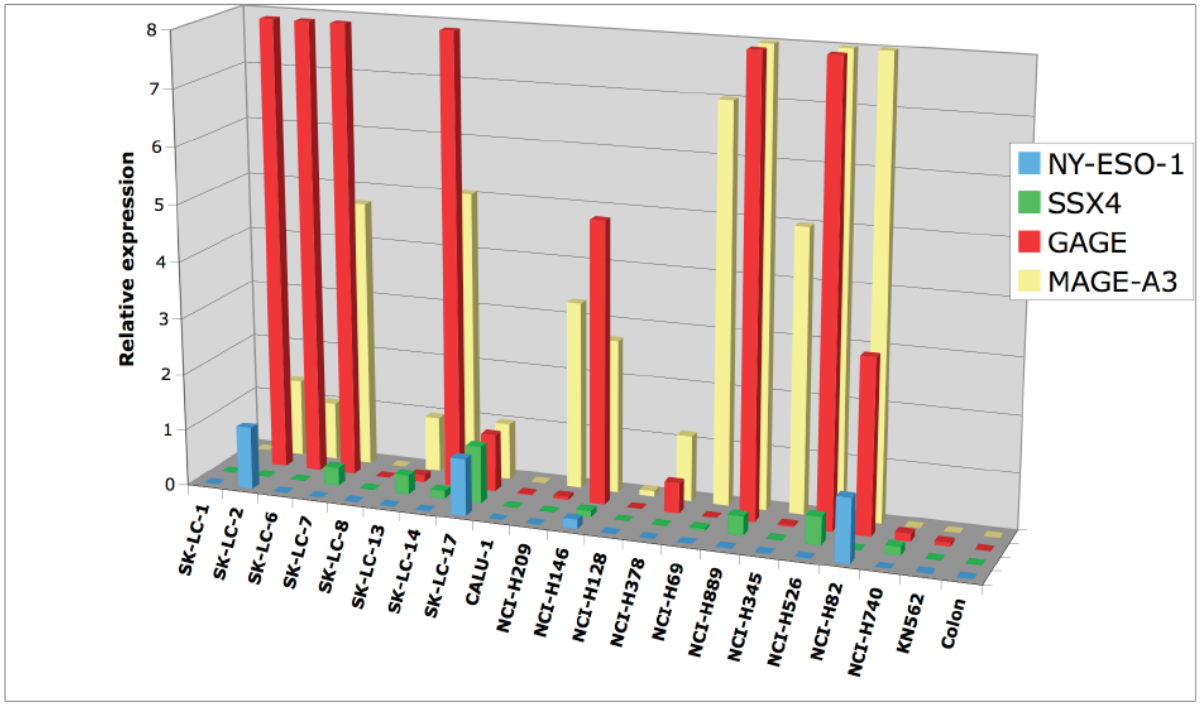


Figure 4.14: Relative expression of SSX4A/SSX4B, CTAG1A/CTAG1B (NY-ESO-1), MAGEA3 and multiple GAGE genes in lung cancer cell lines. Expression of CT-X genes was checked by qPCR using specific Taqman MGB probes in lung cancer cell lines. Relative expression was calculated using ΔΔCt method and CT-X expression in SK-LC-17 was taken as the reference. Colon cDNA was used as a negative control to show that CT-X genes are not expressed in normal tissues.

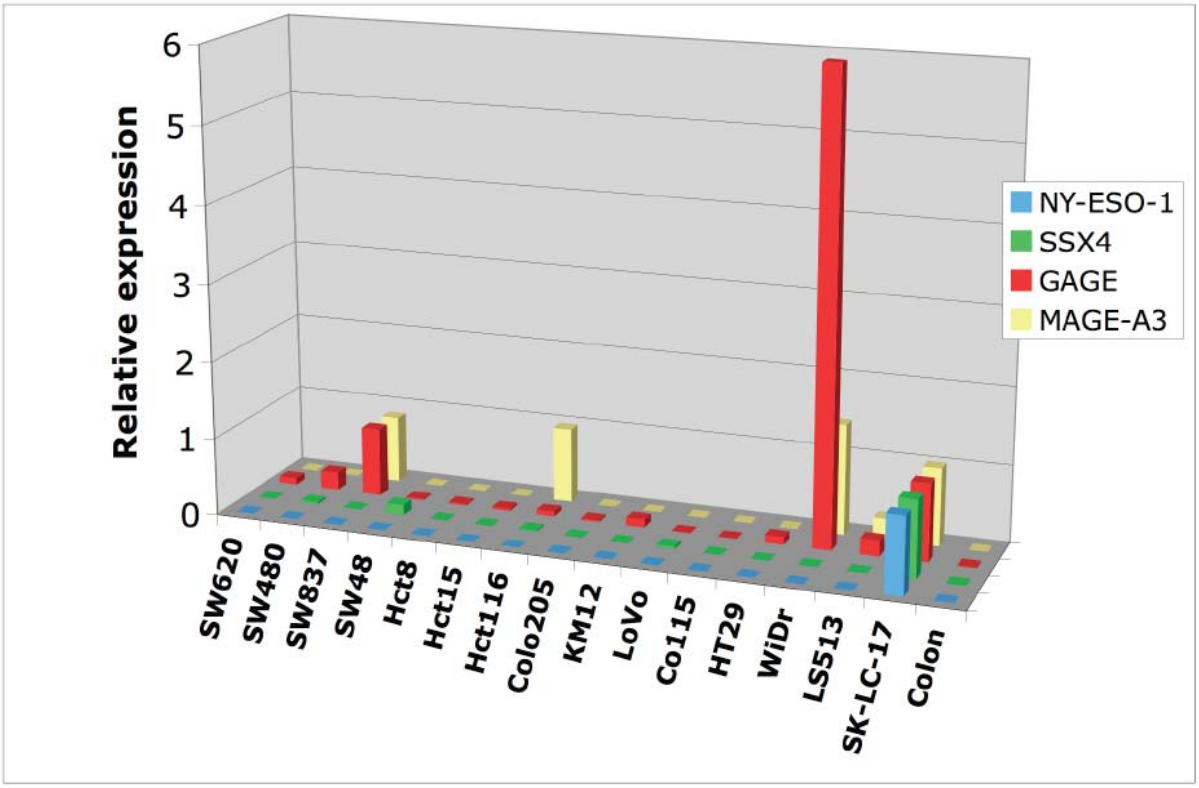


Figure 4.15: Relative expression of SSX4A/SSX4B, CTAG1A/CTAG1B (NY-ESO-1), MAGEA3 and multiple GAGE genes in colon cancer cell lines. Expression of CT-X genes was checked by qPCR using specific Taqman MGB probes in breast cancer cell lines. Relative expression was calculated using ΔΔCt method and CT-X expression in SK-LC-17 was taken as the reference. Colon cDNA was used as a negative control to show that CT-X genes are not expressed in normal tissues.

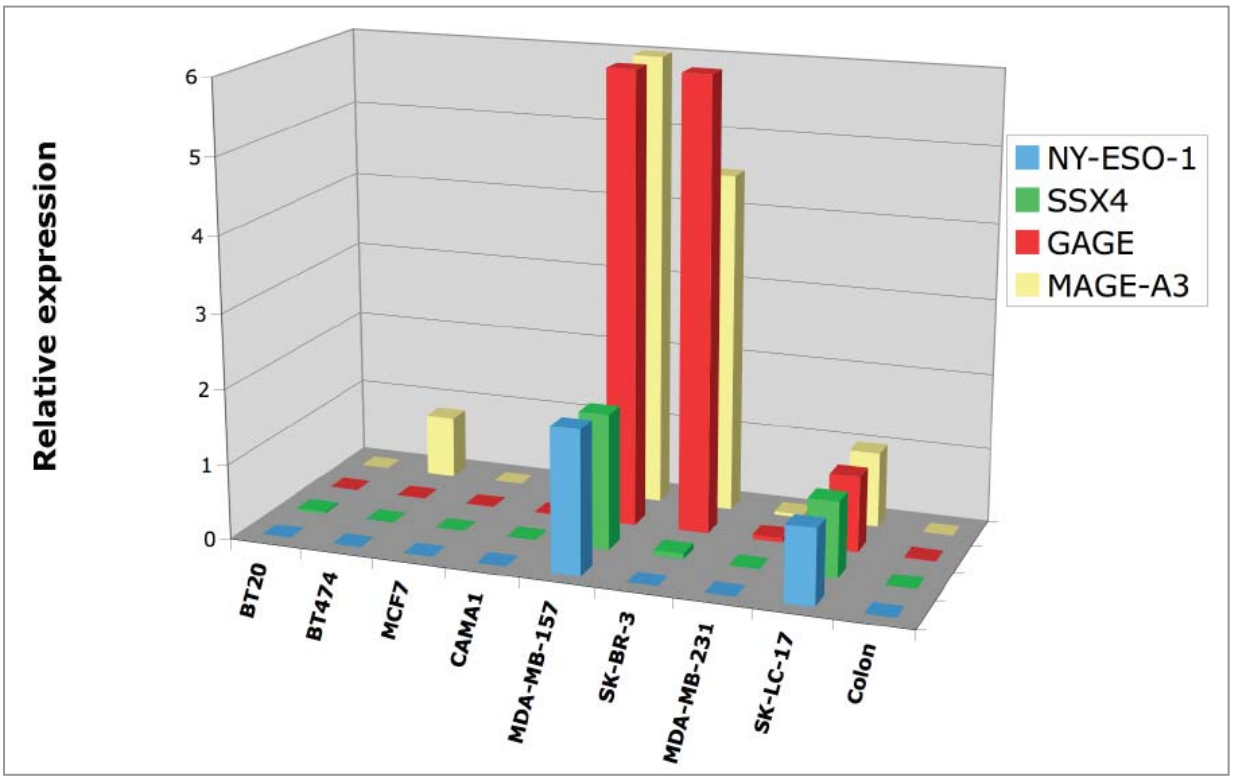


Figure 4.16: Relative expression of SSX4A/SSX4B, CTAG1A/CTAG1B (NY-ESO-1), MAGEA3 and multiple GAGE genes in breast cancer cell lines. Expression of CT-X genes was checked by qPCR using specific Taqman MGB probes in colon cancer cell lines. Relative expression was calculated using ΔΔCt method and CT-X expression in SK-LC-17 was taken as the reference. Colon cDNA was used as a negative control to show that CT-X genes are not expressed in normal tissues.

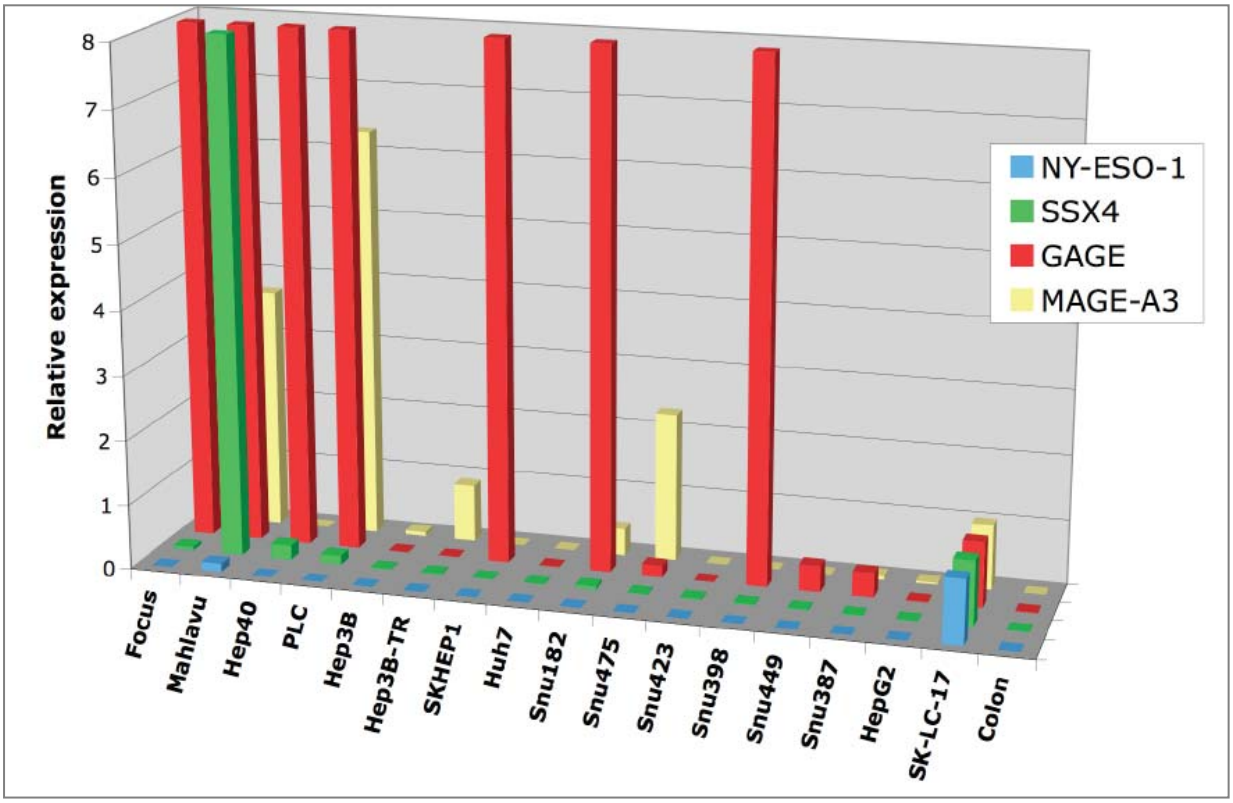


Figure 4.17: Relative expression of SSX4A/SSX4B, CTAG1A/CTAG1B (NY-ESO-1), MAGEA3 and multiple GAGE genes in HCC cell lines. Expression of CT-X genes was checked by qPCR using specific Taqman MGB probes in HCC cell lines. Relative expression was calculated using ΔΔCt method and CT-X expression in SK-LC-17 was taken as the reference. Colon cDNA was used as a negative control to show that CT-X genes are not expressed in normal tissues.

Current evidence indicates that lung cancer and HCC express CT-X genes at high levels whereas breast expresses CT-X genes at moderate levels and colon expresses them at low levels (Hofmann, Caballero et al. 2008). According to our results, CT-X genes were expressed at different frequencies within cell lines of lung, breast, colon cancers and HCC. Lung cancer cell lines expressed CT-X genes at the highest level when compared to the cell lines of other tissues. We found that GAGE, in addition to MAGEA3, was the most frequently expressed CT-X genes in all cancer types tested.

5 DISCUSSION & FUTURE PERSPECTIVES

In this study, we aimed to develop two approaches by which the mechanisms underlying the regulation of CT-X gene expression in cancer could be identified. CT-X genes are unique in that their expression is specific to tumors in adults and is correlated with tumor progression. They are expressed in a coordinate manner in tumors, supporting a common mechanism for their activation. The key event in the activation of CT-X genes in testis and during tumorigenesis is promoter-specific demethylation. However, the exact mechanisms targeting DNA hypomethylation to CT-X promoters remain to be described. Elucidating the mechanisms responsible for the reactivation of CT-X genes during tumorigenesis will not only contribute to the understanding of their role in tumorigenesis but also will bring on with novel therapeutic targets and prognostic indicators.

5.1 Generation of an SSX4 knock-in cell line

We used an SSX4 KI targeting vector in order to generate an SSX4 KI cell line with a GFP reporter gene that is expressed from SSX4 promoter. We first tested expression of GFP from SSX4 promoter cloned into the KI vector by transient transfection of SK-LC-17 cells with the KI vector lacking β-actin DTA (referred to as "Step 6"). Flow cytometry analysis of the Step6 transfected SK-LC-17 cells showed that the SSX4 promoter as contained within the KI vector was functional and expressing GFP. Afterwards, we went on with stable transfection of SK-LC-17 cells with the KI vector and obtained 70 stable clones. Among 70 stable clones, clone #70 could be identified as 100% GFP positive and the rest were either GFP negative and showed partial GFP expression. Among the clones (41 clones) with heterogeneous GFP expression, we chose the ones containing a high percentage of GFP expressing cells. In order to determine correct insertion of the KI vector, we screened the clones by nested PCR using the forward primers that are homologous to the SSX4 5' sequence in the promoter-proximal region corresponding to the genomic DNA of the cell, combined with the reverse primers homologous to EGFP sequence. We could observe neither an amplified product nor a non-specific product in the first run of nested PCR. The 2^{nd} run of PCR reaction using the primary products yielded specific amplicons for 7 of the 42 stable clones. Then, 6 of 7 clones (#14, 21, 70, 77, 85 and 96) were sequence-verified.

We observed a heterogeneous GFP expression for SSX4 KI clones with correctly inserted KI vector except Clone #70 in the initial screening by flow cytometry. We first thought that this partial GFP expression could result from the fact that the KI vector might be inserted into another SSX gene in the SSX gene family. It was highly probable that the KI construct might be integrated into SSX7 since the 3' SSX4 homology sequence was derived from SSX7 which is highly homologous to SSX4 but not expressed in either testis or cancer cells. Then, the clones first might have GFP expression due the activity of inserted SSX4 promoter and they might have lost GFP expression by time. However, in consequent experiments we obtained stable clones with correctly inserted KI vector which had also heterogeneous GFP expression. This challenged our previous idea. We repeated the flow cytometry analyses with these clones in order to see whether these clones were stably expressing GFP. We observed that clone #85 and #21 lost their GFP expression and only minor differences were seen for clones #14, #70, #77 and #96 in terms of percentage of GFP expressing cells and GFP expression intensity. In both flow cytometry analyses, we used parental SK-LC-17 cells as the negative control and determined the cutoff for GFP expression intensity based on that observed for these cells. As it is noticed in **Tables 4.1** and **4.2**, the background fluorescence (M1) of SK-LC-17 cells differed between two flow cytometry analyses and these minor differences in terms of GFP expressing cells and GFP expression intensity might result from this. Even though #85 and #21 were correct clones, we were surprised that they lost their GFP expression. The sequencing of the amplified products with the M4 reverse primer only yielded 152 bp sequence information corresponding to the minimal SSX4 promoter. If there occurred any mutation upstream of this obtained sequence during homologous recombination that would result in the repression of GFP in these clones, we would not be able to see that. Therefore, we would have to obtain more sequence information on the clones which currently seem to have correctly integrated KI vector.

We quantified SSX4 gene expression in these clones by qPCR in order to differentiate whether SSX4 KI construct integrated into one of the SSX4 genes (SSX4; SSX4B) or both of them, which are identical copies of the same gene located head-to-head orientation in the Xp11.23. We thought that if the KI construct integrated into one of them, the clones would have SSX4 expression reduced by %50. According to qPCR results, Clone #21 and #85 had reduced SSX4 expression relative to SK-LC-17 whereas others expressed SSX4 even at higher levels than SK-CL-17. Since these clones were generated by single-cell cloning, the individual clones might have varying SSX4 expression due to the heterogeneity of cell lines.

However, since clones were still expressing SSX4, we could say that the KI construct integrated into one of the SSX4 genes. Interestingly, the clones (#85 and #21) which lost their GFP expression also expressed SSX4 at lower levels as compared to other clones. This led us to think that these clones somehow might lose the expression of one of the SSX4 genes. It should be clarified why and how these clones lost expression of one of the SSX4 genes.

Since clone #70 was classified as 100% GFP positive, we would go on with this clone. We want to transfect clone #70 with a CT-X negative cDNA library and select the clones with repressed GFP expression by flow cytometry. Sequencing of the genomic DNA with the primers, which are designed to obtain the integrated cDNA, would result in identification of the gene that causes repression of SSX4 promoter. In this strategy, our aim is to repress SSX4; thereby theoretically we should obtain some clones with methylated SSX4. However, this would not be so easy considering the fact that cancer cells might lose the factors that target DNMTs to the promoters of CT-X genes (**Figure 1.1**) and if this was the case SSX4 would not be repressed in these clones due to the absence of these factors. On the other hand, we could identify cDNAs that might cause SSX4 repression by other mechanisms that indirectly cause DNA methylation of CT-X genes.

5.2 Meta-analysis of cell line and tumor datasets

Based on the fact that CT-X gene expression occurs coordinately in cancer cells (Gure, Chua et al. 2005; Simpson, Caballero et al. 2005), we utilized a meta-analysis approach to identify differentially expressed genes between CT-X expressing (CT-X positive) and non-expressing (CT-X negative) cancer cells. We developed a methodology, which we made use of R based written functions, in order to classify tumors and tumor cell lines into CT-X positive and negative groups according to the expression pattern of seven CT-X gene families. We observed coordinate CT-X gene expression pattern which was clustered in all cell line and tumor datasets.

We chose to use the rank product method as a meta-analysis approach since it was shown to be more powerful as compared to other meta-analysis methods based on a t-statistics (Hong and Breitling 2008). We validated the reliability and consistency of RankProd analysis by using a different set of tumor datasets generated by HG-U133Plus2 arrays. Our strategy was to employ the same probesets with which RankProd analysis was performed for tumor datasets generated by HG-U133A arrays. This validation resulted in the identification of

almost similar biological outcome of CT-X positive and negative tumors for both analyses revealed by DAVID functional annotation clustering analysis. We found that CT-X positive tumors have higher proliferative and metastatic capacity indicating worse prognosis. We and others have already shown that CT-X gene expression is associated with advanced disease and other variables that indicate worse prognosis in different types of tumors (Gure AO,2005; Condomines M , 2007; Velazquez EF, 2007). In addition, CT-X positive tumors appear to have repressed immune response. In the literature, there are a number publications indicating that T-cell responses against CT-X genes were in general low or they were not sustained, thereby following postvaccination tumor relapsed again (Marchand, van Baren et al. 1999; Jager, Gnjatic et al. 2000; Zendman, Ruiter et al. 2003). Therefore, our findings were in parallel to the current evidence on CT-X genes.

CT-X genes are expressed in proliferating germ cells in testis, which have the capacity to self-renew thereby resembling adult stem cells. Interestingly, NY-ESO-1, SSX, MAGEA3, NRAGE were found to be expressed in human mesenchymal stem cells of the bone marrow and down-regulated after differentiating into osteocyte and adipocyte (Cronwright, Le Blanc et al. 2005). In addition, it was shown by immunohistochemical analysis that CT-X genes including NY-ESO-1 and MAGEA3 are expressed heterogeneously in tumor tissues meaning that all the cells in the tumor population are not expresssing CT-X genes (Scanlan, Gure et al. 2002). All these findings suggest that CT-X gene expression might be a stem cell marker, be it in normal or tumor tissues. According to our meta-analysis, CT-X positive tumors seem to be dividing more as compared to CT-X negative tumors associated to self-renewal, and proliferative capacity of cancer stem cells that maintains the tumor cell mass. Thus, it is probable that CT-X positivity may confer tumor cells with a proliferative capacity and stem cell like phenotype.

Moreover, in human mesenchymal stem cells SSX was shown to be localized in the cytoplasm and co-stained with matrix metalloproteinase 2 (MMP2) and vimentin (Cronwright, Le Blanc et al. 2005). Further investigation revealed that the migration of a melanoma cell line (DFW), which expresses SSX, MMP2, and vimentin, decreases when SSX is down-regulated. In addition, E-cadherin expression increases, mimicking a mesenchymal epithelial transition. These results suggest that SSX might have a functional role in normal stem cell migration and having a potentially similar function in cancer cell metastases (Cronwright, Le Blanc et al. 2005). These findings were also in aggrement with our meta-

analysis of tumors as CT-X positive tumors overexpress genes that are associated with metastasis and angionenesis. In the light of these findings, our approach could infer the functional connection of CT-X gene expression with cancer and stem cell biology and the specific functions of CT-X genes. This would be further clarified by functional experiments.

Although RankProd could generate consistent and reliable data, it might result in identification of tissue-specific gene expression pattern when analyzing tumors originating from different tissues. In addition, it does not take into account the covariance structure of the probeset expression leading to identification of probesets as significant solely based on their fold change value. This could not be very informative or might be false-positive regarding the mechanisms underlying CT-X gene expression in tumors since many of the CT-X genes are coordinately expressed; in fact many probesets in the whole genome might exhibit coordinate expression patterns. We partially verified this in lung cancer cell line dataset by clustering analysis using the probesets generated by meta-analysis and observed that the probesets that were most significant in meta-analysis could not discriminate CT-X positive and negative samples (**section 4.2.5**). We therefore performed class prediction analysis in lung cancer cell line dataset to identify the probesets with more substantial gene expression differences between CT-X positive and negative samples (**section 4.2.6**). We thought that from this type of analysis we can easily select the probesets that could be validated by RT-PCR subsequently. We chose this dataset since we want to perform validation experiments in lung cancer cell lines with known CT-X expression profiles. Although class prediction analysis correctly classified CT-X positive and negative lung cancer cell lines, there appeared to be a subgroup of CT-X negative cell lines that were classified together with those exhibiting CT-X positivity suggesting that the mechanisms underlying CT-X gene expression in this subgroup might be activated leading to a similar gene expression profile of CT-X positive group. It is also possible that the subgroup might express additional markers (other CT-X genes) not considered in the given study leading to its discrimination from other CT-X negative cell lines. This should be verified by additional analyses considering other CT-X genes. A similar class prediction analysis would be performed for the other cancer cell line dataset (GSE5720). If a subgroup similar to the one in lung cancer cell line dataset was generated, the RankProd analysis would be re-performed after excluding these subgroups from cancer cell line datasets.

Lastly, we obtained parallel results between meta-analysis and class prediction analysis. As we expected, the probesets identified by class prediction analysis have higher fold changes

that the probesets identified by meta-analysis. We chose 8 up-regulated and 1 down-regulated genes common to both analyses. The functions of these genes are briefly given below. Interestingly, RAD21, HSP90B1 and LAPTM4B were shown to confer drug resistance in different tumors. This could be associated with CT-X gene expression and we would verify this by testing the effect of chemotherapeutic drugs in CT-X expressing lung cancer cell lines. In contrast, one gene, SSRP1, seemed to confer drug sensitivity and this could also be checked in anti-cancer therapies in lung cancer. Overexpression of TWIST1 and LIMK2 is implicated in metastasis. Interestingly, we found overexpression of TWIST1 whereas downregulation of LIMK2 in our class prediction analysis. NFYC is a transcription factor that induces Gadd45a expression which indirectly causes DNA demethylation. This seems to be relevant to CT-X regulation considering that CT-X genes are activated by promoter-specific DNA demethylation. Therefore, the functional role of NFYC in relation to epigenetic deregulation of CT-X genes should also be confirmed. Lastly, SCD and HSP105 which are expressed in different cell lines should also be verified in CT-X positive cell lines in order to confirm the proliferative consequences.

We will validate these genes in lung cancer cells for which we quantified the expression of four CT-X genes, namely NY-ESO-1, MAGEA3, GAGE and SSX4, by probe-based qPCR and classified them as CT-X expressing and non-expressing cell lines. In the long term, if we would find correlation between their expression and CT-X positivity, further functional analyses would be performed in order to explore their functional relationship to CT-X genes.

5.2.1 Up-regulated genes in CT-X positive lung cancer cell lines

SCD: Stearoyl-CoA desaturase is an iron-containing enzyme that catalyzes a rate-limiting step in the synthesis of unsaturated fatty acids. By globally regulating lipid metabolism, stearoyl-CoA desaturase activity modulates cell proliferation and survival. Simian virus 40-transformed human lung fibroblasts bearing a knockdown of human stearoyl-CoA desaturase exhibited a dramatic decrease in proliferation rate and abolition of anchorage-independent growth (Scaglia and Igal 2005). Remarkably, the reduction of SCD1 expression in lung cancer cells significantly delayed the formation of tumors and reduced the growth rate of tumor xenografts in mice (Scaglia and Igal 2008).

RAD21: The protein encoded by this gene is highly similar to the gene product of Schizosaccharomyces pombe rad21, a gene involved in the repair of DNA double-strand

breaks, as well as in chromatid cohesion during mitosis. RAD21 is a novel target for developing cancer therapeutics that can potentially enhance the antitumor activity of chemotherapeutic agents acting via induction of DNA damage. RAD21 was overexpressed in several human breast cancer cell lines when compared to normal breast tissue. RAD21 depletion by siRNA in breast cancer cell lines MCF-7 and T-47D decreased proliferation and increased apoptosis in these cells. Moreover, MCF-7 cell sensitivity to two DNA-damaging chemotherapeutic agents, etoposide and bleomycin, was increased after inhibition of RAD21 expression (Atienza, Roth et al. 2005).

HSPH1 (HSPH105): Heat shock proteins (HSP) support the folding and function of many proteins, and are important components of the ER stress response. Heat shock protein 105 (hsp105) was overexpressed in human colon and pancreatic adenocarcinoma. In addition thyroid, esophageal, breast and bladder carcinoma and islet cell tumor, gastric malignant lymphoma, pheochromocytoma, and seminoma overexpressed hsp105. On the other hand, hsp105 was evidently overexpressed only in the testis among human adult normal tissues (Kai M, 2003) The expression of HSP105 was related to the invasiveness of the lesions. High expression of HSP105 is associated with malignant melanoma especially advanced and metastatic lesions (Muchemwa, Nakatsura et al. 2008).

HSP90B1 (GRP94): HSP90 proteins normally associate with other co-chaperones and play important roles in folding newly synthesized proteins or stabilizing and refolding denatured proteins after stress . The upregulation of GRP78 and GRP94 can significantly confer the chemoresistance to VP-16 in human lung cancer cell line SK-MES-1 (Zhang, Wang et al. 2009).

LAPTM4B: LAPTM4b was upregulated in solid tumors of lung 88% (23/26) and in colon carcinoma 67% (18/27) patients. In addition an overexpression of LAPTM4b in the majority of carcinomas of the uterus (30/44), breast (27/53) and ovary (11/16). A proposed role for LAPTM4b during disease progression of malignant cells and its putative dual functional involvement in tumour cell proliferation as well as in multidrug-resistance may represent LAPTM4b as a target suitable for development of novel therapeutic agents (Kasper, Vogel et al. 2005).

NFYC: NFYC gene encodes one subunit of a trimeric complex forming a highly conserved transcription factor that binds with high specificity to CCAAT motifs in the promoters of a variety of genes. and both the Oct-1 and NF-Y concertedly participate in TSA-induced activation of the gadd45 promoter. (Hirose, Sowa et al. 2003). Gadd45a (growth arrest and DNA-damage-inducible protein 45 alpha), a nuclear protein involved in maintenance of genomic stability, DNA repair and suppression of cell growth was shown to promote repair-mediate DNA demethylation (Barreto, Schafer et al. 2007)

SSRP1: The protein encoded by this gene is a subunit of a heterodimer that, along with SUPT16H, forms chromatin transcriptional elongation factor FACT. FACT interacts specifically with histones H2A/H2B to effect nucleosome disassembly and transcription elongation. FACT and cisplatin-damaged DNA may be crucial to the anticancer mechanism of cisplatin. This encoded protein contains a high mobility group box which most likely constitutes the structure recognition element for cisplatin-modified DNA (Landais, Lee et al. 2006).

TWIST1: TWIST1 is a basic helix-loop-helix (bHLH) transcription factor. Twist is a master regulator of morphogenesis and plays an essential role in tumor metastasis. Twist contributes to metastasis by promoting an epithelial-mesenchymal transition (EMT) in breast cancer (Yang, Mani et al. 2004). TWIST expression was detected in the large majority of human glioma-derived cell lines and human gliomas examined. Levels of TWIST mRNA were associated with the highest grade gliomas. Overexpression of TWIST protein in a human glioma cell line significantly enhanced tumor cell invasion, a hallmark of high-grade gliomas (Elias, Tozer et al. 2005).

5.2.2 Down-regulated genes in CT-X positive lung cancer cell lines

LIMK2: There are approximately 40 known eukaryotic LIM proteins containing the LIM domains. LIM domains are highly conserved cysteine-rich structures containing 2 zinc fingers. Although zinc fingers usually function by binding to DNA or RNA, the LIM motif probably mediates protein-protein interactions. The LIM kinases have been proposed to play an important role in tumour-cell invasion and metastasis. Increased expression of LIMK2 leads to metastatic phenotype in cancer cells (Shea, Wells et al. 2008).

6 REFERENCES

Almeida, L. G., N. J. Sakabe, et al. (2009). "CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens." Nucleic Acids Res **37**(Database issue): D816-9.

Atienza, J. M., R. B. Roth, et al. (2005). "Suppression of RAD21 gene expression decreases cell growth and enhances cytotoxicity of etoposide and bleomycin in human breast cancer cells." Mol Cancer Ther **4**(3): 361-8.

Bachman, K. E., B. H. Park, et al. (2003). "Histone modifications and silencing prior to DNA methylation of a tumor suppressor gene." Cancer Cell **3**(1): 89-95.

Barker, P. A. and A. Salehi (2002). "The MAGE proteins: emerging roles in cell cycle progression, apoptosis, and neurogenetic disease." J Neurosci Res **67**(6): 705-12.

Barreto, G., A. Schafer, et al. (2007). "Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation." Nature **445**(7128): 671-5.

Barski, A., S. Cuddapah, et al. (2007). "High-resolution profiling of histone methylations in the human genome." Cell **129**(4): 823-37.

Bernstein, E. and C. D. Allis (2005). "RNA meets chromatin." Genes Dev **19**(14): 1635-55.

Bestor, T., A. Laudano, et al. (1988). "Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases." J Mol Biol **203**(4): 971-83.

Bild, A. H., G. Yao, et al. (2006). "Oncogenic pathway signatures in human cancers as a guide to targeted therapies." Nature **439**(7074): 353-7.

Bird, A. (2002). "DNA methylation patterns and epigenetic memory." Genes Dev **16**(1): 6-21.

Boel, P., C. Wildmann, et al. (1995). "BAGE: a new gene encoding an antigen recognized on human melanomas by cytolytic T lymphocytes." Immunity **2**(2): 167-75.

Bourc'his, D. and T. H. Bestor (2004). "Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L." Nature **431**(7004): 96-9.

Bredenbeck, A., V. M. Hollstein, et al. (2008). "Coordinated expression of clustered cancer/testis genes encoded in a large inverted repeat DNA structure." Gene **415**(1-2): 68-73.

Breitling, R., P. Armengaud, et al. (2004). "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments." FEBS Lett **573**(1-3): 83-92.

Breitling, R. and P. Herzyk (2005). "Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data." J Bioinform Comput Biol **3**(5): 1171-89.

Brenner, C., R. Deplus, et al. (2005). "Myc represses transcription through recruitment of DNA methyltransferase corepressor." Embo J **24**(2): 336-46.

Busso, D., D. J. Cohen, et al. (2005). "Human testicular protein TPX1/CRISP-2: localization in spermatozoa, fate after capacitation and relevance for gamete interaction." Mol Hum Reprod **11**(4): 299-305.

Chen, T., S. Hevi, et al. (2007). "Complete inactivation of DNMT1 leads to mitotic catastrophe in human cancer cells." Nat Genet **39**(3): 391-6.

Chen, Y. T., B. Alpen, et al. (2003). "Identification and characterization of mouse SSX genes: a multigene family on the X chromosome with restricted cancer/testis expression." Genomics **82**(6): 628-36.

Chen, Y. T., M. J. Scanlan, et al. (1997). "A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening." Proc Natl Acad Sci U S A **94**(5): 1914-8.
Chiang, D. Y., A. Villanueva, et al. (2008). "Focal gains of VEGFA and molecular classification of hepatocellular carcinoma." Cancer Res **68**(16): 6779-88.
Choi, J. K., U. Yu, et al. (2003). "Combining multiple microarray studies and modeling interstudy variation." Bioinformatics **19 Suppl 1**: i84-90.
Chomez, P., O. De Backer, et al. (2001). "An overview of the MAGE gene family with the identification of all human members of the family." Cancer Res **61**(14): 5544-51.
Cilensek, Z. M., F. Yehiely, et al. (2002). "A member of the GAGE family of tumor antigens is an anti-apoptotic gene that confers resistance to Fas/CD95/APO-1, Interferon-gamma, taxol and gamma-irradiation." Cancer Biol Ther **1**(4): 380-7.
Clark, J., P. J. Rocques, et al. (1994). "Identification of novel genes, SYT and SSX, involved in the t(X;18)(p11.2;q11.2) translocation found in human synovial sarcoma." Nat Genet **7**(4): 502-8.
Condomines, M., D. Hose, et al. (2007). "Cancer/testis genes in multiple myeloma: expression patterns and prognosis value determined by microarray analysis." J Immunol **178**(5): 3307-15.
Cronwright, G., K. Le Blanc, et al. (2005). "Cancer/testis antigen expression in human mesenchymal stem cells: down-regulation of SSX impairs cell migration and matrix metalloproteinase 2 expression." Cancer Res **65**(6): 2207-15.
D'Alessio, A. C., I. C. Weaver, et al. (2007). "Acetylation-induced transcription is required for active DNA demethylation in methylation-silenced genes." Mol Cell Biol **27**(21): 7462-74.
De Backer, O., K. C. Arden, et al. (1999). "Characterization of the GAGE genes that are expressed in various human cancers and in normal testis." Cancer Res **59**(13): 3157-65.
De La Fuente, R., C. Baumann, et al. (2006). "Lsh is required for meiotic chromosome synapsis and retrotransposon silencing in female germ cells." Nat Cell Biol **8**(12): 1448-54.
De Smet, C., C. Lurquin, et al. (1999). "DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter." Mol Cell Biol **19**(11): 7327-35.
Desmedt, C., F. Piette, et al. (2007). "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series." Clin Cancer Res **13**(11): 3207-14.
Eckhardt, F., J. Lewin, et al. (2006). "DNA methylation profiling of human chromosomes 6, 20 and 22." Nat Genet **38**(12): 1378-85.
Elias, M. C., K. R. Tozer, et al. (2005). "TWIST is expressed in human gliomas and promotes invasion." Neoplasia **7**(9): 824-37.
Evans, J. P. (2001). "Fertilin beta and other ADAMs as integrin ligands: insights into cell adhesion and fertilization." Bioessays **23**(7): 628-39.
Fraga, M. F., E. Ballestar, et al. (2005). "Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer." Nat Genet **37**(4): 391-400.
Fuks, F. (2005). "DNA methylation and histone modifications: teaming up to silence genes." Curr Opin Genet Dev **15**(5): 490-5.
Glynn, S. A., P. Gammell, et al. (2004). "A new superinvasive in vitro phenotype induced by selection of human breast carcinoma cells with the chemotherapeutic drugs paclitaxel and doxorubicin." Br J Cancer **91**(10): 1800-7.

Goelz, S. E., B. Vogelstein, et al. (1985). "Hypomethylation of DNA from benign and malignant human colon neoplasms." Science **228**(4696): 187-90.
Grunau, C., C. Sanchez, et al. (2005). "Frequent DNA hypomethylation of human juxtacentromeric BAGE loci in cancer." Genes Chromosomes Cancer **43**(1): 11-24.
Gure, A. O., R. Chua, et al. (2005). "Cancer-testis genes are coordinately expressed and are markers of poor outcome in non-small cell lung cancer." Clin Cancer Res **11**(22): 8055-62.
Gure, A. O., E. Stockert, et al. (2000). "CT10: a new cancer-testis (CT) antigen homologous to CT7 and the MAGE family, identified by representational-difference analysis." Int J Cancer **85**(5): 726-32.
Gure, A. O., O. Tureci, et al. (1997). "SSX: a multigene family with several members transcribed in normal testis and human cancer." Int J Cancer **72**(6): 965-71.
Gure, A. O., I. J. Wei, et al. (2002). "The SSX gene family: characterization of 9 complete genes." Int J Cancer **101**(5): 448-53.
Hall, I. M., G. D. Shankaranarayana, et al. (2002). "Establishment and maintenance of a heterochromatin domain." Science **297**(5590): 2232-7.
Heard, E., J. Chaumeil, et al. (2004). "Mammalian X-chromosome inactivation: an epigenetics paradigm." Cold Spring Harb Symp Quant Biol **69**: 89-102.
Hendrix, N. D., R. Wu, et al. (2006). "Fibroblast growth factor 9 has oncogenic activity and is a downstream target of Wnt signaling in ovarian endometrioid adenocarcinomas." Cancer Res **66**(3): 1354-62.
Hirose, T., Y. Sowa, et al. (2003). "p53-independent induction of Gadd45 by histone deacetylase inhibitor: coordinate regulation by transcription factors Oct-1 and NF-Y." Oncogene **22**(49): 7762-73.
Hofmann, O., O. L. Caballero, et al. (2008). "Genome-wide analysis of cancer/testis gene expression." Proc Natl Acad Sci U S A **105**(51): 20422-7.
Hong, F. and R. Breitling (2008). "A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments." Bioinformatics **24**(3): 374-82.
Hong, F., R. Breitling, et al. (2006). "RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis." Bioinformatics **22**(22): 2825-7.
Hong, J. A., Y. Kang, et al. (2005). "Reciprocal binding of CTCF and BORIS to the NY-ESO-1 promoter coincides with derepression of this cancer-testis gene in lung cancer cells." Cancer Res **65**(17): 7763-74.
Huang da, W., B. T. Sherman, et al. (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat Protoc **4**(1): 44-57.
Imamura, T., S. Yamamoto, et al. (2004). "Non-coding RNA directed DNA demethylation of Sphk1 CpG island." Biochem Biophys Res Commun **322**(2): 593-600.
Irizarry, R. A., D. Warren, et al. (2005). "Multiple-laboratory comparison of microarray platforms." Nat Methods **2**(5): 345-50.
Jager, E., S. Gnjatic, et al. (2000). "Induction of primary NY-ESO-1 immunity: CD8+ T lymphocyte and antibody responses in peptide-vaccinated patients with NY-ESO-1+ cancers." Proc Natl Acad Sci U S A **97**(22): 12198-203.
James, S. R., P. A. Link, et al. (2006). "Epigenetic regulation of X-linked cancer/germline antigen genes by DNMT1 and DNMT3b." Oncogene **25**(52): 6975-85.
Jensen, S., M. P. Gassama, et al. (1999). "Cosuppression of I transposon activity in Drosophila by I-containing sense and antisense transgenes." Genetics **153**(4): 1767-74.
Karpf, A. R. and S. Matsui (2005). "Genetic disruption of cytosine DNA methyltransferase enzymes induces chromosomal instability in human cancer cells." Cancer Res **65**(19): 8635-9.

Kasper, G., A. Vogel, et al. (2005). "The human LAPTM4b transcript is upregulated in various types of solid tumours and seems to play a dual functional role during tumour progression." Cancer Lett **224**(1): 93-103.
Keeney, S., C. N. Giroux, et al. (1997). "Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family." Cell **88**(3): 375-84.
Kim, V. N., J. Han, et al. (2009). "Biogenesis of small RNAs in animals." Nat Rev Mol Cell Biol **10**(2): 126-39.
Kondo, T., X. Zhu, et al. (2007). "The cancer/testis antigen melanoma-associated antigen-A3/A6 is a novel target of fibroblast growth factor receptor 2-IIIb through histone H3 modifications in thyroid cancer." Clin Cancer Res **13**(16): 4713-20.
Kouzarides, T. (2007). "Chromatin modifications and their function." Cell **128**(4): 693-705.
Ladanyi, M. (2001). "Fusions of the SYT and SSX genes in synovial sarcoma." Oncogene **20**(40): 5755-62.
Laduron, S., R. Deplus, et al. (2004). "MAGE-A1 interacts with adaptor SKIP and the deacetylase HDAC1 to repress transcription." Nucleic Acids Res **32**(14): 4340-50.
Landais, I., H. Lee, et al. (2006). "Coupling caspase cleavage and ubiquitin-proteasome-dependent degradation of SSRP1 during apoptosis." Cell Death Differ **13**(11): 1866-78.
Landi, M. T., T. Dracheva, et al. (2008). "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival." PLoS One **3**(2): e1651.
Lehnertz, B., Y. Ueda, et al. (2003). "Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin." Curr Biol **13**(14): 1192-200.
Li, X. and X. Zhao (2008). "Epigenetic regulation of mammalian stem cells." Stem Cells Dev **17**(6): 1043-52.
Liang, G., M. F. Chan, et al. (2002). "Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements." Mol Cell Biol **22**(2): 480-91.
Lim, F. L., M. Soulez, et al. (1998). "A KRAB-related domain and a novel transcription repression domain in proteins encoded by SSX genes that are disrupted in human sarcomas." Oncogene **17**(15): 2013-8.
Lim, J. H., S. P. Kim, et al. (2005). "Activation of human cancer/testis antigen gene, XAGE-1, in tumor cells is correlated with CpG island hypomethylation." Int J Cancer **116**(2): 200-6.
Longo, M. C., M. S. Berninger, et al. (1990). "Use of uracil DNA glycosylase to control carry-over contamination in polymerase chain reactions." Gene **93**(1): 125-8.
Loriot, A., E. De Plaen, et al. (2006). "Transient down-regulation of DNMT1 methyltransferase leads to activation and stable hypomethylation of MAGE-A1 in melanoma cells." J Biol Chem **281**(15): 10118-26.
Losch, F. O., A. Bredenbeck, et al. (2007). "Evidence for a large double-cruciform DNA structure on the X chromosome of human and chimpanzee." Hum Genet **122**(3-4): 337-43.
Madsen, B., M. Tarsounas, et al. (2003). "PLU-1, a transcriptional repressor and putative testis-cancer antigen, has a specific expression and localisation pattern during meiosis." Chromosoma **112**(3): 124-32.
Marchand, M., N. van Baren, et al. (1999). "Tumor regressions observed in patients with metastatic melanoma treated with an antigenic peptide encoded by gene MAGE-3 and presented by HLA-A1." Int J Cancer **80**(2): 219-30.
Martens, J. H., R. J. O'Sullivan, et al. (2005). "The profile of repeat-associated histone lysine methylation states in the mouse epigenome." Embo J **24**(4): 800-12.

Mendenhall, E. M. and B. E. Bernstein (2008). "Chromatin state maps: new technologies, new insights." Curr Opin Genet Dev **18**(2): 109-15.
Mercer, T. R., M. E. Dinger, et al. (2009). "Long non-coding RNAs: insights into functions." Nat Rev Genet **10**(3): 155-9.
Muchemwa, F. C., T. Nakatsura, et al. (2008). "Differential expression of heat shock protein 105 in melanoma and melanocytic naevi." Melanoma Res **18**(3): 166-71.
Nagao, T., H. Higashitsuji, et al. (2003). "MAGE-A4 interacts with the liver oncoprotein gankyrin and suppresses its tumorigenic activity." J Biol Chem **278**(12): 10668-74.
Nicholaou, T., L. Ebert, et al. (2006). "Directions in the immune targeting of cancer: lessons learned from the cancer-testis Ag NY-ESO-1." Immunol Cell Biol **84**(3): 303-17.
Okano, M., S. Xie, et al. (1998). "Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases." Nat Genet **19**(3): 219-20.
Okano, M., S. Xie, et al. (1998). "Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells." Nucleic Acids Res **26**(11): 2536-40.
Park, J. H., G. H. Kong, et al. (2002). "hMAGE-A1 overexpression reduces TNF-alpha cytotoxicity in ME-180 cells." Mol Cells **14**(1): 122-9.
Pfaffl, M. W. (2001). "A new mathematical model for relative quantification in real-time RT-PCR." Nucleic Acids Res **29**(9): e45.
Pivot-Pajot, C., C. Caron, et al. (2003). "Acetylation-dependent chromatin reorganization by BRDT, a testis-specific bromodomain-containing protein." Mol Cell Biol **23**(15): 5354-65.
Pousette, A., P. Leijonhufvud, et al. (1997). "Presence of synaptonemal complex protein 1 transversal filament-like protein in human primary spermatocytes." Hum Reprod **12**(11): 2414-7.
Ramasamy, A., A. Mondry, et al. (2008). "Key issues in conducting a meta-analysis of gene expression microarray datasets." PLoS Med **5**(9): e184.
Rinn, J. L., M. Kertesz, et al. (2007). "Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs." Cell **129**(7): 1311-23.
Rollins, R. A., F. Haghighi, et al. (2006). "Large-scale structure of genomic methylation patterns." Genome Res **16**(2): 157-63.
Roman-Gomez, J., A. Jimenez-Velasco, et al. (2005). "Promoter hypomethylation of the LINE-1 retrotransposable elements activates sense/antisense transcription and marks the progression of chronic myeloid leukemia." Oncogene **24**(48): 7213-23.
Ross, M. T., D. V. Grafham, et al. (2005). "The DNA sequence of the human X chromosome." Nature **434**(7031): 325-37.
Rozen, S. and H. Skaletsky (2000). "Primer3 on the WWW for general users and for biologist programmers." Methods Mol Biol **132**: 365-86.
Sahin, U., O. Tureci, et al. (1998). "Expression of multiple cancer/testis (CT) antigens in breast cancer and melanoma: basis for polyvalent CT vaccine strategies." Int J Cancer **78**(3): 387-9.
Sahin, U., O. Tureci, et al. (1995). "Human neoplasms elicit multiple specific immune responses in the autologous host." Proc Natl Acad Sci U S A **92**(25): 11810-3.
Sayan, A. E., B. S. Sayan, et al. (2001). "Acquired expression of transcriptionally active p73 in hepatocellular carcinoma cells." Oncogene **20**(37): 5111-7.
Scaglia, N. and R. A. Igal (2005). "Stearoyl-CoA desaturase is involved in the control of proliferation, anchorage-independent growth, and survival in human transformed cells." J Biol Chem **280**(27): 25339-49.
Scaglia, N. and R. A. Igal (2008). "Inhibition of Stearoyl-CoA Desaturase 1 expression in human lung adenocarcinoma cells impairs tumorigenesis." Int J Oncol **33**(4): 839-50.

Scanlan, M. J., A. O. Gure, et al. (2002). "Cancer/testis antigens: an expanding family of targets for cancer immunotherapy." Immunol Rev **188**: 22-32.
Schotta, G., M. Lachner, et al. (2004). "A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin." Genes Dev **18**(11): 1251-62.
Shankavaram, U. T., W. C. Reinhold, et al. (2007). "Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study." Mol Cancer Ther **6**(3): 820-32.
Shea, K. F., C. M. Wells, et al. (2008). "ROCK1 and LIMK2 interact in spread but not blebbing cancer cells." PLoS One **3**(10): e3398.
Simpson, A. J., O. L. Caballero, et al. (2005). "Cancer/testis antigens, gametogenesis and cancer." Nat Rev Cancer **5**(8): 615-25.
Stevenson, B. J., C. Iseli, et al. (2007). "Rapid evolution of cancer/testis genes on the X chromosome." BMC Genomics **8**: 129.
Svoboda, P. (2004). "Long dsRNA and silent genes strike back:RNAi in mouse oocytes and early embryos." Cytogenet Genome Res **105**(2-4): 422-34.
Tabara, H., M. Sarkissian, et al. (1999). "The rde-1 gene, RNA interference, and transposon silencing in C. elegans." Cell **99**(2): 123-32.
Tajima, K., Y. Obata, et al. (2003). "Expression of cancer/testis (CT) antigens in lung cancer." Lung Cancer **42**(1): 23-33.
Tan, K., A. L. Shaw, et al. (2003). "Human PLU-1 Has transcriptional repression properties and interacts with the developmental transcription factors BF-1 and PAX9." J Biol Chem **278**(23): 20507-13.
Terranova, R., S. Yokobayashi, et al. (2008). "Polycomb group proteins Ezh2 and Rnf2 direct genomic contraction and imprinted repression in early mouse embryos." Dev Cell **15**(5): 668-79.
Thompson, J. D., T. J. Gibson, et al. (1997). "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." Nucleic Acids Res **25**(24): 4876-82.
Tureci, O., U. Sahin, et al. (1996). "The SSX-2 gene, which is involved in the t(X;18) translocation of synovial sarcomas, codes for the human tumor antigen HOM-MEL-40." Cancer Res **56**(20): 4766-72.
van der Bruggen, P., C. Traversari, et al. (1991). "A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma." Science **254**(5038): 1643-7.
Vatolin, S., Z. Abdullaev, et al. (2005). "Conditional expression of the CTCF-paralogous transcriptional factor BORIS in normal cells results in demethylation and derepression of MAGE-A1 and reactivation of other cancer-testis genes." Cancer Res **65**(17): 7751-62.
Velazquez, E. F., A. A. Jungbluth, et al. (2007). "Expression of the cancer/testis antigen NY-ESO-1 in primary and metastatic malignant melanoma (MM)--correlation with prognostic factors." Cancer Immun **7**: 11.
Vire, E., C. Brenner, et al. (2006). "The Polycomb group protein EZH2 directly controls DNA methylation." Nature **439**(7078): 871-4.
Volpe, T. A., C. Kidner, et al. (2002). "Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi." Science **297**(5588): 1833-7.
Walsh, C. P., J. R. Chaillet, et al. (1998). "Transcription of IAP endogenous retroviruses is constrained by cytosine methylation." Nat Genet **20**(2): 116-7.
Warburton, P. E., J. Giordano, et al. (2004). "Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes." Genome Res **14**(10A): 1861-9.

Weber, J., M. Salgaller, et al. (1994). "Expression of the MAGE-1 tumor antigen is up-regulated by the demethylating agent 5-aza-2'-deoxycytidine." Cancer Res **54**(7): 1766-71.
Weber, M., J. J. Davies, et al. (2005). "Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells." Nat Genet **37**(8): 853-62.
Weber, M., I. Hellmann, et al. (2007). "Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome." Nat Genet **39**(4): 457-66.
Weber, M. and D. Schubeler (2007). "Genomic patterns of DNA methylation: targets and function of an epigenetic mark." Curr Opin Cell Biol **19**(3): 273-80.
Wilson, A. S., B. E. Power, et al. (2007). "DNA hypomethylation and human diseases." Biochim Biophys Acta **1775**(1): 138-62.
Wischnewski, F., K. Pantel, et al. (2006). "Promoter demethylation and histone acetylation mediate gene expression of MAGE-A1, -A2, -A3, and -A12 in human cancer cells." Mol Cancer Res **4**(5): 339-49.
Xiao, J. and H. S. Chen (2004). "Biological functions of melanoma-associated antigens." World J Gastroenterol **10**(13): 1849-53.
Yang, J., S. A. Mani, et al. (2004). "Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis." Cell **117**(7): 927-39.
Zendman, A. J., D. J. Ruiter, et al. (2003). "Cancer/testis-associated genes: identification, expression profile, and putative function." J Cell Physiol **194**(3): 272-88.
Zhang, L., S. Wang, et al. (2009). "Upregulation of GRP78 and GRP94 and its function in chemotherapy resistance to VP-16 in human lung cancer cell line SK-MES-1." Cancer Invest **27**(4): 453-8.
Zhou, B. B., M. Peyton, et al. (2006). "Targeting ADAM-mediated ligand cleavage to inhibit HER3 and EGFR pathways in non-small cell lung cancer." Cancer Cell **10**(1): 39-50.

7 APPENDICES

7.1 APPENDIX A: THE "GROUPING", "PRE-PROCESSING" AND "RANKPROD" SCRIPTS USED IN R

7.1.1 The "grouping" script

```
#this function reads ct-x gene names and their family names from
u133agroups_genenames.txt.
u133a<-read.table("~/duygu/u133agroups_genenames.txt", sep="\t", header=T)
u133a<-as.matrix(u133a)

#this function gets the names of the datasets.
files<-try(system("ls ~/duygu/HG-U133AData_Allsamples_New/",intern=T))

#from names this function reads data.
readgivenfilelist<-function(x,skip) {
for (i in 1:length(x)) {
y=read.table(x[i],header=T, skip=skip, sep="\t")
rownames(y)=y[,1]
assign(x[i],y[,-1], envir=as.environment(pos=1))
}
}
readgivenfilelist(files,4)

#this function combines all read data in one matrix.
zdata<-function(x) {
z=get(ls(pos=1)[x[1]], envir=as.environment(pos=1))
data=as.matrix(z)
for (i in 2:length(x) ) {
y=as.matrix(get(ls(pos=1)[x[i]], envir=as.environment(pos=1)))
data<-cbind(data,y)
}
return(data)
}
#zalldata<-zdata(c(2,3,5:13))

#this function gives an output such that rows are the unique genes and columns are the
columns of the selected data and values are the mean values of each ct-x gene for each
experiment.
zctdata<-function(x,y) {
#x;data
#y;u133aCTgenes
#j is the number of unique ct gene families
#the mean of probesets belonging to one gene.
j=length(genes<-unique(sub('\\s+$', '', y[,2], perl = TRUE)))
print(j)
```

```
for (i in 1:length(x)) {
	z=as.matrix(get(ls(pos=1)[x[i]], envir=as.environment(pos=1)))
	data=matrix(ncol=ncol(z),nrow=j)
	di=1
		for (k in 1:j) {
			probes<-sub('\\s+$', '', y[,1], perl = TRUE)[which(sub('\\s+$', '', y[,2], perl =
TRUE)==genes[k])]
			if (length(probes)>1) {
			lp=length(probes)
			pindex=vector(length=lp)
			for(l in 1:lp) {
			pindex[l]=which(rownames(z)==probes[l])
			}
			v=t(apply(z[pindex,],2,mean))
			colnames(v)=NULL
			rownames(v)=NULL
			}
			else {
			v=t(z[which(rownames(z)==probes),])
			colnames(v)=NULL
			rownames(v)=NULL
			}
		data[k,]<-v
		print(genes[k])
		print(length(rownames(data)))
}
colnames(data)<-colnames(z)
rownames(data)<-as.vector(genes)
	assign(paste("zct_",ls(pos=1)[x[i]],sep=""),data,envir=as.environment(pos=1))
}
}
#zctdata(c(1,2,4:12),u133a)

#this function calculates the mean normalized expression values of probesets belonging to
each ct-x gene family.
zctdataf<-function(x,y) {
j=length(family<-unique(sub('\\s+$', '', y[,3], perl = TRUE)))
print(j)
for (i in 1:length(x)) {
	z=as.matrix(get(ls(pos=1)[x[i]], envir=as.environment(pos=1)))
	data=matrix(ncol=ncol(z),nrow=j)
	di=1
		for (k in 1:j) {
			genes<-sub('\\s+$', '', y[,2], perl = TRUE)[which(sub('\\s+$', '', y[,3], perl =
TRUE)==family[k])]
			if (length(genes)>1) {
			lp=length(genes)
			pindex=vector(length=lp)
			for(l in 1:lp) {
			pindex[l]=which(rownames(z)==genes[l])
```

```
                }
                v=t(apply(z[pindex,],2,mean))
                colnames(v)=NULL
                rownames(v)=NULL
                }
                else {
                v=t(z[which(rownames(z)==genes),])
                colnames(v)=NULL
                rownames(v)=NULL
                }
        data[k,]<-v
        }
rownames(data)<-as.vector(family)
colnames(data)<-colnames(z)
      assign(paste("z",ls(pos=1)[x[i]],sep=""),data,envir=as.environment(pos=1))
}
}
#zctdataf(c(17:27),u133a)

#this single script calculates the mean rank value as the cutoff.
mean(apply(zalldata,2,function(x){sum(x<3.5)+1}))



#this function assigns a number ("0" and "1") for each ct-x gene family.
groupmeans2zeroonep<-function(x,y) {
for (i in 1:length(x)) {
z=as.matrix(get(ls(pos=1)[x[i]], envir=as.environment(pos=1)))
newmatrix<-matrix(nrow=dim(z)[1],ncol=dim(z)[2])
colnames(newmatrix)=colnames(z)
rownames(newmatrix)=rownames(z)
org=as.matrix(get(ls(pos=1)[y[i]], envir=as.environment(pos=1)))
for (j in 1:ncol(z)) {
cutoff<-sort(org[,j])[the mean rank value]
subbso<-which(z[,j]>=cutoff)
subbsu<-which(z[,j]<cutoff)
newmatrix[subbso,j]<-1
newmatrix[subbsu,j]<-0
}
assign(paste("z",ls(pos=1)[x[i]],"01", sep=""), newmatrix, envir=as.environment(pos=1))
}
}
#groupmeans2zeroonep(c(30:40),c(1,2,5:13))

#this function calculates the sum values of 0 and 1 for each ct-x gene family
sumofones<-function(x) {
for (i in 1:length(x)) {
z=as.matrix(get(ls(pos=1)[x[i]], envir=as.environment(pos=1)))
newvector<-vector(length=dim(z)[2])
names(newvector)=colnames(z)
```

```
newvector<-apply(z,2,sum,na.rm=T)
assign(paste("z",ls(pos=1)[x[i]],"sum", sep=""), newvector, envir=as.environment(pos=1))
}
}
#sumofones(c(43:53))

#this function classifies each sample according to the sum values of ct-x gene families.
giveclass<-function(x) {
for (i in 1:length(x)) {
z=get(ls(pos=1)[x[i]], envir=as.environment(pos=1))
print(names(z))
newvector<-vector(length=length(z))
subct<-which(z>1)
subnonct<-which(z==0)
subinter<-which(z==1)
newvector[subct]<-1
newvector[subnonct]<-0
newvector[subinter]<--1
names(newvector)=names(z)
assign(paste("z",ls(pos=1)[x[i]],"class", sep=""), newvector, envir=as.environment(pos=1))
}}
#giveclass(c(55:65))


#this function writes the classification of samples into .txt file.
writeclass<-function(x,filename) {
for (i in 1:length(x)) {
z=get(ls(pos=1)[x[i]], envir=as.environment(pos=1))
write.table(paste("#",ls(pos=1)[x[i]],sep=""),filename,sep="\t",quote=F,append=T,eol="\n",co
l.names=F,row.names=F)
write.table(z,filename,sep="\t",quote=F,append=T,eol="\n")
}}
writeclass(c(67:77),"groups.txt")


for (i in 1:length(x)) {
z=ls(pos=1)[x[i]]
```

7.1.2 “Pre-processing” and “RankProd” scripts

```
#from names this function reads data.
readgivenfilelist<-function(x,skip) {
for (i in 1:length(x)) {
y=read.table(x[i],header=T, skip=skip, sep="\t")
rownames(y)=y[,1]
assign(x[i],y[,-1], envir=as.environment(pos=1))
}
}
readgivenfilelist(files,4)
```

```
#this function combines all read data in one matrix.
zdata<-function(x) {
z=get(ls(pos=1)[x[1]], envir=as.environment(pos=1))
data=as.matrix(z)
for (i in 2:length(x) ) {
y=as.matrix(get(ls(pos=1)[x[i]], envir=as.environment(pos=1)))
data<-cbind(data,y)
}
return(data)
}
#zalldata<-zdata(c(2,3,5:13))

#this function filters out probesets with normalized expression values above the cutoff (y) in
combined array data called x.
filterunder<-function (x,y) {
z=vector(length=dim(x)[1])
z2=vector(length=dim(x)[2])
for (j in 1:dim(x)[2]) {
z2[j]<-sort(x[,j])[y]
}
for (i in 1:dim(x)[1]) {
if (any(x[i,]>z2)) {
z[i]=T
}
else {
z[i]=F
}
}
return(z)
}

#combine the common probesets after data filtering.
zdata<-function(x) {
z=get(ls(pos=1)[x[1]], envir=as.environment(pos=1))
data=as.matrix(z)
for (i in 2:length(x) ) {
y=as.matrix(get(ls(pos=1)[x[i]], envir=as.environment(pos=1)))
data<-cbind(data,y)
}
return(data)
}

#After data-processing, RP analysis was run with the following scripts using data[sub].
library(RankProd)
#the number of ct-x positive (1) and ct-x negative (0) samples in data[sub] is indicated by
data.cl in order.
data.cl<-c(rep(0,52),rep(1,33),rep(0,48),rep(1,5),rep(0,44),rep(1,28),rep(0,42),rep(1,38))
#the origin of each data in data[sub] is indicated by data.origin.
data.origin<-c(rep(1,85),rep(2,53),rep(3,72),rep(4,80))
#RP analysis was run
```

```
RP.adv.out <- RPadvance (data[sub,], data.cl, data.origin, num.perm = 100, logged = TRUE,
gene.names = rownames(data[sub,]), rand = 123)
#generates RP plot (PFP vs the number of identified probesets.
plotRP(RP.adv.out,cutoff=0.05)
jpeg("tumor133a.jpeg")
#generates the output probeset lists, p≤0.05.
table=topGene(RP.adv.out, cutoff = 0.05, method = "pfp", logged = TRUE, logbase=2,
gene.names = rownames(data[sub,]))
up=table$Table1
down=table$Table2
write.csv(down,"tumor133adown.csv")
write.csv(up,"tumor133aup.csv")

#this function gets the probesets used in meta-analyis of hg-u133a based data from hg-
u133plus2 arrays.
forplus2<-rownames(data)[sub]
fdpia<-function(probelist,rows) {
subs=vector()
for (i in 1:length(probelist)) {
subs=c(which(probelist[i]==rows),subs)
 }
return(subs)
}
#RP was run.
RP.adv.out <- RPadvance (data[rev(p2rows),], data.cl, data.origin, num.perm = 100, logged =
TRUE, gene.names = rownames(data[rev(p2rows),]), rand = 123)
table=topGene(RP.adv.out, cutoff = 0.05, method = "pfp", logged = TRUE, logbase=2,
gene.names = rownames(data[rev(p2rows),]))
up=table$Table1
down=table$Table2
write.csv(down,"pdown.csv")
write.csv(up,"pup.csv")
```

7.2 APPENDIX B: THE SEQUENCE OF THE SSX4 KNOCK-IN VECTOR

Sequences confirmed by sequencing analysis are shown in yellow. Transcription start and stop codons of EGFP and HYG are highlighted in red. Sequencing primers for EGFP are shown in brackets. The restriction enzymes used in the construction of the KI vector are in red letters and underlined. The following are how the other sequences are shown: **β-actin DTA**; **SSX4 5' = SSX4 A3-B**; **EGFP**; SV40 PolyA site; **PGK-HYG (Hygromycin)**; **SSX4 3' = SSX7 3'**; β-lac ; f1 ori

GGGAGGGCGGGCAGGGATTGCTGAGTGCTGCACAGACAGCTTCAGCTGTGACTTGTTGATAGTGTGTTTTGTTTGTTTTTGTTTTTGAGATGGA
GTTTCACTCTTCTCGGCCAGGGTAGAGTGCAATAAGGCAATCTCAGCTCACTCCAACTTTCACCTCCTGGGTTCAAGTGATTCTCCTGCCTCAG
CCTCCGGAGTAGCTGGGGTTACGGGTGCCCGCCCCCACACCCAGCTCATTTTTTATTTTTGGTAGAGACGGGGTTTTACCATGTTGACCAGGCT
GGTCTCAAACTTCCTGACCTCAGGTGATCCACCTGCCTCAGCCTCCCAAAGTGCTGGGATTACAAGTGTGAGCTACCATGCCCGGCCTGTTTGT
AGTTTTTCTAATATTGTGAATAAATAAATCAAACCTAGCATAGCTGTGGGGTAATGTTGAGATCCGACTGGACTCAATATTATTCCTCATACTT
TTCTGTGTGTTTGAAATATTTCTTTTTTAGATGACATGTTGTTCTTCCTAAGCACCATTAATGAATCAAAGGACTGCTAAAAAAATGTTACAAG
AGAAAAAAAGAAGAAAGAAAATGTTACAACTGTAGATTCGCCAAAAACTTCCAGAGTTTGTCTCATTAACAGCATGTAGGTATTGGATAGGTAT
CTTAGGAGTGAGGATGATGAACACATTATGTAATAAAGATCGCTGTTTCTCTGTATTTTATCAGAGCCAAATAGTCTTTTCATTCCCAAAGAAC
CCTGATTCTCCGTGATGAGCTTGGAAGAGAGTTTGAAAGAGTGATCCCTCATCCAACACACAGAGAGAGCTTTCCCACTTGCCAGAGAGCAGAG
ATAACATAGGGTGAAAAAAAGACAGGTTCTTCGGTAGAGATCTTTGTACGTTTTGGGAATATAAAGGGGTCACATATGTGTTTCCTTGCTCTTC
TGCTCTGACAACACAATTATAAGACAAGGTCAGAACGTCCAAACTGTCTCCAATAGACCTATTACTCCCCAACTAAAAAGGCCAGATTCTAAAA
TCTCCCACGATCACTGTAAGAGATCTGAAAATCCAGTGCTTGGGTGTCTGCCAAGATTTTGACATTAAAGGACTGTCTTTATACTGAAAATATC
TCAGAGCCACTGGACTAAATCATCCATGGTTCATCACACATTTAACAGCTTAATTCACATACCATAAGATTCACCCATTTGAAGTGTGCAATGA
TTTTCAGTTGTTGCACATCTTGAGTGGATACAGTTCAGATTCCTAACCAATCAATTTAATTATTTGGAAAAAAATAAAAGATATGTAAAGGAAT
AAGATGAGACTGTGATGGGGTTTAGTCCCCATTTGATAAACCATGAGATGGAAAATTCTGAATTGATGCCACAGATAAATGAATCAACCATGAC
TAAACATAATTCAGAAGCAAAACTGAAATAACTCCTCAACAATGAGTGGACTCATAACACTCTGCTGCAGAATACCCTGATGCTACAGAAGTCT
CTAGAGTTTGGAAATCTTTACCAACAAAGAAAAATTCTGATATATTCTCTTTCAGTTGAACGTCCTCAGATGACTTTTTGCAGGCTCCAGAGAA
TCTTCCCGAAGGTGAGTGTCTCTCAGATCTAAAGGACCAGAGAACCTTTGTCCCTCCATGGATGCAAACACTGGTAAGAGTGGGAGAATATAAA
AAATGCCCTCACTGCCTCCTTCTCCCCATGTCTATCACAACACCTGATGTAGCACCAACAGCTTAATAATACTAACAGTTGTGATCCTTAATAC
ATCTTTTGTTTTCATAGTGATGCCAGATACTATTTTAAGCAGTTCACATGGATTAATGTATTTAATCCTTAAGAAGACCTGTATGACTTTGTTT
CTATTATTATCTCCAAATAATAATGAGTCACACACTTCGGTTGTCGTCCATATAAAAGCCATGTGACTTGGCACAAGTCTTCTAAGTTCTCTGA
GCTCCAGATTCCTGGTCCATGAAATGGAAGTAAAGAATCGTAGTTATCAGCAACATAATAATAAAATGAGGCTATTGGGGTATAGATATGTTAA
ATAATTTTCCTGGGGCACAGTGGCAGTGGTAGTCTAATCCAGAGCTCCAAGCCATTTAAAGCTCATTCACGTTTGCATTTGTTTATGAAGTTCA
GATGTTGCTCACTAGGGCTTCACCCCATAGGGCCTGCTGGTGCTTCCATTGAGACACCCATTCTCGCAATAGGAAAGACCAGCTGGCCTCTGCT
CTGTTACTGGGACCACTTGCATGGCTTAGGAATCGTTTTGACTGTTGGCCCCTCCCTACTGTGAGCTCCTTGATGGTCGACCGATGCCCTTGAG
AGCCTTCAACCCAGTCAGCTCCTTCCGGTGGGCGCGGGGCATGACTATCGTCGCCGCACTTATGACTGTCTTCTTTATCATGCAACTCGTAGGA
CAGGTGCCGGCAGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCG
GTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGT
TGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAG
ATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGA
AGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTC
AGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAG
GATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGC
GCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCA
AGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTA
AGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAG
TAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCG
TCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATC
AGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCT
AGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCAT
TCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAG
AAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACT
GGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATA
GCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACC
CACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGA
ATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACA
TATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGCGCCCTGTAGCGGCGCATTAAG
CGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCC
ACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTG
ATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTT
GTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAG
CTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGCTTACAATTTGCCATTCGCCATTCAGGCTGCGCAACTGTTGGGAAGGGC
GATCGGTGCGGGCCTCTTCGCTATTACGCCAGCCCAAGCTACCATGATAAGTAAGTAATATTAAGGTACGGGAGGTACTTGGA

7.3 APPENDIX C: SEQUENCING RESULTS OF SSX4 KI CLONES

BLAST analysis of the sequenced region with M4 reverse primer:

>ref|NG_005575.1| Homo sapiens synovial sarcoma, X breakpoint 4/synovial sarcoma, X breakpoint 4B region (SSX4/SSX4B@) on chromosome X, Length=29447

Score = 246 bits (133), Expect = 3e-63
Identities = 133/133 (100%), Gaps = 0/133 (0%)
Strand=Plus/Plus

Query 1 CTCCTGGGCTCAAGCGATCCTCTCGCCTCGGCCCCGGGACTACAGGCGTGCACCACCCCG 60

```
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  786  CTCCTGGGCTCAAGCGATCCTCTCGCCTCGGCCCCGGGACTACAGGCGTGCACCACCCCG  845

Query  61   CCCAGAGCACCAAAGGTCCTGAGGCTGGAAAGACTCAGGCTGTTTCTCTCGCAGGTGAGA  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  846  CCCAGAGCACCAAAGGTCCTGAGGCTGGAAAGACTCAGGCTGTTTCTCTCGCAGGTGAGA  905

Query  121  CTGCTCCCAGTGC  133
            |||||||||||||
Sbjct  906  CTGCTCCCAGTGC  918


Score =  246 bits (133),  Expect = 3e-63
Identities = 133/133 (100%), Gaps = 0/133 (0%)
Strand=Plus/Minus

Query  1      CTCCTGGGCTCAAGCGATCCTCTCGCCTCGGCCCCGGGACTACAGGCGTGCACCACCCCG  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  28373  CTCCTGGGCTCAAGCGATCCTCTCGCCTCGGCCCCGGGACTACAGGCGTGCACCACCCCG  28314

Query  61     CCCAGAGCACCAAAGGTCCTGAGGCTGGAAAGACTCAGGCTGTTTCTCTCGCAGGTGAGA  120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  28313  CCCAGAGCACCAAAGGTCCTGAGGCTGGAAAGACTCAGGCTGTTTCTCTCGCAGGTGAGA  28254

Query  121    CTGCTCCCAGTGC  133
              |||||||||||||
Sbjct  28253  CTGCTCCCAGTGC  28241
```

BLAST analysis of the sequenced region with A2.1 forward primer:

```
>ref|NG_005851.1|  Homo sapiens ornithine aminotransferase pseudogene (LOC791095)
on chromosome X, Length=3843

GENE ID: 791095 LOC791095 | ornithine aminotransferase pseudogene

Score =  464 bits (251),  Expect = 2e-128
Identities = 254/255 (99%), Gaps = 1/255 (0%)
Strand=Plus/Minus

The Sbjct sequence belonged to SSX4 gene which showed 99% identity to the sequenced region.

Query  1    CCCAAATGTCCAACAATGATAGACTGGATTAAGAAAATGTGGCACATATACACC-TGGAA  59
            |||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||
Sbjct  713  CCCAAATGTCCAACAATGATAGACTGGATTAAGAAAATGTGGCACATATACACCATGGAA  654

Query  60   TACTATGCAGCCATAAAAATTGATGAGTTCATGTCCTTTGTAGGGACATGGATGAAATTG  119
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  653  TACTATGCAGCCATAAAAATTGATGAGTTCATGTCCTTTGTAGGGACATGGATGAAATTG  594

Query  120  GAAATCATCATTCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGAATATTCTCA  179
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  593  GAAATCATCATTCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGAATATTCTCA  534

Query  180  CTCATAGGTGGGAATTGAACAATGAGATCACATGGACACAGGAAGGGGAACATCACACTC  239
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

```
Sbjct  533  CTCATAGGTGGGAATTGAACAATGAGATCACATGGACACAGGAAGGGGAACATCACACTC  474

Query  240  TGGGGACTGTTGTGG  254
            ||||||||||||||
Sbjct  473  TGGGGACTGTTGTGG  459
```

>gb|AF196972.1| Homo sapiens chromosome X multiple clones map p11.23, complete Sequence, Length=122568

Score = 464 bits (251), Expect = 2e-128
Identities = 254/255 (99%), Gaps = 1/255 (0%)
Strand=Plus/Minus

The Sbjct sequence belonged to SSX4 gene which showed 99% identity to the sequenced region.

```
Query  1     CCCAAATGTCCAACAATGATAGACTGGATTAAGAAAATGTGGCACATATACACC-TGGAA  59
             |||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||
Sbjct  7308  CCCAAATGTCCAACAATGATAGACTGGATTAAGAAAATGTGGCACATATACACCATGGAA  7249

Query  60    TACTATGCAGCCATAAAAATTGATGAGTTCATGTCCTTTGTAGGGACATGGATGAAATTG  119
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  7248  TACTATGCAGCCATAAAAATTGATGAGTTCATGTCCTTTGTAGGGACATGGATGAAATTG  7189

Query  120   GAAATCATCATTCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGAATATTCTCA  179
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  7188  GAAATCATCATTCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGAATATTCTCA  7129

Query  180   CTCATAGGTGGGAATTGAACAATGAGATCACATGGACACAGGAAGGGGAACATCACACTC  239
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  7128  CTCATAGGTGGGAATTGAACAATGAGATCACATGGACACAGGAAGGGGAACATCACACTC  7069

Query  240   TGGGGACTGTTGTGG  254
             ||||||||||||||
Sbjct  7068  TGGGGACTGTTGTGG  7054
```

Score = 339 bits (183), Expect = 1e-90
Identities = 229/251 (91%), Gaps = 3/251 (1%)
Strand=Plus/Plus

The Sbjct sequence belonged to SSX7 pseudogene which showed 91% identity to the sequenced region.

```
Query  1      CCCAAATGTCCAACAATGATAGACTGGATTAAGAAAATGTGGCACATATACACC-TGGAA  59
              |||||||||||| ||||||||||||||||||||||||||||||||| |||||| |||||
Sbjct  32271  CCCAAATGTCCATCAATGATAGACTGGATTAAGAAAATGTGGCACATCTACACCATGGAA  32330

Query  60     TACTATGCAGCCATAA-AAATTGATGAGTTCATGTCCTTTGTAGGGACATGGATGAAATT  118
              ||||||||||||||| |||  ||||||||||||||||||||||||||||||||||  |
Sbjct  32331  TACTATGCAGCCATAAGAAA-GGATGAGTTCATGTCCTTTGTAGGGACATGGATGAAGCT  32389

Query  119    GGAAATCATCATTCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGAATATTCTC  178
              ||||| ||||||||| || |||||||||||||| ||| ||||||||||||  ||||||||
Sbjct  32390  GGAAACCATCATTCTGAGCAAACTATCGCAAGGACAGAAAACCAAACACCTCATATTCTC  32449

Query  179    ACTCATAGGTGGGAATTGAACAATGAGATCACATGGACACAGGAAGGGGAACATCACACT  238
              |||||||||||||||||||||||||| ||| ||||||||||  ||||||||||||||
Sbjct  32450  ACTCATAGGTGGGAATTGAACAATGAGAACACTTGGACACAGGGTGGGGAACATCACACA  32509
```

```
Query  239    CTGGGGACTGT  249
              |||| | ||||
Sbjct  32510  CTGGTGCCTGT  32520
```

>emb|AL606490.14| Human DNA sequence from clone RP11-344N17 on chromosome X Contains the 3'end of the SSX1 gene for synovial sarcoma X breakpoint 1, a synovial sarcoma X breakpoint pseudogene, four ornithine aminotransferase (gyrate atrophy)(OAT) pseudogenes, the SSX9 gene for synovial sarcoma X breakpoint 9, a synovial sarcoma X breakpoint 4 pseudogene (psiSSX4), the SSX3 gene for synovial sarcoma X breakpoint 3, a novel gene, the SSX4 gene for synovial sarcoma X breakpoint 4 and the 3' end of a novel gene similar to synovial sarcoma X breakpoint 4, complete sequence, Length=141676

Score = 459 bits (248), Expect = 7e-127
Identities = 253/255 (99%), Gaps = 1/255 (0%)
Strand=Plus/Plus

The Sbjct sequence belonged to SSX4 gene which showed 100% identity to the sequenced region.

```
Query  1       CCCAAATGTCCAACAATGATAGACTGGATTAAGAAAATGTGGCACATATACACC-TGGAA  59
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||
Sbjct  116636  CCCAAATGTCCAACAATGATAGACTGGATTAAGAAAATGTGGCACATATACACCATGGAA  116695

Query  60      TACTATGCAGCCATAAAAATTGATGAGTTCATGTCCTTTGTAGGGACATGGATGAAATTG  119
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  116696  TACTATGCAGCCATAAAAATTGATGAGTTCATGTCCTTTGTAGGGACATGGATGAAATTG  116755



Query  120     GAAATCATCATTCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGAATATTCTCA  179
              ||||||||||||||||||||||||||||||||||||||||||||||||| |||||||||
Sbjct  116756  GAAATCATCATTCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGCATATTCTCA  116815

Query  180     CTCATAGGTGGGAATTGAACAATGAGATCACATGGACACAGGAAGGGGAACATCACACTC  239
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  116816  CTCATAGGTGGGAATTGAACAATGAGATCACATGGACACAGGAAGGGGAACATCACACTC  116875

Query  240     TGGGGACTGTTGTGG  254
              |||||||||||||||
Sbjct  116876  TGGGGACTGTTGTGG  116890
```

Score = 392 bits (212), Expect = 7e-107
Identities = 241/255 (94%), Gaps = 1/255 (0%)
Strand=Plus/Plus

The Sbjct sequence belonged to SSX3 which showed 94% identity to the sequenced region.

```
Query  1      CCCAAATGTCCAACAATGATAGACTGGATTAAGAAAATGTGGCACATATACACC-TGGAA  59
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||
Sbjct  77615  CCCAAATGTCCAACAATGATAGACTGGATTAAGAAAATGTGGCACATATACACCATGGAA  77674

Query  60     TACTATGCAGCCATAAAAATTGATGAGTTCATGTCCTTTGTAGGGACATGGATGAAATTG  119
              |||||||||||||||||| ||||||||||||||||||||||||||||||||||  ||
Sbjct  77675  TACTATGCAGCCATAAAAAATGATGAGTTCATGTCCTTTGTAGGGACATGGATGAAGCTG  77734
```

```
Query  120    GAAATCATCATTCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGAATATTCTCA  179
              |||| |||||||||||| |||||||| |||| |||||||||||||||||| || ||||||
Sbjct  77735  GAAACCATCATTCTCAGCAAACTATCCCAAGGACAAAAAACCAAACACCGCATGTTCTCA  77794

Query  180    CTCATAGGTGGGAATTGAACAATGAGATCACATGGACACAGGAAGGGGAACATCACACTC  239
              |||||||||||||||||||||||||| |||||||||||||||||||||||||||||| |
Sbjct  77795  CTCATAGGTGGGAATTGAACAATGAGAACACATGGACACAGGAAGGGGAACATCACACAC  77854

Query  240    TGGGGACTGTTGTGG  254
               |||| |||||||||
Sbjct  77855  CGGGGCCTGTTGTGG  77869


 Score =  375 bits (203),  Expect = 7e-102
 Identities = 238/255 (93%), Gaps = 1/255 (0%)
 Strand=Plus/Minus
```

The Sbjct sequence belonged to the AL606490.14 genomic clone not to SSX4 showing 93% identity to the sequenced region.

```
Query  1       CCCAAATGTCCAACAATGATAGACTGGATTAAGAAAATGTGGCACATATACACC-TGGAA  59
              |||||||||||||||||||||||||||||||||||||||||||||||||||| ||||| |||||
Sbjct  130847  CCCAAATGTCCAACAATGATAGACTGGATTAAGAAAATGTGGCACATAGACACCATGGAA  130788

Query  60      TACTATGCAGCCATAAAAATTGATGAGTTCATGTCCTTTGTAGGGACATGGATGAAATTG  119
              ||||||| |||||||||| ||||||| |||||||||||||||||||||||| | ||
Sbjct  130787  TACTATGTAGCCATAAAAAATGATGAGCTCATGTCCTTTGTAGGGACATGGATGACACTG  130728

Query  120     GAAATCATCATTCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGAATATTCTCA  179
              |||| |||||||||||| ||||||| ||||| |||||||||||||||||| || ||||||
Sbjct  130727  GAAACCATCATTCTCAGCAAACTATTGCAAGGACAAAAAACCAAACACCGCATGTTCTCA  130668

Query  180     CTCATAGGTGGGAATTGAACAATGAGATCACATGGACACAGGAAGGGGAACATCACACTC  239
              |||||||||||||| |||||||||||||| ||||||||||||||||||||||||||||||
Sbjct  130667  CTCATAGGTGGGACTTGAACAATGAGAACACATGGACACAGGAAGGGGAACATCACACTC  130608

Query  240     TGGGGACTGTTGTGG  254
               |||| |||||||||
Sbjct  130607  CGGGGCCTGTTGTGG  130593
```