Identification and mechanistic characterization of *DLX6-AS1* lncRNA in prostate cancer

DISSERTATION

submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by

Master in Sciences, Mélanie Weiss

born in: Wissembourg, France

Oral-examination:……

Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by

Master in Sciences, Mélanie Weiss

born in: Wissembourg, France

Oral-examination: ……

Identification and mechanistic characterization of *DLX6-AS1* lncRNA in prostate cancer

Referees: PD. Dr. Odilia Popanda

Prof. Dr. Christoph Plass

Dedicated to all persons who offered me their support

on this long (non-coding) and unpredictable road.

CONTRIBUTIONS

Several sections of the introduction (**section 1.2.1**, **1.2.2** and **1.3.3**) as well as **Table 1** are partially based on the review published under the title "Role of lncRNAs in prostate cancer development and progression" by M. Weiss, C. Plass and C. Gerhäuser submitted to the journal of *Biological Chemistry* (Weiss et al., 2014). Therefore, part of the text in these sections might contain suggestions and corrections from co-authors.

The polysome profiling (**Figure 4-27**) was performed by Dr. Johanna Schott (ZMBH, Heidelberg). Treated cells used for this experiment were provided by Dr. Michael Daskalakis (DKFZ, Heidelberg). The description of the corresponding method was written in part with their support (**section 3.9**).

Integration of RNA sequencing and 450k data from the screening sample set (**section 4.1**) was performed in collaboration with Dr. Olga Bogatyrova (DKFZ, Heidelberg).

DECLARATIONS

Declarations according to §8 (3) b) and c) of the doctoral degree regulations:

a) I hereby declare that I have written the submitted dissertation myself and in this process have used no other sources or materials than those expressly indicated.
b) I hereby declare that I have not applied to be examined at any other institution, nor have I used the dissertation in this or any other form at any other institution as an examination paper, nor submitted it to any other faculty as a dissertation.

TABLE OF CONTENTS

SUMMARY

Deregulation of promoter methylation is a well-defined phenomenon in prostate cancer (PCa). This does not only lead to the inactivation of tumor suppressors but also to the activation of oncogenes, enabling the acquisition of the hallmarks of cancer. Accumulating evidence suggests that long non-coding RNAs (lncRNAs) are involved in DNA methylation and demethylation processes through recruitment of DNA methylation modifiers to specific genomic locations.

Since alterations in DNA methylation of protein-coding genes are associated with prostate carcinogenesis and might be regulated by *cis* transcribed RNAs, we aimed to identify lncRNAs involved in DNA (de)methylation mechanisms in PCa. By integrating DNA methylation and RNA sequencing data generated by collaborators from the International Cancer Genome Consortium (ICGC) project, we have identified the lncRNA *DLX6-AS1* as the potential regulator of the *cis* protein-coding gene *DLX6*.

DLX6 and *DLX6-AS1* transcript expression is upregulated and significantly correlates in two independent cohorts of PCa patients, as well as in five different cancer entities. In line with elevated expression levels of *DLX6/DLX6-AS1*, the *DLX6* promoter region is hypomethylated in tumor samples. The impact of elevated *DLX6/DLX6-AS1* levels on patient's prognosis varies with the tumor type.

The clinical data corroborate our hypothesis that *DLX6-AS1* can regulate the *cis* protein-coding gene *DLX6* by decreasing *DLX6* promoter DNA methylation. Nevertheless, we showed through loss- and gain-of-function approaches that *DLX6-AS1* does not regulate *DLX6* at the transcript level by influencing local DNA methylation levels. The sequence features as well as the association to polysomes of the major splice variant of *DLX6-AS1* (*DLX6-AS1 T1*), suggests rather a DLX6 protein synthesis regulatory role for this transcript.

Detailed investigation of the function and expression pattern of all *DLX6-AS1* splice variants in different cancer types will help to clarify the link between *DLX6/DLX6-AS1* expression and patients' prognosis.

ZUSAMMENFASSUNG

Das Phänomen der Deregulation von Promoter-Methylierung wurde im Prostatakarzinom (PCa) ausführlich beschrieben. Diese Deregulation kann einerseits eine Inaktivierung von Tumorsuppressoren und andererseits eine Aktivierung von Onkogenen zur Folge haben, was wiederum die Anreicherung von Krebsmerkmalen ermöglichen kann. Studienergebnisse deuten darauf hin, dass lange, nicht-kodierende RNA (long non-coding RNA, lncRNA) eine Rolle in DNA Methylierungs- und Demethylierungsprozessen durch lokus-spezifische Rekrutierung von DNA Methylierungs-Modifizierern spielt.

Da Veränderungen im DNA Methylierungsmuster protein-kodierender Gene mit der Entstehung von Protstatakrebs in Verbindung stehen und durch *cis* transkribierte RNAs reguliert sein können, war es unser Ziel lncRNAs zu identifizieren, die DNA (De-)Methylierungsmechanismen im PCa deregulieren können. Durch Integration von DNA Methylierungs- und RNA Sequenzierungs-Daten, die von Kollaborationspartnern des International Cancer Genome Consortium (ICGC) Projekts erhoben wurden, konnten wir *DLX6-AS1* als potentiellen Regulator des *cis* protein-kodierenden Gens *DLX6* identifizieren.

In sowohl zwei unabhängigen Kohorten von PCa Patienten, als auch in fünf unterschiedlichen Krebsentitäten, konnte eine Korrelation sowie Hochregulierung der Expression von *DLX6* und *DLX6-AS1* Transkripten beobachtet werden. Im Einklang mit erhöhten Expressionsleveln von *DLX6/DLX6-AS1* stellt sich die *DLX6* Promoterregion als hypomethyliert dar. Der Einfluss der erhöhten *DLX6/DLX6-AS1* Level auf die Prognose der Patienten ist dabei abhängig vom Tumortyp.

Die klinischen Daten bestätigen unsere Hypothese, dass *DLX6-AS1* das *cis* protein-kodierende Gen *DLX6* durch Reduzierung der Promoter DNA-Methylierung reguliert. Wir konnten allerdings durch Überexpressions- und RNA Interferenz-Experimente zeigen, dass *DLX6-AS1* die Transkriptlevel von *DLX6* nicht über die DNA-Methylierung beeinflusst. Sowohl die Sequenzeigenschaften der Haupt-Spleißvariante *DLX6-AS1* (*DLX6-AS1 T1*) als auch die direkte Assoziation mit Polysomen lässt darauf schließen, dass dieses Transkript eher die DLX6 Proteinsynthese reguliert.

Weitergehende Funktions- und Expressionsstudien aller *DLX6-AS1* Spleißvarianten in unterschiedlichen Krebsformen sollten dabei helfen, den Zusammenhang zwischen *DLX6/DLX6*−AS1 Expression und Patientenprognose aufzuklären.

ABBREVIATIONS

Abbreviations	Full Name
450k	HumanMethylation450 BeadChip
5’RACE	Rapid amplification of 5’ cDNA ends
5caC	5-carboxycytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
aa	Amino acids
ALAS1	5'-Aminolevulinate synthase 1
AR	Androgen receptor
BER	Base excision repair
bp	Base pairs
CAGE	Cap analysis of gene expression
CGIs	CpG islands
ChIP-seq	Chromatin immunoprecipitation and sequencing
COAD	Colon adenocarcinoma
DLX5	Distal-less homeobox 5
DLX6	Distal-less homeobox 6
DLX6-AS1	DLX6 antisense RNA 1
DMRs	Differentially methylated regions
DNMT1/3A/3B	DNA methyltransferase 1/3A/3B
DNMT3L	DNA methyltransferase 3-like
EO-PCa	Early onset prostate cancer
ERG-	*TMPRSS2:ERG* fusion-negative prostate cancer patients
ERG+	*TMPRSS2:ERG* fusion-positif prostate cancer patients
Evf2	Mouse homologue of DLX6-AS1 lncRNA
GADD45A	DNA-damage-inducible protein 45A
HDAC	Histone deacetylase
HNSC	Head and neck squamous cell carcinoma
HPRT1	Hypocanthine-guanine phosphoribosyltransferase
ICGC	International Cancer Genome Consortium
KIRC	Kidney renal clear cell carcinoma

lincRNA	Long intergenic non-coding RNA
lncRNA	Long non-coding RNA
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MBD	Methyl-CpG binding domain
mCRPCa	Metastatic castration-resistant prostate cancer
MLN	Myoregulin
mRNA	Messenger RNA
ncRNA	Non-coding RNA
ORF	open reading frame
PCa	Prostate cancer
Pol II	RNA polymerase II
PSA	Prostate-specific antigen
PTMs	Post-translational modifications
RIP-seq	RNA immunoprecipitation followed by deep sequencing
SDHA	Succinate dehydrogenase complex flavoprotein subunit A
SWI/SNF	SWItch/Sucrose NonFermentable complex
TCGA	The Cancer Genome Atlas
TDG	Thymine-DNA glycosylase
TET	Ten-eleven translocation enzyme
TSG	Tumor suppressor gene
TSS	Transcription/Transcriptional start site
TTS	Transcriptional termination site

1. INTRODUCTION

1.1 Epigenetic regulation of gene expression

1.1.1 From *epigenesis* to the modern definition of epigenetics

Until the 1940s the fields of genetics and developmental biology were separated. In 1942, Conrad H. Waddington united the two disciplines by coining the term “epigenetics” as a derivate of the Greek word “*epigenesis*” (theory by which an adult develops from the zygote by a series of steps) combined with the term “genetics”. He used the term “ epigenetics” to describe the process by which a zygote with a certain genotype gives rise to a complex organism with a specific phenotype (Waddington, 2012). Waddington’s notion of epigenetics was however broad, imprecise and not drastically different from the study of embryology.

In 1975, two independent publications by Riggs and by Holliday & Pugh proposed a model for the switching of gene activity by DNA methylation during the developmental process (Holliday et al., 1975, Riggs, 1975), subsequently experimentally proven in the mid-1980s (Doerfler, 1983). Nevertheless, DNA methylation was not linked to epigenetics until the finding that it is heritable, similar to genetic patterns (Holliday, 1987). Consequently, the field of epigenetics evolved to “the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence” (Riggs et al., 1996) translating our 21st century view of epigenetics.

The refinement to the actual definition of epigenetics was mainly the consequence of the discovery of DNA methylation. Epigenetic regulation of gene expression is however achieved through a complex interplay of epigenetic tools and mechanisms, establishing, reinforcing, spreading, transmitting and reversing epigenetic patterns. This includes DNA methylation, histone modifications, histone variants, chromatin remodelling and non-coding RNAs (ncRNAs). The coordinate action of this epigenetic toolbox controls physiological processes such as the

identity of over 200 human cell types or cell proliferation. We focus here on histone modifications, DNA methylation and non-coding RNAs.

1.1.2 Histone modifications and chromatin remodeling

The genome of a typical eukaryotic cell is wrapped into chromatin and folded onto itself to form high-order structures, in order to accommodate the 2 meters of DNA present in the nucleus (Fussner et al., 2011).

The chromatin is a dynamic structure composed of nucleosomes. The main components of the nucleosome are histones, organized in an octamer comprising two copies of H2A/H2B dimer and one H3/H4 tetramer, surrounded by 146 base pairs of DNA and stabilized by histone H1 (Luger et al., 1997). Each individual histone possesses an N-terminal tail subjected to a variety of covalent post-translational modifications (PTMs). So far, more than 100 different histone modifications have been identified; the most widely studied comprise methylation, acetylation and phosphorylation (Kouzarides, 2007). Histone modifications are modulated by two families of enzymes, the histone “writers”, establishing specific modifications (e.g. histone acetyltransferase and methyltransferase), and the histone “erasers” (e.g. histone deacetylase and demethylase) that reverse the process. The combination of PMTs is known as “histone code” and organizes the genome into active regions of euchromatin, where DNA is accessible for transcription, and into inactive heterochromatin regions, where DNA is highly compacted and less accessible for transcription (telomeres, pericentric regions, repetitive sequences) (Strahl et al., 2000).

A deeper look into the most modified histone, histone H3, and its well-studied methylation and acetylation modifications can distinguish functional elements of the genome and determine their active or repressive state. A transcriptionally active gene is decorated by the tri-methylation of lysine 4 (K4) on histone H3 (H3K4me3) in its gene promoter region, while tri-methylation of K36 (H3K36me3) is associated with transcriptionally active gene bodies. Acetylation of lysine residues participate in the binding of transcription factors and are consequently enriched in the promoter

and gene bodies of active genes (**Figure 1-1A**). Promoters of repressed genes are instead marked by tri-methylation of K27 (H3K27me3) or K9 (H3K9me3) (**Figure 1-1B**) (Barski et al., 2007). Histone modifications are interpreted by chromatin "readers" harboring functional recognition domains such as bromodomains and are critical for transcriptional regulation (**Figure 1-1**). PMTs not only regulate chromatin structure, but they also allow the recruitment of remodeling complexes, such as the SWI/SNF (SWItch/Sucrose Non Fermentable) complex leading to nucleosome repositioning (Peterson et al.,2000).

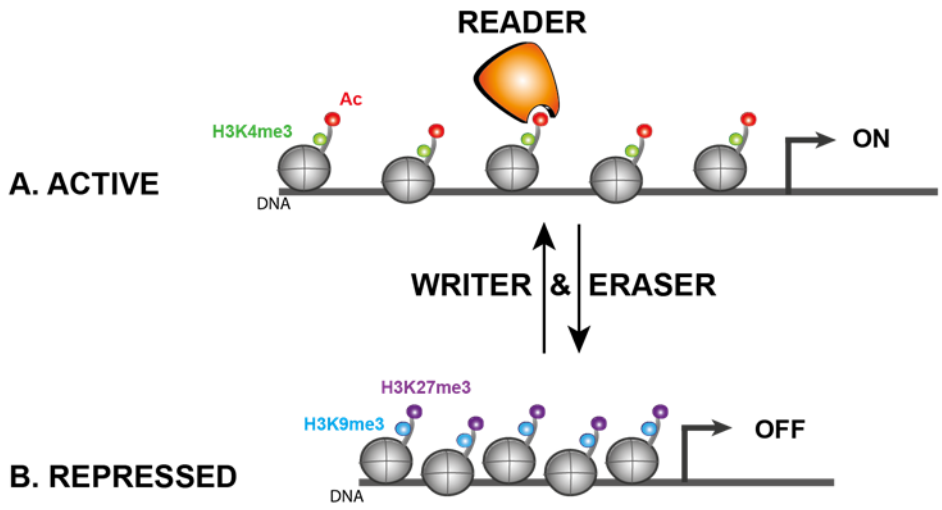


Figure 1-1: Modulation of chromatin marks from an active to a repressed transcriptional state

(**A**) Transcriptionally active genes are marked by chromatin "writers" through acetylation of histone H3 (Ac) and trimethylation of H3 lysine 4 (H3K4me3) in their promoter region. These marks can then be recognized by chromatin "readers". (**B**) Histone acetylation is removed by the histone "eraser" histone deacetylase (HDACs), while repressive trimethylation of K9 (H3K9me3) and K27 (H3K27me3) residues are established by histone methyltransferases. This combination leads to chromatin condensation and subsequent gene silencing.

1.1.3 DNA methylation and demethylation in mammals

DNA methylation is a highly evolutionary conserved epigenetic mark that exists in both prokaryotic and eukaryotic genomes. It is also the first and best-studied epigenetic modification in mammals. In concert with histone modifications, DNA methylation is a crucial player in the

physiological and pathological regulation of gene expression, genomic imprinting and genome stability.

A. Establishment and maintenance of DNA methylation patterns

In humans, the addition of a methyl group from the methyl donor S-adenosylmethionine (SAM) occurs prevalently at the fifth position of cytosine (5-methylcytosine; 5mC) in a cytosine-guanine CpG dinucleotide context.

In somatic cells, DNA methylation patterns are relatively stable compared to histone modifications and are mitotically and meiotically inherited. The faithful restoration of DNA methylation patterns on the newly-replicated DNA strand is ensured through the action of the maintenance DNA methyltransferase 1 (DNMT1) (Hermann et al., 2004).

In contrast, DNA methylation patterns of the pre-implantation embryo are erased genome-wide and subsequently re-established at the blastocyst stage by the *de novo* methyltransferases DNMT3A and DNMT3B (Okano et al., 1999). Both enzymes cooperate with DNA methyltransferase 3-like (DNMT3L), the catalytically inactive member of the family. Indeed, DNMT3L favors *de novo* methylation of gene bodies of housekeeping genes in embryonic stem cells (Neri et al., 2013) and of imprinted genes in germ cells (Bourc'his et al., 2001, Hata et al., 2002, Kareta et al., 2006). *De novo* DNA methyltransferases do not only establish methylation during development but also play a role in somatic cells. Gain of methylation was associated with aging in a fraction of CG rich promoters (Jones et al., 2015) and during tumorigenesis (reviewed in section **1.3.2**). The gain of methylated CpGs in somatic cells is a very slow process, lasting from several weeks in a provirus (Lorincz et al., 2000), to several months in a cell line treated with the DNA methylation inhibitor 5-azacytidine (Flatau et al., 1984).

In addition, *de novo* and maintenance methyltransferases have overlapping functions. Indeed, double DNMT3A/3B knockout of mouse or embryonic stem cells leads to a gradual loss of

methylation as cells divide. This supports a proofreading role for DNMT3A and DNMT3B by methylating CpG sites missed by DNMT1 (Jones et al., 2009, Jeltsch et al., 2014).

B. DNA demethylation

DNA methylation is a dynamic and reversible epigenetic modification, as a result of the interplay between active and/or passive DNA demethylation.

The inhibition, degradation, or nuclear exclusion of DNMT1 result in a progressive dilution of DNA methylation patterns in the process of passive demethylation (**Figure 1-2A**). Passive demethylation is relatively slow and is restricted to dividing cells. This mechanism was exemplified by the replication-dependent dilution of 5mC in the maternal pronucleus during preimplantation development (Rougier et al., 1998).

Research during the past years has identified the dioxygenase family of ten-eleven translocation (TET) enzymes as mediators of active DNA demethylation processes. TET 1, 2 and 3 catalyze the hydroxylation of 5mC, yielding 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxycytosine (5caC) through iterative oxidation (Iyer et al.,2009, Tahiliani et al., 2009). Active DNA demethylation is accompanied by an enzymatic modification of 5mC followed either by passive dilution (**Figure 1-2B**) or by DNA repair of the methyl marks (**Figure 1-2C**). The presence of 5hmC, 5fC or 5caC moieties interferes and prevents the binding of DNMT1 during DNA replication, resulting in a passive dilution of the methylation patterns (**Figure 1-2B**).

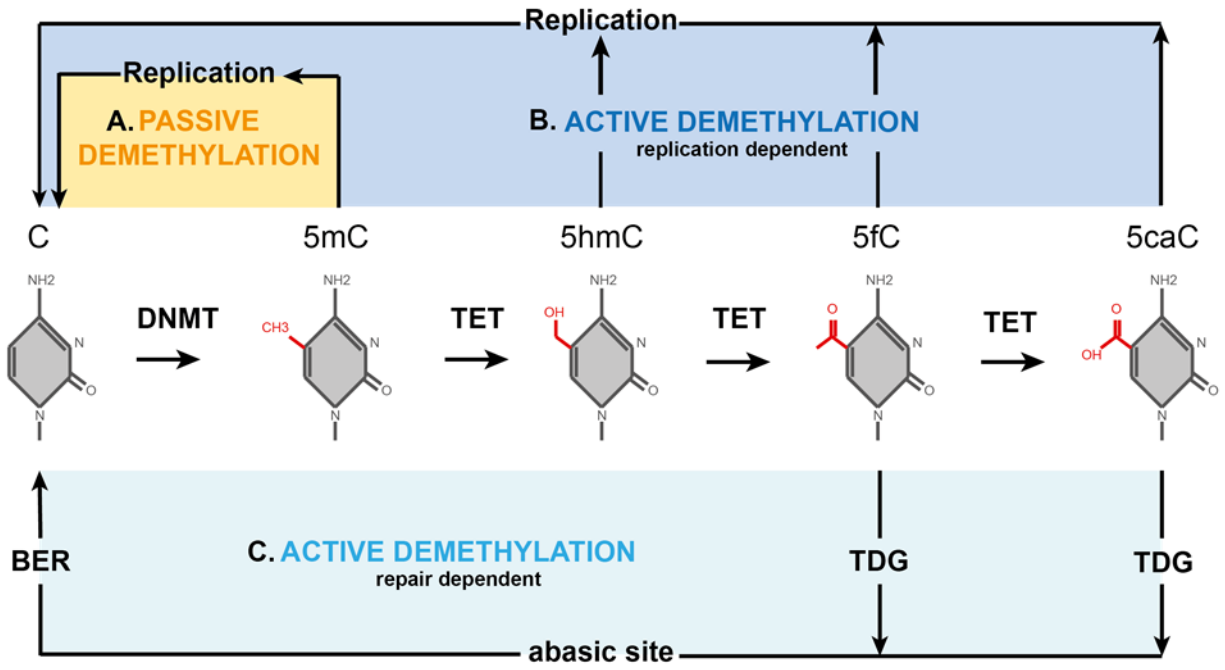


Figure 1-2: Passive and active DNA demethylation

5-Methylcytosine marks (5mC) are established and maintained by DNMT enzymes and then reversed by (**A**) passive or (**B** and **C**) active DNA demethylation. Iterative oxidation of 5-methylcytosine leads to the formation of 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). **(A)** Passive DNA demethylation of 5mC through successive replication. Active DNA demethylation includes (**B**) replication dependent demethylation of 5hmC, 5fC or 5caC residues, (**C**) while 5fC and 5caC can be replaced by an unmodified cytosine after excision by thymine DNA glycosylase (TDG), followed by base excision repair (BER).

Alternatively, 5fC and 5caC can be recognized and excised by thymine-DNA glycosylase (TDG). The generated abasic site is subsequently recognized by the base excision repair (BER) pathway and replaced by an unmodified cytosine (**Figure 1-2C**) (He et al., 2011) (Maiti et al., 2011). Consistent with this model, TDG null mice are lethal around embryonic day 12.5 and exhibit increased 5mC levels in gene promoters (Cortazar et al., 2011, Cortellino et al., 2011). Evidence suggests that this mechanism is used for the global 5mC erasure of the paternal pronucleus a few hours after fertilization as well as in primordial germ cells (Mayer et al., 2000, Oswald et al., 2000). However, this notion has recently been challenged as the formation of 5hmC is observed after demethylation (Amouroux et al., 2016).

Active DNA demethylation is not restricted to development but can also affect specific loci in somatic cells. For example, *p15* (also known as CDKN2B or INK4B) and *SP2* (SP2 transcription factor) promoters are actively demethylated through the interplay of TDG and BER in response to the addition of TGFβ (Thillainadesan et al., 2012) or to the binding of estrogen to the estrogen receptor (Metivier et al., 2008), respectively.

Alternatively, DNA-damage-inducible protein 45A (GADD45A) was reported to stimulate active DNA demethylation by promoting DNA repair via the nucleotide exchange repair (NER) pathway (Barreto et al., 2007). The role of GADD45A in DNA demethylation remains controversial. Evidence suggests a role for GADD45A in active demethylation in mammalian cells (Schmitz et al., 2009, Arab et al., 2014) but conflicting results have also been reported (Jin et al., 2008).

C. Regulation of gene expression by DNA methylation

Of the 29 million CpGs in the human genome, 60 to 80% are methylated in somatic tissues. Most genomic regions are methylated according to their CG content. DNA methylation mainly affects repetitive elements (e.g. retrotransposons) and non-coding regions (e.g. centromere). Hypermethylation of repetitive elements prevents translocations and consequently chromosomal instability (Deaton et al., 2011).

In contrast, regions longer than 200 base pairs (bp) enriched for CpG dinucleotides (GC content >50%), designated as CpG islands (CGIs), are mainly unmethylated in normal cells. Promoter regions remain unmethylated and protected from *de novo* methylation through binding of CFP1 (CXXC finger protein 1) and KDM2A proteins to unmethylated DNA (Blackledge et al., 2010, Thomson et al., 2010, Blackledge et al., 2013). CGIs are mostly located at transcription start sites (Consortium et al.) of gene promoters, flanked by regions termed as shores (up to 2kb away from a CGI) and followed by shelves (located 2 to 4kb from a CGI). Approximatively 70% of genes have a CGI associated to their promoter region. In general, levels of DNA methylation in CGI-associated promoters (including shores) are negatively correlated with the level of gene expression, whereas

in gene body, DNA methylation is positively associated with transcriptional activity (Irizarry et al., 2009, Deaton et al., 2011).

The presence of 5mC in promoter regions is believed to repress gene transcription through interference with the recognition sequence of transcription factors or with the transcriptional machinery. In lieu of interference, DNA methylation can also attract proteins belonging to the methyl-CpG binding domain (MBD) family that establish a repressive chromatin state through recruitment of chromatin remodeling complexes or histone deacetylases (Du et al., 2015).

1.1.4 Interplay between DNA methylation and histone modifications

A coordinated cross-talk between histone modifications and DNA methylation allows precise and dynamic regulation of gene expression. Most proteins responsible for the establishment, removal, or reading of epigenetic marks, leading to a common active or repressive chromatin state, are part of large protein complexes.

MeCP2 (Methyl CpG Binding Protein 2), a member of the MBD family of transcriptional repressors, can be shown as an example that establishes the link between histone and DNA modifications for gene repression. On one side MeCP2 selectively binds to methylated DNA, while simultaneously recruiting histone deacetylases (HDAC) through interaction with the corepressor Sin3A (SIN3 transcription regulator family member A). The MeCP2-Sin3a-HDAC complex establishes a strong transcriptional repression *via* histone deacetylation (Jones et al., 1998, Nan et al., 1998). Unmethylated CGIs, on the other hand, are recognized by zinc finger-CxxC domain containing proteins, such as CFP1. In addition to binding to unmethylated DNA, CFP1 recruits SETD1A (SET Domain containing 1A), a H3K4 methyltransferase, that in turn creates new H3K4me3 marks allowing access of the transcriptional machinery (Thomson et al., 2010).

In 1975, Paul & Duerksen found that in chromatin, RNA is twice more prevalent than DNA, suggesting that RNA may have important roles in establishing chromatin structures (Paul et al., 1975). Since the past years it has become evident that RNA, especially long non-coding RNAs

(lncRNAs), participate in the epigenetic cross-talk by influencing the access or dismissal of chromatin complexes.

1.2 Non-coding RNAs

Protein coding genes account for less than 2% of the human genome; the remaining non-coding space is composed of repetitive elements (de Koning et al., 2011) and was therefore merely considered as "junk DNA" (Comings, 1972, Ohno, 1972). The field of ncRNAs received an increased interest with the advent of next-generation sequencing techniques allowing the finding that 70% of our genome can be transcribed into ncRNAs (Lander et al., 2001, Venter et al., 2001, Consortium, 2012, Djebali et al., 2012).

1.2.1 General characteristics of long non-coding RNAs (lncRNAs)

Non-coding RNAs have been arbitrarily divided into short ncRNAs with a size below 200 nt (e.g. microRNAs, Piwi-associated RNAs, ribosomal RNAs) and lncRNAs being longer than 200 nt (Ponting et al.,2009). LncRNAs have no or limited protein-coding potential but resemble messenger RNAs (mRNAs) in terms of their biogenesis and form. They are generally transcribed by the RNA polymerase II, have a 5' terminal methylguanosine cap, are polyadenylated, often spliced, and share similar chromatin states (Derrien et al., 2012). Alternate transcription by the RNA polymerase III can also produce non-polyadenylated lncRNAs but they are poorly characterized (Kapranov et al., 2007). Compared to their protein-coding counterparts, lncRNAs are more tissue-specific (78% for lncRNAs versus 18% for mRNAs), less expressed, and enriched in the nucleus (Quinn et al., 2016).

Based on their localization relative to nearby protein-coding genes, lncRNAs have been subdivided into i) antisense lncRNAs encoded in opposite direction and overlapping at least partially with the exonic sequence of a protein-coding gene, ii) intronic lncRNAs that are encoded within introns of coding genes, and iii) intergenic lncRNA (lincRNA) that are entirely encoded within the region between successive coding genes (**Figure 1-3**) (Rinn et al., 2012).

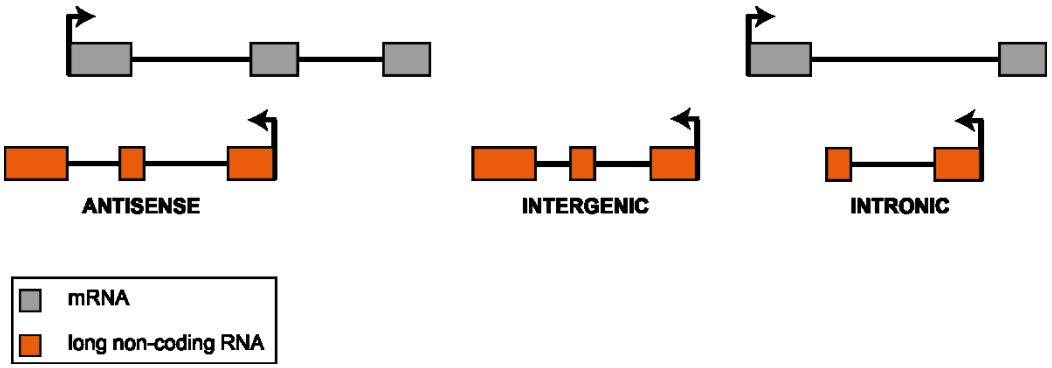


Figure 1-3: Classification of lncRNAs

LncRNAs are defined as antisense, intergenic and intronic lncRNA based on their location relative to neighboring protein-coding genes. Grey or orange rectangles depict exonic regions of an mRNA or lncRNA, respectively, whereas black lines represent intronic regions.

RNA-seq data across 24 human tissues and cell lines identified a catalog of 8195 lncRNAs (Cabili et al., 2011). This number expanded to about 101,700 lncRNA genes in humans according to the NONCODE database (Zhao et al., 2016). Despite their abundance, only 184 lncRNAs (lncRNA database, version from May 2016) are so far functionally validated in humans (Quek et al., 2015). Functional studies have associated lncRNAs with numerous roles, ranging from alternative splicing regulation to mRNA stability and modulation of translation (Rashid et al.,2016). A broad area of their function involves transcriptional regulation, notably through the binding of transcriptional regulators and chromatin modifiers (Maruyama et al., 2012). Experimental evidence suggests that lncRNAs play key roles in the recruitment, scaffolding, sequestration and guiding of such complexes in the genome (Rinn et al., 2012).

Dependent on the expression level, stability, and cellular context of a specific RNA as well as the three-dimensional folding of chromatin, the targeting of a lncRNA occurs either on local genes

(*cis* regulation) or on more distant chromosomal domains or different chromosomes (*trans* regulation) (Guttman et al., 2011). Around 70% of the coding sense transcripts have a reported antisense non-coding counterpart (He et al., 2008) and often share a coordinated expression pattern. In many cases, the antisense RNA has been shown to be involved in the *cis* regulation of the sense transcript (Orom et al., 2010).

1.2.2 Involvement of lncRNAs in epigenetic regulation

A single lncRNA can modulate gene expression through the recruitment of various writers, erasers, and readers of functionally related chromatin marks, and guide or titrate them away to the *cis* or *trans* location of their targets. Initial evidence into how lncRNAs are involved in epigenetic regulation came from the study of the lincRNA *XIST* (X inactive-specific transcript) during X chromosome inactivation. *XIST* is expressed from one of the two female X chromosomes and locally recruits the polycomb repressive complex 2 (PRC2) (Rinn et al., 2007). This recruitment results in H3K27me3 deposition leading to transcriptional silencing of the X chromosome from which *XIST* is transcribed (**Figure 1-4**).

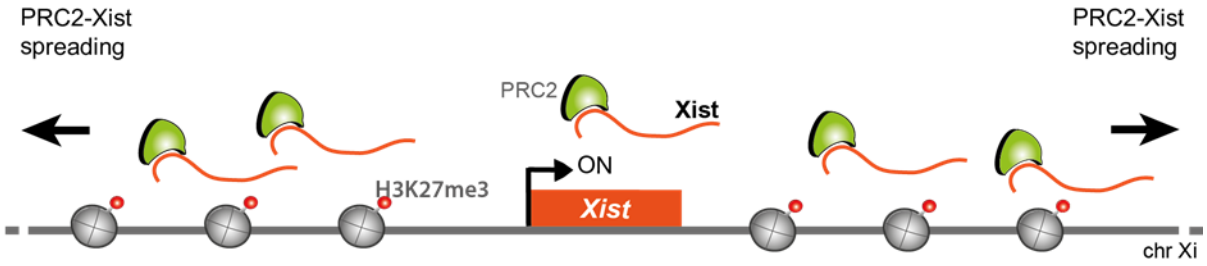


Figure 1-4: Epigenetic repression mediated by the lncRNA *XIST*

XIST lncRNA mediates X chromosome inactivation (chr Xi) through binding to the PRC2 complex and subsequent deposition of the repressive H3K27me3 mark on the entire Xi chromosome.

Interaction of lncRNAs and the PRC2 complex was demonstrated through sequencing of PRC2-bound RNAs to be a common feature for thousands of lncRNAs. Knockdown of selected lncRNAs associated with PRC2 components proved their requirement for proper epigenetic and transcriptional repression of PRC2 targets (Khalil et al., 2009, Zhao et al., 2010).

More recent studies have also demonstrated the recruitment of gene activating activities. For instance, the lncRNA *TARID* (TCF21 antisense RNA inducing demethylation) keeps the *TCF21* gene (transcription factor 21) in a transcriptionally active state through recruitment of the GADD45A-TDG-TET DNA demethylation complex (Arab et al., 2014) (**Figure 1-5A**). In another example, **ecCEBPA** (extra-coding CEBPA) binds to DNMT1 protein and in a decoy mechanism prevents the methylation of *CEBPA* gene promoter (**Figure 1-5B**). This notion has been extended by sequencing the RNAs that co-immunoprecipitate with DNMT1, suggesting that local lncRNA-DNMT1 interaction is a prevalent mechanism in regulating the methylation pattern of genes located *in cis* (Di Ruscio et al., 2013).

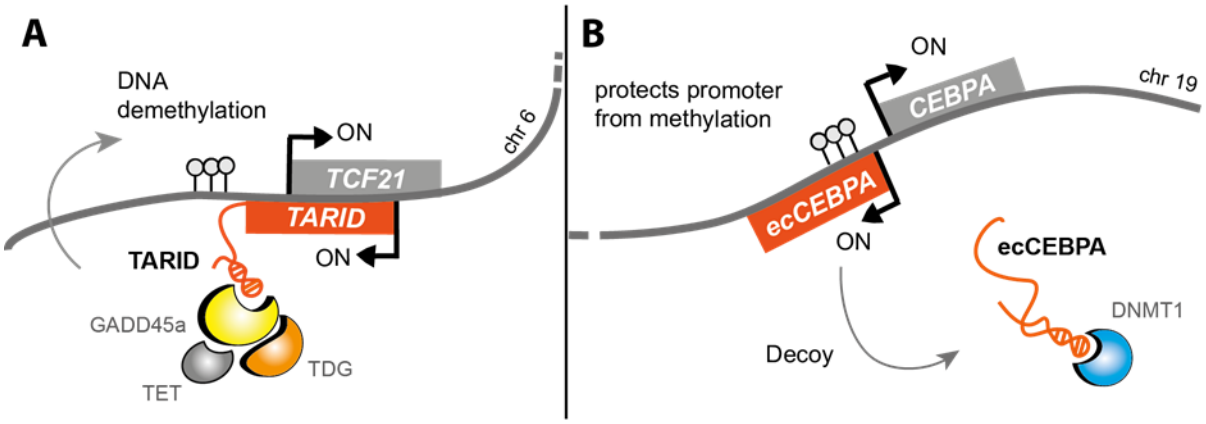


Figure 1-5: Epigenetic activation mediated by lncRNAs

(A) *TARID* activates TCF21 transcription through binding and recruitment of the GADD45A-TDG-TET DNA demethylation complex to the *TCF21* promoter. **(B)** The lncRNA *ecCEBPA* binds and sequesters DNMT1 to prevent *CEBPA* promoter methylation.

These reports imply that lncRNAs can bind to histone and DNA methylation modifiers, which in turn can modulate the chromatin status of specific target genes, *via* either specific targeting or the sequestration of epigenetic modifiers.

1.2.3 Do lncRNAs code for proteins?

Among the plethora of identified lncRNAs, only a minority has been linked to a function, which addresses the question whether some of them may actually encode for short functional peptides.

LncRNAs and mRNAs are generated by the same transcriptional machinery and thus share many features (**section 1.2.1**). LncRNAs differ from mRNAs by their lack of an extended open reading frame (ORF) encompassing a minimum of 100 amino acids (aa). The detection of ORFs is therefore fundamental for the definition of lncRNAs and has been assessed by searching for evolutionary conservation, matches to proteomic or domain databases (e.g, BLASTP, Swissport, PFAM) and for codon occurrence patterns. These tools, however, are designed to identify conserved and larger ORFs (>100 aa) enriched for these features (Andrews et al., 2014). Ribosomal profiling has been used to assess the number of lncRNAs bound and scanned by ribosomes. Interestingly, several studies found that many lncRNAs engage to ribosomes, and their footprints correspond to potential short ORFs (Ingolia et al., 2011, Bazzini et al., 2014, van Heesch et al., 2014). Others however argued that the ribosome patterns seen for lncRNAs are similar to the ones from known non-coding RNAs, suggesting that lncRNAs are not translated (Guttman et al., 2013).

The interpretation of the association of lncRNAs to ribosomes is therefore controversial, and lncRNA translation does not guaranty the production of a functional and stable peptide. In concordance with this observation, proteomic analysis by mass spectrometry indicated that 92% of lncRNAs annotated in GENCODE are not translated in the leukemic K562 and in the B−lymphocytic GM12878 cell lines (Banfai et al., 2012). This is in line with another study where only eight lncRNAs harbored a micropeptide in K562 cells (Slavoff et al., 2013). These reports are in contrast with a study by Kuster and colleagues that delineated a set of 430 peptides from 404 lncRNAs detected in 16,857 mass spectrometry data from human tissues, cell lines and body fluids (Wilhelm et al., 2014). Reanalysis of these data by Tress and colleagues concluded, however, that the results contain many false positives due to poor data filtering (Ezkurdia et al., 2014). Genome wide identification of lncRNAs encoding peptides is limited to a handful. This is possibly due to the limited resolution to peptide smaller than 50 amino acids of the current mass spectrometry technology along with a low stability of potential peptides (Bazzini et al., 2014).
Olson and colleagues have recently identified two peptides MLN (myoregulin) (Anderson et al., 2015) and DWORF (dwarf open reading frame) (Nelson et al., 2016) encoded in two transcripts

annotated as lncRNA. These reports demonstrate that lncRNAs can indeed code for stable and functional peptides.

In addition to their coding potential, lncRNAs may function as RNA molecules. The *SRA* (steroid receptor RNA activator) gene is a well characterized lncRNA, but also produces a protein-coding variant; both are independently involved in the co-activation of nuclear receptors (Lanz et al., 1999, Chooniedass-Kothari et al., 2004, Cooper et al., 2011). Future studies and advances in the peptidomic field may reveal other lncRNAs that encode functional small ORFs.

1.3 Prostate cancer

1.3.1 Prostate Cancer: Clinical and Genetic features

Prostate adenocarcinoma (PCa) is the most common type of cancer and the third cause of cancer-related deaths among men in developed countries (Society, 2015). The incidence of PCa has been rising, since the approval of PCa detection in blood by prostate-specific antigen (PSA) testing by the Food and Drug Administration in 1987 (Potosky et al., 1995). With 758,700 men diagnosed and an estimated 142,000 death cases in developed countries in 2012, PCa is the cancer with the highest incidence in men (Society, 2015). The incidence increases with age, with a median age of about 70 years at diagnosis. About 2% of PCa patients are however detected earlier at the age of 50 years or below and are defined as early-onset prostate cancer (EO-PCa) patients. This subset of patients is believed to represent a different cancer entity compared to the late onset PCa cancer cases (LO-PCa) (Weischenfeldt et al., 2013). Since the 1990s, most men diagnosed with PCa have an improved survival rate over 5 years, reaching 99% in developed countries (Siegel et al., 2016). Notably, EO-PCa was associated with slightly lower survival rates (98%) compared to elderly patients in other studies (Salinas et al., 2014).

Normal prostate epithelium is composed of a balanced amount of basal and luminal cells with rare neuroendocrine cells (Abate-Shen et al., 2000). The prostate carcinogenic tissue was instead shown by lineage marking to preferentially stem from luminal cells (Wang et al., 2014). PCa

develops in a series of steps from prostatic intraepithelial neoplasia (PIN) to localized tumors, which can further progress to metastatic castration-resistant prostate cancer (mCRPCa). The staging of a prostate tumor is assigned by the TNM (Tumor Node Metastasis) classification, combining the information about the extent of the primary tumor (T category) and the spread to nearby lymph nodes (N category) or to other body parts (M category), further subdivided in function of their aggressiveness from T0 to T4, NX to N1 and M0 to M1c, respectively. The aggressiveness of the tumor is, in addition, measured by PSA levels and through evaluation of the cancer architecture by the Gleason score ranging from 6 (low grade, lowly aggressive) to 10 (high grade, highly aggressive) (N. Mottet, 2015).

The etiology underlying PCa development and progression remains poorly understood. The risk factors include, but are not limited to, increasing age, family history of PCa or breast cancer, single nucleotide polymorphisms in the 8q24 region (Al Olama et al.,2009) and African decent (Pienta et al., 1993).

Genome-wide studies have improved our understanding of the genetic alterations contributing to the malignant transformation and progression of PCa. Overall, only a few genes are mutated in PCa with a somatic mutation rate of 1.4 per Megabase (Barbieri et al., 2012) compared to 3.8 mutations per Megabase in lung adenocarcinoma (Kan et al., 2010). Whole-exome sequencing enabled the detection of twelve recurrently mutated genes, with a prevalence for alterations in *SPOP* (Speckle-type POZ Protein), *TP53* (Tumor Protein P53) and *PTEN* (Phosphatase and Tensin Homolog) genes (Barbieri et al., 2012). In mCRPCa, the most frequently mutated gene is the androgen receptor (AR). Alterations of AR through mutations, expression of splice variants, focal gene amplification (Xq12) and/or upregulation (Taylor et al., 2010, Grasso et al., 2012), enable the cells to overcome AR deprivation therapy, and result in the constant activation of the receptor or its activation by alternative ligands than steroids (Mills, 2014).

Gene fusions dominate in PCa, involving members of the *ETS* transcription factor family and the androgen-regulated *TMPRSS2* (transmembrane protease serine 2) gene. In approximately 50% of all the PCa cases, *TMPRSS2* promoter is fused to *ERG* (v-ets erythroblastosis E26 oncogene homolog) gene (*TMPRSS2:ERG*), leading to androgen-dependent ERG overexpression (Tomlins et

al., 2005). The impact of the *TMPRSS2:ERG* fusion on PCa outcome is controversial, reported to be linked with good prognosis in early-onset patients (Steurer et al., 2014), bad prognosis or not linked to any clinical outcome in other reports (Kumar-Sinha et al., 2008). Nevertheless, this fusion gene along with a second oncogenic event such as *PTEN* (phosphatase and tensin homolog gene) deletion on 10q23 or PI3-kinase (phosphoinositide 3-kinase) pathway activation was shown to be necessary for the formation of PIN lesions as precursors of PCa (Carver et al., 2009, King et al., 2009). Focal deletion of the region spanning *FOXP1* (Forkhead box protein P1), *RYBP* (Ring andYY1 binding protein) and *SHQ1* genes on 3p14, and *TP53* on 17p31 were also associated with *TMPRSS:ERG* fusions (Taylor et al., 2010). Other common rearrangements are the recurrently deleted regions targeting in 30 to 50% of cases; *NKX3.1* (NK3 Homeobox 1) on 8q21, *RB1* (Retinoblastoma 1) on 13q14, and the amplification of 8q chromosome harboring, notably *c-Myc* (V-Myc Avian Myelocytomatosis Viral Oncogene Homolog) gene in 20 to 40% of the cases (Tomlins et al., 2005, Taylor et al., 2010). Compared to elderly patients, structural rearrangements in EO-PCa patients affect mainly androgen-driven genes, especially with a higher frequency of *TMPRSS2:ERG* fusion. The authors suggest that this phenomenon is the cause of elevated androgen levels in younger men (Weischenfeldt et al., 2013).

1.3.2 Epigenetic alterations in prostate cancer

Cancer genomes display frequent alterations in epigenetic enzymes (Plass et al., 2013). A number of chromatin/histone remodelers exhibit single nucleotide variations in primary tumors (Barbieri et al., 2012) and more prominently in castration-resistant PCa (Grasso et al., 2012). This supports the existence of a close interplay between genetic and epigenetic mechanisms during carcinogenesis. Integration of epigenetic and genetic profiles demonstrated a global increase in DNA methylation in *TMPRSS2:ERG* fusion negative patients, in comparison to patients harbouring the fusion gene (Kim et al., 2011, Borno et al., 2012). The link between genetic and epigenetic alterations was also demonstrated through the concomitant acquisition of CNVs along with changes in DNA methylation patterns during PCa tumor evolution (Brocks et al., 2014). Investigation of the three-dimensional chromatin organization by Hi-C (Genome-wide

Chromosome Conformation Capture) in LNCaP and PC3 prostate cancer cell lines, compared to the normal cell line PrEC, mechanistically linked the coevolution of genetic and epigenetic aberrations. Indeed, Clark and colleagues demonstrated that the presence of CNVs in cancer cells coincide with a different genome spatial organization. The authors linked the acquisition of novel topological domain boundaries to an altered chromatin and transcriptional state (Taberlay et al., 2016).

Alteration of the normal epigenetic landscape is known to participate in tumor development by enabling the acquisition of the hallmarks of cancer (Hanahan et al., 2011). Cancer genomes are characterized by a global loss of 5mC (hypomethylation) mainly affecting gene poor regions such as repetitive elements, centromeres and telomeres as well as gene bodies. The degree of hypomethylation was shown to increase from benign to metastatic stage in PCa cell lines (Yegnasubramanian et al., 2008). Decondensation of the chromatin structure through DNA hypomethylation leads to chromosome instability, allowing the reactivation of transposable elements (Ehrlich, 2002). In addition to genomic instability, loss of methylation in repeats was associated with the activation of not only alternative transcription (Wolff et al., 2010), but also oncogenes (Lamprecht et al., 2010). Hypomethylation of promoters is less common, but was reported in metastatic PCa cell lines and tissue to affect predominantly a large fraction of genes belonging to the family of cancer-testis antigens (Yegnasubramanian et al., 2008).

Hypermethylation is instead rather restricted to promoter regions of genes, especially in CGIs, and is associated with tumor suppressor gene (TSG) repression and gain of repressive histone marks (Jones et al., 2007). Characterization of genome-wide DNA methylation patterns in prostate tissue and cell lines delineated hundreds of differentially methylated CGI-associated genes (Kron et al., 2009, Kim et al., 2011, Kobayashi et al., 2011, Borno et al., 2012). These studies revealed that various cancer-relevant cellular pathways are deregulated by a gain of DNA methylation in PCa, such as cell cycle regulation (e.g, *p16* also known as *CDKN2A* or *INK4a*), hormonal response (e.g, Retinoic Acid Receptor β (*RARB*)) or DNA repair (e.g, glutathione S-transferase Pi (*GSTP1*)). However, the role of DNA methylation in gene repression and PCa

progression has recently been challenged. Indeed, Pellacani and colleagues demonstrated that among all genes reported to be hypermethylated and repressed in PCa, 5% of them including *GSTP1* are not repressed as a consequence of DNA methylation gain, but as a result of loss of active histone marks during the differentiation process of normal luminal cells (Pellacani et al., 2014).

Epigenetic inactivation is not only restricted to single genes but can also encompass multiple adjacent CGIs in PCa. Indeed, gain of DNA methylation, repressive histone marks and histone deacetylation were associated with the epigenetic silencing of multiple TSGs across domains of several Megabases (Coolen et al., 2010).

Intriguingly, CGIs associated with active transcription were more frequently hypermethylated than hypomethylated in PCa cells. Clark and colleagues found that gain of DNA methylation across CGI or in CGI borders along with gain of active and loss of repressive histone marks is associated with long range epigenetic activation. The authors proposed that hypermethylation in CGI borders could hinder the binding of transcriptional repressor, while methylation throughout the CGI is related to gene activation by H3K4me3 gain of alternate gene promoter usage (Bert et al., 2013).

DNA methylation or chromatin modifications are collectively and significantly altered in PCa. Nevertheless, the specific mechanism by which epigenetic marks are established and removed at specific loci remains unclear. In the past years, accumulating evidence has suggested that non-coding RNAs might be involved in this process.

1.3.3 Epigenetic regulation by lncRNAs in prostate cancer

Recent large scale RNA profiling projects have identified a multitude of novel lncRNAs dysregulated in PCa. About 121 lncRNAs dysregulated in PCa termed the PCAT family (Prostate Cancer Associated intergenic non-coding RNA Transcripts) were discovered by Prensner and colleagues (Prensner et al., 2011). Transcriptional profiling of fourteen tumors derived from

Chinese PCa patients identified 406 differentially expressed lncRNAs during prostate carcinogenesis (Ren et al., 2012). Independent of RNA-sequencing, Liu and colleagues have developed a reannotation pipeline of the Affymetrix microarray probes that map to lncRNAs. With this methodology and based on available expression data, they characterized a set of 102 novel lncRNAs upregulated in PCa (Du et al., 2013).

Despite the growing number of lncRNAs identified in PCa, only a few have been functionally characterized so far. Emerging mechanisms that involve lncRNAs trigger PCa development and progression by influencing PTEN/AKT pathway, AR signaling, and transcriptional regulation. These roles are predominantly accomplished in a lncRNA-dependent recruitment and guiding of chromatin modifiers to specific loci, and can be related to the altered chromatin patterns reported in cancer cells leading to oncogene activation and TSG inactivation (Gregory et al., 2004).

The current knowledge of lncRNAs that are deregulated in PCa and implicated in the targeting of chromatin remodeling complexes with an activation and/or repressive transcriptional outcome is summarized in **Table 1-1**.

Table 1-1 LncRNAs mediating chromatin remodeling in prostate cancer

lncRNA	Epigenetic interactors	Type of regulation	Target gene(s)	Role in carcinogenesis	Reference
TRANSCRIPTIONAL ACTIVATORS					
PCGEM1 (*LINC00071*)	PYGO2 (Pygopus 2)	*Trans*	Multiple AR regulated genes	Oncogene ↑ proliferation ↓ apoptosis	(Srikantan et al.,2000) (Petrovics et al.,2004) (Fu et al.,2006) (Yang et al.,2013) (Prensner et al.,2014)
PRNCR1 (*PCAT8*)	DOT1L (DOT1-like histone H3K79 methyltransferase)	*Trans*	Multiple AR regulated genes	Oncogene ↑ cell viability	(Chung et al.,2011) (Yang et al.,2013) (Prensner et al.,2014)
TRANSCRIPTIONAL REPRESSORS					
CTBP1-AS (*PCAT10*)	HDAC/Sin3A (histone deacetylase and sin3A complex)	*Cis* and *trans*	*CTBP1* and multiple AR regulated genes	Oncogene ↑ tumor growth ↑ cell cycle progression ↑ proliferation	(Takayama et al.,2013)
PCAT-1	PRC2 (Polycomb complex 2)	*Trans*	**BRCA2**, *CENPE*, *CENPF*	Oncogene ↑ proliferation	(Prensner et al.,2011)
ANRASSF1	PRC2	*Cis*	*RASSF1A*	↑ proliferation ↓ apoptosis	(Beckedorff et al.,2013)
ANRIL (*p15AS*)	PRC1 (Polycomb complex 1) PRC2	*Cis*	*CDKN2A, CDKN2B, ARF*	↑proliferation ↓ senescence	(Yap et al.,2010) (Kotake et al.,2011)
TRANSCRIPTIONAL ACTIVATOR AND REPRESSOR					
SChLAP1 (*PCAT114*)	SWI/SNF (SWItch/Sucrose NonFermentable)	*Trans*	multiple targets	Oncogene ↑ invasiveness and metastasis	(Prensner et al.,2013)

Abbreviations: ↑ enhanced activity; ↓ reduced activity; *CTBP1* (C-terminal binding domain protein 1); *BRCA2* (breast cancer 2, early onset); *CENPE* and *CENPF* (centromere-associated proteins E and F); *RASSF1A* (Ras Association Domain family 1, isoform A); *CDKN2A* (known as *p16* or *INK4a*); *CDKN2B* (known as *p15* or *INK4b*); *ARF* (alternate reading frame of the *INK4a/ARF* locus).

Only two lncRNAs, ***PCGEM1*** (prostate cancer gene expression marker 1) and ***PRNCR1*** (prostate cancer non-coding RNA 1), both overexpressed in PCa (Petrovics et al., 2004) (Chung et al., 2011), have so far been associated with chromatin-mediated gene activation. They have been shown to activate the AR even in the absence of its ligand. RNA immunoprecipitation experiments demonstrated the interaction between *PRNCR1* and the AR. This binding leads to the recruitment of DOT1L (DOT1-like histone H3K79 methyltransferase) methyltransferase, which methylates AR, resulting in its subsequent interaction with *PCGEM1*. In turn, *PCGEM1* recruits PYGO2 (Pygopus 2), which enables the interaction of AR to H3K4me3 chromatin marks in the promoter regions of AR-regulated genes and leads to their activation (Yang et al., 2013). However, neither the reported *PRNCR1* overexpression in PCa, nor the interaction of both lncRNAs with the AR could be confirmed or recapitulated by Chinnaiyan and colleagues (Prensner et al., 2014).

CTBP1 (C-terminal binding domain protein 1) is a transcriptional co-repressor of the *AR*. The interaction of its antisense transcript ***CTBP1-AS*** (C-terminal binding protein 1 antisense RNA) with the transcriptional repressor PSF (PTB-associated splicing factor) and HDAC/Sin3A complex leads to the repression of the sense transcript *CTBP1*. Indirectly, *via* the repression of *CTBP1*, *CTBP1-AS* leads to AR target gene reactivation and cell cycle promotion. Moreover, *CTBP1-AS* was suggested to directly repress the expression of other genes *in trans* (Takayama et al., 2013).

Epigenetic silencing through the coupling of lncRNAs with the Polycomb complex was demonstrated for three lncRNAs in PCa. The nuclear fraction of the lincRNA ***PCAT-1*** (prostate cancer associated transcript-1) interacts with the PRC2 complex and mediates the *trans* repression of *BRCA2* (breast cancer 2, early onset) and the centromere-associated proteins E and F (CENPE and CENPF) genes. Consequently, *PCAT-1* upregulation promotes cell proliferation and impairs homologous repair in PCa (Prensner et al., 2011). ***ANRASSF1*** (antisense intronic non-coding RASSF1A) (Beckedorff et al., 2013) and ***ANRIL*** (antisense non-coding RNA in the INK4 locus) (Kotake et al., 2011) were also shown to mediate transcriptional repression through their binding to the Polycomb complex. Elevated expression of *ANRASSF1* and *ANRIL* in PCa was associated with increased proliferation and impaired cell death, as a result of the repression of

the TSG *CDKN2A/B* (Kotake et al., 2011) and *RASSF1A* (Ras association domain family 1, isoform A) (Beckedorff et al., 2013), respectively.

Gene expression is regulated by chromatin modifiers in concert with the remodeling of nucleosomes along the chromatin. The lincRNA ***SChLAP1*** (second chromosome locus associated with prostate 1, also called *LINC00913*), upregulated in about 25% of aggressive PCa, is involved in nucleosome remodeling by interacting with the SWI/SNF subunit SNF5/INI1 (also known as SMARCB1). Upregulation of this lincRNA in PCa cells was associated with decreased SNF5 recruitment to DNA, thus antagonizing the genome-wide localization and regulatory functions of the SWI/SNF complex. *SChLAP1* silencing in castration-resistant 22Rv1 cells have reduced propensity to form secondary lesions *in vivo*, further supporting its role in cancer progression to metastasis (Prensner et al., 2013).

Overall, these data have demonstrated that lncRNAs are involved in the activation and/or repression of genes participating in prostate carcinogenesis by binding to diverse epigenetic regulators, in particular the Polycomb complex.

2. AIM OF THE THESIS

Dysregulation of promoter methylation is a well-defined phenomenon in PCa, resulting in the malfunction of cellular pathways (**section 1.3.2**). The actors of DNA methylation establishment, maintenance, and reversal are well documented (**section 1.1.3**). Nevertheless, the specific mechanism by which DNA methylation modifiers are recruited in a spatial and temporal manner remains unclear. Accumulating evidence suggests that lncRNAs are involved in the modulation of the DNA methylation status of specific target genes *via* either scaffolding, sequestration and/or targeting of DNA methylation modifiers to specific loci (**section 1.2.2**). Despite the growing number of lncRNAs identified and deregulated in prostate tissue, our knowledge regarding their mechanism of action is limited to a handful, and none of them have so far been documented to be implicated in DNA (de)methylation mechanisms (**section 1.3.3, Table 1**).

Since alterations in DNA methylation of protein-coding genes are associated with prostate carcinogenesis and might be regulated by RNAs transcribed in *cis*, we aimed to investigate lncRNAs as potential mediators of DNA (de)methylation in PCa through:

- **Identification of lncRNAs involved in *cis* DNA (de)methylation mechanisms in PCa.**

By integrating DNA methylation and RNA sequencing data generated by collaborators from the International Cancer Genome Consortium (ICGC) project on EO-PCa patients, we planned to identify lncRNAs regulating the transcription of protein-coding gene *in cis*, through modulation of their promoter DNA methylation patterns (simplified as mRNA/lncRNA pairs).

- **Investigation of the clinical relevance of a candidate lncRNA in cancer.**

The dysregulation of a selected mRNA/lncRNA pair will be evaluated in PCa validation cohorts, as well as in other tumor entities. The association between the expression of the pair with clinical and genetic features will be evaluated to select a pair crucial for carcinogenesis.

- **Characterization of the expression of the lncRNA *in vitro***

Detailed investigation of the lncRNA variant being deregulated and the specific chromatin marks associated with its expression in various cancer cell lines will be investigated.

- **Investigating if the lncRNA mediates *in cis* DNA (de)methylation *in vitro***

Interdependency between the expression of the lncRNA and the methylation status of the gene *in cis* will be evaluated. This will be achieved by comparing the methylation and expression levels of the corresponding transcripts, by Mass Array and RT-qPCR, after modulation of the lncRNA expression through knockdown or overexpression.

- **Investigating if the coding potential of the lncRNA**

We tested first *in silico* and then *in vitro* whether DLX6-AS1 could exert its function through the production a small protein or peptide.

3. MATERIAL AND METHODS

3.1 Patient material and utilized datasets

All ICGC EO-PCa samples originate from radical prostatectomy of patients diagnosed with PCa and obtained from the Martini Clinics at the University Medical Center Hamburg-Eppendorf. Genome-wide DNA methylation was analyzed on Illumina HumanMethylation450 beadChip arrays (450k) (Illumina). PolyA+ RNA-sequencing was carried out on the Illumina HiSeq2000 instrument and mapped to the reference human transcriptome Gencode version 21. Both RNA-seq and 450k data of EO-PCa patients were generated and processed by collaborators of the ICGC project on EO-PCa.

RNA-seq and 450k data generated and processed as part of The Cancer Genome Atlas (TCGA) project on late-onset PCa, lung squamous cell carcinoma (LUSC), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), lung adenocarcinoma (LUAD) and kidney renal clear cell carcinoma (KIRC) were downloaded from the TCGA Data Matrix (https://tcga-data.nci.nih.gov/).

For the genome-wide screening of mRNA/lncRNA pairs (**section 4.1**), we employed three datasets referred to as the "screening set" in **Table 3-1**. Among them are two different RNA-seq sample sets of ICGC EO-PCa patients analyzed at the MPI in Berlin named "EO-PCa Berlin" and at the DKFZ in Heidelberg termed "EO-PCa Heidelberg", as well as one combined dataset of ICGC EO-PCa and TCGA PCa 450k data (TCGA data release 2013) named as "EO-PCa +TCGA".

The datasets designed as "validation set" in **Table 3-1** were used to confirm the result of the screening strategy (**section 4.2.1**). At the time of the screening for mRNA/lncRNA pairs, the initial ICGC cohort for EO-PCa patients was restricted to 6 normal and 25 PCa tissue samples (Berlin+Heidelberg). Major sequencing efforts extended the EO-PCa set to 10 normal and 91 tumors samples, after exclusion of samples failing to pass RNA-seq quality metrics based on strand specificity, ribosomal RNA contamination, sequence duplicates, 5' and 3' coverage bias of

transcripts, gene body coverage, and transcript integrity number. The corresponding 450k profiles from patients with both transcriptome and DNA methylation profiles were used in this validation set.

An independent cohort of late-onset PCa patients available from the TCGA database (TCGA data release 2015) was used to validate the result from the ICGC cohort. In the PCa TCGA dataset, we restricted our analysis to patient-matched tumor/normal samples. Corresponding 450k data from patients with both RNA-seq and 450k data were used for this validation. The clinical characteristics of each cohort within both the screening and the validation datasets are summarized in **Table 3-1**.

Table 3-1 Clinical characterization of PCa sample sets

		SCREENING SET			VALIDATION SET			
Type of data		**RNA-seq**		**450k**	**RNA-seq**	**450k**	**RNA-seq**	**450k**
Cohort		**EO-PCa Berlin**	**EO-PCa Heidelberg**	**EO-PCa + TCGA**	**EO-PCa**	**EO-PCa**	**TCGA matched**	**TCGA matched**
Number of Normal/ Tumor		2/12	4/13	60/342	10/91	8/89	52/52	34/34
Age	Median	48	47	60	48	48	61	62
	Range	38-51	40-50	32-72	32-52	32-52	43-72	44-72
Gleason score (n)	N/A	0	0	5	4	3	0	0
	6	0	1	49	12	12	5	2
	7	11	12	238	63	62	40	27
	8	0	0	8	1	1	3	2
	9	1	0	42	11	11	4	3
Pathological T stage (n)	N/A	0	0	8	3	3	0	0
	T2	7	10	214	60	58	29	19
	T3	5	3	103	24	24	21	13
	T4	0	0	17	4	4	2	2

To characterize DLX6/DLX6-AS1 expression across multiple cancers (**section 4.2.2**), we used TCGA data on LUSC, COAD HNSC, LUAD and KIRC (TCGA data release 2015). Corresponding 450k data from patients with both RNA-seq and 450k data were employed. The clinical characteristics of each cohort are summarized in **Table 3-2**.

Table 3-2 Clinical characterization of additional TCGA cancer datasets

Dataset		RNA-seq					450k				
		LUSC	COAD	HNSC	LUAD	KIRC	LUSC	COAD	HNSC	LUAD	KIRC
Number of Normal/ Tumor		50/ 503	41/ 286	43/ 519	58/ 511	72/ 533	8/ 359	19/ 246	20/ 448	21/ 416	24/ 291
Age	Median	68	67	61	66	61	69	66	61	67	61
	Range	39-90	31-90	19-90	38-88	26-90	40-90	31-90	19-90	40-88	26-90
Pathological T stage (n)	T1	113	6	33	169	272	85	6	29	139	136
	T2	287	43	149	273	69	196	35	135	224	38
	T3	69	193	131	47	180	57	174	117	36	109
	T4	22	39	179	19	11	11	30	146	15	8
	Tx	0	0	27	2	0	0	0	21	2	0
	Tis	0	1	0	0	0	0	1	0	0	0
	N/A	12	4	0	1	1	10	0	0	0	0

Abbreviations: Tx=primary tumor cannot be evaluated, Tis=carcinoma in situ, N/A=non-available

To characterize the impact of DLX6/DLX6-AS1 overexpression on overall and disease-free survival time (**section 4.2.3B**), we used the TCGA cohorts on PRAD, LUSC, COAD HNSC, LUAD and KIRC described as the "RNA-seq" datasets in **Table 3-2**. Each cohort was separated into "LOW" and "HIGH" subgroup based on the median expression value of DLX6-AS1 lncRNA (median value for PRAD=2.99, LUSC=4.79, COAD=1.18, HNSC=3.21, LUAD=0.609, KIRC=1.77). The numbers of events for both subgroups in each dataset are summarized in **Table 3-3**.

Table 3-3 Clinical characterization of "LOW" and "HIGH" expression subgroups in the TCGA cancer datasets

		PRAD		LUSC		COAD		HNSC		LUAD		KIRC	
Subgroup		LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH	LOW	HIGH
Number of Tumor (n)		250	248	251	250	142	143	260	259	256	255	266	267
Overall survival (n)	Living	244	243	135	149	104	109	146	151	165	161	167	190
	Deceased	5	5	116	100	38	31	113	108	91	93	99	76
	N/A	1	0	0	1	0	3	1	0	0	1	0	1
Disease Free Survival (n)	Disease Free	198	201	119	126	87	89	124	123	127	123	146	162
	Recurred	48	44	66	64	37	34	58	66	92	90	68	58
	N/A	4	3	66	60	18	20	78	70	37	42	52	47

Abbreviation: N/A=non-available

Table 3-4 Clinical characterization of "LOW" and "HIGH" expression subgroups in LUAD

		LUAD					
DLX6 probe ID		242940_x_at		239309_at		221289_at	
Subgroup		LOW	HIGH	LOW	HIGH	LOW	HIGH
Number of Tumor (n)		338	335	339	334	387	333
Overall survival	Living	217	189	219	187	260	173
	Deceased	121	146	120	147	127	160

In addition, an independent and larger cohort of LUAD cases was retrieved from the Kaplan Maier online survival analysis software (Gyorffy et al., 2013). Similarly to TCGA datasets, this cohort was divided into "LOW" and "HIGH" groups based on the mean expression value of DLX6 measured by expression arrays with the three following probes 242940_x_at, 239309_at and 221289_at. The numbers of events for every probe in both subgroups are summarized in **Table 3-4**.

3.2 Human cell lines and cell culture

Table 3-5 Cell culture material

Material	Manufacturer
Trypsin EDTA (1x) phenol red free	GE Healthcare
DMEM, RPMI-1640 and F12-K Nutrient Mix medium with phenol red	Gibco, Invitrogen, Life Technologies
RPMI-1640 Medium without phenol red and $NaHCO_3$ -powder	Sigma-Aldrich
Dulbecco's Phosphate-Buffered Saline (D-PBS)	Gibco, Invitrogen, Life Technologies
Fetal Bovine Serum (FBS)	Biochrom, Millipore
Androstan-17β-ol-3-one (testosterone)	Sigma-Aldrich
Hydrocortisone (≥98% purity)	Sigma-Aldrich
Insuline solution, human	Sigma-Aldrich
Transferrin (≥98% purity)	Sigma-Aldrich
D-(+)-Glucose (≥99.5% purity)	Sigma-Aldrich
$NaHCO_3$ ACS reagent (≥99.7% purity)	Sigma-Aldrich
HEPES (≥99.5% purity)	Sigma-Aldrich
Sodium Pyruvate, 100 mM	Gibco, Invitrogen, Life Technologies
Venor®GeM Classic	Minerva Biolabs GmbH

The benign immortalized prostate cell lines BPH1 and PNT2 as well as the cancerous cells originating from prostate (LNCaP, PC3, DU145, VCaP and 22RV1), colon (HCT116 and HCT116 DKO), lung (A549 and H1299) and embryonal kidney (HEK293T) tissue were cultured in their respective medium supplemented with 10% FBS (**Table 3-6**). All cells were grown under S1 biosafety level with the exception of VCaP cells cultured under S2 biosafety level. The cell lines were grown in humidified atmosphere with 5% CO_2 at 37°C. Cell lines were passaged every three to four days by dissociating D-PBS-washed cells with trypsin-EDTA and reseeding the cells at a dilution ranging from 1:3 for the slow growing cells VCaP and LNCaP and 1:10 for the remaining cell lines. All cell lines were regularly confirmed to be free of mycoplasma contamination by using the Venor®GEM Classic kit.

Table 3-6 Human cell lines and culture conditions

Tissue	Material	Medium	Origin
PROSTATE	**BPH1**	RPMI-1640 25 µg/ml testosterone, 0.5 mg/ml hydrocortisone, 100 µg/ml insulin, 25 mg/ml transferrin	(Hayward et al.,1995)
	PNT2	RPMI-1640	Prof. Dr. W. Schulz, Düsseldorf, Germany
	LNCaP	RPMI-1640 medium without phenol red and $NaHCO_3$ 4.5 g/L D-glucose, 1.5 g/L $NaHCO_3$, 2.283 g/L HEPES, 0.11 g/L sodium pyruvate pH 7.2	ATCC clone FGC CRL-1740
	PC3	RPMI-1640	ATCC CRL-1435
	DU145	RPMI-1640	ATCC HTB-81
	VCaP	DMEM	ATCC CRL-2876
	22RV1	45% DMEM +45% RPMI-1640	DSMZ ACC438
COLON	**HCT116**	DMEM	Dr. B. Vogelstein, Baltimore, USA
	HCT116 DNMT1-/-, DNMT3B-/- (HCT116 DKO)	DMEM	Dr. B. Vogelstein, Baltimore, USA
LUNG	**A549**	F12-K Nutrient mix	ATCC CCL-185
	H1299	RPMI-1640	Prof. Dr. Sültmann, DKFZ Heidelberg
KIDNEY	**HEK293T**	DMEM	Prof. Dr. Boutros, DKFZ, Heidelberg

3.3 Knockdown experiments

3.3.1 Cloning of shRNAs into shRNA expression vector

Table 3-7 Material used for cloning

Material	Manufacturer
shRNA template oligonucleotides, desalted	Sigma-Aldrich
pRSI9-U6-(sh)-HTS3-UbiC-TagRFP-2A-Puro	Cellecta, Inc
BbsI (10 u/µl)	New England Biolabs Inc.
Shrimp alkaline phosphatase (1 u/µl) (rSAP)	New England Biolabs Inc.
T4 Polynucleotide Kinase (PNK, 10 u/µl)	New England Biolabs Inc.
T4 DNA ligase (40 u/µl)	New England Biolabs Inc.
One Shot TOP10 chemically competent *E. coli*	Thermo Fisher Scientific

shRNA sequences used for cloning and targeting either human DLX6 (targeting DLX6 3'UTR or second exon), DLX6-AS1 (targeting the common last exon of both DLX6-AS1 T1 and T2 variants) or luciferase transcripts as control have been designed according to the DECIPHER Human Module 1-3 database and are listed in the Appendix (**Table 7-1**).

To clone shRNA template into the lentiviral expression vector pRSI9-U6-(sh)-HTS3-UbiC-TagRFP-2A-Puro plasmid, 2 µg of this plasmid was digested with 10 units BbsI enzyme and dephosphorylated with 1 µl rSAP overnight at 37°C in a final volume of 50 µl. Digestion product was loaded on an agarose gel, before gel extraction and purification of the digested plasmid. 2.5 µl of corresponding single stranded forward and reverse shRNA oligonucleotides (10 µM each) were simultaneously annealed and phosphorylated by adding 0.2 µl PNK and 1 µl of T4 ligase buffer to a final volume of 10 µl and incubated under following conditions: phosphorylation for 1 h at 37°C, inactivation of the enzyme for 20 minutes (min) at 65°C and annealing by gradual cooling (-0.1°C/sec) from 95°C to 4°C. Annealed oligonucleotides (diluted 1:10) were ligated with about 25 ng of the digested plasmid by using 0.4 µl of T4 DNA ligase in a final volume of 10 µl. The ligation reaction was incubated overnight at 16°C, followed by 10 min enzyme inactivation at 65°C. One Shot TOP10 *E.coli* (10 µl) were transformed with 1 µl of the ligation reaction by chemical transformation according to the manufacturer's instructions. Positive clones were identified by Sanger sequencing (GATC) using the "pRS" primer indicated in the Appendix (**Table 7-7**).

3.3.2 Lentivirus production and target cells transduction

Table 3-8 Material for Lentivirus production and target cells transduction

Material	Manufacturer
psPAX2 vector	Addgene
pMD2.G vector	Addgene
pRSI9-u6-sh-UbiC-TagRFP-2A-Puro plasmid	Cellecta
TransIT-LT1	Mirus Bio LLC
Puromycin	Sigma-Aldrich
RPMI-1640 medium with phenol red	Gibco, Invitrogen, Life Technologies
Puradisc FP30 Cellulose Acetate Syringe filter, 0.2µm, sterile	GE Healthcare Life Sciences

HEK293T were seeded in 6 well plates at a concentration of $5.25x10^5$ cells per well in 1.5 ml medium 24 h prior to transfection. To produce lentiviral particles, HEK293T cells were co-transfected with 1 µg pRSI9-U6-(sh)-HTS3-UbiC-TagRFP-2A-Puro shRNA vector along with 667 ng and 333 ng of the lentiviral packaging vectors psPAX2 and pMD2.G, respectively, using 7.5 µl TransIT-LT1 transfection reagent according to manufacturer's recommendation. 24 h after transfection 1 ml of medium was added to the cells. After one additional day, lentiviral particles present in the cell medium were collected and filtered through Puradisc filters to remove any contaminating cells. Supernatant containing lentivirus was either directly used for transduction or aliquoted before storage at -80°C.

VCaP cells were seeded in 6 well plates at a concentration of $5.25x10^5$ cells per well in 1.5 ml medium 24 h prior to lentiviral transduction. Virus titer was tested by qPCR after treating the cells with volumes ranging from 60 to 250 µl of supernatant containing virus particles. 24 h after transduction, virus containing media was exchanged. Stable clones were selected with puromycin 48 h after transduction. VCaP cells were harvested 8 days after the start of transduction.

3.3.3 siRNA and LNA (locked nucleic acids) transfection

Table 3-9 Material for siRNA or LNA transfection

Material	Manufacturer
TransIT-TKO transfection reagent	Mirus Bio LLC
RPMI-1640 medium	Gibco, Invitrogen, Life Technologies
Locked nucleic acids (LNA)	Exiqon
ON-TARGETplus siRNA	GE Dharmacon
Lincode siRNA	GE Dharmacon

siRNA or LNA directed against DLX6, DLX6-AS1 or non-targeting control listed in the appendix (**Table 7-2**) were supplied by GE Dharmacon as single or pool of 4. HEK293T cells were seeded in 6 well plates at a concentration of $5.1x10^5$ cells per well in 1.5 ml medium 24 h prior to transfection. LNA or siRNA at a final concentration of 25 nM were transfected with 10 µl TransIT-TKO according to the manufacturer's instructions. 72 h after transfection, cells were harvested and stored at -80°C until nucleic acids extraction.

3.4 RNA expression analysis

3.4.1 RNA fractionation

Table 3-10 Material for RNA fractionation

Material	Manufacturer
Dulbecco's Phosphate-Buffered Saline (D-PBS)	Gibco, Invitrogen, Life Technologies
RNasin ribonuclease inhibitors (40 u/µl)	Promega Corporation
Trypan Blue solution 0,4%	Thermo Fisher Scientific
2-Mercaptoethanol (≥99.0% purity)	Sigma-Aldrich
RLT buffer (RNeasy Mini kit)	Qiagen
Ethanol absolute (≥99.8% purity)	Sigma-Aldrich

Partition of cellular RNA into cytoplasmic and nuclear fractions was performed with VCaP, DU145 and HEK293T cells, based on Gagnon and colleagues protocol (Gagnon et al.,2014). For each cell line, two cell pellets of equal cell number (minimum $2.5x10^6$ cells) were collected and washed once with D-PBS. One of the cell pellets was prepared for RNA extraction by direct addition of 600 µl RLT lysis buffer supplement with 0.6 µl β-mercaptoethanol and represents the whole cell RNA content of the cell, whereas the second pellet was subjected to cellular fractionation. The second cell pellet was resuspended by gentle pipetting in 380 µl of ice-cold HLB hypotonic buffer (10 mM Tris pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 0.3% NP-40 and 10% glycerol) supplemented with 100 units of RNasin and incubated on ice for 10 min. After centrifugation at 1,000 g for 3 min at 4°C, the first supernatant and the first HLB-wash were retained as the cytoplasmic fraction. The cytoplasmic fraction was immediately resuspended with 1 ml ice-cold RNA precipitation solution (0.5 ml 3 M NaAc (pH 5.5) and 9.5 ml 100% ethanol) and stored at -20°C for a minimum of 1 h. The nuclear pellet was subjected to two additional washes in ice-cold HLB buffer and was then centrifuged at 200 xg for 3 min at 4°C. To assess the integrity and purity of nuclei preparation by microscopic inspection, 10 µl of the nuclei suspension was mixed with the same amount of Trypan blue. After the last centrifugation step and removing remaining HLB buffer, nuclei were resuspended in 600 µl RLT buffer supplemented with 0.6 µl β-mercaptoethanol and stored at -20°C. RNA from the previously collected cytoplasmic fraction was pelleted by centrifugation at full speed for 20 min and subsequently washed with ice-cold 70% ethanol, before air-drying the

pellet. The cytoplasmic pellet was then resuspended in 600 µl RLT buffer supplemented with 0.6 µl beta-mercaptoethanol and stored at -20°C until RNA extraction.

Whole cell, nuclear and cytoplasmic fraction in RLT buffer were thawed at room temperature before incubation at 37°C for 10 min. RNA was extracted from all fractions using the RNeasy kit with the procedure detailed in the next section. The relative enrichment of RNA species in the nuclear and cytoplasmic fraction measured by RT-qPCR was normalized to the whole cell RNA content of the cell by using the following formula 2ˆ(CT whole cell RNA)-(CT fraction).

3.4.2 RNA isolation and quantification of RNA by Reverse Transcription-quantitative PCR (RT-qPCR)

Table 3-11 Material for RNA isolation and RT-qPCR

Material	Manufacturer
RNeasy Mini kit or All Prep DNA/RNA Mini kit	Qiagen
RNasin ribonuclease inhibitors(40 u/µl)	Promega Corporation
RNase-free DNase set	Qiagen
Random primers	Thermo Fisher Scientific
SuperScript II Reverse Transcriptase kit	Thermo Fisher Scientific
Lightcycler®480 Real-Time PCR system	Roche Diagnostics
Lightcycler®480 Probes Master	Roche Diagnostics
Universal probe Library	Roche Diagnostics
Nanodrop ND-1000	Thermo Scientific

Total RNA was extracted either by using the RNeasy Mini kit or the All Prep RNA/DNA extraction kit according to manual instructions. Residual genomic DNA was digested on-column with DNase for 15 min. After RNA quantification by Nanodrop ND-1000, RNA was reverse transcribed. First, 2 µg total RNA was incubated at 65°C for 5 min with 200 ng random primers in a final volume of 46 µl. Then, first-strand synthesis was carried out by adding 16µl 5x first-strand buffer, 8 µl DTT, 4 µl dNTPs, 1 µl SuperScript II reverse transcriptase and 1 µl RNasin to a final volume of 80 µl under the following cycling conditions: 3 min at 25°C, 50 min at 42°C followed by a linear amplification at 70°C for 15 min.

Relative expression levels were determined by quantitative real time PCR, carried out with 2.5 µl diluted cDNA (dilution 1:2.5) along with 3.5 µl probes master, 1 µl gene specific primer mix (10

µM) and 0.05 µl universal probe to a final volume of 7 µl per well. The reaction was run on the Lightcycler 480 system under following conditions: 10 min initial denaturation at 95°C, 45 cycles of 10 sec at 95°C, 20 sec at 55°C and 10 sec at 72°C. All qPCRs were run in duplicates, *HPRT1* (hypocanthine-guanine phosphoribosyltransferase), *ALAS1* (5'-aminolevulinate synthase 1) (ALAS1) and *SDHA* (succinate dehydrogenase complex flavoprotein subunit A) housekeeping genes were used for normalization using the ΔΔCT-method (Livak et al.,2001). All gene-specific primers listed in the appendix (**Table 7-3**) were designed with the Roche Universal Probe Library Assay Design Center, synthesized and desalted by Sigma-Aldrich.

3.5 Rapid amplification of 5’ cDNA ends (5’RACE)

Table 3-12 5’RACE material

Material	Manufacturer
dNTPs (mix of 10 mM each)	Fermentas
ThermoScript™ Reverse Transcriptase	Thermo Fisher Scientific
RNasin ribonuclease inhibitors (40 u/µl)	Promega Corporation
RNase H (5 u/µl)	Thermo Fisher Scientific
Agencourt® RNAClean® XP	Beckman Coulter
Terminal Deoxynucleotidyl Transferase kit	Thermo Fisher Scientific
dATPs, 100 mM	Thermo Fisher Scientific
Expand High Fidelity PCR system enzyme	Roche Diagnostics
TOPO-TA cloning kit	Thermo Fisher Scientific

To determine the transcription start site of DLX6-AS1 transcript in VCaP cells we employed the 5’RACE assay adapted from Frohman and colleagues (Scotto-Lavino et al., 2006). Three DLX6-AS1 gene-specific primers (GSP1 to 3) were designed in the know sequence of the second exon of DLX6-AS1 and are indicated in the appendix (**Table 7-4**). The region between GSP3 and the 5’ end of DLX6-AS1 second exon was selected by Blast in order to be able to discriminate the sequence of this lncRNA from other transcripts (Altschul et al., 1990). In addition, three primers QT, Q0 and Q1 recognizing the poly-A tail added to the 3’ end of the cDNA as described below, serve as binding site upstream of the unknown 5’ sequence of DLX6-AS1 and are indicated in the appendix (**Table 7-4**).

After denaturation for 3 min at 80°C of 5 µg of DNA-free RNA, 10 pmol of GSP1 was annealed for 5 minutes at 65°C with 1 µg of denatured RNA and 1.7 mM dNTPs mix in a final volume of 12 µl. cDNA specific for DLX6-AS1 5'end was then reverse transcribed using Thermoscript™ Reverse Transcriptase supplemented with 40 units of RNasin. To obtain complete DLX6-AS1 5' cDNA ends the reverse transcription mix was incubated for 1 hour at 65°C, followed by 10 min at 70°C. After heat-inactivation of the reverse transcriptase for 5 min at 85°C, the template RNA engaged in cDNA-RNA hybrids was digested by adding 2 units of RNase H and incubated for 20 min at 37°C with subsequent inactivation of the enzyme for 20 min at 65°C. In order to avoid the presence of excess of GSP1 primers and non-incorporated dNTPs, cDNA was purified using 40 µl Agencourt RNA clean XP magnetic beads following manufacturer instructions. To tail the 3'end of the newly synthesized cDNA a poly-A adapter was added via the addition of dATPs (130 pmol) with 30 units of the terminal deoxynucleotidyl transferase to 1 pmol of ssDNA and incubated at 37°C for 15 min, followed by heat-inactivation at 70°C for 10 min and dilution by addition of 20 µl H_2O. The poly(A) tail provides a binding site upstream of the unknown 5' sequence of the target mRNA for the QT primer during the first round of amplification. The first round of PCR amplification was carried out with the Expand High fidelity PCR system supplemented with 1.5 µl of the diluted 5'end tailed cDNA, 25 pmol of GSP2 and Q0 primers and 2 pmol of QT primer under the following cycling conditions: 2 min denaturation at 95°C, 2 min annealing of QT primer at 48°C followed by a linear amplification at 68°C for 40 min. This step was directly followed by an exponential amplification carried out first for 10 cycles under following conditions: 15 sec denaturation at 95°C, 30 sec annealing of Q0 primer at 60°C (-0.5°C per cycle) and extension for 2 min at 68°C followed by 30 cycles at an annealing temperature of 55°C pursued with a 10 min final extension at 72°C.

A second round of PCR was used to increase the specificity of the amplification for our target through the use of GSP3 and Q1 primers. Expand High fidelity PCR system supplemented with 1.5 µl of 1:20 diluted PCR product, 10 pmol of GSP3 and Q1 primers were used under identical cycling conditions as for the previous exponential amplification. The PCR was loaded on a 1.5% agarose gel before gel purification of the appropriate PCR product, cloned into the pCR2.1 TOPO-TA

vector and identified by Sanger sequencing (GATC Biotech) with a primer binding to the M13 promoter (Appendix **Table 7**-**7**"M13") included in this vector.

3.6 Quantitative DNA methylation analysis

Table 3-13 Material for DNA methylation analysis

Material	Manufacturer
QIAmp DNA Mini kit or All Prep DNA/RNA Mini kit	Qiagen
EZ DNA methylation kit	Zymo Research
HotStarTaq DNA Polymerase kit	Qiagen
dNTPs (mix of 10 mM each)	Fermentas
RNaseA	Sequenom
Shrimp alkaline phosphatase (Goldstein et al.)	Sequenom
"T" cleavage MassCLEAVE reagent kit	Sequenom
RNaseA	Sequenom
MassARRAY Nanodispenser	Sequenom
Sequenom MassARRAY MALDI-TOF Mass Spectrometer	Sequenom
Sequenom EpiTYPER software 1.2	Sequenom
Mastercycler 384	Sequenom

DNA was extracted using the All Prep RNA/DNA extraction kit following manual instructions. To distinguish methylated from unmethylated cytosine by converting unmethylated cytosine into uracil, 1 µg of DNA was bisulfite-converted using the EZ DNA methylation kit and eluted twice in 30 µl M-Elution buffer.

DLX6 promoter regions were PCR amplified by adding to 1µl of bisulfite converted DNA, 0.056 µl HotStar Taq DNA polymerase, 0.14 µl dNTPs (10 mM) and 0.14 µl of 10 µM primers specific for bisulfite converted DNA to a final volume of 7 µl. Reactions were hot-started at 95°C for 5 min and followed for 45 cycles of 94°C for 30 sec, 56°C for 30 sec, 72°C for 1 min with a final 5 min extension step at 72°C. PCR amplification specificity and efficiency was verified by gel electrophoresis with 1.5 µl of the PCR product. Unincorporated dNTPs were dephosphorylated from the PCR reaction by adding 0.3 µl SAP diluted with 1.7 µl H_2O and incubated for 20 min at 37°C, followed by enzyme inactivation at 85°C for 5 min. 2 µl of dephosphorylated PCR product was simultaneously *in vitro* transcribed and base-specifically cleaved by RNAseA for 3h at 37°C

by adding to a final volume of 5 µl 0.22 µl "T" cleavage MassCLEAVE reagent, 0.22 µl DTT, 0.4 µl T7 RNA polymerase and 0.06 µl RNaseA. After *in vitro* transcription, uracil residues are converted into adenine, whereas cytosines are transformed into guanine. Cleavage reactions were desalted for 30 min on a rotator after the addition of 20 µl water and 6 mg resin. The unique RNA fragments created by digestion were further analyzed by matrix-assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF). DNA methylation levels were quantified based on the mass spectra peak shift of 16 Da between adenine (former unmethylated) and guanine (former methylated) using the Epityper software.

Bisulfite specific primers were designed manually and appended with a 10 nt mass tag on the 5'end of the forward primer (5'- aggaagagag-3') and with a T7 RNA polymerase promoter on the 5' end of the reverse primer ('5-cagtaatacgactcactatagggagaaggct-3'), and are indicated in the appendix (**Table 7-5**).

3.7 RNA synthesis and transfection

3.7.1 PCR amplification and cloning of DLX6-AS1 cDNA

Table 3-14 Material for PCR amplification and cloning of DLX6-AS1 cDNA

Material	Manufacturer
ThermoScript™ Reverse Transcriptase	Invitrogen
RNasin ribonuclease inhibitors (40 u/µl)	Promega Corporation
dNTPs (mix of 10 mM each)	Fermentas
oligodT	Thermo Fisher Scientific
Rnase H (5 u/µl)	Thermo Fisher Scientific
Expand Long Template PCR system	Roche Diagnostics
MluI (10 u/µl)	New England Biolabs Inc.
HindIII-HF (20u/µl)	New England Biolabs Inc.
QIAquick PCR purification kit	Qiagen
T4 Polynucleotide Kinase (10 u/µl)	New England Biolabs Inc.
T4 DNA ligase (40 u/µl)	New England Biolabs Inc.
One Shot TOP10 chemically competent E. coli	Thermo Fisher Scientific
MEGAscript® T7 Transcription kit	Ambion, Life technologies

High quality cDNA was generated from 4 µg VCaP RNA with 1 µl oligodT (50 µM), 2 µl dNTPs (10 mM) in a final volume of 12 µl and annealed for 5 minutes at 65°C. Reverse transcription was

performed after addition of 1 µl ThermoScript™ reverse transcriptase supplemented with 40 units RNasin to a final volume of 20 µl. This reaction was incubated for 1 h at 65°C. After heat-inactivation of the reverse transcription reaction for 5 minutes at 85°C, template RNA engaged in cDNA-RNA hybrids were digested for 20 minutes at 37°C by adding 2 units of RNaseH with subsequent inactivation of the enzyme for 20 minutes at 65°C.

DLX6-AS1 was PCR amplified from VCaP cells cDNA using a forward primer harboring the T7 RNA polymerase promoter and MluI restriction site, while the reverse primer harbored a HindIII restriction site (Appendix **Table 7-6**). DLX6-AS1 was PCR amplified from high quality cDNA using 0.75 µl Expand Long Template enzyme mix along with 5 µl PCR buffer 1.350 µM dNTPs, 300 µM forward and reverse primers and 100 ng cDNA in a final volume of 50 µl. The reaction was incubated under following conditions: initial denaturation at 94°C for 2 min, 10 cycles at 94°C for 10 sec, annealing for 30 sec at 65°C (-0.1°C per cycle), extension at 68°C for 4 min, followed by 30 cycles under same conditions but at a stable annealing temperature of 60°C and increased extension time of 20 sec per cycle with a final extension step at 68°C for 15 min.

After gel extraction and purification of the appropriate PCR products, they were overnight digested at 37°C with 20 units MluI and HindIII-HF enzymes in 50 µl. Subsequent to heat-inactivation of the enzymes at 80°C for 20 min, digested PCR products were purified with the QIAquick PCR purification kit and eluted in 30 µl water. Similarly, 3 µg of pcDNA3.1(-) was overnight digested at 37°C with 20 units of MluI and HindIII-HF enzymes in a final volume of 50 µl, before enzyme inactivation at 80°C for 20 min. After gel extraction of the digested plasmid, 25 ng digested plasmid was ligated with 25 ng digested cDNA with 0.5 µl T4 DNA ligase in 25 µl and incubated overnight at 16°C. One Shot TOP10 *E.Coli* (15 µl) were then transformed by chemical transformation with 2 µl of the ligation product. Positive clones harbouring the pcDNA3.1(-)ΔMluI-HindIII_DLX6-AS1 plasmid were identified by Sanger sequencing (GATC) employing a primer binding to the T7 promoter (**Table 7-7** "T7").

3.7.2 *In vitro* transcription

Table 3-15 Material for *in vitro* transcription

Material	Manufacturer
HindIII-HF (20 u/µl)	New England Biolabs Inc.
AgeI-HF (20 u/µl)	New England Biolabs Inc.
MEGAscript® T7 Transcription kit	Ambion, Life technologies
Formamide (≥99.5% purity)	Sigma-Aldrich
Formaldehyde solution 37%	Sigma-Aldrich
10x MOPS buffer	Maxim Biotech, Inc
Low range ssRNA ladder	New England Biolabs Inc.
Glycogen from oyster	Sigma-Aldrich
Qubit RNA HS Assay kit	Thermo Fisher Scientific

5 µg of pcDNA3.1(-)ΔMluI-HindIII_DLX6-AS1 or pT-Rex-DEST30-LACZ plasmid was prepared for *in vitro* transcription by linearization of the plasmids in a final volume of 50 µl with 20 units of HindIII-HF or AgeI-HF enzyme, respectively. 1 µg of linearized and phenol/chloroform purified plasmid was *in vitro* transcribed using the MEGAscript kit with T7 RNA polymerase incubated for 4 h at 37°C. After the DNA template got digested with 1 µl DNase for 15 min at 37°C, 4 µl of RNA supplemented with 6 µl RNA loading buffer (250 µl 100% formamide, 83 µl 37% formaldehyde, 50 µl 10x MOPS, 50 µl 100% glycerol, 5 µl 1mg/ml ethidium bromide, 20 µl 1.25% bromophenol blue and 45 µl H_2O) was loaded on a denaturing agarose/formaldehyde gel (1.8 g agarose, 142 ml 1x MOPS buffer and 8.1 ml 37% formaldehyde) along with 4 µl of low range ssRNA ladder. RNA was extracted from agarose gel through centrifugation of the agarose piece through Whatman filter paper. To the flow-through were added 0.5 volumes LiCl (7.5 M), 2.5 volumes 100% ethanol and 20 µg glycogen for overnight RNA precipation at -20°C. The next day, RNA was collected by full speed centrifugation for 30 min at 4°C, followed by a wash with 1 ml ice-cold 70% ethanol and air-dried at room temperature. RNA pellet was then resuspended in 20 µl water and quantified by Qubit using the Qubit RNA HS Assay kit.

3.7.3 RNA transfection

Table 3-16 Material for RNA transfection

Material	Manufacturer
TransIT-mRNA transfection kit	Mirus Bio LLC
RPMI-1640 medium	Gibco, Invitrogen, Life Technologies
EGFP mRNA	Trilink Biotechnologies

A549 and PC3 cells were seeded 24h prior to transfection in 6 well plates at a concentration of $1.5x10^5$ and $2x10^5$ cells per well, respectively, in a final volume of 1.5 ml. 400 fmol of *T1*, *T2S*, *T2L* *DLX6-AS1* RNA or *LACZ* control RNA supplemented with *EGFP* mRNA to a final concentration of 1 µg were transfected with 5 µl TransIT-mRNA and 9.9 µl mRNA Boost reagent according to manufacturer's instructions. 48 h after transfection, cells were harvested and stored at -80°C until nucleic acids extraction.

3.8 *In vitro* translation

Table 3-17 Material used for the *in vitro* translation

Material	Manufacturer
Phusion® High-Fidelity PCR kit	New England Biolabs
dNTPs (mix of 10 mM each)	Fermentas
QIAquick PCR purification kit	Qiagen
RNasin ribonuclease inhibitors (40 u/µl)	Promega Corporation
TnT®T7 coupled reticulocyte lysate systems	Promega
16.5% Mini-Protean® Tris-Tricine Gel, 10 well, 50 µl	Biorad
Amino acid mixture, minus leucine 1mM	Promega
L-[^{14}C(U)]-Leucine, Specific Activity: >300 mCi (11.1GBq)/mMol, >97%, 250 µCi (9.25 MBq)	Perkin Elmer
Spectra Multicolor Low range protein ladder	Thermo Fisher Scientific
Tricine ≥99%	Sigma-Aldrich
BioMax® MS film	Kodak
BioMax TranScreen LE intensifying screen	Kodak

pcDNA3.1(-)ΔMluI-HindIII_DLX6-AS1 plasmid with *DLX6-AS1 T1*, *T2* or *T3* transcript variants generated by cloning as described in section 3.7.1 were used as template for the amplification of the different transcript variants with a forward primer harboring the T7 RNA polymerase

promoter. Similarly, CS2_T7_MLN and CS2_T7_MLN (FS) plasmid provided by Dr. Douglas M. Anderson served as template to amplify the mouse lncRNA *2310015B20RiK* harboring the peptide MLN with (FS) or without a frameshift mutation (Anderson et al., 2015). All primers used for amplification can be found in the appendix (**Table 7-6**). 50 pg of plasmid was used to PCR amplify the different transcripts with 0.5 µl Phusion® High-Fidelity DNA polymerase, 0.5 µM of each primer, 10 µl Phusion HF buffer and 1 µl dNTPs in 50 µl. The reaction was initially denaturated at 98°C for 2 min, then amplified for 40 cycles of 10 sec at 98°C, 30 sec at 60°C and 2 min at 72°C with a final extension step at 72°C for 10 min. PCR reactions were purified with the QIAquick PCR purification kit, eluted with 20 µl water supplemented with 40 units RNasin and incubated for 30 min at 37°C.

Coupled *in vitro* transcription/translation assay of 850 fmol purified PCR product was performed using 12.5 µl TnT rabbit reticulocyte lysate, 2.5 µl ^{14}C labelled leucine amino acid, 0.5 µl amino acid mixture, 1 µl TnT reaction buffer, 0.5 µl RNasin, 0.7 µl TnT T7 RNA polymerase in a final volume of 25 µl and incubated for 90 min at 37°C. One tenth of the coupled *in vitro* transcription/translation reaction was denatured for 10 min at 70°C, after addition of 2.5µl SDS loading buffer (200 mM Tris-HCl, pH 6.8, 2% SDS, 40% glycerol, 0.04% Coomassie Brilliant Blue G-250 and 2% β-mercaptoethanol). Proteins and peptides were separated on a 16.5% Tris-Tricine gel in 1x Tris-Tricine buffer (10 mM Tris, 10 mM Tricine, 1% SDS at pH 8.3) at a constant voltage of 100V along with 10 µl of low range protein ladder. After electrophoresis, the gel was immediately immersed in fixative solution (10% glacial acetic acid, 50% methanol and 40% H_2O) for 30 min, to prevent the diffusion of peptides. The gel was then soaked for 5 min in a solution (7% acetic acid, 7% methanol, 1% glycerol) preventing gel cracking, which can occur during the drying process under vaccum performed on a gel dryer for 40 min at 70°C. The gel was then exposed for two weeks to a Biomax® MS film with a BioMax® TranScreen LE intensifying screen.

3.9 Polysome fractionation

Table 3-18 Material used for polysome fractionation

Material	Manufacturer
Dimethylsulfoxid ≥99.5% purity bioscience-grade (DMSO)	Carl-Roth
Harringtonine from *Cephalotaxus harringtonia* ≥90% (HPLC)	Sigma-Aldrich
Cycloheximide from microbial, ≥94% (TLC)	Sigma-Aldrich
Phenol non stabilized:chloroform:isoamyl alcohol 25:24:1	AppliChem GmbH
Complete™,Mini Protease Inhibitor Cocktail	Sigma-Aldrich
Glycogen from oyster	Sigma-Aldrich
TURBO DNA-free™ kit	Ambion, Life Technologies
SuperScript® III Reverse Transcriptase	Thermo Fisher Scientific
Random primers	Thermo Fisher Scientific

H1299 cells were treated by Dr. Michael Daskalakis (DKFZ, Heidelberg) for 48 h with dimethyl sulfoxide (DMSO). To distinguish coding versus non-coding transcripts by inhibition of translation initiation, we treated one part of the cells with harringtonine for 15 min at a final concentration of 10 µg/ml, whereas another part remained untreated. Both treated and untreated cell fractions were washed in ice-cold D-PBS containing 100 µg/ml cycloheximide to arrest translation elongation. After washing the cells, polysomal fractionation was performed by Dr. Johanna Schott (ZMBH, University of Heidelberg). In brief, cells were lysed in 200 µl polysome lysis buffer (15 mM Tris-HCl pH 7.4, 15 mM $MgCl_2$, 300 mM NaCl, 100 µg/ml cycloheximide , 1% Triton-X-100, 0.1% β-mercaptoethanol, 200 u/ml RNAsin, 1 complete Mini Protease Inhibitor Tablet per 10 ml). Nuclei were removed by centrifugation at 9,300 xg for 10 min at 4°C, while the cytoplasmic lysate was subjected to fractionation on a sucrose density gradient (17.5–50% in 15 mM Tris-HCl pH 7.4, 15 mM $MgCl_2$, 300 mM NaCl). After ultracentrifugation (2.5 h, 35 000 rpm at 4°C in a SW60Ti rotor), gradients were eluted with a Teledyne Isco Foxy Jr. system into 14 fractions of similar volume and directly supplemented with 300 µl solution II (10 mM Tris pH 7.5, 350 mM NaCl, 10 mM EDTA, 1% SDS, 7 M urea), and 300 µl Phenol/Chloroform/Isoamyl alcohol, before storage at -20°C. 25 fmol of a rabbit *HBB2* (hemoglobin subunit beta 2) *in vitro* transcript was added to each fraction as a spike-in control.

Before RNA extraction, samples were thawed at 65°C for 10 min and centrifuged for 20 min at full speed. The upper aqueous phase was then collected and supplemented with 20 µg glycogen and 600 µl isopropanol before overnight RNA precipitation at -20°C. On the next day, RNA was collected by full speed centrifugation for 20 min at 4°C, followed by a wash with 1 ml ice-cold 70% ethanol. RNA pellets were dried at room temperature and resuspended at 65°C for 15 min in 20 µl water. Residual genomic DNA was digested with the TURBO™ DNase kit. 4 µl of DNase-treated RNA was reverse transcribed using the Superscript™ III reverse transcriptase, with 200 ng random primers, and 200 nM gene specific primers following manufacturer's instructions. qPCR was performed with the same protocol as described previously in section 3.4.2.

3.10 Software and statistical analysis

Figures and statistical analysis were generated either with GraphPad Prism 5 software or Microsoft Excel 2010. Heatmaps were created with the Multi Experiment Viewer. Multiple sequence alignments were generated with Jalview version 2-a (Waterhouse et al., 2009) by using JABAWS 2.1 web services (Troshin et al., 2011).

To test the significant difference of patient-matched tumor/normal samples data, the non-parametric Wilcoxon signed-rank test was used. The non-parametric Mann-Whitney U test was applied to compare the statistical significance of data between unpaired normal/tumor samples. All experimental data were compared with unpaired two-tailed Student's t-test. The following levels of p-value apply in all figures: ns= non-significant, * = $p<0.05$; ** = $p<0.01$; *** = $p<0.001$ and **** = $p<0.0001$.

4. RESULTS

4.1 Genome-wide screens identifies DLX6/DLX6-AS1 pair

The aim of the first part of the thesis was to identify lncRNAs involved in the regulation of DNA methylation patterns, which in turn influence the expression of mRNAs *in cis* (simplified as mRNA/lncRNA pairs). To this end, we integrated RNA-seq and 450k data referred to as the "screening set" in **Table 3.1** through the strategy explained below and depicted in **Figure 4-1**.

The first step was to retrieve from GENCODE release v17 the positions of lncRNAs overlapping the genomic region of a coding gene. For the analysis we included all genes annotated either as lincRNA or antisense RNA based on GENCODE descriptions (n= 10,609 lncRNA genes). The overlapping region of a given coding gene was defined as its genomic region expanded by 1 kb upstream and downstream of the transcriptional start site (TSS) and transcriptional termination site (TTS) of the gene, respectively. An overlap between an mRNA and an lncRNA was considered valid when the lncRNA gene overlapped the mRNA region by at least 1 nucleotide. Through this approach, we identified 413 mRNA/lncRNA pairs of transcripts (**Figure 4-1 A**).

To select mRNA/lncRNA pairs associated with prostate carcinogenesis, we used high-throughput RNA sequencing data from EO-PCa ICGC patients (screening set, RNA-seq -**Table 3-1**) to select pairs with significant differential expression (p-value ≤ 0.05 and -0.5≤ Log2 fold change ≥0.5) in tumor versus normal prostate samples (**Figure 4-1 B**). RNA sequencing data from the ICGC consortium were generated in two different centers, namely Berlin and Heidelberg (EO-PCA Berlin and Heidelberg-**Table 3-1**). To avoid possible batch effects, differentially expressed genes were identified independently in the two datasets. Only mRNA/lncRNA pairs significantly differentially expressed in both set of samples were selected.

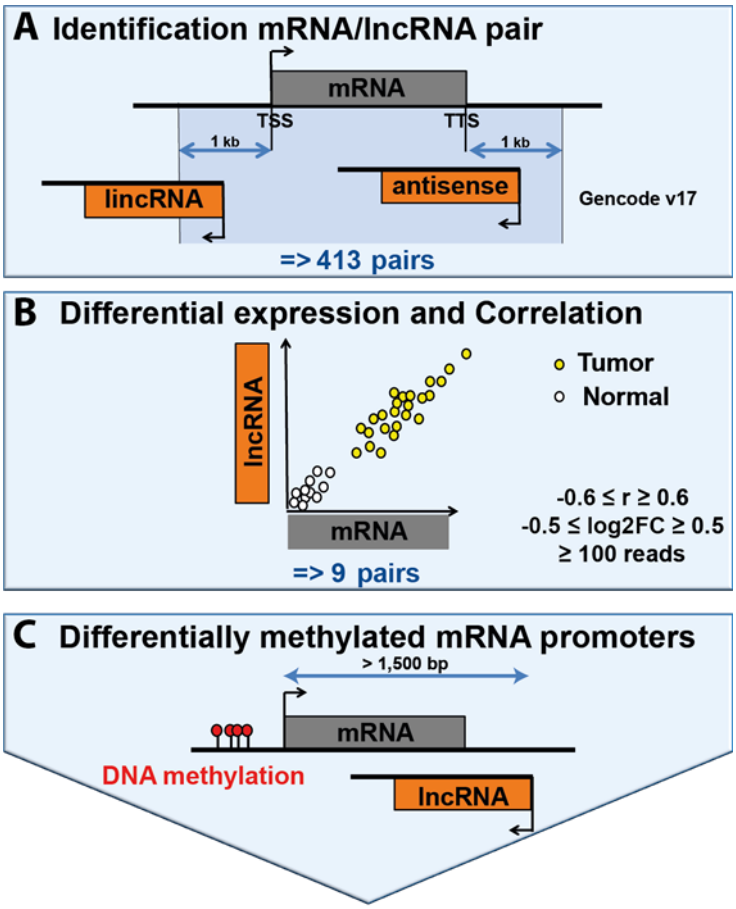


Figure 4-1 Integrative analysis strategy for the identification of candidate mRNA/lncRNA pairs

Schematic overview of the methodological strategy for the identification of mRNA/lncRNA pairs in prostate cancer. Abbreviations: TSS= transcriptional start sites, TTS= transcriptional termination site, log2FC= log2 transformed expression fold changes.

The selection was further restricted to mRNA/lncRNA pairs expressed with a significant Spearman correlation coefficient ranging between -0.6 ≤ r ≥ 0.6, indicating that the lncRNA might regulate the expression of the protein-coding gene in *cis*.

Additionally, to exclude lncRNAs not sufficiently expressed and thus not detectable in an experimental setting, only pairs reaching a minimum differential expression threshold of 100 reads in the tumor samples were considered as candidates. Overall, with the aforementioned criteria we found nine mRNA/lncRNA pairs of transcripts (**Table 4-1**).

Table 4-1 Differentially expressed candidate mRNA/lncRNA pairs

mRNA							lncRNA							
Symbol	ID	Coordinates	HD FC	HD p-value	Berlin FC	Berlin p-value	Symbol	ID	Coordinates	HD FC	HD p-value	Berlin FC	Berlin p-value	r
Upregulated mRNA/lncRNA pairs														
DLX6	ENST00000518156.2	chr7:96634860-96640351	3.11	7.26E-04	2.12	5.36E-03	*DLX6-AS1*	ENST00000430027.2	chr7:96594839-96643377	3.29	2.02E-04	2.68	5.98E-03	0.94
FBXL19	ENST00000562798.1	chr16:30934376-30937252	1.22	3.74E-02	1.68	1.05E-02	*FBXL19-AS1*	ENST00000563777.1	chr16:30930640-30934590	1.8	3.37E-02	2.89	2.62E-05	0.74
MCF2L	ENST00000375608.3	chr13:113548692-113751175	1.2	3.06E-02	1.6	1.25E-02	*MCF2L-AS1*	ENST00000446789.2	chr13:113621798-113623138	1.29	3.94E-02	1.9	4.63E-03	0.69
PCSK6	ENST00000558864.1	chr15:101840818-101853607	2.6	1.38E-05	2.72	1.19E-04	*RP11-299G20.2*	ENST00000558838.1	chr15:101835624-101845942	3.16	1.23E-03	2.09	2.25E-03	0.65
THBS4	ENST00000513310.1	chr5:79287134-79351729	2.56	7.87E-04	3.62	7.92E-03	*CTD-2201I18.1*	ENST00000503007.1	chr5:79348197-79378227	1.92	1.92E-03	2.32	1.66E-02	0.69
Downregulated mRNA/lncRNA pairs														
JAZF1	ENST00000283928.5	chr7:27870192-28220362	-2.06	1.49E-06	-2.94	4.27E-08	*JAZF1-AS1*	ENST00000455963.1	chr7:28219941-28280383	-2.16	3.45E-07	-2.59	3.60E-06	0.96
MKX	ENST00000375790.4	chr10:27961803-28034989	-1.61	2.74E-04	-2.31	1.46E-05	*RP11-360I20.2*	ENST00000419777.1	chr10:28033715-28056723	-2.98	2.26E-13	-3.02	4.71E-08	0.89
RBPMS	ENST00000538486.1	chr8:30241944-30429734	-1.37	1.79E-03	-1.34	1.24E-02	*CTD-3107M8.4*	ENST00000517521.1	chr8:30239635-30242917	-1.05	2.05E-02	-1.5	9.66E-03	0.73
UBXN10	ENST00000375099.3	chr1:20512578-20522541	-1.48	6.67E-04	-1.55	3.55E-03	*RP3-340N1.5*	ENST00000442226.1	chr1:20510735-20512979	-2.73	4.93E-11	-3.23	4.66E-09	0.87

Result of the screening strategy after integration of RNA-seq data. In each row are indicated for one mRNA/lncRNA pair the gene symbol and Ensembl identification number of the mRNA and lncRNA along with their respective chromosomal coordinates (hg19). Also indicated are the log2 transformed expression fold changes (FC) in tumor versus normal tissue in Heidelberg (Goldberg et al.) and Berlin datasets along with their corresponding p-values. The combined Spearman correlation (r) calculated for Heidelberg and Berlin datasets between an mRNA and the corresponding lncRNA is also indicated.

Among the five overexpressed lncRNAs, ***DLX6-AS1*** (also named *Evf2* in mice or DLX6 antisense RNA 1) and ***FBXL19-AS1*** (*FBXL19* antisense) have been previously linked to the regulation of DNA methylation patterns. *DLX6-AS1* has been reported to alter the methylation level of an ultraconserved enhancer region leading to ***DLX6*** (distal-less homeobox 6) and *DLX5* (distal-less homeobox 5) transcriptional activation in mice during brain development (Berghoff et al., 2013). In addition, LongTarget algorithm predicted *DLX6-AS1* binding in *DLX6* promoter region through the formation of a triplex structure by Hoogsten base-pairing, suggesting a regulatory role for the lncRNA on *DLX6* expression (He et al., 2015). In concordance with this finding, *DLX6-AS1* silencing by siRNA (small interfering RNA) in A549 and H1650 lung adenocarcinoma cell lines was shown to decrease DLX6 RNA and protein expression (Li et al., 2015).

RNA immunoprecipitation followed by deep sequencing (RIP-seq) analysis of DNMT1-associated transcripts, identified RNA partners matching the ***FBXL19*** (F-Box and leucine-rich repeat protein 19) /***FBXL19-AS1*** locus (Di Ruscio et al., 2013). From this report it remains unclear if the coding or non-coding RNA was bound to DNMT1, but DNMT1 binding was associated with *FBXL19* transcription and hypomethylation of *FBXL19* promoter in HL-60 cells. The authors propose that the RNA-DNMT1 complex protects *FBXL19* promoter from DNA methylation by a decoy mechanism. Thus, *FBXL19-AS1* upregulation could potentially lead to hypomethylation of *FBXL19* promoter.

All the identified downregulated mRNAs were linked to polycomb-mediated repression. ***JAZF1*** (JAZF Zinc Finger 1), ***MKX*** (Mohawk Homeobox)(Nuytten et al., 2008) and ***UBXN10*** (UBX Domain Protein 10) (Kondo et al., 2008) were shown to be upregulated after knockdown of PRC2 components, whereas ***RBPMS*** (RNA Binding Protein With Multiple Splicing promoter) was marked by increased H3K27me3 levels in hepatocellular carcinoma compared to normal liver tissue (Acevedo et al., 2008). In line with a regulation of the aforementioned coding genes by the PRC2 complex, the transcripts located on the reverse strand of *JAZF1*, *RBPMS* and *UBXN10* mouse homologs were found by RIP-seq to be associated with the PRC2 complex in mouse embryonic stem cells (Zhao et al., 2010). This raises the possibility that ***JAZF1-AS1***, ***CTD-3107M8.4*** and ***RP3-***

34ON1.5 mouse lncRNA homologs binding to PRC2 prevents in a decoy mechanism the epigenetic repression of the corresponding coding gene in *cis*.

In line with our result in EO-PCa patients (**Table 4-1**), the ***MCF2L*** (MCF.2 cell line derived transforming sequence-like)/***MCF2L-AS1*** (MCF2 antisense RNA 1) pair was reported to be recurrently upregulated among a set of 4055 different tumor types from the TCGA (Kaczkowski et al., 2016). The regulation of *MCF2L* expression is however not documented. In contrast, ***PCSK6*** (also named PACE4, proprotein convertase subtilisin/kexin type 6) was demonstrated to be upregulated in response to several stimuli, such as miR-124 downregulation in PCa (Kang et al., 2014), DNA demethylation by 5-azacytidine in ovarian cancer (Fu et al., 2003), and by upregulation of the transcription factor E2F (Yuasa et al., 2007). Last, ***THBS4*** (Thrombospondin-4) promoter was shown to be marked by higher methylation levels in colorectal cancer, but could not be associated with altered *THBS4* expression levels (Greco et al., 2010).

Based on these reports, six of the lncRNA candidates are potential epigenetic regulators (*DLX6-AS1*, *FBXL-19AS1*, *JAZF1-AS1*, *CTD-3107M8.4*, *RP11-360I20.2* and *RP3-340N1.5*), whereas *MCF2L-AS1*, *RP11-299G20.2* and *CTD-3107M8.4* have not been previously studied. In order to identify lncRNAs regulating mRNA by influencing DNA methylation *in cis*, we further restricted our selection of pairs to protein-coding genes presumably regulated by a differentially methylated promoter region.

A combined set of HumanMethylation450k Beadchip data from ICGC and the TCGA consortium (screening set, 450k, EO-PCa + TCGA - **Table 3-1**) was used to identify differentially methylated regions (DMRs) in the promoter regions of the aberrantly expressed protein-coding genes (**Figure 4-1 C**). Promoter regions covered the genomic region located 2kb upstream and 2kb downstream of the gene TSS. These regions were separated in 300 bp non-overlapping windows and defined as a DMR when significantly (FDR<0.01, p-value<0.05) enriched for differentially methylated CpGs (methylation difference ≥5%) in tumor samples. After addition of these criteria the

following mRNA/lncRNA pairs remained; *DLX6/DLX6-AS1*, *MKX/RP11-360I20.2* and *UBXN10/RP3-340N1.5* (**Table 4-2**).

Table 4-2 Differentially expressed and methylated candidate mRNA/lncRNA pairs

mRNA	lncRNA	Distance between TSS
DLX6	***DLX6-AS1***	8517 bp
MKX	***RP11-360I20.2***	1274 bp
UBXN10	***RP3-340N1.5***	401 bp

Result of the screening strategy after integration of DNA methylation data. In each row are indicated for one mRNA/lncRNA pair the distance between the transcriptional start sites (Consortium et al.) of each respective pair of transcripts.

About 10% of human protein-coding genes share their promoter with another gene transcribed on the opposite DNA strand and are defined as bidirectional promoters (Trinklein et al.,2004). To exclude mRNA/lncRNA pairs concomitantly expressed in response to a common bidirectional promoter, only pairs were considered as candidates when the TSS of both RNA components were separated by more than 1,500 bp (**Figure 4-1 C**). Among the three identified mRNA/lncRNA pairs, *MKX* and *UBXN10* coding genes are separated from the TSS of the adjacent lncRNA gene by 1274 bp and 401 bp, respectively. In contrast, *DLX6* and *DLX6-AS1* TSSs are separated by 8517 bp and seemed therefore not to be regulated by a common promoter (**Table 4-2**).

In summary, this screening approach allowed the identification of a single mRNA/lncRNA pair, *DLX6/DLX6-AS1*, overexpressed and positively correlated in EO-PCA patients. Based on the existing evidence *DLX6-AS1* was selected as promising lncRNA candidate for the regulation of the expression of *DLX6* mRNA *in cis* through alteration of the DNA methylation pattern of *DLX6* promoter region.

4.2 Clinical Characterization of *DLX6/DLX6-AS1* pair

In order to corroborate the *in silico* identification of *DLX6/DLX6-AS1* pair, we investigated in more details the RNA expression and DNA methylation levels and their interdependency in two independent cohort of PCa patients (**section 4.2.1**) and in other tumor entities (**section 4.2.2**) before this pair *in vitro* (**section 4.3**).

4.2.1 Clinical characterization of *DLX6/DLX6-AS1* expression in PCa patients

In the validation cohort of EO-PCa patients (validation set, RNA-seq, EO-PCa - **Table 3-1**), *DLX6-AS1* along with *DLX6* are highly and significantly overexpressed compared to normal prostate (p-value<0.0001) (**Figure 4-3A**). In line with the hypothesis that the pair of transcripts is coordinately expressed, expression levels of *DLX6* and *DLX6-AS1* are associated by a positive, linear and significant Spearman correlation (r=0.8256, p-value<0.0001) (**Figure 4-3B**).

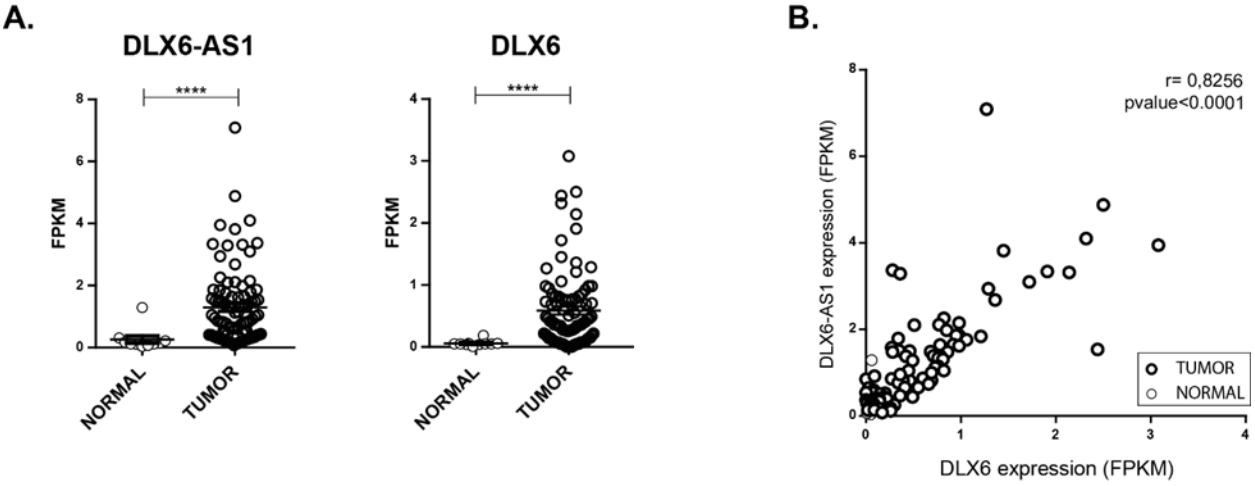


Figure 4-3 Coordinated upregulation of *DLX6* and *DLX6-AS1* in EO-PCA cohort

(A) *DLX6* and *DLX6-AS1* RNA expression levels measured by RNA sequencing in tumors compared to normal tissue (NORMAL=10, TUMOR=91) from ICGC EO-PCa cohort. The horizontal bar represents the mean ± SEM. Mann-Whitney U test ****p<0.0001 **(B)** Corresponding Spearman correlation plot of *DLX6-AS1* and *DLX6* expression measured by RNA sequencing in the ICGC EO-PCa cohort.

An independent cohort of patient-matched tumor/normal samples of late-onset PCa patients from the TCGA was used to validate the result from the ICGC cohort (Validation cohort, RNA-seq, TCGA matched - **Table 3-1**). Similar to the ICGC cohort, *DLX6-AS1* expression is highly and significantly overexpressed (p<0.001), whereas *DLX6* expression is not significantly different compared to their matched normal tissues (**Figure 4-4A**). Indeed, DLX6 expression separates into two groups with expression values from 0 to 5.7 RPKM and thus seemed to be upregulated in a subset of prostate tumors. We therefore divided the TCGA dataset in "LOW" and "HIGH" expression groups according to the median expression value of DLX6-AS1 (median= 4.11). When comparing "TUMOR-HIGH" PCa cases with their matched normal prostate tissues ("NORMAL-HIGH"), *DLX6-AS1* along with *DLX6* were significantly overexpressed (p<0.0001). *DLX6-AS1* expression in the "TUMOR-LOW" subgroup was not significantly different compared to matched normal tissue ("NORMAL-LOW"), whereas *DLX6* expression in the same subgroup was significantly downregulated in comparison to its matched normal counterpart (**Figure 4-4A**). Similarly to the ICGC dataset, the expression of *DLX6-AS1* and *DLX6* was significantly and positively correlated (Spearman r=0.7848, p-value<0.0001) (**Figure 4-4B**).

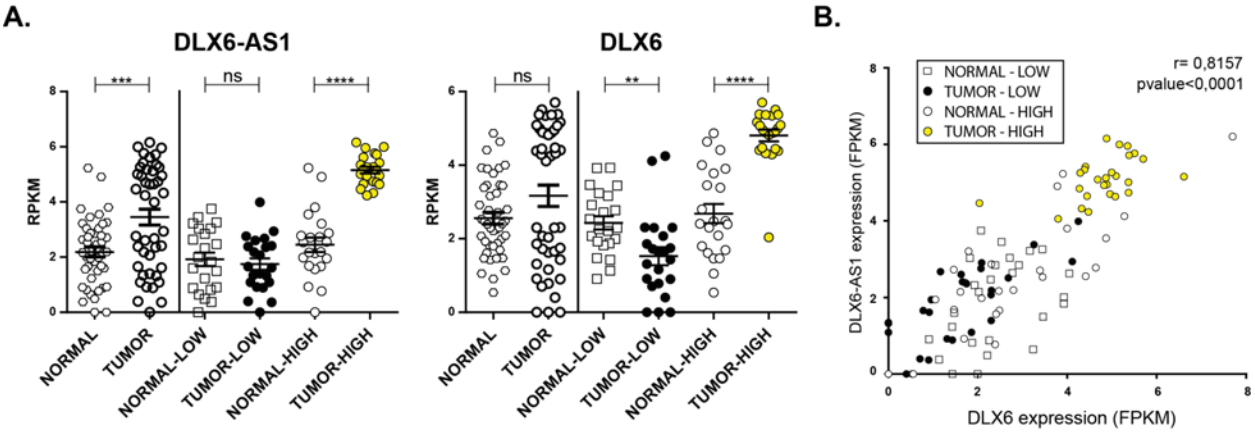


Figure 4-4 Coordinated upregulation of *DLX6* and *DLX6-AS1* in TCGA PCa cohort

(A) *DLX6-AS1* and *DLX6* RNA expression levels measured by RNA sequencing separated into LOW and HIGH tumor subgroups compared to matched normal tissue (NORMAL-LOW=26 , TUMOR-LOW=26, NORMAL-HIGH=26 and TUMOR-HIGH=26) from TCGA late-onset PCa cohort. Patients were divided into LOW and HIGH expression groups according to the median expression value of *DLX6-AS1* in tumor tissue. The horizontal bar represents the mean ± SEM. Wilcoxon signed-rank test ****p<0.0001, ***p<0.001, **p<0.01, ns= non-significant. **(B)** Spearman correlation plot of *DLX6-AS1* and *DLX6* expression measured by RNA sequencing in the TCGA PCa cohort.

Overall, *DLX6/DLX6-AS1* pair is overexpressed in a subset of PCa patients and the expression of both genes is positively associated independently of the age at diagnosis of patients.

To examine whether *DLX6* expression is regulated by DNA methylation in both datasets, corresponding 450k data were used. Direct comparison of the normal versus tumor samples from the ICGC cohort did not reveal a difference at the DNA level (**Figure 4-5B**). Since *DLX6* overexpression was restricted to a subgroup in the TCGA dataset, we also divided the cancerous patients of the ICGC cohort into "LOW" and "HIGH" subgroups according to *DLX6-AS1* median expression value (median= 0.89) (**Appendix Figure 7-1**). In agreement with elevated *DLX6* expression in the "HIGH" PCa samples (**Appendix Figure 7-1**) the *DLX6* promoter region is significantly hypomethylated in the "HIGH" versus "LOW" patients in the ICGC cohort (**Figure 4-5B**). DNA methylation changes are even more pronounced when comparing the "TUMOR-HIGH" and "TUMOR-LOW" subgroup in the TCGA cohort. Indeed, DNA methylation in samples of patients belonging to the "TUMOR-HIGH" subgroup shows a similar profile at the *DLX6* promoter as the matched ("NORMAL-HIGH") and non-matched ("NORMAL-LOW") normal samples. In contrast, in the subset of patients associated with low *DLX6-AS1* expression ("TUMOR-LOW") the *DLX6* promoter is hypermethylated compared to matched "NORMAL-LOW" or to the "TUMOR-HIGH" subgroup of samples (**Figure 4-5C**).

Noteworthy, in both cohort the most pronounced DNA methylation differences are observed in eight CpG probes (cg09327602, cg01229860, cg05663341, cg04599026, cg02898094, cg20479774, cg07598549 and cg20703729) (**Figure 4-5A** probes highlighted with blue), suggesting that the regulatory elements controlling DLX6 expression overlap with these regions.

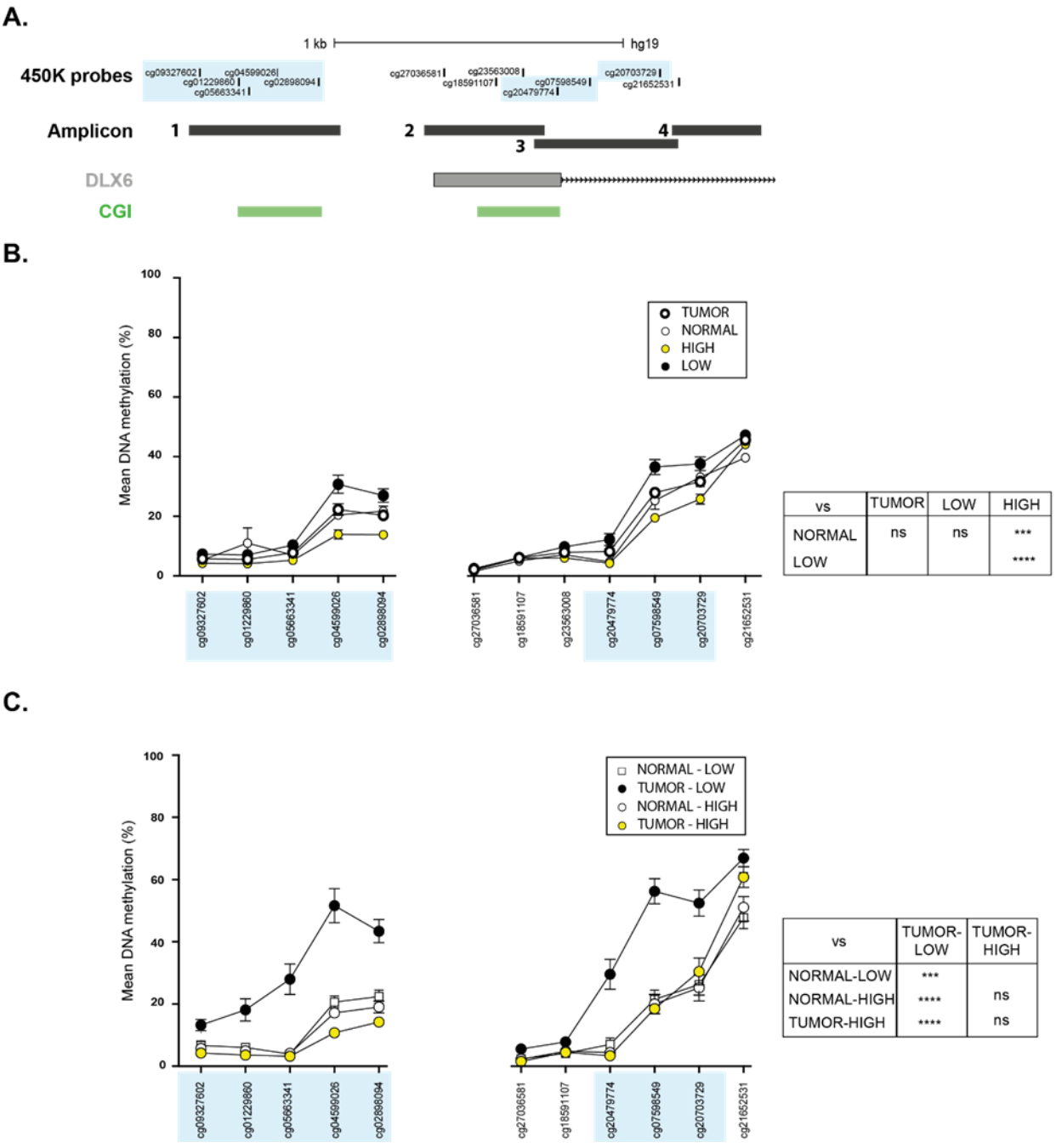


Figure 4-5 Hypomethylation of *DLX6* promoter in "HIGH" subset of PCa patients

(A) Representation of HumanMethylation450 beadchip probes and Mass-Array amplicons overlapping with *DLX6* promoter region and CGIs. (**B** and **C**) DNA methylation levels are displayed on the y-axis as a percentage ranging from 0% to 100%. **(B)** Each point represents mean methylation levels of single CpG sites in normal prostate, tumors as well as tumors separated into the HIGH or LOW subgroups of EO-PCa patients from the ICGC cohort. **(C)** Each point represents mean methylation levels of single CpG sites in matched normal/tumor samples subdivided into HIGH and LOW subgroups of late-onset patients from TCGA. **(B)** Mann-Whitney U test or **(C)** Wilcoxon signed-rank test was used to calculate statistical

significances between each group by comparing the mean DNA methylation levels across the eight CpG sites highlighted in light blue. ****p<0.0001, ***p<0.001, ns= non-significant.

We further tested whether *DLX6* expression levels are inversely proportional to *DLX6* promoter methylation levels. Based on the combined average DNA methylation of the eight CpG probes correlated with *DLX6* expression, we could show that *DLX6* expression and *DLX6* promoter methylation are significantly and negatively correlated in both ICGC (r= -0.7720, p-value<0.0001) (**Figure 4-6A**) and TCGA datasets (r= -0.5337, p-value<0. 0001) (**Figure 4-6B**).

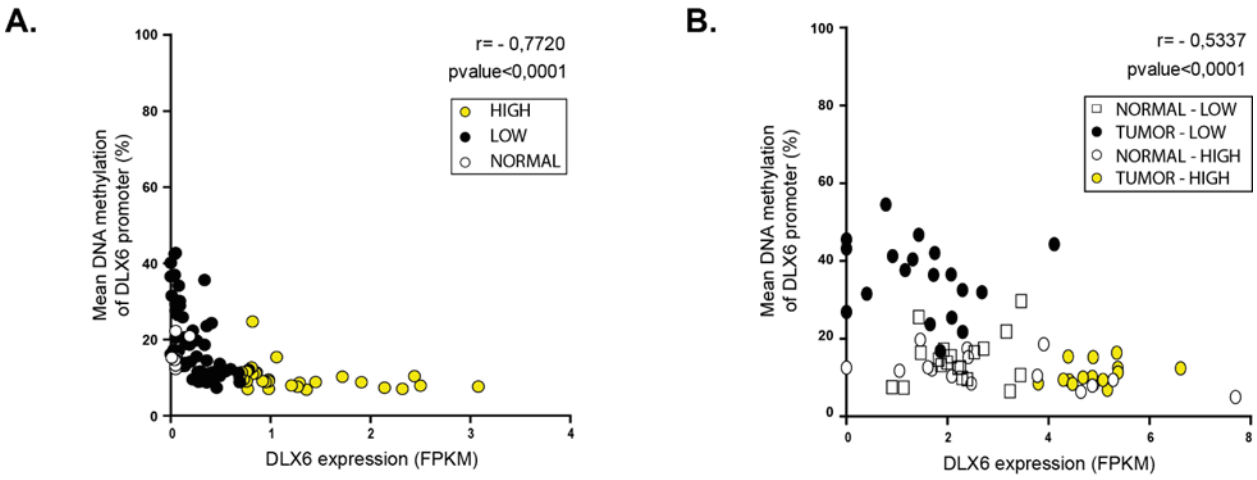


Figure 4-6 *DLX6* expression is inversely correlated with *DLX6* promoter DNA methylation

Spearman correlation plot of *DLX6* expression measured by RNA-sequencing and the average DNA methylation across the eight CpG sites for each sample in **(A)** ICGC or **(B)** TCGA datasets.

Collectively, these results show that elevated *DLX6/DLX6-AS1* expressions along with *DLX6* promoter hypomethylation are linked to a subset of cancer patients. The positive correlation between *DLX6* and *DLX6-AS1* suggests that the expression of these two genes is linked by a common mechanism. Our hypothesis is that *DLX6-AS1* expression regulates *DLX6* mRNA expression through hypomethylation of the *DLX6* promoter region in a mechanism that might involve DNA demethylation.

4.2.2 Clinical characterization of *DLX6/DLX6-AS1* expression across cancers

Comprehensive analysis of lncRNAs in seven tumor types analysed by the TCGA consortium, revealed that 60% (n=4512) of dysregulated lncRNAs are expressed in a cancer-type specific manner. The remaining lncRNAs are expressed in two or more cancers and only 29 were significantly upregulated in at least six tumor entities, among them *DLX6-AS1*. Besides PCa, *DLX6-AS1* was reported to be overexpressed in LUSC, COAD, HNSC, LUAD and KIRC (Yan et al.,2015). Independent differential expression enrichment analysis of lncRNAs in 25 cancer types confirmed the upregulation of *DLX6* and *DLX6-AS1* in LUSC and LUAD (Iyer et al., 2015).

To analyse if the observed relationship between *DLX6* and *DLX6-AS1* in PCa (**section 4.2.1**) can be confirmed in the five above mentioned cancer entities from TCGA, we compared the expression of both transcripts (**Figure 4-7A**) and analysed the involvement of DNA methylation in *DLX6* regulation (**Figure 4-7B**). Since the number of patient-matched tumor/normal samples with both RNA-seq and 450k data was too low, we included for this analysis all available samples (**Table 3-2**). Across all five cancer types, *DLX6-AS1* expression significantly and positively correlated with *DLX6* expression (p-value<0.0001), with the highest correlation observed for LUSC (r=0.8952). KIRC had the lowest Spearman correlation (r=0.6524), and the expression range of 0 to 6.9 RPKM of the pair is also the lowest among the five cancer types (**Figure 4-7A**).

To examine whether *DLX6* expression is regulated by DNA methylation in the five datasets, corresponding 450k data were used (**Table 3-2 - 450k**). In agreement with elevated *DLX6* expression in LUSC, COAD, HNSC and LUAD tumor samples, the *DLX6* promoter region was consistently hypomethylated. Indeed, *DLX6* expression and *DLX6* promoter methylation are significantly and negatively correlated in the four datasets (r=-0.4801 p-value<0.0001 for LUSC, r=-0.5192 p-value<0.0001 for COAD, r=-0.4543 p-value<0.0001 for HNSC and r=-0.4694 p-value<0.0001 for LUAD).

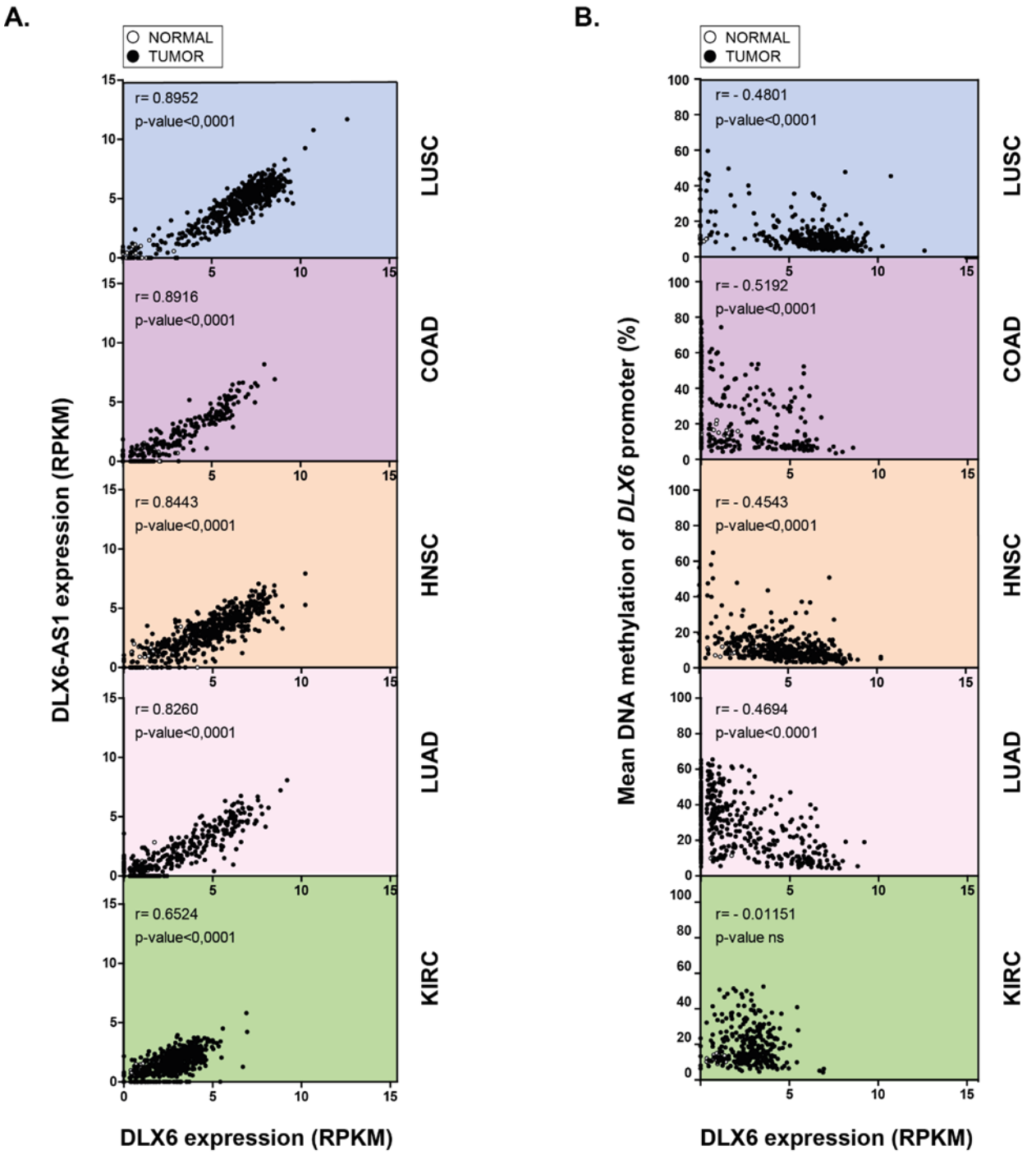


Figure 4-7 Coordinated upregulation of *DLX6* and *DLX6-AS1* in cancer and assessment of *DLX6* regulation by DNA methylation

(**A**) Spearman correlation plot between *DLX6-AS1* and *DLX6* expression measured by RNA sequencing from multiple TCGA datasets detailed in Table 3-2. (**B**) Spearman correlation plot between *DLX6* expression measured by RNA-sequencing and the average DNA methylation across the eight CpG sites for each sample in multiple TCGA datasets. Abbreviations: r= Spearman correlation, ns= non-significant.

Among the four datasets, the correlation coefficients in COAD and LUAD are affected by the presence of a distinct subgroup of tumors where *DLX6* expression is not dependent on DNA methylation levels. Indeed, in COAD a subgroup of tumors cluster with normal samples and are characterized by concomitant low *DLX6* expression and DNA methylation levels. In LUAD a fraction of tumors are instead defined by low *DLX6* expression with different ranges of *DLX6* DNA methylation (**Figure 4-7B)**.

In contrast, *DLX6* promoter DNA methylation levels in KIRC were not significantly correlated with expression. The correlation coefficient was very low and statistically not significant, that is, there is negligible or no relationship between *DLX6* expression levels and the methylation levels at its promoter in KIRC (**Figure 4-7B**).

In summary, *DLX6-AS1* overexpression correlates with *DLX6* expression in five different cancer types. The significant relationship between *DLX6* expression and promoter methylation in LUSC, COAD, HNSC and LUAD suggests that *DLX6* can be regulated through DNA methylation in these cancers. In contrast, *DLX6* seems not to be regulated by altered DNA methylation levels in KIRC and in a subgroup of samples in COAD and LUAD.

4.2.3 Clinical evaluation of *DLX6/DLX6-AS1* overexpression

A. Clinical significance of *DLX6/DLX6-AS1* overexpression in PCa

DLX6/DLX6-AS1 pair is overexpressed and negatively correlates with *DLX6* DNA methylation levels in a subset of PCa patients designed previously as the “HIGH” expression group in ICGC or TCGA datasets (**section 4.2.1**). We sought to determine whether this expression subgroup is associated with distinct clinical or genomic parameters compared to the “LOW” expression subset of patients (**Table 4-3**).

Table 4-3 Correlation of "LOW" and "HIGH" *DLX6-AS1* expression subgroups with clinical parameters

		ICGC EO-PCa			TCGA Late-onset PCa			
All samples		**LOW (n=46)**	**HIGH (n=45)**	**P-VALUE**		**LOW (n=250)**	**HIGH (n=248)**	**P-VALUE**
Age	**Median**	48	49	0.644[T]		62	61	0.1234[T]
Gleason score n=88	**6**	7(15.6)	5(11.6)	0.2406[CT]	**n= 497**	20(8)	24(9.7)	0.2642[C]
	7	29(64.4)	35(81.4)			116(46.6)	131(52.8)	
	8	1(2.2)	0			31(12.5)	33(13.3)	
	9	8(17.8)	3(7)			79(31.7)	59(23.8)	
	10	0	0			3(1.2)	1(0.4)	
Pathological T stage n=88	**T1**	0	0	0.5520	**n=490**	0	0	0.3805
	T2	31(68.9)	29(67.5)			98(40)	88(35.9)	
	T3	11(24.4)	13(30.2)			140(57.2)	153(62.5)	
	T4	3(6.7)	1(2.3)			7(2.8)	4(1.6)	
ERG fusion n=41	**ERG +**	6(27.3)	17(89.5)	<0.0001	**n=333**	24(15.8)	128(70.7)	<0.0001
	ERG -	16(72.7)	2(10.5)			128(84.2)	53(29.3)	
Deletion 3p13 n=57	**Deleted**	3(14.3)	16(44.4)	0.0230				
	Non deleted	18(85.7)	20(55.6)					
Deletion 5q21.1 n=57	**Deleted**	4(19.1)	2(5.6)	0.7250				
	Non deleted	17(80.9)	34(94.4)					
Deletion 10q23.31 n=57	**Deleted**	5(23.8)	12(33.3)	0.5549				
	Non deleted	16(76.2)	24(66.7)					
Deletion 17p31 n=57	**Deleted**	5(23.8)	16(44.4)	0.1589				
	Non deleted	16(76.2)	20(55.6)					

For each feature the number of patients affected is indicated and within brackets the respective percentage is given. To calculate the statistical differences among groups the following test were applied: [T]= Student's t-test, [CT]=chi-squared test trend, [C]=chi-squared test and if not otherwise stated Fisher's exact test was employed.

There was no statistical significant association between DLX6-AS1 expression and age, Gleason score or the pathological T stage in both ICGC and TCGA cohorts (**Table 4-3**). Indeed, *DLX6/DLX6-AS1* overexpression is not dependent on the age at diagnosis of patients and did not differ significantly between the two groups (48 versus 49 years old) in ICGC EO-PCa patients and in TCGA late-onset PCa patients (62 versus 61 years old). Alterations in *DLX6/DLX6-AS1* levels were also independent of Gleason scores with a prevalence of patients reaching a Gleason score of 7 in both the "LOW" and "HIGH" subgroups in ICGC and TCGA datasets. Similarly, the two

subgroups could not be separated by distinct pathologic T stages, with preponderance for the T2 and T3 stage in the ICGC and TCGA cohorts, respectively.

In contrast, *DLX6/DLX6-AS1* overexpression ("HIGH") is highly and significantly associated with tumors harbouring the *TMPRSS2:ERG* fusion gene (ERG+) in ICGC (89.5% versus 27.3%) and TCGA (70.7% versus 18.8%) datasets (p-value<0.0001). Consistently the "LOW" subgroup is linked to patients not harbouring the fusion gene (ERG-) in ICGC (72.7% versus 10.5%) and TCGA (84.2% versus 29.3%) cohorts.

The abundance of ERG+ patients within the "HIGH" expression subgroup prompted us to investigate the absence of 5q21 deletion and the presence of deletions on 3p13, 10q23 and 17p31 described to frequently co-occur in patients harbouring the fusion gene (**section 1.3.1**). We found that the absence of deletions on the 5q21 locus (p-value 0.7050) and the presence of deletions on 10q23 (p-value 0.5549) and 17p31 (p-value 0.1589) were not significantly enriched in the "HIGH" compared to the "LOW" expression subgroup. In contrast, predominance of patients (85.7%) without a deletion on the 3p13 locus is significantly linked to the "LOW" expression group (p-value 0.0230) (**Table 4-3**).
Overall, our comparison of *DLX6/DLX6-AS1* expression with clinical and genetic parameters revealed, that high-level of *DLX6/DLX6-AS1* expression are particularly linked with the occurrence of both *TMPRSS2:ERG* fusion gene and 3p13 deletion.

B. Impact of *DLX6/DLX6-AS1* overexpression on prognosis

Two clinical endpoints, overall (**Figure 4-8A**) and disease-free survival (**Figure 4-8B**) were evaluated by Kaplan-Meier survival analysis to assess the relationship between *DLX6-AS1* expression and survival times in the TCGA datasets on PRAD, LUSC, COAD, HNSC, LUAD and KIRC. To study the clinical significance of *DLX6-AS1* expression, we divided each dataset into two groups (LOW and HIGH) with the median *DLX6-AS1* expression value serving as the cut-off point (**Table 3-3**).

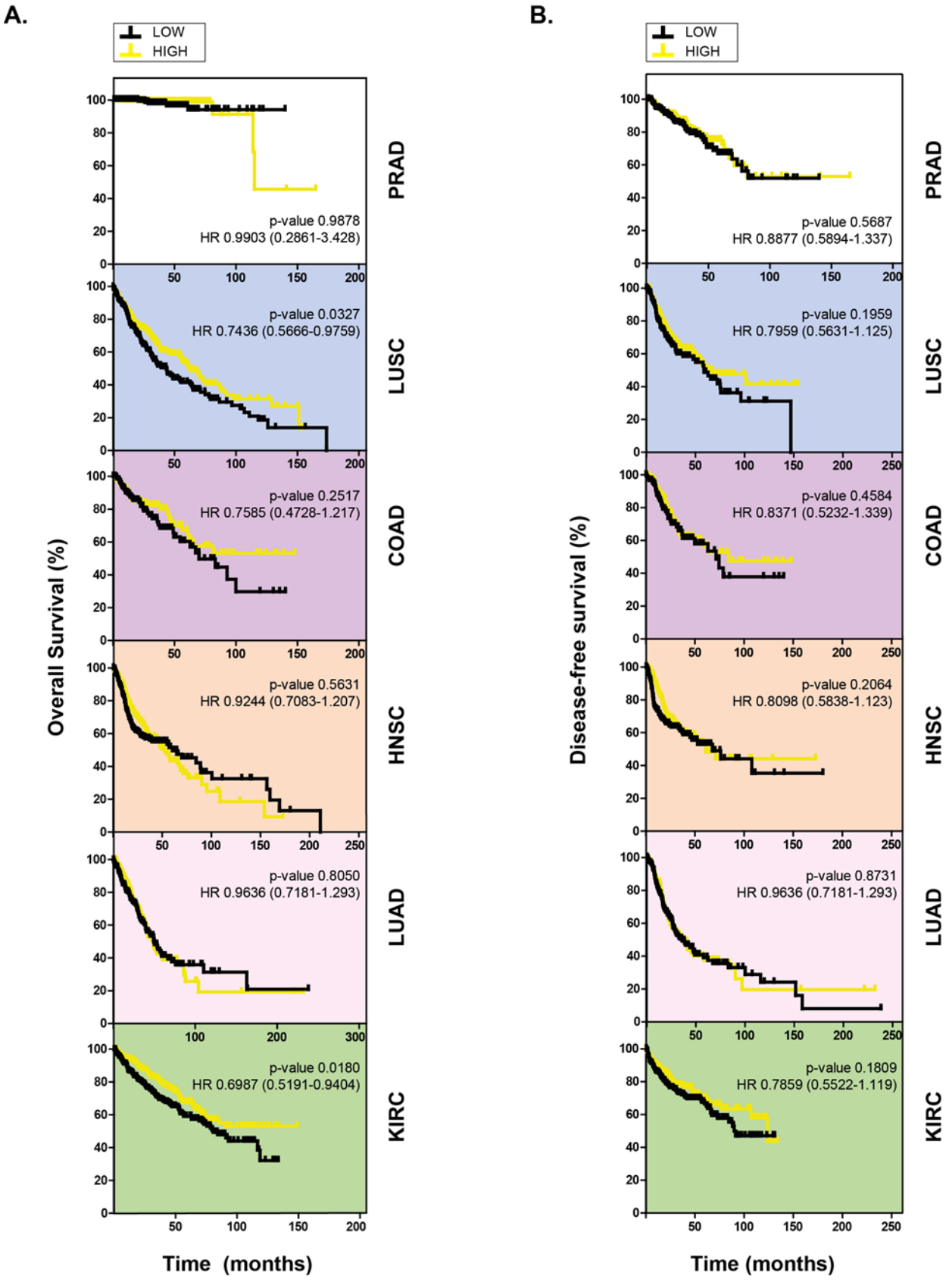

A.
LOW
HIGH
Overall Survival (%)
PRAD
p-value 0.9878
HR 0.9903 (0.2861-3.428)
LUSC
p-value 0.0327
HR 0.7436 (0.5666-0.9759)
COAD
p-value 0.2517
HR 0.7585 (0.4728-1.217)
HNSC
p-value 0.5631
HR 0.9244 (0.7083-1.207)
LUAD
p-value 0.8050
HR 0.9636 (0.7181-1.293)
KIRC
p-value 0.0180
HR 0.6987 (0.5191-0.9404)
Time (months)
B.
LOW
HIGH
Disease-free survival (%)
PRAD
p-value 0.5687
HR 0.8877 (0.5894-1.337)
LUSC
p-value 0.1959
HR 0.7959 (0.5631-1.125)
COAD
p-value 0.4584
HR 0.8371 (0.5232-1.339)
HNSC
p-value 0.2064
HR 0.8098 (0.5838-1.123)
LUAD
p-value 0.8731
HR 0.9636 (0.7181-1.293)
KIRC
p-value 0.1809
HR 0.7859 (0.5522-1.119)
Time (months)

Figure 4-8 *DLX6-AS1* expression subgroups association with overall and disease-free survival in TCGA datasets

Kaplan-Meier analyses of the overall survival time (**A**) or disease-free time (**B**). Subjects were stratified into "LOW" and "HIGH" subgroups according to *DLX6-AS1* expression levels in PRAD, LUSC, COAD, HNSC, LUAD and KIRC from the TCGA (Table 3-3). P-values were determined using the log-rank test. Also shown is the Hazard ratio (HR) with their respective 95% confidence intervals indicate within brackets.

Elevated *DLX6-AS1* expression ("HIGH") was not significantly associated with altered overall survival time in PRAD, COAD, HNSC and LUAD (**Figure 4-8A**). This result should however, be interpreted with caution for PCa patients, since only a limited number of events (n= 10) are available to assess the difference in overall survival times in this cohort (**Table 3-3** - PRAD). In contrast, high *DLX6-AS1* expression was associated with longer overall survival times in LUSC (p-value 0.0327) and KIRC (p-value 0.0180), respectively. In addition, for the six cohorts, higher *DLX6-AS1* expression was not associated with increased disease-free survival (**Figure 4-8B**).

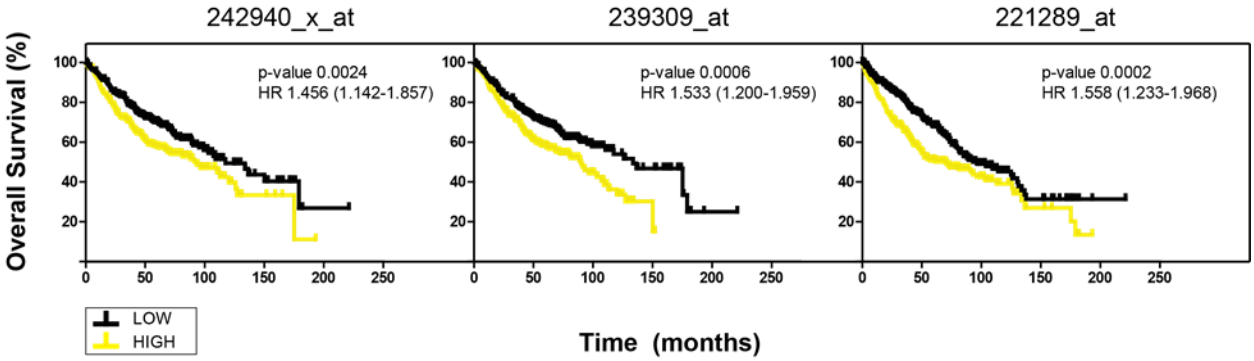


Figure 4-9 Elevated *DLX6-AS1* expression predicts overall survival in LUAD

Kaplan-Meier analyses of the overall survival time. Subjects were stratified into "LOW" and "HIGH" subgroups according to DLX6 expression levels in LUAD extracted from the KM plotter (Table 3-4). P-values were determined using the log-rank test. Also shown is the Hazard ratio (HR) with its respective confidence interval indicate in brackets.

In the TCGA cohort for LUAD patients, high *DLX6-AS1* expression did not significantly impact the overall survival time of patients (**Figure 4-8A**). We extended our analysis to an independent and larger cohort of LUAD patients (n=673 for 242940_x_at and 239309_at and n=720 for 221289_at) extracted from the Kaplan-Meier plotter software (**Table 3-4**). These data are based on expression microarrays not including probes mapped to the lncRNA *DLX6-AS1*. Since *DLX6* and *DLX6-AS1* expression are significantly and positively correlated in LUAD (**Figure 4-7A**), we used

the *DLX6* median expression value to stratify the patients into “LOW” and “HIGH” expression subgroups. In this cohort, elevated *DLX6* expression levels in the three probes for *DLX6* present on the array (242940_x_at, 239309_at and 221289_at) are significantly associated with decreased overall survival (for 242940_x_at p-value=0.0024, for 239309_at p-value=0.0006 and for 221289_at p-value=0.002) (**Figure 4-9**).

In summary, elevated *DLX6-AS1* levels are associated with a favourable survival in LUSC and KIRC, whereas in LUAD the overexpression is linked to worse overall survival.

4.3 *In vitro* characterization of *DLX6/DLX6-AS1* expression

4.3.1 Identification of the *DLX6-AS1* major transcript variant

The lncRNA ***DLX6-AS1*** is an intergenic RNA located on chromosome 7q22 antisense to the protein coding gene *DLX6*. *DLX6* is a homeobox transcription factor, located tail-to-tail with *DLX5* (GENCODE **- Figure 4-10**).

Of the transcript variants that have been described for *DLX6-AS1*, RNA sequencing reads of EO-PCa patients (RNA-seq TUMOR - **Figure 4-10**) used for the screening approach mapped most accurately to the second and the beginning of the third exon of the *ENST00000430027* transcript, whereas no peak could be detected for its first exon lying within the intergenic region between *DLX6* and *DLX5* genes (Gencode and RNA-seq TUMOR- **Figure 4-10**). *Ab initio* assembly of RNA sequencing data from 27 human tumor or tissue types recently has expanded the landscape of *DLX6-AS1* transcripts to variants encompassing alternative 5’ends (MiTranscriptome **- Figure 4-10**) (Iyer et al., 2015) . The first exon of the two alternative variants depicted in Figure 4-2 corresponds to a RNA-seq peak lying next to a peak corresponding to DLX6 first exon.

We identified the transcriptional start site of *DLX6-AS1* by 5’RACE (Rapid Amplification of 5’ cDNA Ends) using nested PCR-primers located within the *DLX6-AS1* second exon (5´RACE primers **- Figure 4-10**) in the VCaP prostate cancer cell line (5´RACE VCaP **- Figure 4-10**) where the pair of transcripts is highly expressed (**Figure 4-13**). The TSS was mapped to the first intron of *DLX6* and

fitted best to *DLX6-AS1* transcript variants annotated as G077079|T328744 (renamed as *DLX6-AS1 T1*) and G077079|T328743 (renamed as *DLX6-AS1 T2*) from the MiTranscriptome catalog (**Figure 4-10**).

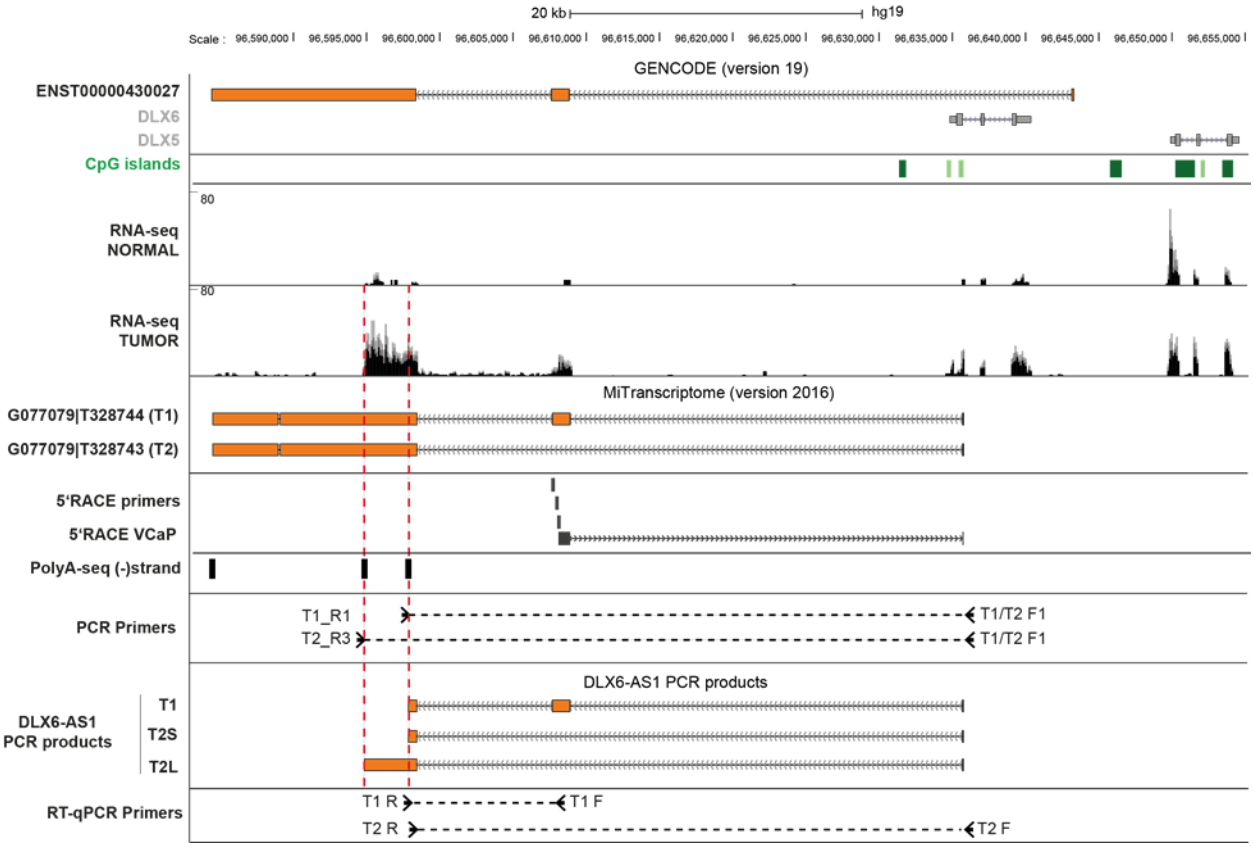
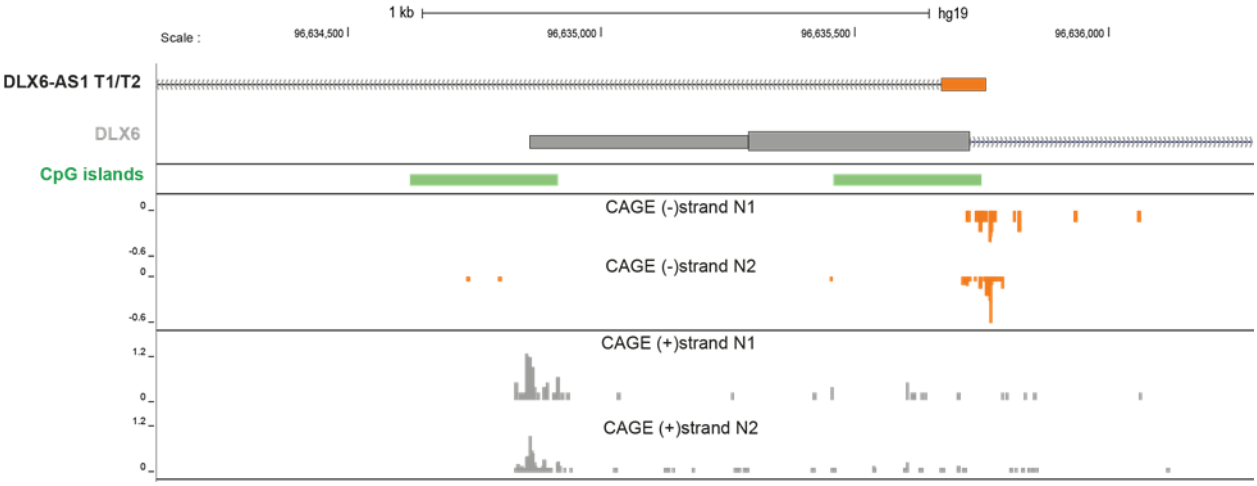


Figure 4-10 Characterization of *DLX6-AS1* transcript variants and transcription start site

Transcript structure of *DLX6-AS1* variants from GENCODE version 19 as well as determined by the *de novo* assembly in the MiTranscriptome catalog. Representative RNA sequencing reads mapping to *DLX6-AS1*, *DLX6* and *DLX5* from one representative normal and one early-onset ICGC prostate cancer patient. The mapping of *DLX6-AS1* TSS was determined by 5'RACE and depicted is the sequencing result along with the position of primers used for the nested PCR in VCaP cells. Also shown are the location of the primers used to amplify *DLX6-AS1* cDNA by PCR and the resulting genomic structure of *DLX6-AS1* PCR products identified by sequencing of gel extracted PCR products. In the last panel is also shown the location of primers use to detect *DLX6-AS1 T1* and *T2* variants by RT-qPCR.

Figure 4-11 Confirmation of the localization of *DLX6-AS1* TSS



Enlargement of the region overlapping the first exon of *DLX6* and *DLX6-AS1 T1/T2* transcripts. Depicted are the sequencing profiles of the CAGE data for the (-) and (+) DNA strand in duplicates (N1 and N2) generated in H1299 cells treated with dimethyl sulfoxide (data generated by Dr. C. Schmidt).

This result was confirmed by strand specific CAGE (Cap analysis of gene expression) followed by sequencing (data provided by Dr. C. Schmidt) (CAGE-seq(-)strand N1 and N2 - **Figure 4-11**), performed in the lung cancer cell line H1299 also exhibiting high expression levels of *DLX6* and *DLX6-AS1* (**Figure 4-13**).

To confirm the existence of *DLX6-AS1 T1* and *T2* variants we PCR-amplified in VCaP cells both variants based on the annotation described in the MiTranscriptome database, 5'RACE result, and on the RNA-sequencing profiles in EO-PCa patients. The forward primer T1/T2_F1 was localized at the 5' end of the first exon common for the *T1* and *T2* transcripts. Two reverse primers, T1_R1 and T2_R3 correspond to alternative polyadenylation sites overlapping with the RNA-seq reads matching *DLX6-AS1* lncRNA (**Figure 4-10**). We identified three different *DLX6-AS1* products with the two sets of primers. Indeed, the reverse primer T1_R1 allowed the amplification of two products corresponding to *T1* and *T2* transcripts, while with T2_R3 only a single PCR product coinciding with an elongated form of the *T2* transcript could be detected (**Figure 4-12**). These results suggest that *DLX6-AS1 T1* and *T2* variants have different 3' ends that are formed by polyadenylation at distinct sites. *DLX6-AS1 T2* transcript with an extended 3' end was named *T2L* (long), whereas the shorter form was referred to as *T2S* (*DLX6-AS1* PCR products-**Figure 4-10**).

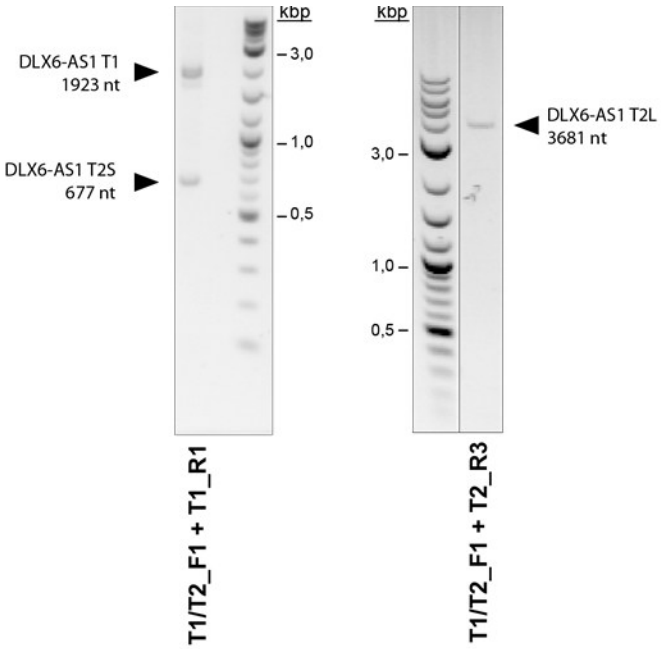


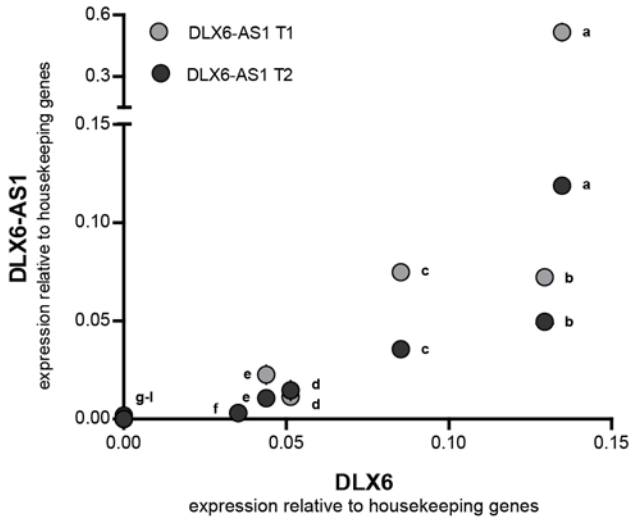
Figure 4-12 *DLX6-AS1* codes for two splice variants

DLX6-AS1 PCR product generated from cDNA using the two indicated set of primers.

4.3.2 Characterization of *DLX6/DLX6-AS1* expression in cell lines

To confirm the differential and coordinated expression of *DLX6* with *DLX6-AS1* observed in patients, we measured and correlated their expression in two benign prostate epithelial cell lines (BPH1 and PNT2) and ten cancerous cell lines from different tissue types (**Figure 4-13**). The two *DLX6-AS1* transcripts *T1* and *T2* (*T2S* and *T2L*) differ by the presence of an additional exon in the *T1* variant. This unique sequence information allowed us to determine the levels of both splice variants with the two sets of RT-qPCR primers depicted in **Figure 4-10**. RT-PCR analysis revealed that both *DLX6-AS1* variants are expressed, with a predominant expression of the *T1* variant in all cell lines expressing the pair of transcripts. In concordance with our screening strategy, the expression of *DLX6* and *DLX6-AS1* variant *T1* or *T2* are highly and significantly correlated in the twelve different screened cell lines (Spearman correlation =0.9841 for *T1* and 0.9714 for *T2*) (**Figure 4-13**). Moreover, the *DLX6/DLX6-AS1* pair exhibits higher expression levels in the PCa cell lines VCaP and DU145 compared to the benign prostate epithelial cells BPH1 and PNT2. The

expression of the sense and antisense transcript is not restricted to prostate, but can also be detected in the non-prostate cancerous cell lines HEK293T, H1299 and HCT116 DKO. In comparison to the DNMT1 and DNMT3B double knock out cell line HCT116 DKO, the sense/antisense pair is not expressed in HCT116 cells, suggesting a role for DNA methylation in *DLX6/DLX6-AS1* expression regulation.



ID	Cell lines	DLX6 mean	DLX6-AS1 T1 mean	DLX6-AS1 T2 mean
a	VCaP	0.1348	0.5161	0.1187
b	HEK293T	0.1294	0.0722	0.0495
c	H1299	0.0852	0.0748	0.0355
d	DU145	0.0513	0.0113	0.0145
e	HCT116 DKO	0.0438	0.0225	0.0106
f	BPH1	0.0352	0.0031	0.003
g	A549	0	0	0
h	22RV1	0	0	0
i	LNCaP	0	0	0
j	PNT2	0	0	0
k	PC3	0	0	0
l	HCT116	0	0	0

Transcript(s)	Spearman correlation	pvalue
DLX6-AS1 T1	**0,9841**	**<0,0001**
DLX6-AS1 T2	**0,9714**	**<0,0001**

Figure 4-13 *DLX6/DLX6-AS1* are coordinately expressed in cell lines

Spearman correlation plot between *DLX6-AS1 T1* or *T2* variants and *DLX6* expression measured by RT-qPCR and normalized to *HPRT1*, *SDHA* and *ALAS1* housekeeping genes. Each point in the correlation plot is attributed an identifier (ID) ranging from **a** to **l** and corresponding to a cell line indicated in the table. Mean expression values for each transcript in each cell lines is also reported in the table. Data are depicted as the mean, n= 3 biological replicates.

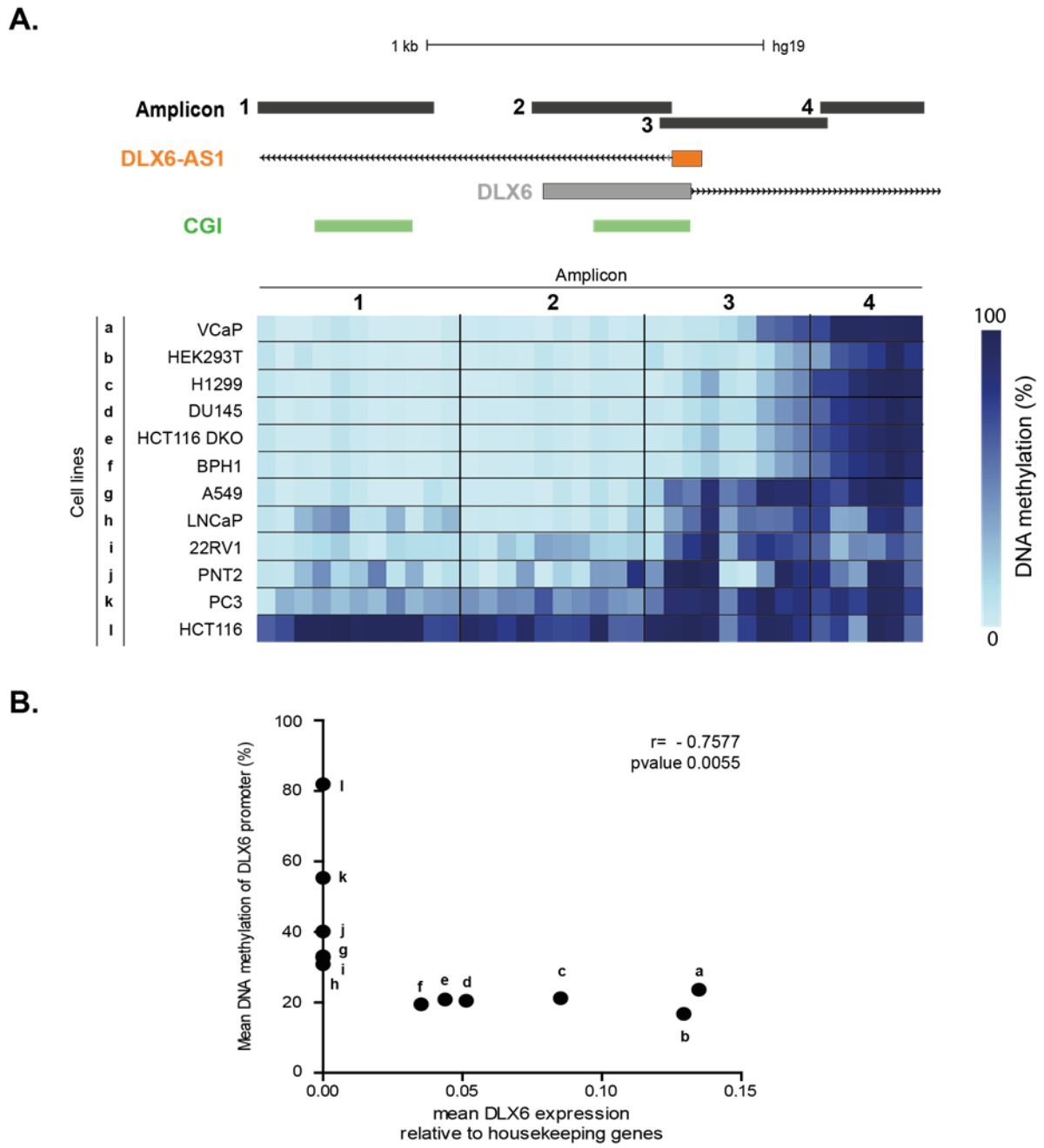


Figure 4-14 *DLX6* promoter is differentially methylated in cell lines

(A) Representation of MassArray amplicons overlapping with *DLX6* and *DLX6-AS1* promoter region and CpG islands. Quantitative DNA methylation levels are displayed as a heatmap at the *DLX6* and *DLX6-AS1* promoter region in VCaP (a), HEK293T (b), H1299 (c), DU145 (d), HCT116 DKO (e), BPH1 (f), A549 (g), LNCaP (h), 22RV1 (i), PNT2 (j), PC3 (k) and HCT116 (l) cell lines. Each column represents the mean methylation value of single or multiple CpG sites in the corresponding MassArray amplicon for every sample in each row. Mean DNA methylation levels of n= 3 biological replicates are depicted as a percentage ranging from 0% (light blue) to 100% (dark blue). **(B)** Spearman correlation plot between DLX6 expression measured by RT-qPCR and the mean DNA methylation of all amplicons for each cell line. Each

point on the correlation plot is attributed an identifier number (ID) ranging from **a** to **l** corresponding to a cell line indicated in the upper panel.

In line with *DLX6/DLX6-AS1* expression patterns in cell lines, the region of *DLX6* promoter overlapping Amplicon 1 - 3 is lowly methylated in cell lines expressing the pair of transcripts, while higher methylation levels are measured in cell lines **g**-**l** not expressing *DLX6/DLX6-AS1*. In contrast, Amplicon 4 located in *DLX6* gene body exhibits higher DNA methylation levels with increasing *DLX6/DLX6-AS1* expression (**Figure 4-14 A**). In addition, *DLX6* expression levels are significantly and inversely correlated with *DLX6* promoter methylation levels in all cell lines (r=−0.7577, p-value=0.0055) (**Figure 4-14 B**).

4.3.3 Characterization of *DLX6* promoter

Inverse correlation of *DLX6* and *DLX6-AS1* expression and methylation of the *DLX6* promoter region in all cell lines corroborated our initial hypothesis that the lncRNA regulates the coding mRNA by modulating DNA methylation levels. Nevertheless, the proximity of *DLX6-AS1* TSS to *DLX6* transcription start (distance of 900 bp) identified by 5'RACE implicate that both RNAs could be regulated by this shared genomic region (**Figure 4-15**). Publicly available chromatin immunoprecipitation and sequencing (ChIP-seq) data for RNA polymerase II (Pol II) binding and H3K4 trimethylation indicating active transcription (Yu et al., 2010) were inspected for a common regulatory region of the sense and antisense RNAs in VCaP and LNCaP cells. In VCaP cells, active expression of the *DLX6/DLX6-AS1* transcript pair is supported by the presence of H3K4me3 along with a single Pol II binding site in the *DLX6* promoter (in the region of the CGI covered by MassArray Amplicon 1), both characterizing active promoters. This result is in line with the idea that a common promoter regulates *DLX6* and *DLX6-AS1* transcription. By contrast, neither active histone marks nor Pol II binding were identified in the region between *DLX6* and *DLX6-AS1* TSSs in LNCaP cells (**Figure 4-15**).

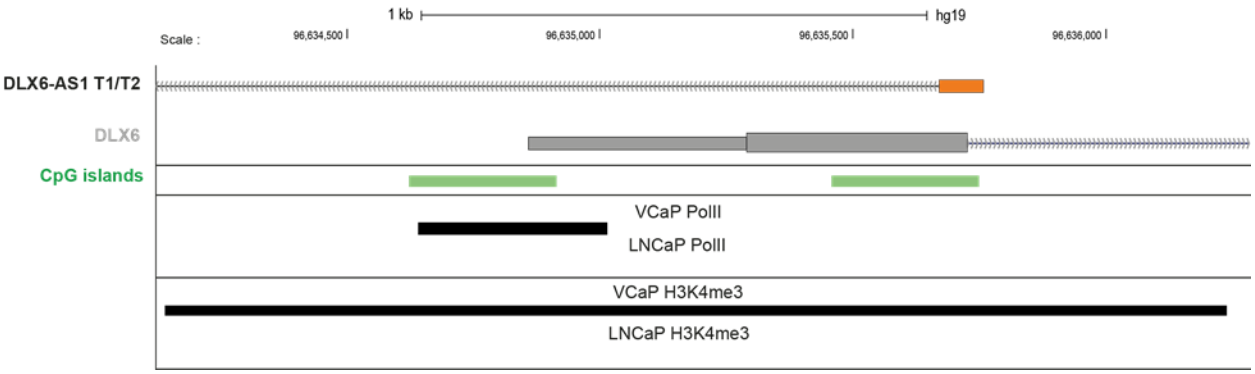


Figure 4-15 *DLX6/DLX6-AS1* share a common promoter

Profiles of H3K4me3 and RNA polymerase II ChIP-seq data in VCaP and LNCaP cell lines from Yu and colleagues (Yu et al.,2010) overlapping with *DLX6* promoter region.

In summary, *DLX6* and *DLX6-AS1* are coordinately expressed, and *DLX6* expression inversely correlates with DNA methylation levels in the *DLX6* promoter region. The coordinated expression of the sense and antisense transcript could be the sign for interdependency of their expression, or the result of a shared promoter.

4.4 Study of the influence of *DLX6-AS1* on *DLX6* expression

The following set of experiments was designed to test the screening hypothesis and delineate the interrelationship between the expression of *DLX6* and *DLX6-AS1* either by knockdown or overexpression of the antisense transcript in experimental conditions where the genes are expressed or lowly expressed, respectively. The potential epigenetic regulation between the sense and antisense transcripts was determined by RT-qPCR and DNA methylation analysis by MassArray.

4.4.1 Downregulation of *DLX6-AS1* does not influence *DLX6* expression

We observed a positive correlation between *DLX6* and *DLX6-AS1* expression and, through knockdown of *DLX6-AS1*, wanted to investigate whether *DLX6* is directly regulated by the antisense lncRNA *DLX6-AS1*.

First, we downregulated *DLX6-AS1* by stable expression of shRNAs in VCaP cells and by transient siRNA or LNA transfection in HEK293T cells. Even though *DLX6-AS1 T1* and *T2* transcripts expression were drastically reduced in VCaP cells with a respective minimum knockdown efficiency of 80% and 65%, *DLX6* transcript levels remained unaffected (**Figure 4-16A**-upper panel). RNA interference (RNAi)-based techniques such as shRNA and siRNA are believed to target mainly cytoplasmic RNAs (Bassett et al., 2014). Our hypothesis relies however on a nuclear role of *DLX6-AS1*. For this reason, we employed additionally LNA GapmeRs. Their chimeric RNA-DNA nature triggers RNase H dependent degradation of nuclear RNAs more efficiently than RNAi-based techniques (Bassett et al., 2014). *DLX6-AS1* knockdown by siRNA or LNA in HEK293T cells did however, not reduce *DLX6* expression (**Figure 4-16B**- upper panel).

While *DLX6-AS1* does not regulate *DLX6* at the RNA level, we additionally tested whether knockdown of the transcription factor *DLX6* could influence *DLX6-AS1* expression. Downregulation of *DLX6* mRNA levels by shRNAs, although not as effective as for *DLX6-AS1*, lead

to a weak, non-significant decrease of *DLX6-AS1 T1* levels by about 20% in VCaP cells (mean rel. expression level=0,83 for sh#6 and =0,64 for sh#2) without affecting the *DLX6-AS1 T2* variant (**Figure 4-16A -** lower panel). This effect could not be observed with a pool of four different siRNAs targeting *DLX6* mRNA in HEK293T cells, although the knockdown was more effective than by shRNA in VCaP cells (mean rel. expression =0,29 in HEK293T versus mean rel. expression=0,49 after transfection of VCaP with sh#2) (**Figure 4-16B -** lower panel). This suggests that *DLX6-AS1 T1* alteration after *DLX6* knockdown in VCaP cells is not related to *DLX6* function.

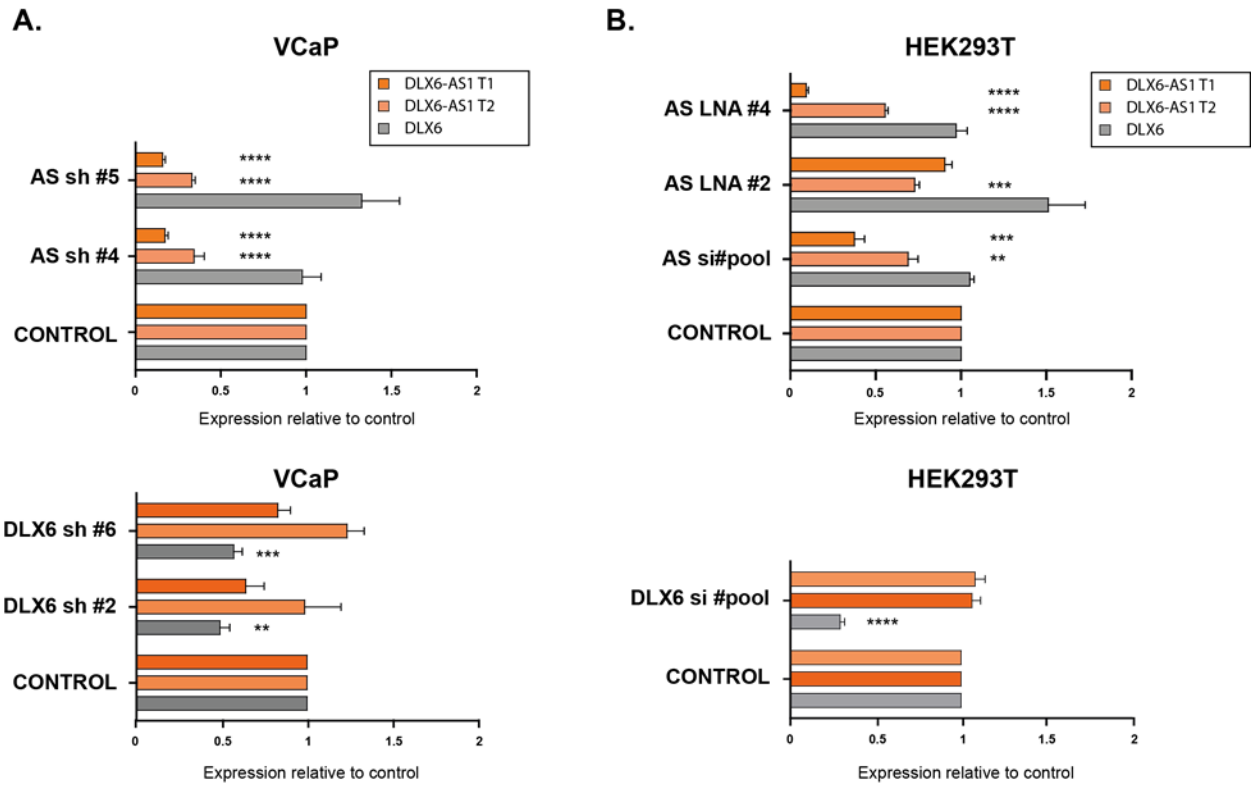


Figure 4-16 Downregulation of *DLX6-AS1* lncRNA does not impact *DLX6* expression

(A) Relative expression levels of *DLX6-AS1* and *DLX6* 8 days after knockdown of *DLX6-AS1* (AS) (upper panel) or *DLX6* (lower panel) relative to control scrambled shRNA in VCaP prostate cancer cells. Data are depicted as the mean +/- s.e.m., n= 4 biological replicates. **(B)** Relative expression levels of *DLX6-AS1* and *DLX6* 72h after knockdown of *DLX6-AS1* (upper panel) or *DLX6* (lower panel) with 25nM siRNA or LNA relative to control scrambled shRNA or LNA in HEK293T cells. Data are depicted as the mean +/- s.e.m., n= 3 biological replicates. Two-tailed unpaired t-test was used to calculate statistical significance in comparison with the control group and annotated as follow: ****p<0.0001, ***p<0.001, **p<0.01. If not otherwise indicated there is no significant difference between samples.

Consistent with a lack of changes in *DLX6* transcript levels, *DLX6-AS1* downregulation did not influence DNA methylation levels in the region around the TSS of DLX6 gene covered by four MassArray Amplicons in VCaP (**Figure 4-17B**) and HEK293T cells (**Figure 4-17C**).

If according to our initial hypothesis *DLX6-AS1* would affect *DLX6* mRNA expression *via* targeting of DNA demethylating enzymes to the *DLX6* promoter, after *DLX6-AS1* silencing we would expect a gain in *DLX6* promoter DNA methylation resulting in lower *DLX6* expression. However, gain in DNA methylation can be dependent on the cell cycle, and in slow cycling cell lines (see introduction **section 1.1.3 A**) such as VCaP (doubling time 4-5 days) this process might take longer than the time frame of the knockdown.

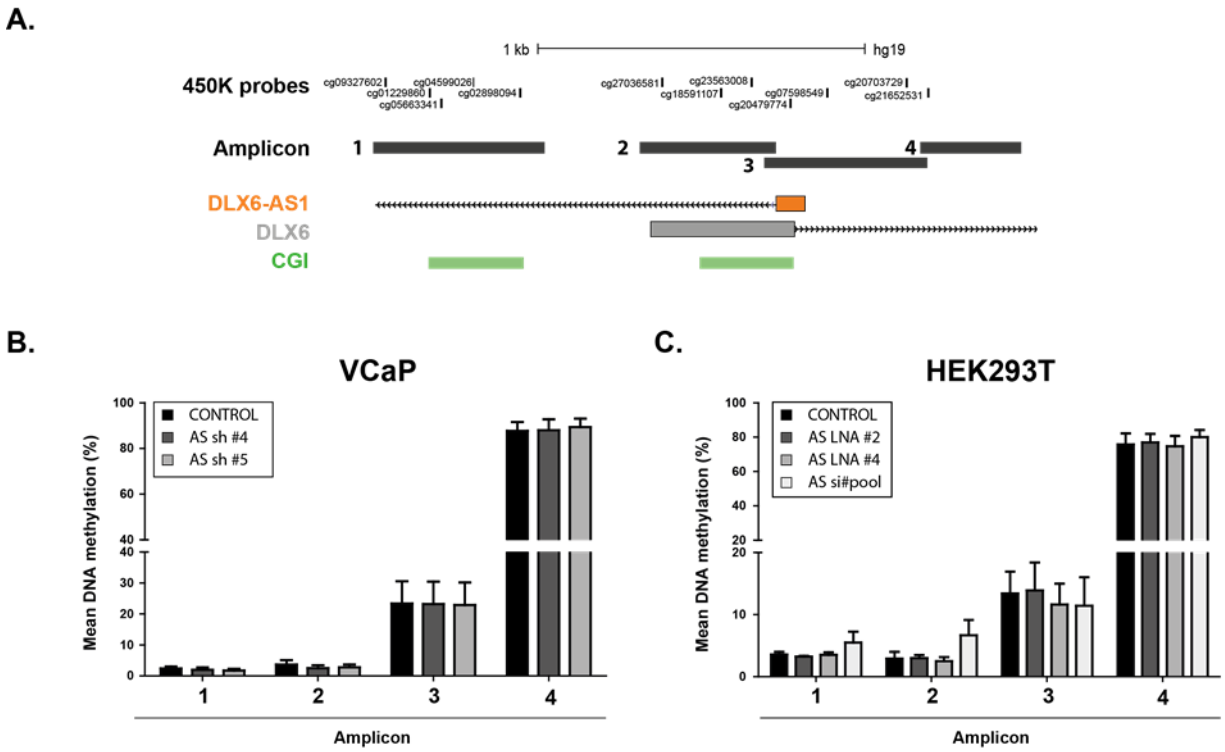


Figure 4-17 Downregulation of *DLX6-AS1* lncRNA does not impact *DLX6* promoter methylation

(A) Representation of MassArray amplicons overlapping with the *DLX6* promoter region. **(B** and **C)** DNA methylation levels at the *DLX6* promoter region are displayed on the y-axis as a percentage ranging from 0% to 100% in VCaP **(B)** or HEK293T **(C)** cells after knockdown of *DLX6-AS1*. Each depicted value represents the mean methylation level of all CpG sites within one amplicon. Data are depicted as the mean +/- s.e.m., n= 3 biological replicates. We detected no statistically significant differences between the control group and each knockdown experiment using two-tailed unpaired t-test.

4.4.2 Upregulation of *DLX6-AS1* does not influence *DLX6* expression

Since *DLX6-AS1* downregulation did not influence *DLX6* expression or promoter methylation, we next tested the reverse strategy by elevating levels of *DLX6-AS1* in cells lines with low antisense expression.

DLX6-AS1 RNA variants *T1*, *T2S* and *T2L* (**Figure 4-10**) were synthesized by *in vitro* transcription and transfected in A549 lung adenocarcinoma or PC3 prostate cancer cell lines. Both cell lines are characterized by absent expression of the *DLX6/DLX6-AS1* pair of transcripts. Expression is repressed by moderate (A549 cells) to high levels (PC3 cells) of DNA methylation in the region covered by MassArray Amplicons 1 - 3 when compared to VCaP cells expressing both transcripts (**Figure 4-14**). *DLX6-AS1 T1*, *T2S* and *T2L* transcript expression was consistently upregulated in the two cell lines in comparison to the constitutive expression levels of *T1* and *T2* transcripts in VCaP cells (**Figure 4-18A**). *DLX6-AS1* transfection did however not lead to *DLX6* re-expression (no expression detectable).

In a similar manner, overexpression of different amounts of a plasmid encoding for *DLX6* cDNA in LNCaP cells (**Figure 4-18B**) did not cause *DLX6-AS1* re-expression (no expression detectable).

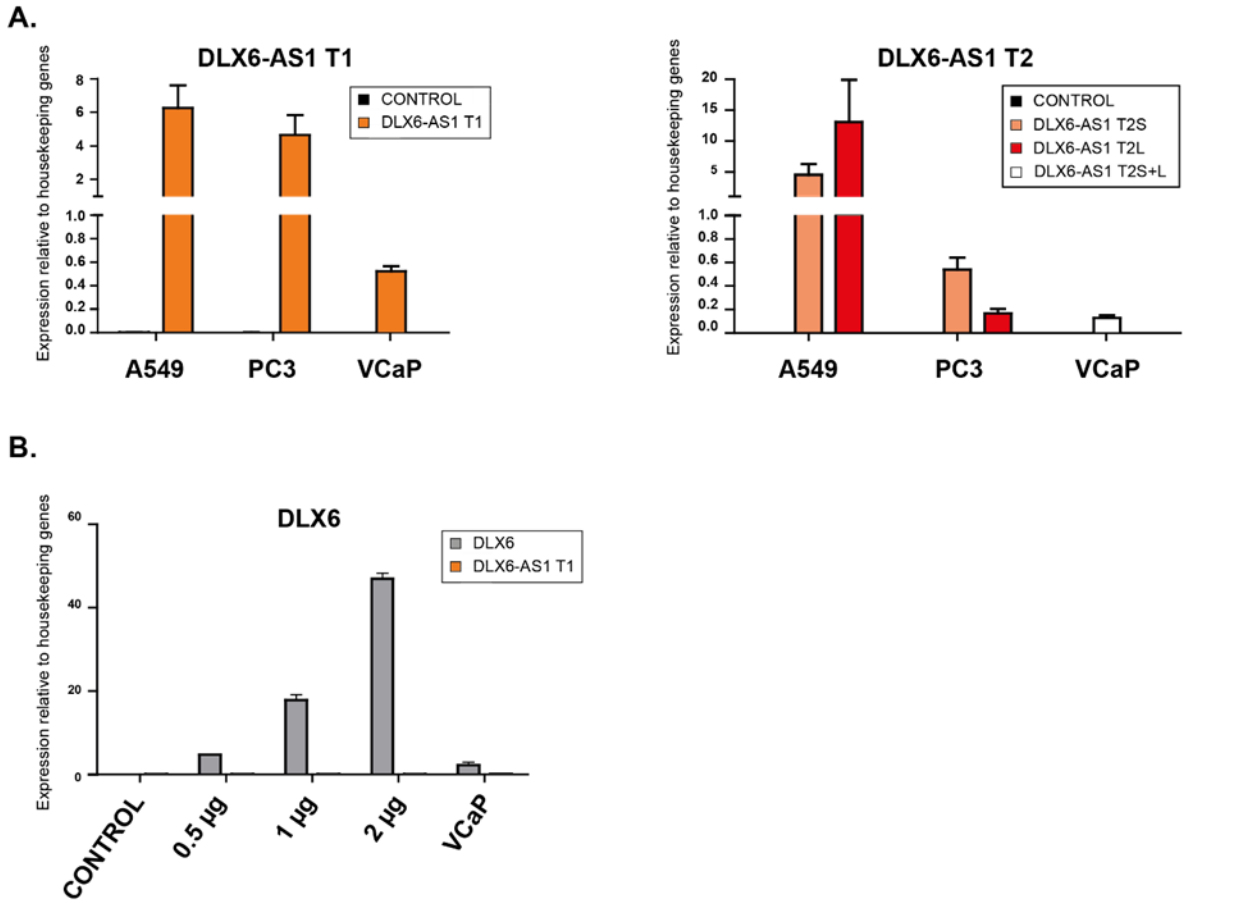


Figure 4-18 Elevated levels of *DLX6-AS1* does not impact *DLX6* expression

(A) Relative expression levels of *DLX6-AS1*, 48h after transfection of 400 fmol of *LACZ* (control), *T1*, *T2S* or *T2L DLX6-AS1* transcript in A549 or PC3 cells. *DLX6-AS1 T1* and *T2* transcripts levels are compared to the expression measured in VCaP cells. Data are depicted as the mean +/- s.e.m., n= 3 biological replicates. **(B)** Relative expression levels of *DLX6*, 72h after transfection of 0.5, 1 or 2µg of plasmid encoding for *DLX6* cDNA in LNCaP cells. Data are depicted as the mean +/- s.e.m., n= 2 biological replicates.

In agreement with unchanged *DLX6* mRNA expression levels after *DLX6-AS1* overexpression, DNA methylation in *DLX6* promoter region was not consistently changed across all four Amplicons in A549 (**Figure 4-19A**) and PC3 cell lines (**Figure 4-19B**). However, following transfection of the *T1* variant, we observed a weak but significant 2.4% decrease in DNA methylation of Amplicon 4 in A549 cells and a 7% increase in methylation of Amplicon 2 in PC3 cells. These weak effects were specific for the respective cell lines, and were not associated with DNA methylation alterations in the other tested Amplicons.

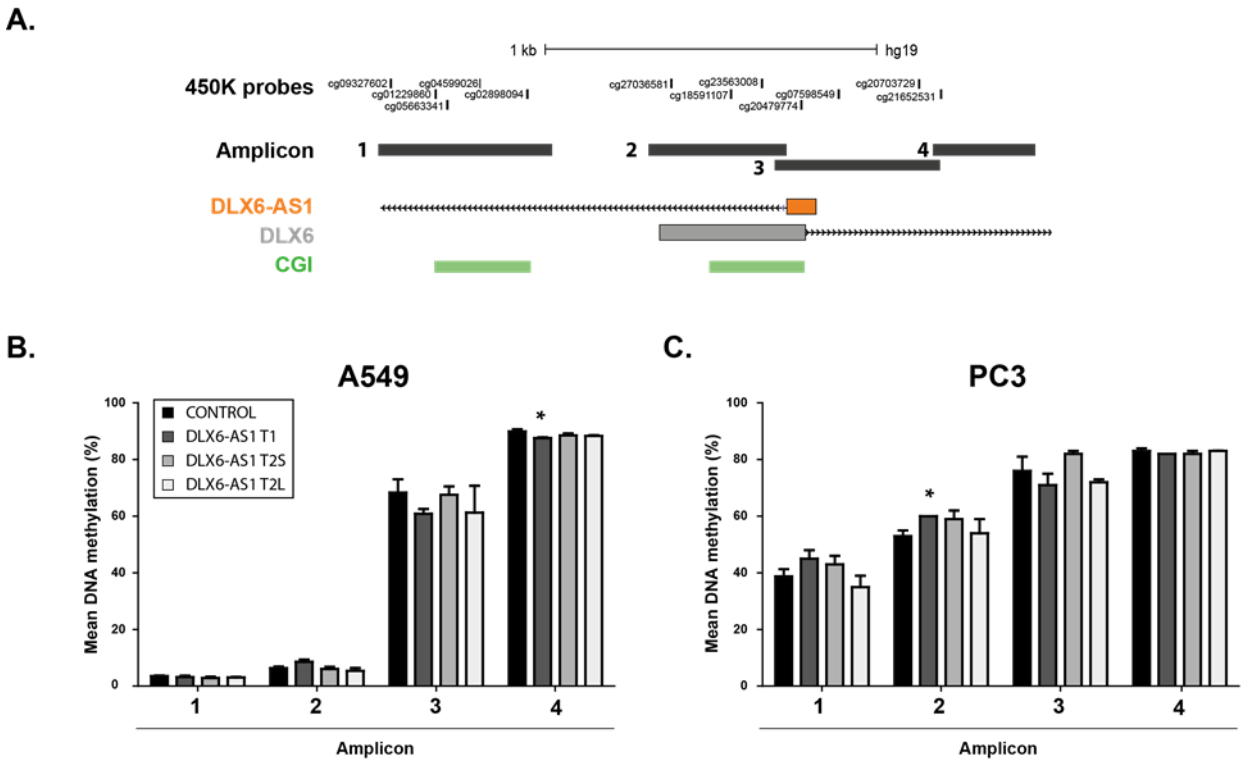


Figure 4-19 Elevated levels of *DLX6-AS1* does not impact *DLX6* promoter methylation

(A) Representation of MassArray amplicons overlapping with *DLX6* promoter region. **(B** and **C)** DNA methylation levels at the *DLX6* promoter region are displayed on the y-axis as a percentage ranging from 0% to 100% in A549 **(B)** or PC3 **(C)** cells after knockdown of *DLX6-AS1*. Each depicted value represents the mean methylation level of all CpG sites within one amplicon. Data are depicted as the mean +/- s.e.m., n= 3 biological replicates. Two-tailed unpaired t-test was used to calculate statistical significances between control versus each overexpression experiment.

Collectively, these results show that the lncRNA *DLX6-AS1* does not regulate expression of *DLX6* at the transcriptional level by influencing DNA methylation. In turn modulation of *DLX6* expression levels did as well not affect *DLX6-AS1* expression. This suggests that the coordinated expression of the pair of transcripts is the consequence of a shared regulation by differential methylation of the bidirectional *DLX6* promoter region.

4.5 Does *DLX6-AS1* have coding potential?

Accumulating evidence suggests that lncRNAs may encode functional micropeptides (**section 1.2.3**) (Anderson et al., 2015, Nelson et al., 2016). For this reason, we investigated by *in silico* and *in vitro* analyses whether *DLX6-AS1* might exert a biological function through translation into a small protein or peptide.

4.5.1 *DLX6-AS1 T1* has coding potential *in silico*

To identify potential small proteins or peptides in *DLX6-AS1 T1* (**Figure 4-20A**), *T2S* and *T2L* variants (**Figure 4-20B**) we employed ORFinder (NCBI). Translation efficiency is influenced by the sequence context surrounding a potential ORF. Therefore, we complemented the search for hypothetical proteins with the evaluation of potential translation start sites and optimal Kozak sequences by NETStart (Pedersen et al., 1997) and ATGpr (Salamov et al., 1998) algorithms, respectively. With this approach we identified four predicted ORFs in *DLX6-AS1 T1*. Among them, an ORF of 65 aa long referred as ORF1 was located in the second exon of *DLX6-AS1 T1* and was the only predicted ORF associated with a potential translation start site (**Figure 4-20A**). ORF1 starts at the first AUG codon in the *DLX6-AS1 T1* transcript sequence and therefore adheres to the first-AUG rule stating that the AUG codon is the exclusive initiation site in eukaryotes (Kozak, 1999).

DLX6-AS1 T2S and *T2L* share the first and third exon of *DLX6-AS1 T1* variant. Consequently, the two ORFs predicted for *T1* and not overlapping with the second exon of the latter are shared by *T2S* and T2L variants. For *DLX6-AS1 T2S* or *T2L*, none of the predicted ORFS were associated with both a reliable Kozak sequence and a potential translation start site. The highest translational start site score (0.71) was associated with an ORF of 34 aa localized in the second exon of *DLX6-AS1 T2L*, which we named ORF2 (**Figure 4-20B**).

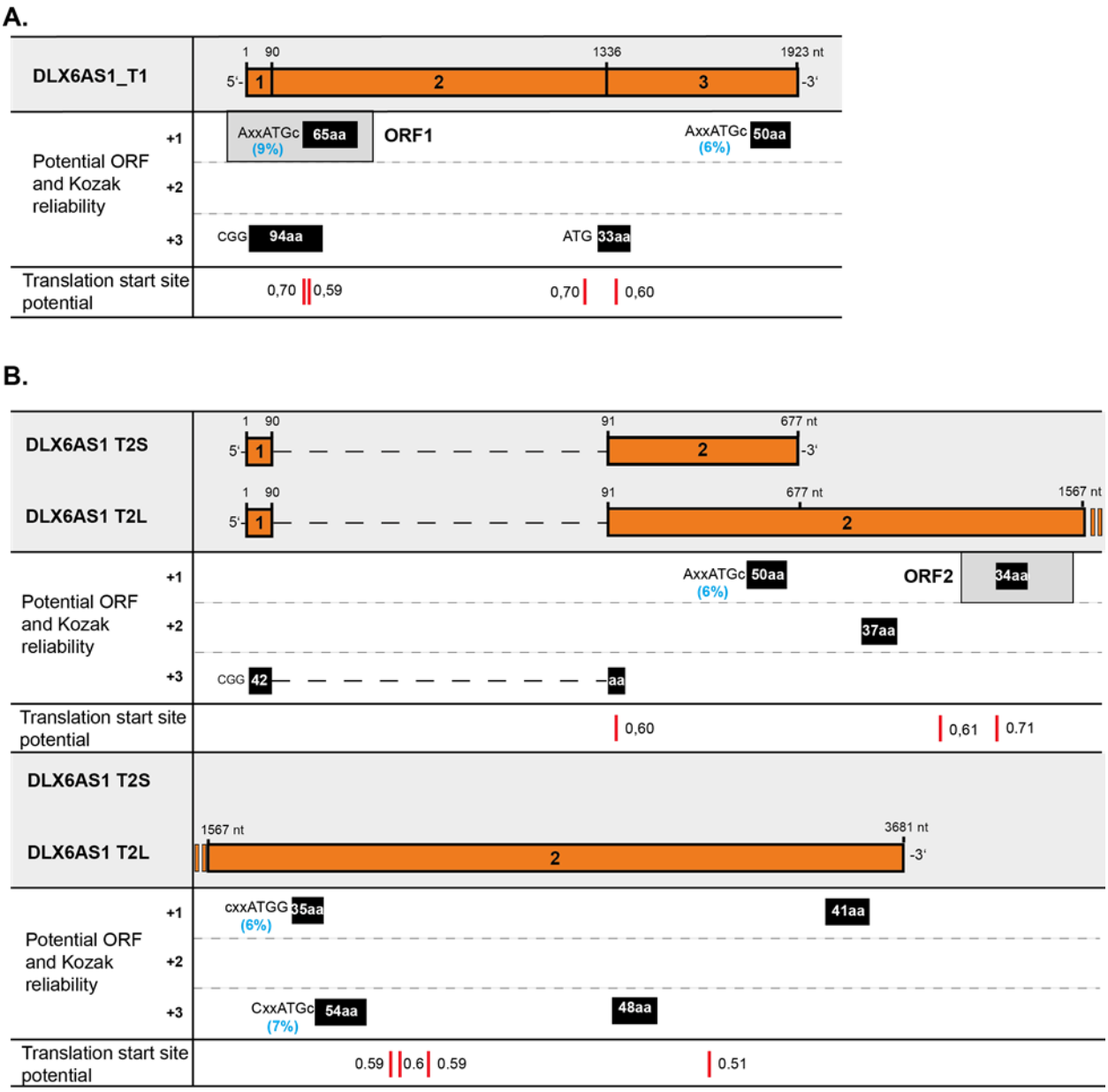


Figure 4-20 *DLX6-AS1* transcript coding potential

Representation of all ORFs predicted with ORFinder (version January 2016) overlapping *DLX6-AS1* **(A)** *T1* or **(B)** *DLX6-AS1 T2S* and *T2L* transcripts. For each ORF, the sequence matching the consensus Kozak sequence is given, within brackets the percentage of reliability of this motif calculated by ATGpr program. Depicted with red lines are all positively scored translation start sites predicted by NETStart with the corresponding reliability score ranging from 0 to 1 with increasing translation initiation site potential.

To assess whether ORF1 and/or ORF2 are biologically relevant, first we examined the conservation of both ORFs among primates by cross-species comparison, then we searched for the existence of similar proteins.

First, to assess the conservation of ORF1 and ORF2, we predicted with ORFinder proteins or peptides encoded in the cDNA homologues of *DLX6-AS1 T1* and *T2S/L* identified by nucleotide BLAST (NCBI). The protein coding sequence with the best match for ORF1 or ORF2 in each species was used for the multiple sequence alignment presented in **Figure 4-21**. The seven primate ORF1 homologues were highly similar with 75.7% overall sequence identity (**Figure 4-21A**). In contrast, ORF2 is poorly conserved with 11.8% sequence identity across six primates (**Appendix Figure 7-2**) but reaches 73.5% sequence identity when the comparison is restricted to the three ORF2 homologues predicted from *Nomascus leucogenesys* (northern white cheeked gibbon), *Rhynopithecus roxellana* (golden snub-nosed monkey) and *Collobus angolensis* (Angola colobus) (**Figure 4-21B**).

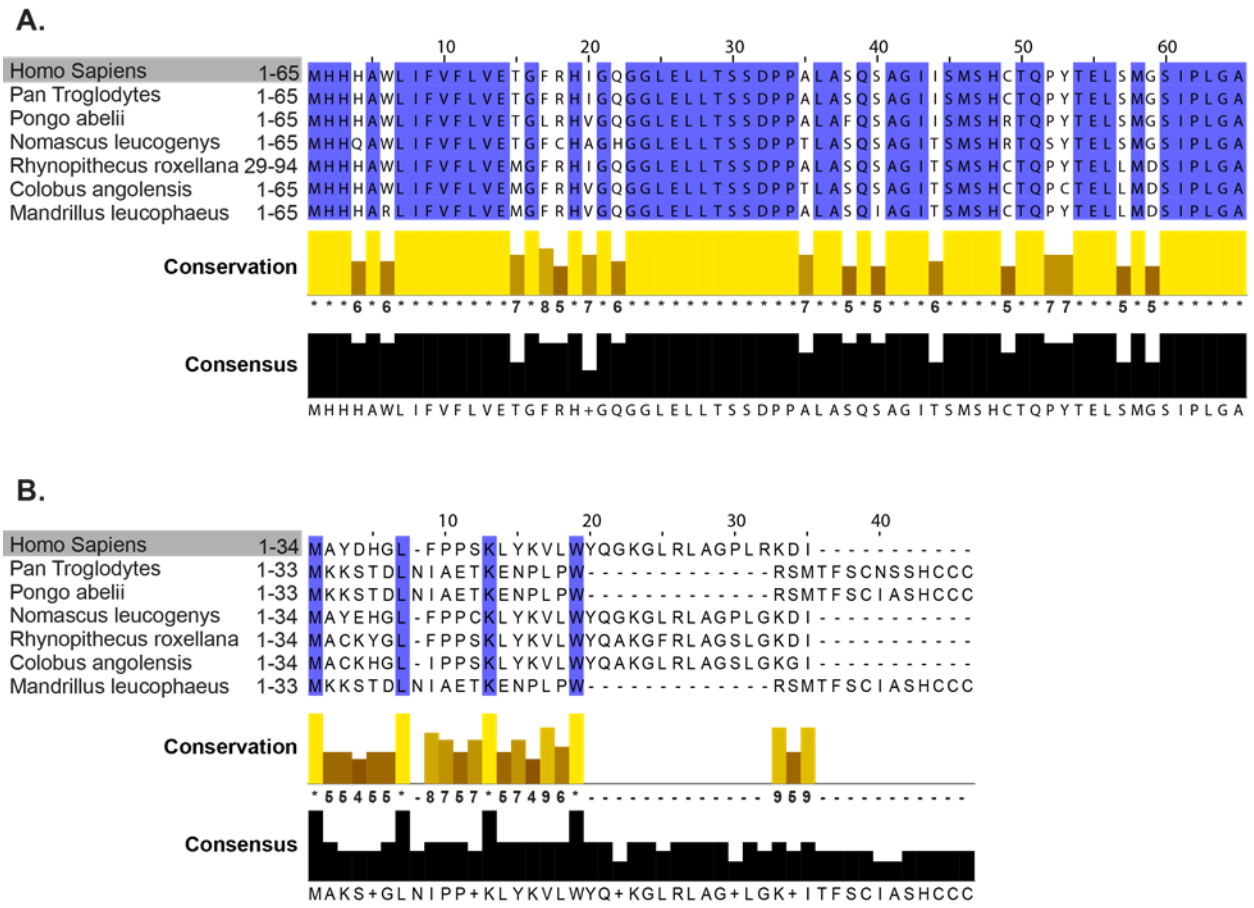


Figure 4-21 *DLX6-AS1 T1* ORF1 is conserved in primates

Multiple sequence alignment of *Homo sapiens* **(A)** ORF1 and **(B)** ORF2 homologs in primates using Mafft version 7. ORF1 and ORF2 homologs were predicted by ORFinder (NCBI) in *DLX6-AS1* cDNA homologs found by nucleotide BLAST (NCBI) of human *DLX6-AS1* cDNA sequence (Appendix 7.8). All the cDNA sequences were retrieved from Genbank with the following accession numbers; *Pan troglodytes* (LOC104007329), *Pongo abelii* (LOC103891166), *Nomascus leucogenys* (LOC100579968), *Rhynopithecus roxellana* (LOC104662289), *Colobus angolensis* (LOC105512765) and *Mandrillus leucophaeus* (LOC105535579). Identical amino acid residues are highlighted in blue, and below the conservation plot is a key denoting conserved sequence (*) and conservative mutations (+). Also indicated is the consensus sequence.

Next, to gain insights into a probable functional and biological relevance of both ORFs, we assessed whether ORF1 or ORF2 shares identity with other proteins or peptides.

For ORF2, a single homologous protein WP_020429213.1 could be identified by protein BLAST (**Figure 4-22**). This protein was annotated in the evolutionary distant bacterium *Paenibacillus riograndensis,* is of unknown function, and shares relatively low sequence identity of 33% with ORF2 with an expected value (e-value) inferred by protein BLAST of 3 (**Appendix Table 7-9**). Altogether, ORF2 is lowly conserved across primates and the low identity score with WP_020429213.1 suggests that this peptide is more likely to be a random sequence not encoding a functional peptide.

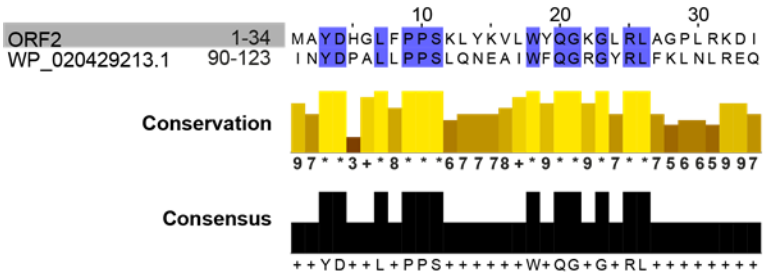


Figure 4-22 ORF2 shares low identity with a bacterial protein

Pairwise sequence alignment created with Clustal Omega of ORF2 amino acid sequence (Appendix 7.4) with WP_020429213.1 protein identified by protein BLAST. WP_020429213.1 amino acid sequence was retrieved from GenPept. Identical amino acid residues are highlighted in blue and below the conservation plot, is a key denoting conserved sequence (*) and conservative mutations (+). Also indicated is the consensus sequence.

ORF1 is the only one of the predicted ORFs for which we obtained by protein BLAST (NCBI) numerous identity matches belonging exclusively to the order of Primates (n=72 with identity ≥70%) (**Appendix Table 7-8**). In addition, the alignment of ORF1 with the proteins reaching a minimum bit-score of 80 and revealed a shared conserved domain spanning the 52 first aa in ORF1 (**Figure 4-23**). The search for functional domains and/or motifs in this sequence by comparison with Pfam (Finn et al., 2016), PROSITE (Sigrist et al., 2013), and the conserved domains databank (CDD) (Marchler-Bauer et al., 2015) was, however, negative.

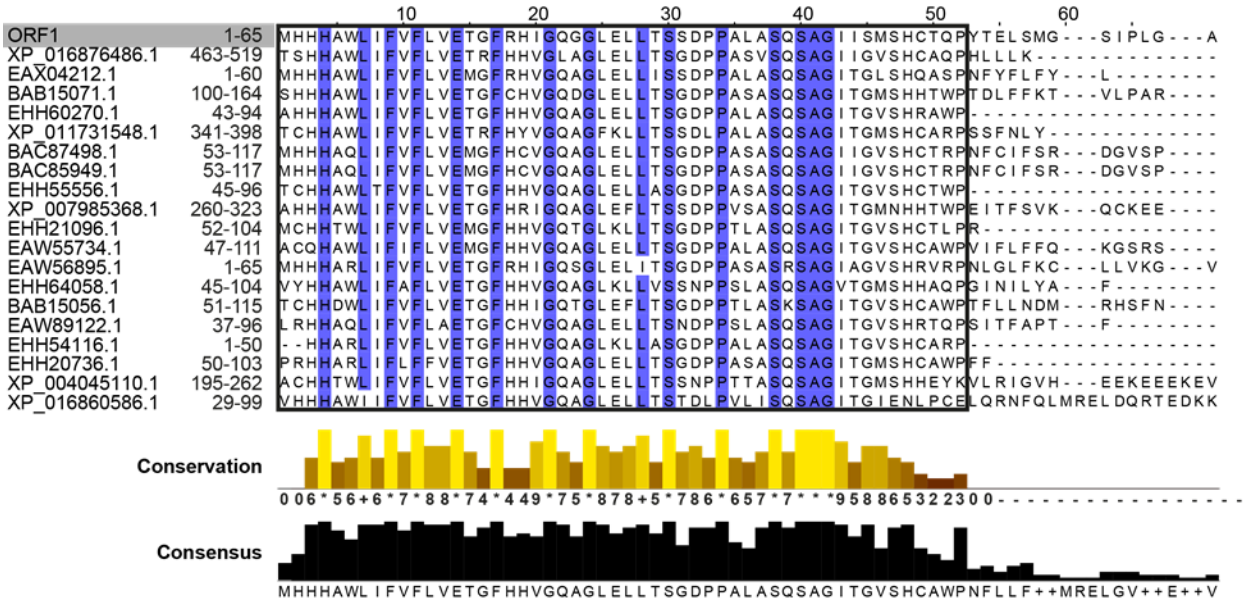


Figure 4-23 ORF1 sequence shares identity with multiple proteins

Multiple sequence alignment created with Clustal Omega of ORF1 amino acid sequence (Appendix 7.4) with proteins identified by protein BLAST (NCBI) with a minimum bit-score of 80. All cDNA sequences were retrieved from GenPept (NCBI) with the indicated accession numbers. Identical amino acid residues are highlighted in blue and below the conservation plot is a key denoting conserved sequence (*) and conservative mutations (+). Also indicated is the consensus sequence and the conserved domain of 52 aa surrounded by a black box.

With the exception of the last 44 nt, the ORF1 coding sequence overlaps with and is almost entirely encoded within the repetitive element AlusZ (**Figure 4-24A**). Alu elements are primate-specific repeats and also the most common repetitive sequences in humans comprising 11% of our genome (Deininger, 2011). Due to the abundance of Alu elements and the high number of proteins sharing similarity with ORF1 by protein BLAST exclusively found in the order of Primates (**Figure 4-23**), we examined if this is the consequence of the presence of an Alu element in these proteins. To test this hypothesis, we compared the sequence of the consensus AlusZ element (**Appendix 7-3**- sequence AlusZ) with the coding nucleotide sequence of all nineteen proteins identified by BLAST with a minimum bit-score of 80 (**Figure 4-24B**). Multiple alignments revealed that the coding sequences of all ORFs are highly similar to the consensus AlusZ sequence starting from nucleotide 128 to 283, the end of the AlusZ sequence.

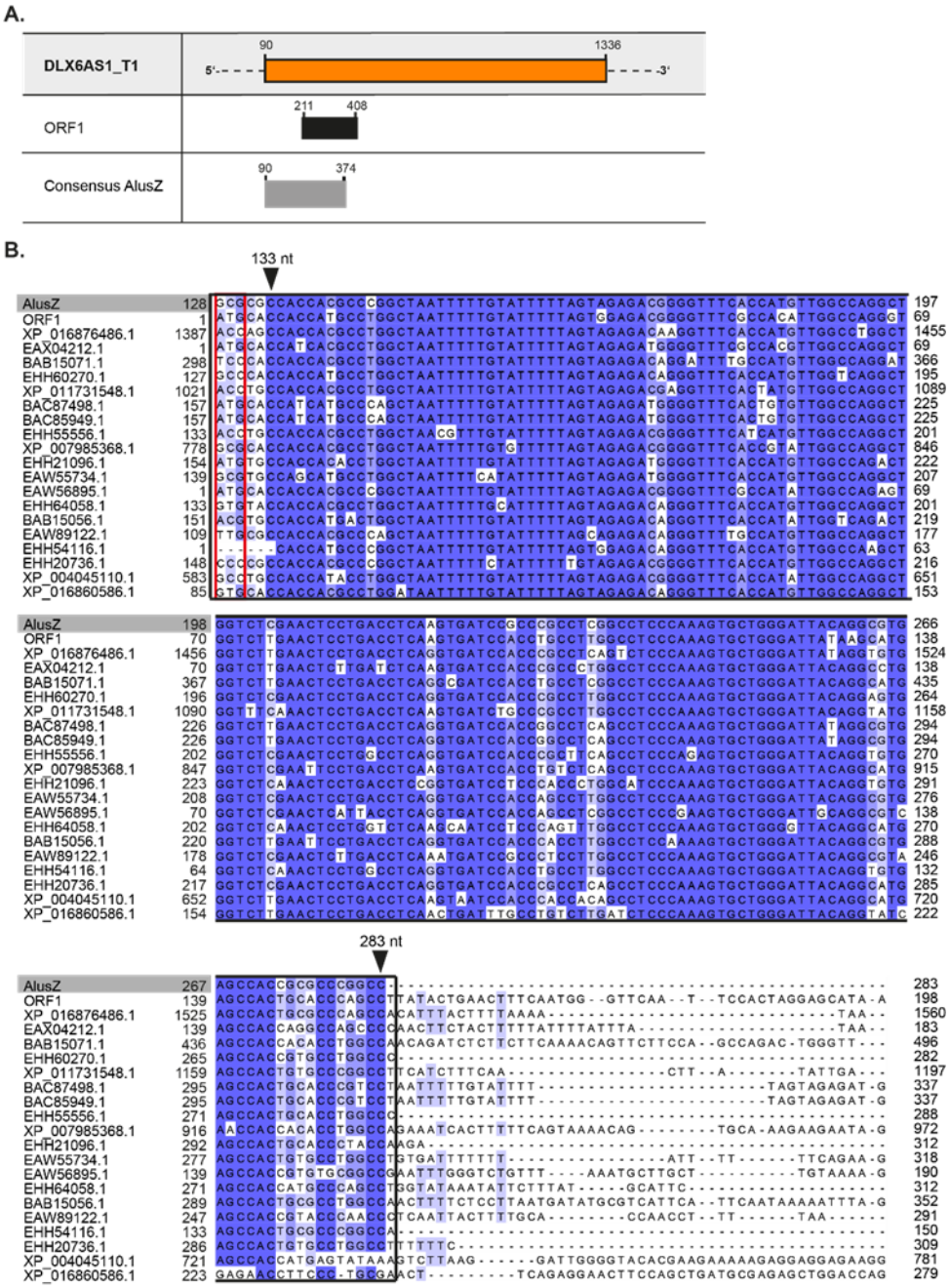


Figure 4-24 ORF1 originates from the AlusZ repetitive element

(A) Representation of the second exon of *DLX6-AS1 T1* encoding ORF1 and overlapping with the consensus AlusZ repetitive element retrieved from Repbase(Bao et al.,2015). **(B)** Multiple sequence alignments created with Clustal Omega of ORF1 coding sequence with the coding sequence of proteins identified by protein BLAST (NCBI) with a minimum bit-score of 80 (**Appendix - Table 7.8**). All coding DNA sequences were retrieved from Genbank with the accession numbers and coordinates indicated in the Appendix in **Table 7-10**. The alignment is colored according to the degree of sequence conservation ranging from 50 to 100 percent. Also indicated is the conserved domain surrounded by a black box and the AlusZ sequence derived into the ATG codon surrounded by a red box.

This sequence (with an additional nucleotide to complete the last codon) corresponds to the previously identified conserved domain of 52 aa for ORF1 and shared with the other nineteen ORFs (**Figure 4-23**). The nucleotide sequence not covered by the AlusZ element corresponds to the amino acids not conserved between ORF1 and the nineteen other ORFs in **Figure 4-23**.
The first three nucleotides of ORF1 matching with AlusZ from position 128-132 have diverged from GCG to the ATG start codon present in ORF1, EAX04212.1, BAC87498.1, BAC85949.1, EHH21096.1 and EAW56895.1 (**Figure 4-24B**).

Although ORF1 possess a high number of protein homologues, the data suggest that this is not the result of a conserved function, but the consequence of the insertion of Alu elements in the ORFs identified by protein BLAST. In addition, the majority of ORFs were annotated not based on experimental evidences, but using gene prediction models and are consequently described as either predicted, hypothetical or unnamed proteins (**Appendix Table 7-8**) and might therefore not be biologically relevant.

Nonetheless, due to a sequence identity of 32% with the characterized human extracellular protein C16orf89 (chromosome 16 Open Reading Frame 89), ORF1 protein was predicted with the LocTree3 algorithm to produce a secreted peptide with a confidence score of 50 on a 0 to 100 scale (**Figure 4-25**) (Goldberg et al., 2014). The localization of ORF1 to the extracellular compartment is in line with a predicted SignalP-4.1 signal peptide cleaved after amino acid number eighteen (Petersen et al., 2011).

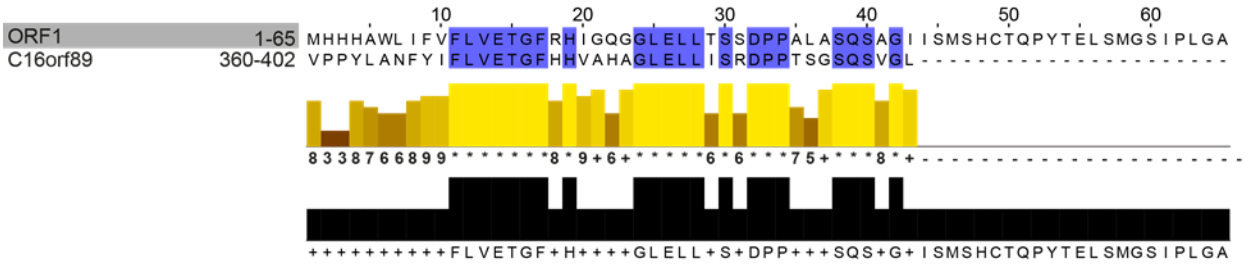


Figure 4-25 ORF1 shares identity with the extracellular protein C16orf89

Pairwise sequence alignment created with Clustal Omega of ORF1 amino acid sequence (Appendix 7.4) with C16orf89 protein identified by LocTree. C16orf89 sequence was retrieved from UniProt with the accession number Q6UX73. Identical amino acid residues are highlighted in blue and below the

conservation plot is a key denoting conserved sequence (*) and conservative mutations (+). Also indicated is the consensus sequence.

In summary, *DLX6-AS1 T1* was predicted to encode for ORF1. ORF1 shares a high level of identity with numerous predicted proteins likely due to the presence of an Alu element. It remains therefore unclear whether ORF1 is biological relevant.

4.5.2 *DLX6-AS1* does not code for a protein *in vitro*

Since coding potential for *DLX6-AS1 T1* into ORF1 was predicted based on an *in silico* approach, it is probable but not certain. To ascertain whether ORF1 is translated and thus biologically relevant we complemented the *in silico* approach with experimental methods.

A. *DLX6-AS1 T1* variant is a cytoplasmic RNA

RNA fractionation of VCaP, DU145 and HEK293T cells showed an enrichment of *DLX6-AS1* transcript *T1* in the cytoplasmic compartment (mean=66, 51 and 43% in VCaP, DU145 and HEK293T, respectively), which is higher than for the sense protein-coding gene *DLX6* (mean=48, 49 and 39% cytoplasmic enrichment in VCaP, DU145 and HEK293T, respectively) in all analyzed cell lines (**Figure 4-26**). *DLX6-AS1* transcript *T2* is instead enriched in the nuclear compartment, particularly in DU145 (mean = 74%) and HEK293T (mean=74%) cell lines. This differential localization implies different biological roles for both transcripts and is in line with the potential existence of ORF1 protein encoded in *DLX6-AS1 T1* variant.

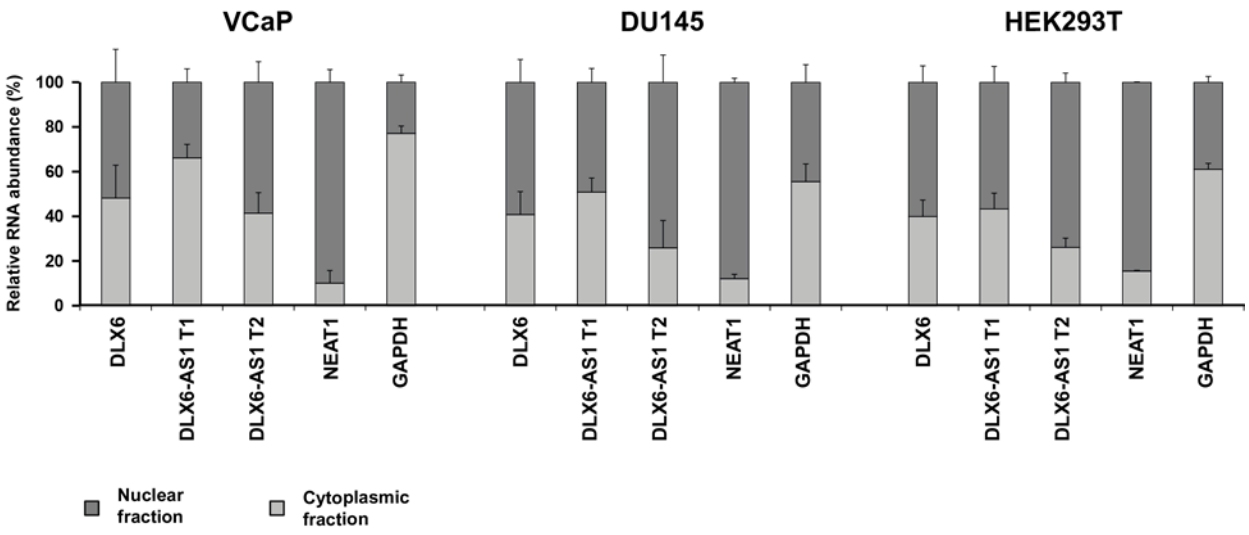


Figure 4-26 *DLX6-AS1 T1* transcript is enriched in the cytoplasmic fraction

RNA fractionation of VCaP, DU145 and HEK293T cell lysates followed by RT-qPCR analysis of *DLX6*, *DLX6-AS1 T1* and *T2* transcript variants. *NEAT1* lncRNA and *GAPDH1* transcripts served as a positive control for nuclear and cytoplasmic enrichment, respectively. Data are depicted as the mean +/- s.e.m., n= 3 biological replicates for VCaP and HEK293T and n=2 biological replicates for DU145.

B. *DLX6-AS1 T1* binds to polysomes

Since *DLX6-AS1 T1* is enriched in the cytoplasm, we investigated by polysomal fractionation whether this lncRNA is associated with polysomes. In addition, we tested if this association is impaired through immobilization of initiating ribosomes at start codons with harringtonine, thereby depleting elongating ribosomes from coding RNAs. After polysome fractionation of H1299 cells into fourteen fractions treated with harringtonine or without (Control), we measured by RT-qPCR the relative enrichment of *HPRT1*, *DLX6-AS1 T1* and *T2* in the different fractions (**Figure 4-27**). In line with a nuclear role for *DLX6-AS1 T2*, this variant could not be detected in the monosomic nor polysomic fractions of H1299 cells. Inhibition of translation initiation by harringtonine causes the accumulation of *HPRT1*-bound 80S monosomes, while decreasing the fraction of *HPRT1*-bound polysomes due to polysome run-off. Concordant with a functional role for *DLX6-AS1 T1* in the cytoplasmic compartment, analysis by RT-qPCR of all fractions demonstrates that *DLX6-AS1 T1* RNA binds to polysomes (**Figure 4-27**). This would point toward the active translation of *DLX6-AS1 T1* RNA. Treatment with harringtonine did however not shift

DLX6-AS1 T1 enrichment towards monosomic fractions such as observed for *HPRT1* (**Figure 4-27**). This result implies that *DLX6-AS1 T1* variant binds to ribosomes without being translated.

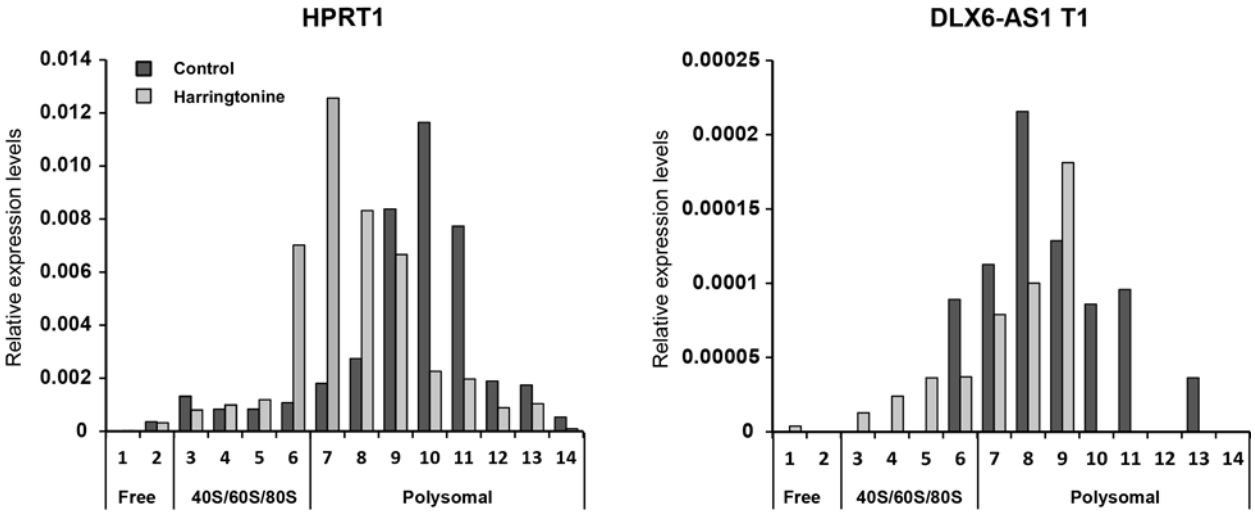


Figure 4-27 *DLX6-AS1 T1* transcript is bound to polysomes but not translated

Relative expression of *HPRT1* or *DLX6-AS1 T1* transcripts measured by RT-qPCR in all fourteen fractions obtained by polysomal fractionation of H1299 cells treated with DMSO only or supplemented with harringtonine. Data are depicted as n= 1. *(Treated cells were provided by Dr. Michael Daskalakis and mono-/polysomal fractions were provided by Dr. Johanna Schott)*

A recent study demonstrated that a single 80S monosome is sufficient for the translation of ORFs smaller than 590 nt in *Saccharomyces cerevisae* (Heyer et al., 2016). Consistent with this, the 198 nt encoding for ORF1 could be translated by a single monosome. The treatment with harringtonine would in consequence not impact the distribution of the translated transcript. We therefore employed *in vitro* translation as a direct approach to detect potential *DLX6-AS1* translation products.

C. *DLX6-AS1 T1* does not code for a protein

To determine if *DLX6-AS1* is translated as a small protein, we generated by PCR the full length cDNA for *DLX6-AS1 T1*, *T2S* and *T2L* transcripts (**Appendix 7-3**-cDNA sequences) flanked in 5' with the T7 RNA polymerase promoter. Similarly, the mouse lncRNA *2310015B20Rik* reported to

encode myoregulin (MLN), a 5kDa micropeptide (Anderson et al.,2015), along with a frameshift mutated version of MLN (MLN(FS)) were used as positive and negative controls, respectively. Both cDNA sequences were flanked by PCR with a T7 RNA polymerase promoter (**Figure 4-28A**). An equimolar amount of each PCR products was used for the coupled *in vitro* transcription and translation reaction using rabbit reticulocytes lysates in the presence of radiolabeled ^{14}C-leucine. The positive control lncRNA encoding for MLN produced a 5kDa peptide visible by autoradiography, while the PCR product with a frameshift mutation (MLN(FS)) abolished the translation. Among the three *DLX6-AS1* transcript variants, only a faint band was detected when we used *DLX6-AS1* transcript *T2S* (**Figure 4-28B**). This result is unexpected, since *T2S* has no unique sequence not shared by *T1* or *T2L* transcripts that could be translated into a peptide. The observed peptide is about 6 to7 kDa and does not match the size of the two predicted ORFs for the *T2S* variant (**Figure 4-20B**).

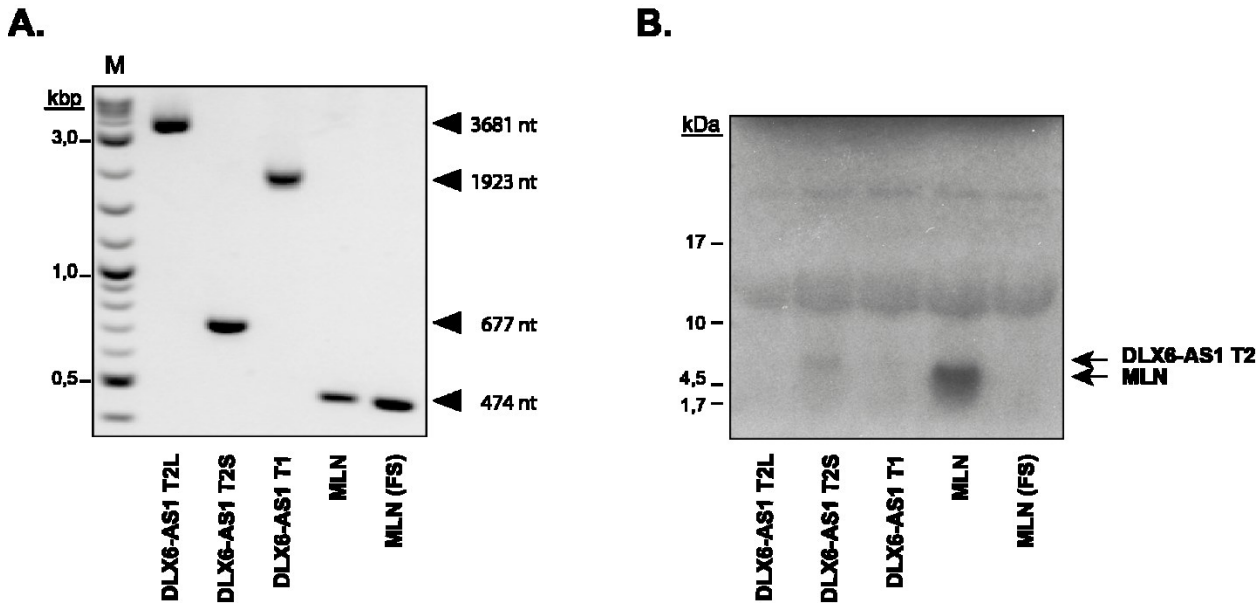


Figure 4-28 *DLX6-AS1 T1* transcript is non-coding

(A) Gel electrophoresis picture of PCR products generated from *MLN*, *MLN(FS)*, *DLX6-AS1 T1*, *T2S* and *T2L* cDNA flanked in 5' with a T7 RNA polymerase promoter. **(B)** Coupled *in vitro* transcription and translation reactions of three *DLX6-AS1* transcript variants. *MLN* and *MLN* with a frameshift mutation (*MLN(FS)*) transcripts served as positive and negative controls, respectively. ^{14}C-leucine labelled peptides were resolved on a 16,8% Tris-Tricine gel and exposed for four weeks to X-ray film. Autoradiography picture is representative of three biological replicates.

Taken together, DLX6-AS1 T1 variant is enriched in the cytoplasm and can interact with the translation machinery but does not code for proteins *in vitro*. This implies that this variant might have regulatory functions of the translational process. In contrast, DLX6-AS1 T2 transcript is enriched in the nuclear fraction and is not bound to ribosome. The differential cellular localization of both variants might be related to their specific function.

5. DISCUSSION

Increasing evidence supports the crucial role for lncRNAs in guiding or titrating away epigenetic modifiers to or from specific genomic loci. In this thesis, we employed an integrative approach to identify lncRNAs that mediate DNA (de)methylation of protein-coding gene promoters *in cis* in PCa (**section 4.1**). We identified the lncRNA *DLX6-AS1* as being overexpressed and positively correlated with *DLX6* expression in human tumor samples (**section 4.2**) and cancer cell lines (**section 4.3**), and investigated a correlation with clinical features in various human tumor entities. Modulation of *DLX6-AS1* transcript levels by knockdown or overexpression in various cell lines did neither influence *DLX6* transcription nor *DLX6* promoter DNA methylation levels (**section 4.4**). In line with our initial hypothesis, we expected a nuclear function for *DLX6-AS1*. However, the major variant of this lncRNA, *DLX6-AS1 T1*, is cytoplasmic, binds to polysomes, and does not encode for a protein (**section 4.5**). These results suggest that *DLX6-AS1 T1* may have a regulatory role at the translational level.

5.1 Identification of *DLX6/DLX6-AS1* pair

5.1.1 LncRNA screening strategy: a unique approach

Our first goal was to identify lncRNAs that regulate close-by protein-coding genes through modulation of DNA methylation patterns in PCa on a genome-wide scale. For this purpose, we employed a unique strategy based on the integration of transcriptome and DNA methylation data from PCa patients (**section 4.1**).

To date, only two studies have systematically identified lncRNAs as potential regulators of DNA (de)methylation patterns, through deep sequencing of lncRNAs that interact with DNMT1 (RIP-seq) in HCT116 (Merry et al., 2015) and HL60 cell lines (Di Ruscio et al., 2013). Although this approach identified DNMT1-interacting lncRNAs, the technique has the disadvantage of being cell and protein-specific. Moreover, RIP-seq is based on immunoprecipitation, a technique likely

to detect non-specific RNA-proteins interactions, and therefore does not recapitulate the real *in vivo* interactome (Mili et al., 2004).

Our strategy has the advantage of reflecting the true *in vivo* situation in human tissue through integration of high throughput data from PCa patients, without being limited to specific protein partners or cell lines.

5.1.2 Main limitation: reference human transcriptome and filtering for bidirectional promoters

The lncRNA transcriptome has been described to be highly tissue (Cabili et al., 2011) and cancer (Yan et al., 2015) type specific. GENCODE v17 annotations were used to both define mRNA/lncRNA pairs and map the RNA-seq data. This reference transcriptome includes only 13,333 lncRNAs, compared to the 101,700 lncRNA genes annotated so far in the NONCODE database, collected by literature mining and from multiple databases including GENCODE (Zhao et al., 2016). GENCODE annotations therefore represent an incomplete reference long non-coding transcriptome.

Ideally, transcript reconstruction through *ab initio* assembly of RNA-seq reads would allow the study of the complete set of transcripts expressed in EO-PCa samples. This strategy has been employed by Chinnaiyan and colleagues and allowed the identification of 1,859 unannotated lincRNAs in late-onset PCa (Prensner et al., 2011). Nevertheless, the RNA-seq Genome Annotation Assessment Project consortium evaluated that transcript reconstruction methods achieve at best 60% accuracy, with the additional drawback of inaccurately defining TSS and TTS (Steijger et al., 2013).

In addition, the current sequencing approach from the ICGC consortium restricted analysis to transcripts with a poly(A) tail, and therefore narrows the analysis to polyadenylated RNAs.

With these limitations, we identified nine mRNA/lncRNA pairs, but a screening approach based on an updated reference transcriptome would certainly increase the number of positive hits. Although it may be beneficial to use the Mitranscriptome assembly (Iyer et al., 2015), the fact that this transcript reconstruction has detected only 31% of RefSeq annotated splicing patterns may undermine the analysis (Zhao et al., 2016).

It is also important to note that our search is biased for *cis* coding lncRNAs co-expressed with a neighbouring protein-coding gene. Therefore the data are enriched for potentially co-regulated sense/antisense pairs by a bidirectional promoter rather than regulating each other. For this reason we included the distance filter of 1,500bp between both TSSs, which is only valid in the case of a reliable and accurate reference genome. Nevertheless, there are known examples of lncRNAs regulating the mRNAs *in cis* while the distance of their TSS is shorter than 1,500bp (Di Ruscio et al., 2013).

5.2 *DLX6/DLX6-AS1* pair is overexpressed in several cancers

5.2.1 Interdependency between *DLX6/DLX6-AS1* and *TMPRSS2:ERG* fusion in PCa

With the aforementioned screening approach we identified the *DLX6/DLX6-AS1* pair as a candidate for further clinical and functional characterization. We revealed that *DLX6* and *DLX6-AS1* transcripts are overexpressed in a subset of PCa samples and expression levels positively correlate in two independent PCa datasets. Elevated *DLX6* levels were accompanied by loss of DNA methylation in *DLX6* promoter region in both early-onset and late-onset PCa patients. In line with this result, *DLX6-AS1* expression levels were not correlated with the age at diagnosis of patients, their pathological tumor T-stage, and Gleason score. In contrast, patients with tumors harbouring the *TMPRSS2:ERG* fusion gene were highly and significantly enriched in the subset of patients characterized by high *DLX6-AS1* levels.

ICGC and TCGA cohorts were divided into “LOW” and “HIGH” expression subgroups based on *DLX6-AS1* median expression, and were associated with differential methylation levels in the

DLX6 promoter region (**Figure 4-5**). A similar result can be obtained by separating the cohort based on the presence of the *TMPRSS2:ERG* fusion-gene (**Appendix Figure 7-3 and 7-4**). Indeed, *DLX6/DLX6-AS1* transcripts are overexpressed (**Appendix Figure 7-3**) and *DLX6* promoter is hypomethylated in patients expressing the ERG fusion gene (ERG+) compared to fusion-negative (ERG-) patients (**Appendix Figure 7-4**). This result is in line with the global increase in DNA methylation as previously reported in ERG- patients (Borno et al., 2012). Additionally, preliminary analyses indicate that the expression levels of *DLX6* and *DLX6-AS1* correlate significantly with *ERG* mRNA expression in the PCa cohorts from ICGC and TCGA (Pearson's correlation with *DLX6-AS1* r=0,7131, p-value<0.0001 and with *DLX6* r=0,7586, p-value<0,0001) (**Appendix Figure 7-5**). The positive correlation between *ERG*, *DLX6* and *DLX6-AS1* and low DNA methylation levels in ERG+ samples suggests that the expression of these three genes is linked by a common mechanism.

Preliminary results suggest that neither *DLX6* nor *DLX6-AS1* are regulated by *ERG* after overexpression of this gene in *TMPRSS2:ERG* fusion-negative LNCaP cells (**Appendix Figure 7-6A**). In contrast, *DLX6* knockdown induced a significant reduction of *ERG* transcript levels in the *TMPRSS2:ERG* fusion-positive VCaP cells (**Appendix Figure 7-6C**). The promoter of the androgen regulated *TMPRSS2* gene is fused to *ERG* coding sequence in *TMPRSS2:ERG* fusion positive patients. This rearrangement leads to an androgen-dependent increase in *ERG* expression (Kumar-Sinha et al., 2008). Downregulation of *AR* mRNA expression after *DLX6* silencing (**Appendix Figure 7-6C**) suggests that *DLX6* regulates *ERG* expression by modulating *AR* transcriptional regulation. The specific effect of *DLX6* on *ERG* and *AR* signalling remains however to be evaluated by luciferase reporter assays with constructs comprising the androgen receptor promoter upstream of a reporter gene in wild-type or *DLX6* knockdown VCaP cell lines. These experiments could answer the question whether *DLX6* overexpression may facilitate *ERG* activation.

In addition to *TMPRSS2:ERG* fusion, elevated *DLX6-AS1* levels are correlated with the presence of 3p13 deletions in PCa (**Table 4-3**). The concomitant rearrangement of these two chromosomal loci was previously associated with an aggressive subset of PCa tumors characterized with early

tumor recurrence (Krohn et al., 2013). In our set of samples, higher *DLX6-AS1* expression levels were in contrast, not linked to adverse clinicopathological parameters including pathological stage, Gleason score, overall and disease-free survival (**Table 4-3** and **Figure 4-8**). The discrepancy is most likely due to the prevalence of ERG+ tumors (89.5%) in the "HIGH" expression subgroup (**Table 4-3**). Indeed, the clinical impact associated with the presence of this fusion gene is subject to controversy, but is mostly unrelated to disease-specific death and disease recurrence in a meta-analysis including 11,092 men in 30 studies (reviewed in (Xu et al., 2014)).

5.2.2 *DLX6/DLX6-AS1* overexpression is not restricted to PCa

Differential expression analysis between tumor and normal tissues conducted by the TCGA consortium identified *DLX6-AS1* as being upregulated in all analysed cancer types with the exception of breast carcinoma (Yan et al., 2015). We could confirm the overexpression of the lncRNA and its correlation to *DLX6* levels in PRAD, LUSC, HNSC, COAD, LUAD, and KIRC (**Figure 4-7A**).

The approach used by the TCGA however, fails to identify overexpressed genes in a subset of patients, as exemplified by *DLX6/DLX6-AS1* upregulation in a restricted number of breast carcinoma tissue samples from the TCGA (**Appendix Figure 7-7**). A search for outlier samples such as with the COPA (cancer outlier profile analysis) algorithm (Wang et al., 2012), would allow the exhaustive identification of tumors characterized by *DLX6-AS1* overexpression.

We also showed that *DLX6* expression is inversely correlated with *DLX6* promoter DNA methylation levels in all studied cancer types with the exception of KIRC and subsets of samples in COAD and LUAD cohorts (**Figure 4-7B**). *DLX6* expression is therefore dependent on DNA methylation in the majority of tumors. It apparently depends on alternative regulatory mechanisms in KIRC and the subset of samples in LUAD and COAD cohorts, such as copy number changes or altered expression of potential microRNA or transcription factors. To identify the

regulatory mechanism in these samples, expression levels could be integrated with whole genome sequencing data, or differentially expressed transcription factors could be investigated.

5.2.3 *DLX6/DLX6-AS1* expression influences patient prognosis

The common upregulation of this sense/antisense pair in at least six cancer types suggests that they hold a conserved function in cancer initiation and/or progression. In line with a role for *DLX6/DLX6-AS1* in carcinogenesis, elevated *DLX6* expression was reported to be linked to increased invasive potential in bone and lung metastatic cells derived from a breast cancer xenograft model with MDA-MB-231 cells (Morini et al., 2010). In addition, we showed that elevated levels of the *DLX6-AS1/DLX6* pair could predict patient's prognosis in LUSC, KIRC and LUAD. Indeed, high *DLX6-AS1* expression was associated with favourable survival in LUSC and KIRC, whereas elevated *DLX6* levels were linked to bad prognosis in LUAD (**Figure 4-8A**). Similarly, high expression levels of the lncRNA in neuroblastoma tissue were reported to be associated with bad overall survival (Olsson et al., 2016). The dual behaviour of *DLX6-AS1* on patient's prognosis seems to be dependent on the tissue type. Variations in the type of transcript variant being expressed and genetic aberrations affecting either the lncRNA or components crucial for its function are inherent to a specific tumor type. The presence of such aberrations may alter the function of the lncRNA and in consequence influence patient prognosis.

5.3 Functional role of *DLX6-AS1* overexpression in cancer

5.3.1 *DLX6-AS1* (*Evf2*) regulates *DLX5/6* transcription in mouse

Although *DLX6/DLX6-AS1* upregulation seems to play an important role during carcinogenesis; nonetheless, the nature of its function in cancer etiology is largely unexplored. The role of *DLX6/DLX6-AS1* in normal development is, on the other hand, better understood. Homozygous deletion of the locus comprising *DLX5*, *DLX6* and *Evf2* (mouse homologue of *DLX6-AS1* also known

as *Dlx6os1*) is associated with craniofacial and skeletal abnormalities in mouse embryo, leading to perinatal lethality (Robledo et al., 2002).

Transcriptional arrest of *Evf2* through knock-in of a polyA termination signal downstream of the *Evf2* TSS results in viable and fertile mice, but is coupled with a drastic reduction of the GABAergic interneurons in the forebrain, suggesting a role for this lncRNA in brain development (Bond et al.,2009). Detailed analysis of the *Evf2* mutant mouse forebrain at embryonic day 13.5 revealed an increased expression of *DLX6* and *DLX5* transcripts in response to *Evf2* transcriptional repression (Bond et al., 2009). In addition, *Evf2* repression was associated with increased DNA methylation levels at two CpG sites located within or near *DLX5/-6* enhancer (named ei enhancer) (Berghoff et al., 2013). Rescue of the *Evf2* mutant phenotype through expression of the lncRNA *in trans*, recapitulated *DLX5* transcript (Bond et al., 2009) and enhancer DNA methylation levels (Berghoff et al.,2013), while *DLX6* expression remained unchanged (Bond et al., 2009). Therefore, the authors suggested that *Evf2* regulates *DLX6 in cis* by the simple act of transcription, whereas *DLX5* and DNA methylation is regulated by *Evf2* RNA product *in trans* (Bond et al., 2009, Berghoff et al., 2013).

5.3.2 *DLX6-AS1* does not regulate *DLX5/6* transcription and *DLX6* promoter methylation in human cell lines

In contrast to *Evf2*, manipulation of *DLX6-AS1* expression did not affect local transcription and DNA methylation levels. To study the impact of *DLX6-AS1 in trans*, we employed shRNAs, siRNAs, LNAs and RNA transfections. We demonstrated that all the different strategies did neither affect *DLX6* (**Figure 4-16 and 4-18**) nor *DLX5* transcript levels (**Appendix – Figure 7-8**). In addition, DNA methylation levels of the differentially methylated *DLX6* promoter (**Figure 4-17 and 4-19**) as well the ei enhancer region (**Appendix - Figure 7-9**) remained unchanged. However, Kohtz and colleagues supported a model where *Evf2* can act through *trans* and *cis* mechanisms (Bond et al., 2009, Berghoff et al., 2013). To interrogate the *cis* regulatory mode of action of *DLX6-AS1*, alteration of the lncRNA transcription rather than solely modulating its RNA levels is required. Genomic engineering tools can allow insertion of transcriptional terminator sequences or

deletion of the entire lncRNA loci by TALE nucleases (transcription activator-like effector) (Boch et al., 2009) or the CRISPR clustered regularly interspersed palindromic repeats)/Cas9 (CRISPR-associated protein 9) systems (Jinek et al., 2012). Alternatively, an endonuclease-deficient Cas9 can allow the recruitment of transcriptional activators or repressors to defined loci (Gilbert et al., 2013, Cheng et al., 2013). These techniques are, however, not suited for the study of *DLX6-AS1* since its promoter proximal region intersects with *DLX6*. Insertion of exogenous sequences as well as the targeting of programmable nucleases to *DLX6-AS1* TSS would inevitably perturb *DLX6* transcription (Goyal et al., 2016).

Although we were hampered by the complexity of the genomic locus to study the *in cis* effect of *DLX6-AS1* on *DLX6*, all evidence currently point towards the co-regulation of both transcripts. Indeed, 5'RACE demonstrated that *DLX6-AS1* TSS is localized within the first intron of *DLX6* (**Figure 4-10**). Consequently, the region that separates the transcription start sites of both sense and antisense transcripts is small enough to be defined as a bidirectional promoter (TSS distance 811nt). This assumption is supported by the positive correlation between *DLX6* and *DLX6-AS1* as well as a single RNA polymerase II peak in VCaP compared to LNCaP cells around the *DLX6* promoter (**Figure 4-15**). The existence of a bidirectional promoter could be confirmed through probing the promoter activity by luciferase assay of the fragment separating both TSSs in sense and antisense direction.

5.3.3 Alternative transcripts, alternative functions

Alignments of CAGE-seq data obtained from 975 human and 399 mouse tissues, primary cells, and cell lines, revealed that only 57% of human TSSs are conserved in the mouse genome (Consortium et al., 2014). In line with this result, only the second exon of *Evf2*, which correspond to the region overlapping with ei enhancer, is conserved between human, mouse, rat, zebrafish, and chicken (Feng et al., 2006). Serial deletion of *Evf2* transcript revealed that the region including the second exon (nucleotide 117 to 395) is essential for *DLX5/-6* ei enhancer activation (Feng et al., 2006).

Through the *in silico* screening approach (**section 4.1.1**) we identified the human *DLX6-AS1* transcript ENST00000430027.3. Experimental investigation of the dominant *DLX6-AS1* transcript expressed in PCa tissue and in immortalized cell lines revealed, however, two *DLX6-AS1* splice variants *T1* and *T2,* generated by alternate usage of the second exon. *DLX6-AS1 T1* and *T2* differ from ENST0000430027.3 and *Evf2* mouse transcript by an alternative TSS, and consequently first exon. Indeed, *Evf2* is transcribed across the ei enhancer element located between *DLX6* and *DLX5*, while *DLX6-AS1 T1/T2* TSSs overlap with the *DLX6* first exon. In consequence, *DLX6-AS1 T1/T2* transcript variants do not contain the evolutionary conserved and functional second exon of *Evf2* associated with ei enhancer activation and gene repression.

The structural difference between *Evf2* and *DLX6-AS1* might underlie their contrasting result on neighbouring gene regulation. Alternative TSS usage might produce transcripts with different functional roles. For example in human cells, *PTENpg1as*, a pseudogene-derived antisense transcript, is alternatively transcribed into α and β variants harbouring distinct functions. *PTENpg1as* α was shown to operate *in trans* by recruiting PRC2 and DNMT3 to repress *PTEN* promoter. Conversely, the variant β duplexes with *PTENpg1* sense transcript thanks to their overlapping and complementary sequences. The formation of this RNA-RNA duplex enables *PTENpg1* to exert its function as a microRNA sponge by ensuring its stability (Johnsson et al., 2013).

5.3.4 Does *DLX6-AS1* function by enhancing *DLX6* translation?

The formation of an RNA-RNA duplex is a common phenomenon among mRNA/lncRNA pairs, such as *UCHL1* (ubiquitin carboxyl-terminal hydrolase L1)/*AS UCHL1* (antisense UCHL1) in mice (Carrieri et al., 2012) and *PPP1R12A* (protein phosphatase 1 regulatory subunit 12A)/ *R12A-AS1* in humans (Schein et al., 2016). These lncRNAs are, similar to *DLX6-AS1*, co-regulated and share an overlapping region with their anti-sense protein-coding gene. Through heterodimer formation with their respective mRNA, *UCHL1-AS* and *R12A-AS1* lncRNAs stimulate the protein synthesis of their sense counterpart without affecting their mRNA levels (Carrieri et al., 2012; Schein et al., 2016). Protein synthesis activation is dependent on the presence of a SINE B2 or Alu element

embedded within the non-overlapping region of *UCHL1-AS* (Carrieri et al., 2012) or *R12-AS1* (Schein et al., 2016), respectively. Similar RNA features could be found in 31 additional mouse (Carrieri et al., 2012) and 129 human (Schein et al., 2016) antisense lncRNAs, termed the SINEUPs (SINE element-containing translation UP-regulators) family. Design of synthetic SINEUPs that comprise of an effector and a binding domain targeting an mRNA of interest was shown to successfully enhance translation of the target (Carrieri et al., 2012) (Zucchelli et al., 2015).

Similarly to *UCHL1-AS* and *R12A-AS1*, *DLX6-AS1* depletion did not affect *DLX6* mRNA levels (**Figure 4-16**), while DLX6 protein levels could not be examined due to the limited specificity of available antibodies raised against DLX6 protein.

In contrast to *DLX6-AS1 T2*, the *T1* variant comprises the two essential SINEUPs RNA features. Indeed, both *DLX6-AS1* splice variants share a 55-nucleotide overlapping region with the first exon of *DLX6*, while only the second exon of the *T1* variant embeds an AlusZ repeat. Dimerization of both transcripts would explain the enrichment of the *DLX6-AS1* transcript *T1* in the cytoplasmic compartment (**Figure 4-26**), where this variant could influence *DLX6* translation. This in turn might explain polysomes binding of *DLX6-AS1 T1* lncRNA, although *DLX6-AS1 T1* is non-coding. The absence of ribosome binding of the nuclear *DLX6-AS1 T2* transcript allows rule out a role for this RNA on translation regulation. With the help of the overlapping region shared with *DLX6* RNA, *DLX6-AS1 T2* could, similar to *PTENpg1as β*, ensure the stability of its partner by escaping ribonuclease, or miRNA-mediated degradation.

All together, we envision that *DLX6-AS1 T1* does not regulate *DLX6* at the RNA level, but so far, we do not exclude a role in post-transcriptional regulation. Concomitant expression of *DLX6* and *DLX6-AS1 T1* RNAs could favour their dimerization and subsequent *DLX6* translational regulation.

5.3.5 Does *DLX6-AS1* function *in trans*?

Based on the co-expression pattern of *DLX6* and *DLX6-AS1* RNAs we suggested in this study that both transcripts serve a common and complementary function. We therefore focused our attention on the *cis* and local action of the lncRNA on its sense protein counterpart. The majority of lncRNAs were, however shown to exert their function on more distant genomic locations (Guttman et al., 2011).

DLX6-AS1 T2 could act *in cis* by stabilizing *DLX6* through heterodimer formation. But the regulation of more distant loci through, for instance, recruitment of chromatin modifiers cannot be excluded. In turn, the presence of an Alu element within *T1* RNA sequence points towards the regulation of genes other than *DLX6* by alternative mechanisms. Indeed, Alu elements within cytoplasmic lncRNAs can duplex with complementary Alu elements located within the 3'-untranslated region of an mRNA. The double-stranded RNA binding protein STAU1 (Staufen 1) recognizes and binds to the formed duplex before initiating mRNA decay (Gong et al., 2011). Alternatively, Alu/Alu paring between *5S-OT* (5S rRNA overlapped transcript) lncRNA and a subset of mRNAs was recently reported to create a binding site for splicing factors. The formation of such heteroduplexes was shown to regulate the splicing of the respective mRNA partner *in trans* (Hu et al., 2016). This mechanism was, however, shown to be restricted to the nuclear compartment, and could therefore only involve *DLX6-AS1 T1* before its export to the cytoplasm.

Differences in the RNA sequence and localization seems to underlie a different functional role for the two *DLX6-AS1* variants. Based on their respective features, both RNAs may work *in cis* or *trans* by involving different regulatory pathways. To infer the function of each transcript variant we could envision identifying the protein partners of biotin labelled DLX6-AS1 T1 or T2 variants by mass spectroscopy (Feng et al., 2014). Alternatively, a bioinformatics approach termed guilt-by-association employs transcriptome data to infer genes and pathways co-expressed with a given lncRNA. Based on the known function of the coregulated protein-coding genes this method allows to infer the function of a lncRNA (Guttman et al., 2009). This strategy could be employed

on the RNA-sequencing datasets employed in this study to infer the function of *DLX6-AS1* and potential genes regulated by this lncRNA. Nevertheless, to distinguish the function of both splice variants such approach should be applied to RNA-seq data generated after modulation of the expression of a single variant.

6. CONCLUSION AND FUTURE ASPECTS

Our initial goal was to i) identify lncRNAs that potentially regulate a close-by protein-coding gene by DNA (de)methylation mechanism in PCa and ii) unravel the regulatory mechanism of action of the most promising lncRNA candidate.

Through an *in silico* screening approach, we identified the *DLX6/DLX6-AS1* pair. We showed that the expression of both RNA species is correlated and that the DLX6 promoter is differentially methylated in tumor compared to normal tissue in two independent prostate cancer datasets. *DLX6/DLX6-AS1* overexpression and differential *DLX6* promoter methylation is a common phenomenon in cancer, and predicts good or bad patient prognosis in function of the tumor type. The overexpression of both transcripts in PCa correlates with the presence of *TMPRSS2:ERG* fusion gene, and preliminary results suggest that *DLX6* modulates *ERG* levels by influencing androgen receptor signaling.

Under certain environmental conditions our genome creates multiple versions of a gene transcript through the process of alternative splicing. The mouse homologue of *DLX6-AS1* (*Evf2*) was shown to regulate *DLX6* and *DLX5* expression by influencing ei enhancer DNA methylation in the developing mouse brain. Through alternative transcriptional start site usage, *DLX6-AS1 T1* and *T2* variants are expressed in human cancerous cells. In contrast with *Evf2*, the human *DLX6-AS1* splice variants do not regulate *DLX6* or *DLX5* transcription by mediating local DNA methylation changes. The proximity of the coding and non-coding transcript start sites in humans, suggests however, a coordinated regulation of *DLX6/DLX6-AS1* expression through differential methylation of their common promoter.

The exclusion of the exon overlapping with ei enhancer in human *DLX6-AS1* and the inclusion of an Alu element in the *T1* variant, participated in the diversification of the lncRNA biological function. This is supported by the differential cellular localization and ribosome binding of *DLX6-AS1* transcript variants. Based on their respective features, both RNAs may work *in cis* or *trans* by involving different regulatory pathways. To infer the function of each transcript variant we could

envision the direct identification of *DLX6-AS1 T1* and *T2* protein partners by mass spectrometry or infer the function of the lncRNA through the guilt by association approach.

In PCa we showed that *T1* and *T2* are overexpressed, where *T1* transcript predominated. Differences in patient's prognosis in LUAD and LUSC might reflect the expression of different splice variants or the disruption of the balance between *T1* and *T2* transcripts. Detailed investigation of the splice variants expressed in LUAD and LUSC based on TCGA raw RNA-seq data could reveal which transcript is expressed and linked to favourable and worse patient's prognosis.

7. APPENDIX

7.1 Oligonucleotide sequences

Table 7-1 shRNA template oligonucleotides

Name	Primer	Sequence 5'-3'	Target
DLX6 sh#2	FORWARD	ACCGGGCTTCCTTAGGATTGACATAAGTTAATATTCATAGCTTGTGTCAGTCCTAAGGAAGCTTTT	DLX6 SECOND EXON
	REVERSE	CGAAAAAAGCTTCCTTAGGACTGACACAAGCTATGAATATTAACTTATGTCAATCCTAAGGAAGCC	
DLX6 sh#6	FORWARD	ACCGAAGAATCTGCACAAACTTGGCGTTAATATTCATAGCGCCAAGTTTGTGCAGATTCTTTTTT	DLX6 3'UTR
	REVERSE	CGAAAAAAAAGAATCTGCACAAACTTGGCGCTATGAATATTAACGCCAAGTTTGTGCAGATTCTT	
AS sh#4	FORWARD	ACCGGGGTCAGATTAAACACAAAGTTAATATTCATAGCTTTGTGTTTAATCTGACCCTTTT	DLX6-AS1 THIRD EXON
	REVERSE	CGAAAAAAGGGTCAGATTAAACACAAAGCTATGAATATTAACTTTGTGTTTAATCTGACCC	
AS sh#5	FORWARD	ACCGGGGTTGGAAGTAATGATTTGGTTAATATTCATAGCCAAATCATTACTTCCAACCCTTTT	DLX6-AS1 THIRD EXON
	REVERSE	CGAAAAAAGGGTTGGAAGTAATGATTTGGCTATGAATATTAACCAAATCATTACTTCCAACCC	
LUC5	FORWARD	ACCGGATCACAGAATCGTCGTATGTAGTTAATATTCATAGCTGCATACGACGATTCTGTGATTTTT	N/A
	REVERSE	CGAAAAAAATCACAGAATCGTCGTATGCAGCTATGAATATTAACTACATACGACGATTCTGTGATC	

Table 7-2 siRNA and LNA sequences

ID	RNA target	Name	Sequence 5'-3'	Manufacturer
si NT1	none	Lincode Non-targeting siRNA #1 5nmol	UGGUUUACAUGUCGACUAA	GE Dharmacon
si NT2	none	Lincode Non-targeting siRNA #2 5nmol	UGGUUUACAUGUUGUGUGA	GE Dharmacon
AS si #1	DLX6-AS1	Lincode Human DLX6-AS1 siRNA	GCUCACUCAACCAAGAAUA	GE Dharmacon
AS si #2	DLX6-AS1	Lincode Human DLX6-AS1 siRNA	GCUAGAUUGUUUAUGACCA	GE Dharmacon
AS si #3	DLX6-AS1	Lincode Human DLX6-AS1 siRNA	ACAUGUCAAUAGUAGGCUA	GE Dharmacon
AS si #4	DLX6-AS1	Lincode Human DLX6-AS1 siRNA	GCGUAGGAGCUUCAAAAUU	GE Dharmacon
si NT	none	ON-TARGETplus Non-Targeting pool	UGGUUUACAUGUCGACUAA UGGUUUACAUGUUGUGUGA	GE Dharmacon

			UGGUUUACAUGUUUUCUGA UGGUUUACAUGUUUUCCUA	
si DLX6 #pool	DLX6	ON-TARGET plus Human DLX6 siRNA	CGAACUGGCAGCUUCCUUA GGAAAUCAGGUUCAAUGGA UACGAUGGCUGACGGCUUG GGACGACACAGAUCAACAA	GE Dharmacon
LNA #2	DLX6-AS1	LNA™ longRNA GapmerR, in vitro standard, 5 nmol	GTGAATGCATGAGAGT	Exiqon
LNA #4	DLX6-AS1	LNA™ longRNA GapmerR, in vitro standard, 5 nmol	ACTTTACTAGCCTCAT	Exiqon
Negative **Control A**	none	LNA™ longRNA GapmerR, negative control A, in vitro standard, 5 nmol	AACACGTCTATACGC	Exiqon

Table 7-3 Primers used for RT-qPCR analysis

cDNA Target	Forward primer (5'-3')	Reverse primer (5'-3')	UPL probe
DLX6	CATGCCTGGCTATTCTCACTG	ACGTTTCCCTAGGGTGTTCC	55
DLX6-AS1_T1	TGATTCCTGTATGTATGGCAGCTA	GGTTTTCCTTTGTCTCAGCAAT	63
DLX6-AS1_T2	TGCTGTTGTGGTAGGACTGG	TGGGAAGAATTACAGGAAAAGG	25
HPRT1	TGACCTTGATTTATTTTGCATACC	CGAGCAAGACGTTCAGTCCT	73
ALAS1	CAGTAATGACTACCTAGGAATGAGTCG	CCATGTTGTTTCAAAGTGTCCA	43
SDHA	TCCACTACATGACGGAGCAG	TCCACTACATGACGGAGCAG	70
ERG	AACGAGCGCAGAGTTATCGT	CGTCTGGAAGGCCATATTCT	19
Rabbit β-globin	GAAGGCTCATGGCAAGAAGG	ATGATGAGACAGCACAATAACCAG	N/A
GAPDH	AGCCACATCGCTCAGACAC	GCCCAATACGACCAAATCC	60
NEAT1	AGGCCTGGTCTTGTGGAAC	AGCGCCAAACCTAGAGAAAA	87

Table 7-4 Primers used for the 5'RACE assay

Primer Name	Primer Sequence (5'-3')
GSP1	TAGCTGCCATACATACAGGAATCA
GSP2	CCTTTTGGTCTATCTGCATGG
GSP3	CCAAGGATTTCCCTTTCCAT
Q0	CCAGTGAGCAGAGTGACG
Q1	GAGGACTCGAGCTCAAGC
QT	CCAGTGAGCAGAGTGACGAGGACTCGAGCTCAAGCTTTTTTTTTTTTTTTT

Table 7-5 Primers used for PCR amplification of bisulfite treated DNA

ID	Target region	Forward primer (5'-3')	Reverse primer (5'-3')	Amplicon coordinates (hg19)
DLX6 Region 1	DLX6 promoter	AGGAAGAGAGGTGTAGTAGGATTAGAGTGGTAG	CAGTAATACGACTCACTATAGGGAGAAGGCTCTTTCCCCCCTTAAATTTTTTAAAAAC	chr7 :97005141 -97005668
DLX6 Region 2	DLX6 promoter	AGGAAGAGAGTTGTTGTTGTGGTAGGATTGGAGGTAAGGGTT	CAGTAATACGACTCACTATAGGGAGAAGGCTACCCCCAAAATTTTTAATAATAACC	chr7 :97005961 -97006381
DLX6 Region 3	DLX6 promoter	AGGAAGAGAGGTTTTTATTTTTAGTTTTATTATAATAGTAG	CAGTAATACGACTCACTATAGGGAGAAGGCTCAAAATAAACCACTACCTACCCAA	chr7 :97006350 -97006850
DLX6 Region 4	DLX6 promoter	AGGAAGAGAGTTGGGTAGGTAGTGGTTTATTTTG	CAGTAATACGACTCACTATAGGGAGAAGGCTCAAAACACAACCAACAACTC	chr7 :97006826 -97007139
DLX6 Region 5	Enhancer ei	AGGAAGAGAGTTTTTTATTGTGAAATTTTGGGTT	CAGTAATACGACTCACTATAGGGAGAAGGCTAACAAAACCCCACTACCAACTATAC	chr7 :96641276 -96641698
DLX6 Region 6	Enhancer ei + extension	AGGAAGAGAGTGGAAGAGGTTGTAGAAGTATAGTTGG	CAGTAATACGACTCACTATAGGGAGAAGGCTACCCTTAAAATTCAAAAAATCCAAA	chr7 :96641661 -96641987
DLX6 Region 7	Extension enhancer ei	AGGAAGAGAGGGTTAATATTTTTATTTTTTAGAATG	CAGTAATACGACTCACTATAGGGAGAAGGCTCTATAAATAACTCTATTTCCTAAC	chr7 :96641985 -96642432

In blue is highlighted the T7 RNA polymerase promoter, in green additional nucleotides as a mass tag and in black the sequence complementary to the target DNA.

Table 7-6 Primers used for DLX6-AS1 and MLN cDNA cloning and PCR amplification

Target cDNA	Primer	Primer Name	Primer Sequence (5'-3')
DLX6-AS1 T1	FORWARD	DLX6AS1_T1/T2_F1_T7_MluI	**CCG**ACGCGTTAATACGACTCACTATAGGGGGCGGGAGAAGCGAGCTG
	REVERSE	DLX6AS1_T1_R1_T7_HindIII	**CCC**AAGCTTTACATTCAACCTGGATTCAAAGATGTC
DLX6-AS1 T2L	FORWARD	DLX6AS1_T1/T2_F1_T7_MluI	**CCG**ACGCGTTAATACGACTCACTATAGGGGGCGGGAGAAGCGAGCTG
	REVERSE	DLX6AS1_T2_R3_T7_HindIII	**CCC**AAGCTTTTGTTCAATCTTTATTAGAGGAAAGG
DLX6-AS1 T2S	FORWARD	DLX6AS1_T1/T2_F1_T7_MluI	**CCG**ACGCGTTAATACGACTCACTATAGGGGGCGGGAGAAGCGAGCTG
	REVERSE	DLX6AS1_T1_R1_T7_HindIII	**CCC**AAGCTTTACATTCAACCTGGATTCAAAGATGTC
2310015B20RiK	FORWARD	2310015B20RiK_T7_F	**GCTGAAAT**TAATACGACTCACTATAGGGGAGCTTTTCGTCCATGGAGA
	REVERSE	2310015B20RiK_R	TCAGATTAAAATGTAGATCTTTATTTGC

In blue is highlighted the T7 RNA polymerase promoter, in violet the respective restriction sites for MluI or HindIII and in bold are the additional nucleotides allowing not only restriction digest close to the end of the PCR product, but ensures as well efficient transcription.

Table 7-7 Primers used for Sanger sequencing

Target	Primer Sequence (5'-3')
pRS	CAAGGCTGTTAGAGAGATAATTG
M13	CAGGAAACAGCTATGACC
T7	TAATACGACTCACTATAGGG

7.2 Supplementary results

7.2.1 Supplementary figures

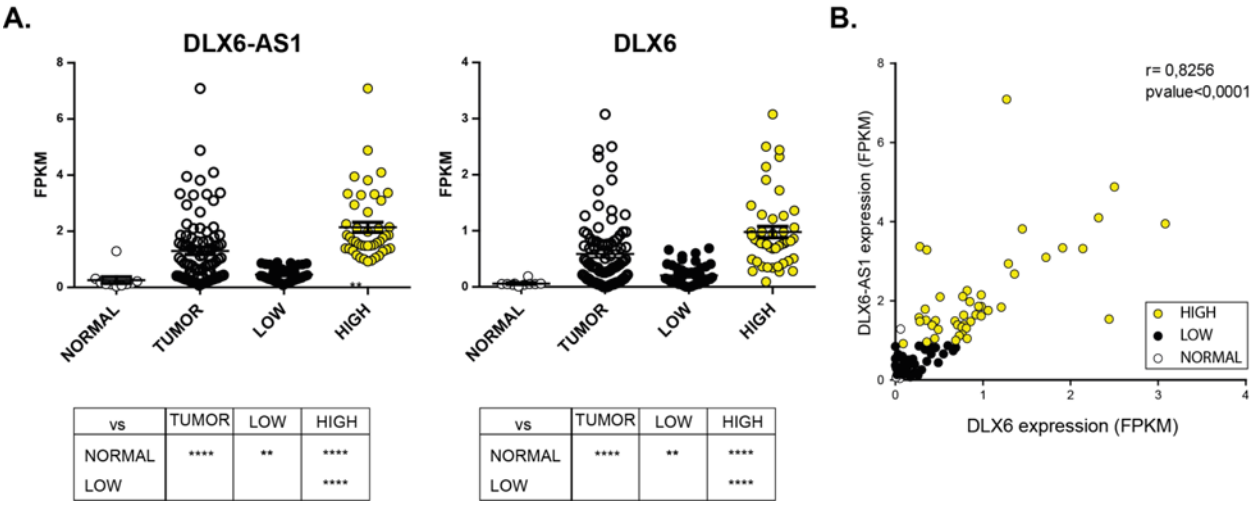


Figure 7-1 Coordinated upregulation of *DLX6* and *DLX6-AS1* in ICGC EO-PCA cohort

(A) *DLX6* and *DLX6-AS1* RNA expression levels measured by RNA sequencing in all tumors as well as separated into LOW and HIGH tumor subgroups compared to normal tissue (NORMAL=10 ,LOW=46 and HIGH=45) from ICGC EO-PCa cohort. Patients were divided into "LOW" and "HIGH" expression groups relative to the median expression value of *DLX6-AS1*. Mann-Whitney U test **** =p<0.0001, **=p<0.01. **(B)** Corresponding Spearman correlation plot of *DLX6-AS1* and *DLX6* expression measured by RNA sequencing in the ICGC EO-PCa cohort.

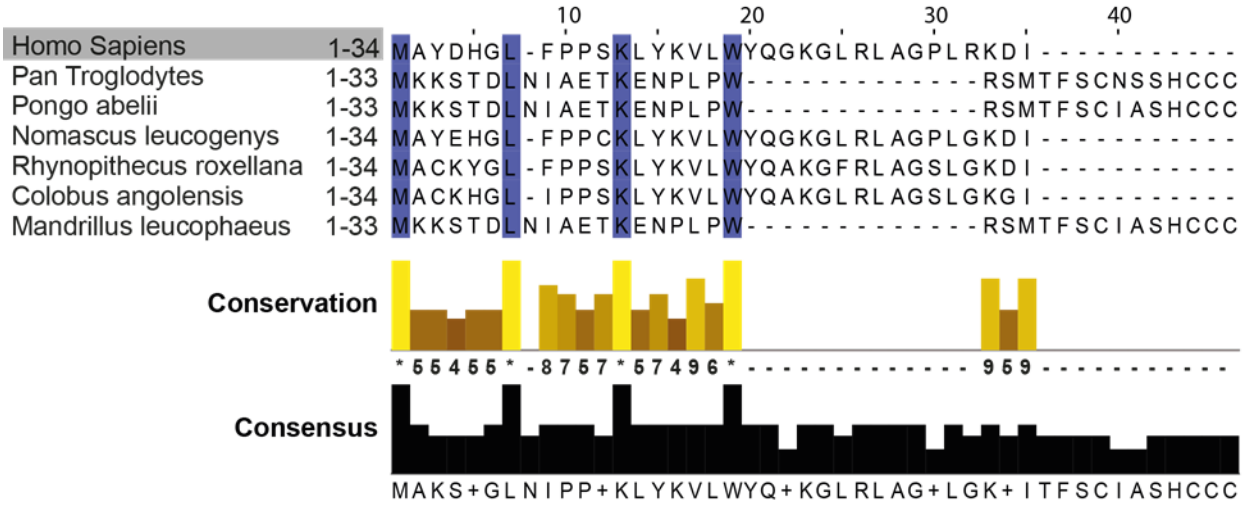


Figure 7-2 *DLX6-AS1* ORF2 is not conserved in primates

Multiple sequence alignment created with Mafft version 7 of ORF2 homologs predicted by ORFinder (NCBI) in *DLX6-AS1 T2L* homologs found by nucleotide BLAST (NCBI) from seven primate species. All the cDNA sequences were retrieved from Genbank with the following accession numbers; *Pan troglodytes* (LOC104007329), *Pongo abelii* (LOC103891166), *Nomascus leucogenys* (LOC100579968), *Rhynopithecus roxellana* (LOC104662289), *Colobus angolensis* (LOC105512765) and *Mandrillus leucophaeus* (LOC105535579). Sequence identity across the seven species is indicated in dark blue and the number of sequences with identical amino acids residues are indicated in the conservation plot.

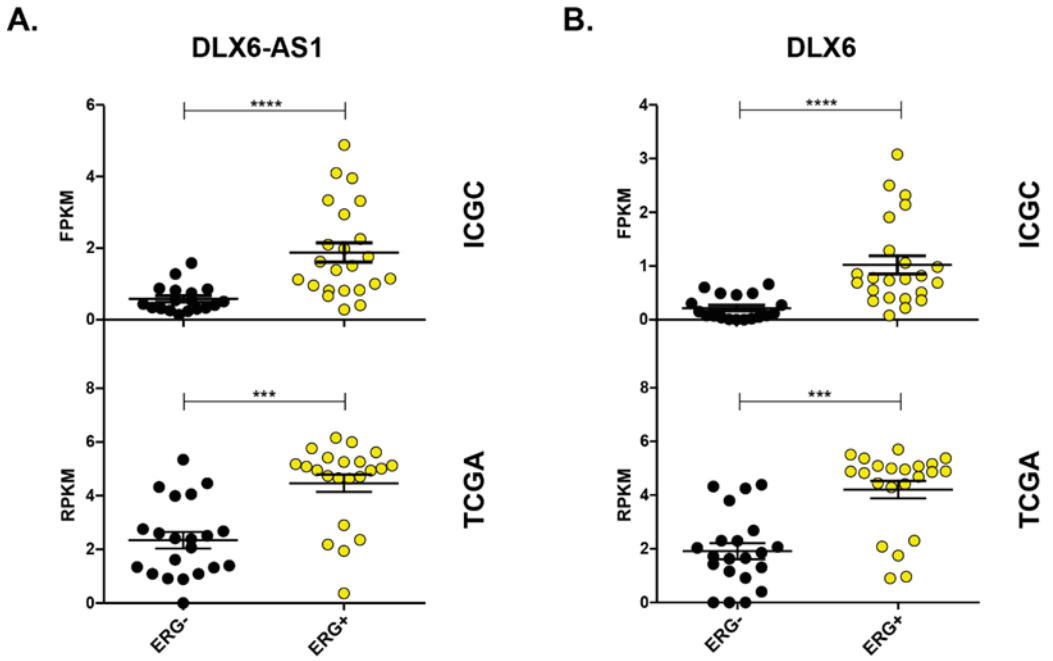


Figure 7-3 *DLX6/DLX6-AS1* pair is overexpressed in ERG fusion-positive PCa

DLX6 **(A)** and *DLX6-AS1* **(B)** RNA expression levels measured by RNA sequencing in ERG- tumors compared to ERG+ tumors from ICGC EO-PCa (ERG-=18, ERG+=23) and TCGA (ERG-=21, ERG+=22) cohort. Mann-Whitney U test (ICGC) or Wilcoxon signed-rank (TCGA) tests were used to calculate statistical significances between each group. ****p<0.0001, ***p<0.001.

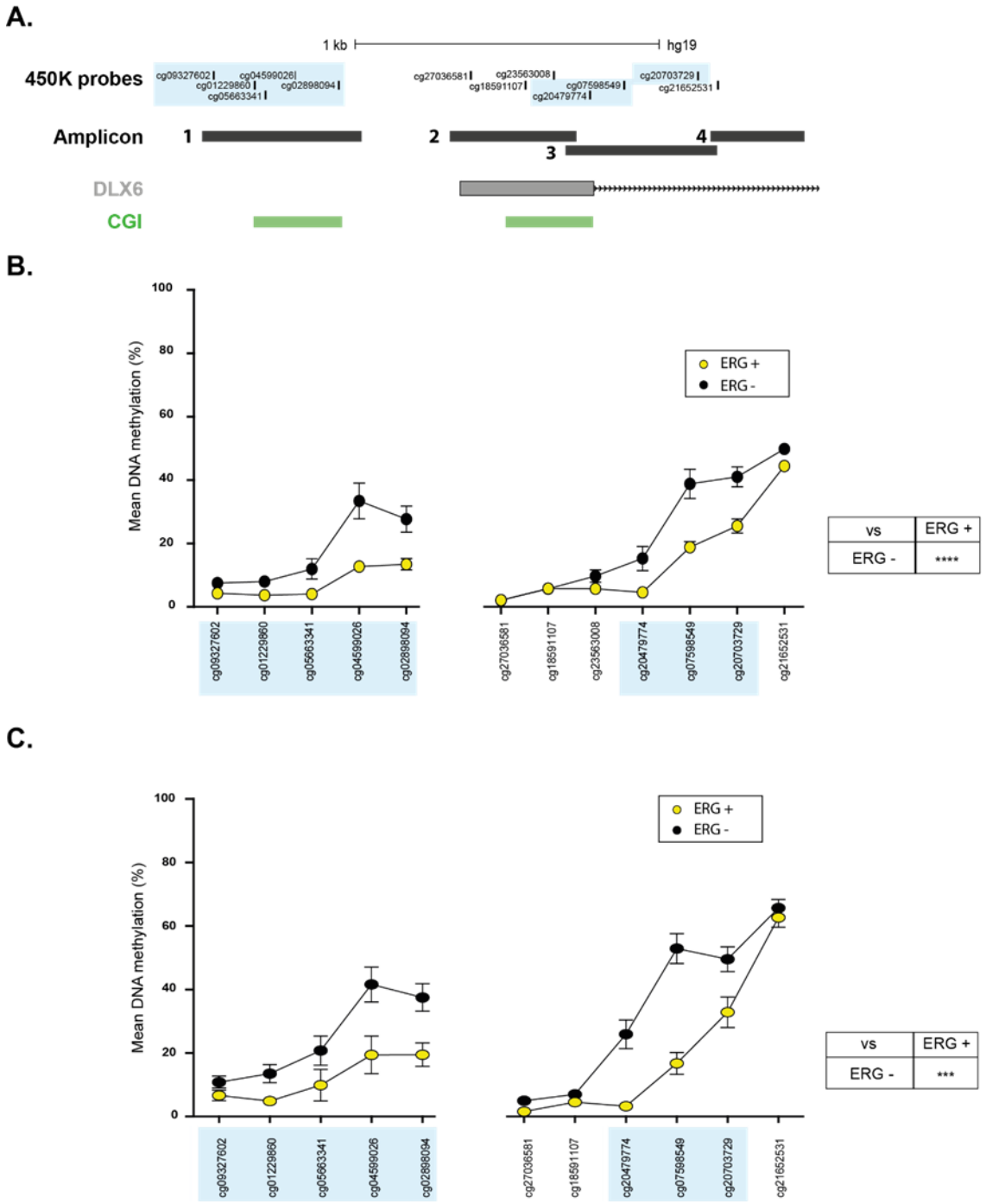


Figure 7-4 Hypomethylation of *DLX6* promoter in *ERG* fusion-positive PCa

(A) Representation of HumanMethylation450 beadchip probes and Mass-Array amplicons overlapping with *DLX6* promoter region and CGIs. (**B** and **C**) DNA methylation levels are displayed on the y-axis as a percentage ranging from 0% to 100%. Each point represents mean methylation levels of single CpG sites in *TMPRSS2:ERG* positive (ERG+) and negative (ERG-) tumor samples from the **(B)** ICGC or **(C)** the TCGA cohort. **(B)** Mann-Whitney U test or **(C)** Wilcoxon signed-rank test were used to calculate statistical significances between each group by comparing the mean DNA methylation levels across the eight CpG sites highlighted in light blue. ****p<0.0001, ***p<0.001.

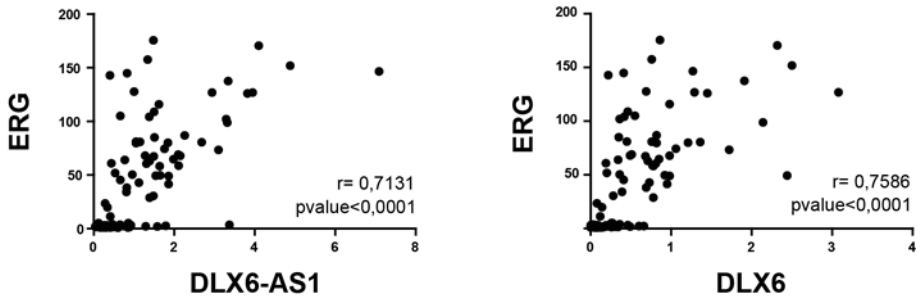
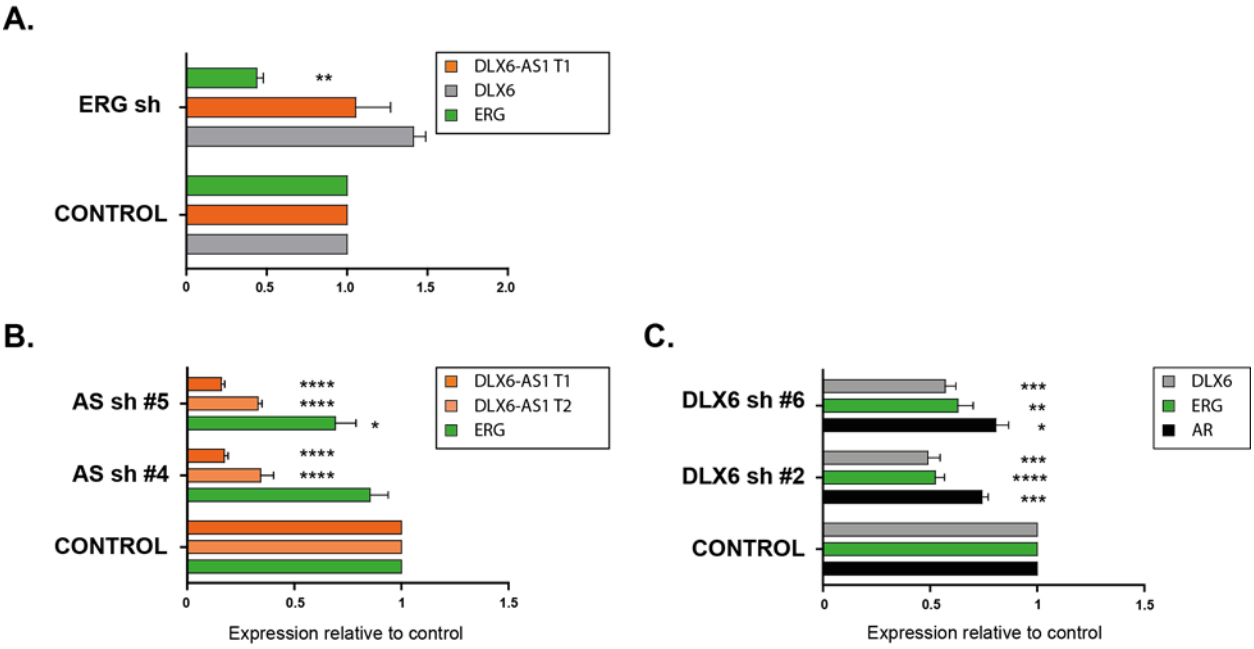


Figure 7-5 Coordinated upregulation of *DLX6/DLX6-AS1* with *ERG*

Spearman correlation plots between *ERG* and *DLX6-AS1* or *DLX6* expression measured by RNA sequencing in the ICGC cohort of EO-PCa patients (normal= 10 and tumor= 91).

Figure 7-6 Downregulation of *DLX6* leads to *ERG* downregulation



Relative expression levels of *DLX6-AS1 T1* and *T2*, *DLX6*, *ERG* and *AR* after knockdown of **(A)** *ERG*, **(B)** *DLX6-AS1* (AS) or **(C)** *DLX6* in fusion-positive VCaP cells. Data are depicted as the mean +/- s.e.m., n= 2 for **(A)** and mean +/- s.e.m., n=3 for **(B)** and **(C)** biological replicates.

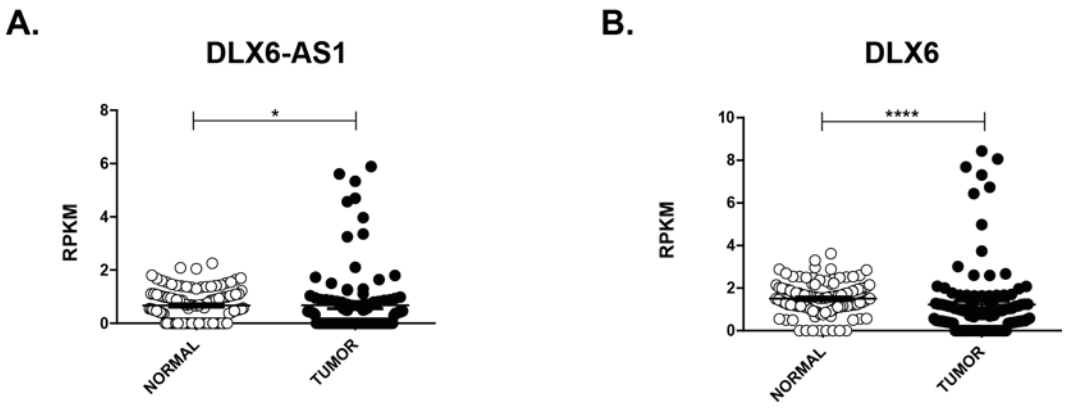


Figure 7-7 *DLX6/DLX6-AS1* expression in breast carcinoma

DLX6-AS1 **(A)** and *DLX6* **(B)** RNA expression levels measured by RNA sequencing in tumors compared to matched normal tissue (NORMAL=111 TUMOR=111) from TCGA breast carcinoma cohort. Wilcoxon signed-rank test ****p<0.0001, *p<0.05.

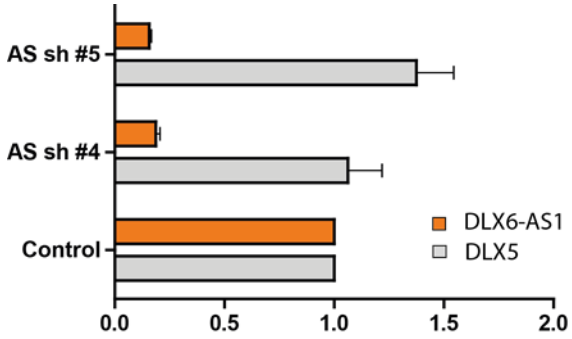


Figure 7-8 Downregulation of *DLX6-AS1* does not impact *DLX5* RNA levels

Relative expression levels of *DLX6-AS1* and *DLX5* after knockdown of *DLX6-AS1* (AS) in fusion-positive VCaP cells. Data are depicted as the mean +/- s.e.m., n= 2 biological replicates. We detected no statistically significant differences between the control group and each knockdown experiment using two-tailed unpaired t-test.

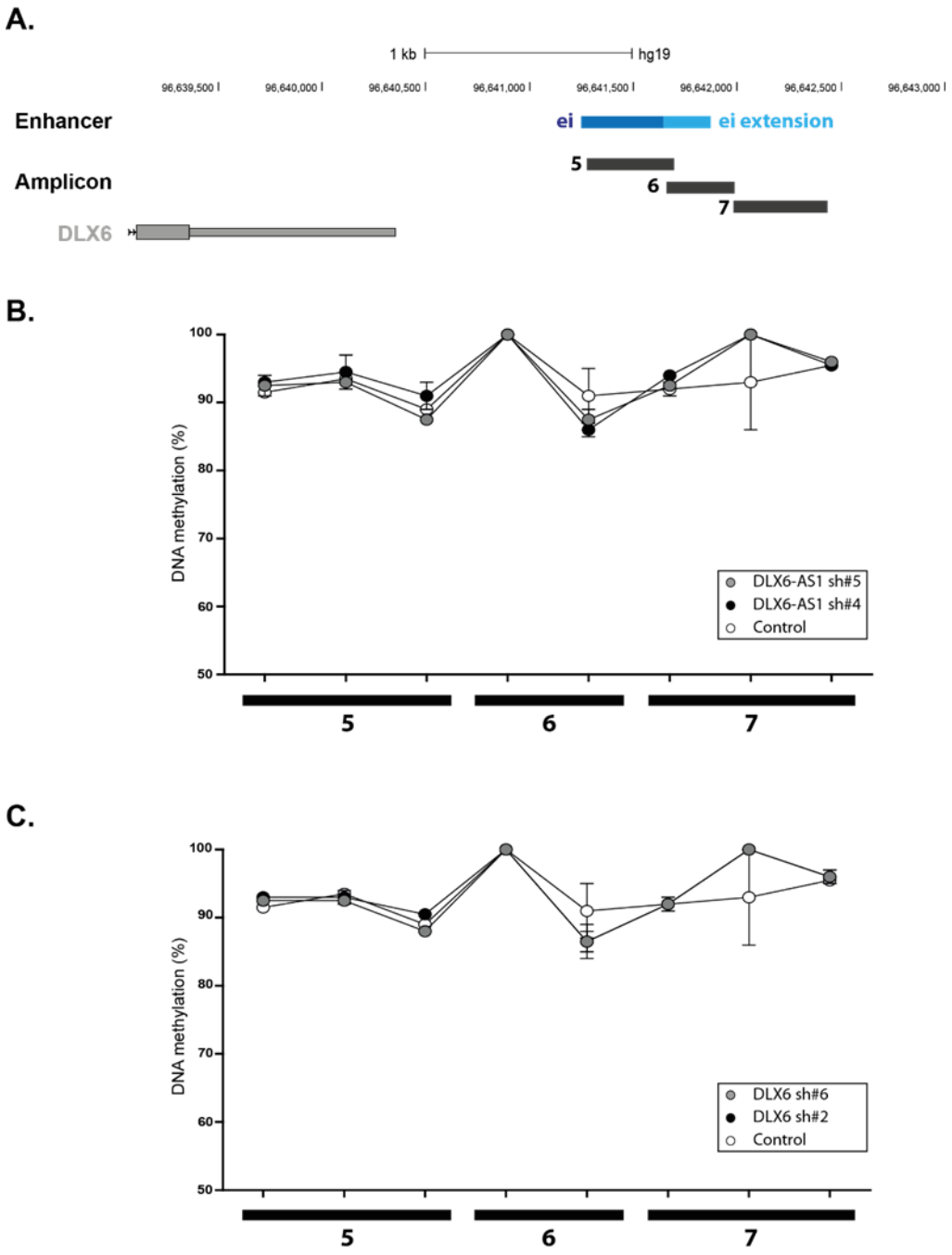


Figure 7-9 *DLX6-AS1* or *DLX6* downregulation does not impact ei enhancer DNA methylation

(A) Representation of Mass-Array amplicons overlapping with the conserved ei enhancer region and its extension. (**B** and **C**) DNA methylation levels are displayed on the y-axis as a percentage ranging from 50% to 100%. Each point represents mean methylation levels of single CpG sites in VCaP cells after knockdown of DLX6-AS1 **(B)** or DLX6 **(C)** for 8 days. Data are depicted as the mean +/- s.e.m., n= 3 biological replicates. We detected no statistically significant differences between treatment groups and the control using Mann-Whiney U test.

7.1.2 Supplementary tables

Table 7-8 ORF1 identity matches

ID	Description	Species	Bit-score	Identity (%)	e-value	Evidence
XP_016876486.1	PREDICTED: mirror-image polydactyly gene 1 protein isoform X1	Homo sapiens	90.5	76.471	6.41E-20	Prediction
EAX04212.1	polymerase (DNA directed), eta, isoform CRA_c	Homo sapiens	89.4	80.769	2.22E-22	Conceptual translation
EAW62471.1	unnamed protein product	Homo sapiens	85.5	82.353	1.08E-19	Conceptual translation
EHH60270.1	hypothetical protein EGM_11598	Macaca fascicularis	85.5	80.392	1.94E-20	Conceptual translation
XP_011731548.1	PREDICTED: sphingosine 1-phosphate receptor 2 isoform X1	Macaca nemestrina	83.6	76	1.62E-17	Prediction
BAC87498.1	unnamed protein product	Homo sapiens	83.2	78.846	4.07E-19	Unknown
BAC85949.1	unnamed protein product	Homo sapiens	83.2	78.846	4.16E-19	Unknown
EHH55556.1	hypothetical protein EGM_04788	Macaca fascicularis	83.2	78	1.61E-19	Conceptual translation
XP_007985368.1	PREDICTED: uncharacterized protein LOC103229289	Chlorocebus sabaeus	83.2	71.93	1.32E-17	Prediction
EHH21096.1	hypothetical protein EGK_04085	Macaca mulatta	82.8	75	2.53E-19	Conceptual translation
EAW55734.1	hCG2038067	Homo sapiens	82.8	72.222	4.79E-19	Conceptual translation
EAW56895.1	ribosomal protein S16, isoform CRA_d	Homo sapiens	82.4	76.923	2.97E-19	Conceptual translation
EHH64058.1	hypothetical protein EGM_17177	Macaca fascicularis	82.4	71.154	4.19E-19	Conceptual translation
BAB15056.1	unnamed protein product	Homo sapiens	82	76	1.03E-18	Unknown
EAW89122.1	hCG2039054	Homo sapiens	82	75	3.73E-19	Conceptual translation
EHH54116.1	hypothetical protein EGM_14878	Macaca fascicularis	81.6	76	1.61E-19	Conceptual translation
EHH20736.1	hypothetical protein EGK_03652	Macaca mulatta	80.5	74.51	1.73E-18	Conceptual translation
XP_004045110.1	PREDICTED: uncharacterized protein LOC101129799	Gorilla gorilla gorilla	80.1	71.93	1.12E-16	Prediction
XP_016860586.1	PREDICTED: inhibitor of growth protein 5 isoform X1	Homo sapiens	80.1	70.833	1.39E-16	Prediction

Listed are all proteins or peptides identified by protein BLAST sharing similarity with ORF1 protein with a bit-score of at least 80.

Table 7-9 ORF2 identity match

ID	Description	Species	Bit-score	Identity (%)	e-value	Evidence
WP_020429213.1	hypothetical protein	Paenibacillus riograndensis	33.1	50	3	Unknown

Table 7-10 DNA accession numbers and coordinates corresponding to ORF1 identity matches

Protein accession number	DNA accession number	DNA strand	DNA coordinates corresponding to the coding sequence
XP_016876486.1	XM_017020997.1	+	513-2072
EAX04212.1	CH471081.1	+	16602455-16602637
EAW62471.1	AK025116.1	+	387-902
EHH60270.1	CM001295.1	-	27174454-27174627 + 27265213-27265320
XP_011731548.1	XM_011733246.1	+	128-1324
BAC87498.1	AK128554.1	+	1228-1644
BAC85949.1	AK124786.1	+	1099-1515
EHH55556.1	CM001288.1	-	46780689-46780888 + 46902969-46903056
XP_007985368.1	XM_007987177.1	+	1263-2234
EHH21096.1	CM001263.1	-	100171355-100171666
EAW55734.1	CH471145.2	-	5597461-5597850
EAW56895.1	CH471126.1	-	2133982-2134272
EHH64058.1	CM001283.1	-	25977917-25978195 + 25978215-25978247
BAB15056.1	AK025047.1	+	734-1117
EAW89122.1	CH471099.1	+	8361188-8361478
EHH54116.1	CM001280.1	+	183247767-183247916
EHH20736.1	CM001263.1	+	48486496-48486804
XP_004045110.1	XM_004045062.1	+	1-825
XP_016860586.1	XM_017005097.1	+	914-1816

7.3 Nucleic acid sequences

> DLX6-AS1 T1

GGCGGGAGAAGCGAGCTGCCCCAGCGGCCTCTCACCTGTGTCGTCCCCTCGCGTCTGGGCGGCTGCGCTGCTGTTGTGGTAGGACTGGAGACAGAGTCTTGCT
CTATTGCCCAGGCTGCAACTGGTGTGATCTCGGCTCACTACAACCTCTGCCTCCTGGGTTCAAGCGATTCTCCTGCCTCAAGCTCCCAAGTAGCTGGGATTACAA
GCATGCACCACCATGCCTGGCTAATTTTTTGTATTTTTAGTGGAGACGGGGTTTCGCCACATTGGCCAGGGTGGTCTTGAACTCCTGACCTCAAGTGATCCACCTG
CCTTGGCCTCCCAAAGTGCTGGGATTATAAGCATGAGCCACTGCACCCAGCCTTATACTGAACTTTCAATGGGTTCAATTCCACTAGGAGCATAAAGGCCACTGC
ATATGAGTTGTGGAAAGAAGAGATTAGAAGAAGGAAGAACTTGAGATGAGTTCCTCCCTTCAACATTCTGTCTCCTCCTACCTAGCATCTTCTTTCTTTTAGTCTT
TCTAGAATGTCCATCTGTTTTTGGCCATTGCGGAGAGAGAAGCTGAGCTTTAAAGGAGTAGGAGCTTCAAAGGCGTAGGAGCTTCAAAATTCTTGTTTCTTCATG
TTTGATCACCCTTCTAAACCTGTCTTCTGTTCCTTCTGCTATTCTTTTTTCTTAGAGCATAGGAAAGGGGAGCTTTTAAATTAATACTTAAAGCATGGAAAAAAAGA
ACTTGAGAAGAAAGTAAAACAAGGGAGATGAGGCTAGTAAAGTAAGGAAAATGAAGAGGAAGAGGAGGAAGGGTTAGCTTCTAAATTCCAAGTCAAATTGAT
ATGGAACAGGCAAGCCGCTTGTCTTACTTAAACTTCAGAAAAGGATCTGCTGAAACTTGATAGAAATGGAAAGGGAAATCCTTGGGGTGGGGAACCTCCAAAC
ATTAGTAATGATATTGAACAACTCAAAGTATTGAGGAAATCTGCAGGCTACATGCCTGAAGATTACCCATGCAGATAGACCAAAAGGATTAGAATTATCTGTTG
ATATTAGTAATATTTATTGACATCTAGCTAGTATTGGTAATTTTAAGTTTTAGATTAATTTCTTTGGTAATAGCTATGATATATTTTATAGACAAGAATTATATCTAT
AGGCTTGCTATCATAGGCTCTTTTAATCAGCATTAATTTAGTCTACTGATTTTTAGCACATTTGAATCATTCACTTATGCTAGGTAACTCATTGCAAAATAAAAAGA
TGATTCCTGTATGTATGGCAGCTATACATTAAGGAGGAGTCTACCAGAATATGAAAAAGTCAGCTGACCTAAATATTGCTGAGACAAAGGAAAACCCACTCCCT
TGGAGGAGCATGACCTTTTCCTGTAATTCTTCCCACTGCTGTTGTTGAGCTCCTTGGATCCTGGCTCCTGGACACCATCATCAAGAAGACTTTATGGATGGGCTGT
CCACCCACTGAGAGAAGAGGAGCATCAGCTACAGTTTCTCTCTAGATTGCCTTCTTCATTTTGAGTAATGACTGTCAGCAGGGTCAGATTAAACACAAAACAACT
GGACAATTGCTTGGAGGACTAAACTATAAGGGCACTAACATGTCAATAGTAGGCTAACACATCCATGGAAAATATATTTACCAGCTCTTCTCTCAGGGAGGATTC
TGTGTGGGGTTGGAAGTAATGATTTGTTAAATTCCTTAGGGGTAGAAAGTAGGGCATAATCAGAATATAGAGGAATATGCTGTTTGACTTCAGGGTTTCTGTTT
TTCTTACTAGGATATATAAAACAGGGACTCTAGCTAGATTGTTTATGACCACAGAGGGTAGGCTGAGTGCTCCCATGATCTTCCTGCTTGGTTCTTGCCCATACA
GAGGTCAGCCTTTCCTCTAATAAAGATTGAACAA

> DLX6-AS1 T2S

GGCGGGAGAAGCGAGCTGCCCCAGCGGCCTCTCACCTGTGTCGTCCCCTCGCGTCTGGGCGGCTGCGCTGCTGTTGTGGTAGGACTGGAGTTGCTGAGACAAA
GGAAAACCCACTCCCTTGGAGGAGCATGACCTTTTCCTGTAATTCTTCC**C**ACTGCTGTTGTTGAGCTCCTTGGATCCTGGCTCCTGGACACCATCATCAAGAAGAC
TTTATGGATGGGCTGTCCACCCACTGAGAGAAGAGGAGCATCAGCTACAGTTTCTCTCTAGATTGCCTTCTTCATTTTGAGTAATGACTGTCAGCAGGGTCAGAT
TAAACACAAAACAACTGGACAATTGCTTGGAGGACTAAACTATAAGGGCACTAACATGTCAATAGTAGGCTAACACATCCATGGAAAATATATTTACCAGCTCTT
CTCTCAGGGAGGATTCTGTGTGGGGTTGGAAGTAATGATTTGTTAAATTCCTTAGGGGTAGAAAGTAGGGCATAATCAGAATATAGAGGAATATGCTGTTTGAC
TTCAGGGTTTCTGTTTTTCTTACTAGGATATATAAAACAGGGACTCTAGCTAGATTGTTTATGACCACAGAGGGTAGGCTGAGTGCTCCCATGATCTTCCTGCTTG
GTTCTTGCCCATACAGAGGTCAGCCTTTCCTCTAATAAAGATTGAACAA

> DLX6-AS1 T2L

GGCGGGAGAAGCGAGCTGCCCCAGCGGCCTCTCACCTGTGTCGTCCCCTCGCGTCTGGGCGGCTGCGCTGCTGTTGTGGTAGGACTGGAGTGCTGAGACAAAG
GAAAACCCACTCCCTTGGAGGAGCATGACCTTTTCCTGTAATTCTTCCCACTGCTGTTGTTGAGCTCCTTGGATCCTGGCTCCTGGACACCATCATCAAGAAGACT
TTATGGATGGGCTGTCCACCCACTGAGAGAAGAGGAGCATCAGCTACAGTTTCTCTCTAGATTGCCTTCTTCATTTTGAGTAATGACTGTCAGCAGGGTCAGATT
AAACACAAAACAACTGGACAATTGCTTGGAGGACTAAACTATAAGGGCACTAACATGTCAATAGTAGGCTAACACATCCATGGAAAATATATTTACCAGCTCTTC
TCTCAGGGAGGATTCTGTGTGGGGTTGGAAGTAATGATTTGTTAAATTCCTTAGGGGTAGAAAGTAGGGCATAATCAGAATATAGAGGAATATGCTGTTTGACT
TCAGGGTTTCTGTTTTTCTTACTAGGATATATAAAACAGGGACTCTAGCTAGATTGTTTATGACCACAGAGGGTAGGCTGAGTGCTCCCATGATCTTCCTGCTTGG
TTCTTGCCCATACAGAGGTCAGCCTTTCCTCTAATAAAGATTGAACAAGTAGTGGTCTGAGGGAGACACCAATTCATTACCCTACATGTCTCTTCTCTGCACTCCA
GGGCTTTGATAATAAAGACACTGGCAGACTATCTATCTTCCATTTCTATAATGTGAGCCCTTAGGGAGTCTTCGTTCACTTGGGGGTGAGGGTCATTGCTCACAG
AGTAGTTCAAGTCAAATGGAACTTGAACTCTTTGCCTATGGGCCTGGTGGTCAGACTCTGTGTTGAGTTCATTAGATTATTGGAGACACAAGGTAGAGCTGGAT
GCTTCAAAAATATTTGGCTAAAGGATGACATTGCTGGTTATTTGTAGATAAAGCCATGATGGAACCTGCTTGGAATCATGAAATATGGAACTGGTGGTCATGTTT
AAAAATACAACTAATAGTTAAGTACCTACTGGACACTGTAGAGACTTAGGGGCTAGACAGACATGGTTCCTGCCCCCTTGGAGCTTACACTGTAGCTTCCCCTTA
GGTATGAAGAACAGTGGCTACAACTAACAAATGGCCACAAAGATATAATTGAGCCAGTGTTCCAATTATTAGGGTAATTCCTATTTCCTTAATCATTCCTATTGAC
CATGTTCTATAAGCCTGCATTCTATAAATGGCGTATGACCATGGGCTGTTTCCCCCCAGCAAGTTGTACAAAGTTCTGTGGTACCAGGGAAAGGGCTTAAGGTTA
GCAGGGCCTCTGCGGAAGGACATATGAAGTGACTTGGGTTAGGAAACAGGAAGGGATAGGATTCAAGAACAGCTATTGCTTCTGTTCTATATAGGAAACTGCA
GCGTGAAAAATGCTGGGCTGGGAATTTTGAGACCTGGGTTTTAGTTTGTGTTCTAATACTAACAAGCTATGTGACTGTGGGTAAGTCATTTCACATTCCATTTGG
ATGCCTCTTGAGTGACTCCAGGCCTCTCCAGCTCTAAAACATTAAGATCAGGCCCTACGCTACAGCTGGCCAGTGTGTAATTCTTCTGTTTCTATGCTGTTAGGTC
AAATAGATCTTCAATAGTTACTTGATTGTTATTACTTTTTTCTGAAGTGGGTGTTTTATCAATGTTTTAGGATACAGTGAGTCTGCTTCTCCCCTTTGGAGTTAGGA
AGGTTGTAGGAATATACACTGTAGAGCATATGGGAGCTTTATCCCCTCCTTTTTTCCCCGCTACCTTTCTCCCTCTCCTTCCCTTCATCACATTTCTATTGAGCATAT
GCCATTGCCGTGTACCAGGTGTGTGCTAGGTTTGGAGATACAAGGTAAACCTGAGACTTTCCCAGTCTCCAGGAGACAAAACCCTAATTTCTTTCATCTGCTGTC
TTTCTCTTTGGAAAGAATCAACGATATCCCAGGGGAATGTGCCCATGTCCCAGGGTAACCAACTACAGACAGATGCCCCATTCTACTCAAGCAACTTTTAGAGTG
CCTTGAGATACACATCAGATAATTATGCAGGGCCAGGCATGGTGCTGCACACTTGTAATCACAGCACTTTGGGAGGCCGAGGCAGCAGATTGCTTGAGCCCAG
GAATTCAAGTGTAGCCTAGGGAACATGGCAAAACCCCAGCTCTACAAAAAAATACACAAATTAGCTGCGTGTGGTGGCCTATGACAGGAGGCTGAGATGGGAG
GATTGCTTGAGCCTGTGAGGTCGAGGCTGCAGTGAGCCGAGATCATGCCATTACTCCAGCCTGGGTGACATAGGGAGACCCTGTCTTGAAAAAAAAAAAAAAA

AGATAATTATTCAGCCCTAGAGTCATTGTGAAAAGATCTATCTTCAGATATAAGGAAGAAACAATCTTTTATTTCTTAGGATAAATCTGTAGAAGGACCTCCAGA
CAGTGAAGGCCACTGACTACTTTATACTCTGTAAGCCATCCCTCCCTGGTAGGAAGGACTATTTCCAATCTTACAGAGTACCTCTCAGCAAATAGACGTTTTCACA
TATACTGTGATTCATACATCCCTATGGCTGGTGACCTCTTTAAAAAGGAAAGGAAAAAGCCTAATCAAACAAAAAGATGCTGCTAGTAATTCTTACCCTATTGTG
AATCCTATATAAGCAAATTTGTATCTTTGTTTTTTCCTACATTAGCAGATCTATTTGATATATCTCTGAGTGCAGAAAATATTTTATGGAAAAATCAATATATGGAA
TTTCAAATTCAGAATTGCTGATACACACTATTTGGTTTCACAATTTTATCCTAGGAAATAGTATTAGAGATTTCAATTTCTGGCTTAAATGGTAGAATTAATTACTC
TTAACTCTTAATTTTACTTCTGAGTTGAGGTCAAGGAACAGGCAGACACCTGCAGTTAACGTCTATACCTCTCCATGGCCAAGAGTTTTAATTTTCTCGTCTTCAAT
TTTGTAGATGTTCATCATTACTAAATGGATTGATTAGTATTTTATCTCCTCTCCTTGTCCTTACTTTCCCTCTGGTAAATATGTTATAAACAGTGTAAGGCTCCTAAG
ATAGAGTAGCTGGTAGGACTTAGAAGAGAAACAAAGGGCACTGATAACTCACATAAATGGAAAATTGGCTCTGGAATAACTGACAACATATTCAAGTATTTTAG
TGCAGTGTCACTCTCATTAAGAAGAAGAGAATCAGTAAATCTATGTGACTCTAAACATTCTAATGAAAAAAGGAATATTCTGCCAATTATCTCACATTTCTAAATA
TCTGGATATTGGCCATTGTAAAGACAAAACATACAGATGATGGACTTGTCTTTCCACCTCTCATTTGCATGGTTTGGAGCATTGTACCTCCAGCCATAGACTCTAA
GGCAATTTATATTTGCTTCCTCTTCCCTCTTGAGAGAAAACGAAAATCTTATTTTTCCAAGCAATTAAAACTCTTCTGCTTCAGCTAGGATGAAAGAATTAGGAGT
TCTGTCTCCT**T**GTATCTAATTGCATGTTTCATCTTTCTTGTTTTAATGATTGACAGAAAACTAATAAACTGAGACATCTTTGAATC**C**AGGTTGAATGTA

>ALUSZ (reverse complement)

GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACTTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACATGG
TGAAACCCCGTCTCTACTAAAAATACAAAAATTAGCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAAC
CCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTCA

7.4 DLX6-AS1 amino acid sequences

> ORF1

MHHHAWLIFVFLVETGFRHIGQGGLELLTSSDPPALASQSAGIISMSHCTQPYTELSMGSIPLGA*

> ORF2

MAYDHGLFPPSKLYKVLWYQGKGLRLAGPLRKDI*

8. REFERENCES

Abate-Shen, C. and M.M. Shen, *Molecular genetics of prostate cancer.* Genes Dev, 2000. **14**(19): p. 2410-34.

Acevedo, L.G., et al., *Analysis of the mechanisms mediating tumor-specific changes in gene expression in human liver tumors.* Cancer Res, 2008. **68**(8): p. 2641-51.

Al Olama, A.A., et al., *Multiple loci on 8q24 associated with prostate cancer susceptibility.* Nat Genet, 2009. **41**(10): p. 1058-60.

Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

Amouroux, R., et al., *De novo DNA methylation drives 5hmC accumulation in mouse zygotes.* Nat Cell Biol, 2016. **18**(2): p. 225-33.

Anderson, D.M., et al., *A micropeptide encoded by a putative long noncoding RNA regulates muscle performance.* Cell, 2015. **160**(4): p. 595-606.

Andrews, S.J. and J.A. Rothnagel, *Emerging evidence for functional peptides encoded by short open reading frames.* Nat Rev Genet, 2014. **15**(3): p. 193-204.

Arab, K., et al., *Long Noncoding RNA TARID Directs Demethylation and Activation of the Tumor Suppressor TCF21 via GADD45A.* Mol Cell, 2014. **55**(4): p. 604-14.

Banfai, B., et al., *Long noncoding RNAs are rarely translated in two human cell lines.* Genome Res, 2012. **22**(9): p. 1646-57.

Bao, W., K.K. Kojima, and O. Kohany, *Repbase Update, a database of repetitive elements in eukaryotic genomes.* Mob DNA, 2015. **6**: p. 11.

Barbieri, C.E., et al., *Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer.* Nat Genet, 2012. **44**(6): p. 685-9.

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

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

Bassett, A.R., et al., *Considerations when investigating lncRNA function in vivo.* Elife, 2014. **3**: p. e03058.

Bazzini, A.A., et al., *Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation.* EMBO J, 2014. **33**(9): p. 981-93.

Beckedorff, F.C., et al., *The intronic long noncoding RNA ANRASSF1 recruits PRC2 to the RASSF1A promoter, reducing the expression of RASSF1A and increasing cell proliferation.* PLoS Genet, 2013. **9**(8): p. e1003705.

Berghoff, E.G., et al., *Evf2 (Dlx6as) lncRNA regulates ultraconserved enhancer methylation and the differential transcriptional control of adjacent genes.* Development, 2013. **140**(21): p. 4407-16.

Bert, S.A., et al., *Regional activation of the cancer genome by long-range epigenetic remodeling.* Cancer Cell, 2013. **23**(1): p. 9-22.

Blackledge, N.P., J.P. Thomson, and P.J. Skene, *CpG island chromatin is shaped by recruitment of ZF-CxxC proteins.* Cold Spring Harb Perspect Biol, 2013. **5**(11): p. a018648.

Blackledge, N.P., et al., *CpG islands recruit a histone H3 lysine 36 demethylase.* Mol Cell, 2010. **38**(2): p. 179-90.

Boch, J., et al., *Breaking the code of DNA binding specificity of TAL-type III effectors.* Science, 2009. **326**(5959): p. 1509-12.

Bond, A.M., et al., *Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry.* Nat Neurosci, 2009. **12**(8): p. 1020-7.

Borno, S.T., et al., *Genome-wide DNA methylation events in TMPRSS2-ERG fusion-negative prostate cancers implicate an EZH2-dependent mechanism with miR-26a hypermethylation.* Cancer Discov, 2012. **2**(11): p. 1024-35.

Bourc'his, D., et al., *Dnmt3L and the establishment of maternal genomic imprints.* Science, 2001. **294**(5551): p. 2536-9.

Brocks, D., et al., *Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer.* Cell Rep, 2014. **8**(3): p. 798-806.

Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes Dev, 2011. **25**(18): p. 1915-27.

Carrieri, C., et al., *Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat.* Nature, 2012. **491**(7424): p. 454-7.

Carver, B.S., et al., *Aberrant ERG expression cooperates with loss of PTEN to promote cancer progression in the prostate.* Nat Genet, 2009. **41**(5): p. 619-24.

Cheng, A.W., et al., *Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system.* Cell Res, 2013. **23**(10): p. 1163-71.

Chooniedass-Kothari, S., et al., *The steroid receptor RNA activator is the first functional RNA encoding a protein.* FEBS Lett, 2004. **566**(1-3): p. 43-7.

Chung, S., et al., *Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility.* Cancer Sci, 2011. **102**(1): p. 245-52.

Comings, D.E., *The structure and function of chromatin.* Adv Hum Genet, 1972. **3**: p. 237-431.

Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

Consortium, E.P., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.

Consortium, F., et al., *A promoter-level mammalian expression atlas.* Nature, 2014. **507**(7493): p. 462-70.

Coolen, M.W., et al., *Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity.* Nat Cell Biol, 2010. **12**(3): p. 235-46.

Cooper, C., et al., *Steroid Receptor RNA Activator bi-faceted genetic system: Heads or Tails?* Biochimie, 2011. **93**(11): p. 1973-80.

Cortazar, D., et al., *Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability.* Nature, 2011. **470**(7334): p. 419-23.

Cortellino, S., et al., *Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair.* Cell, 2011. **146**(1): p. 67-79.

de Koning, A.P., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet, 2011. **7**(12): p. e1002384.

Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription.* Genes Dev, 2011. **25**(10): p. 1010-22.

Deininger, P., *Alu elements: know the SINEs.* Genome Biol, 2011. **12**(12): p. 236.

Derrien, T., et al., *The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression.* Genome Res, 2012. **22**(9): p. 1775-89.

Di Ruscio, A., et al., *DNMT1-interacting RNAs block gene-specific DNA methylation.* Nature, 2013. **503**(7476): p. 371-6.

Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.

Doerfler, W., *DNA methylation and gene activity.* Annu Rev Biochem, 1983. **52**: p. 93-124.

Du, Q., et al., *Methyl-CpG-binding domain proteins: readers of the epigenome.* Epigenomics, 2015. **7**(6): p. 1051-73.

Du, Z., et al., *Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer.* Nat Struct Mol Biol, 2013. **20**(7): p. 908-13.

Ehrlich, M., *DNA methylation in cancer: too much, but also too little.* Oncogene, 2002. **21**(35): p. 5400-13.

Ezkurdia, I., et al., *Analyzing the first drafts of the human proteome.* J Proteome Res, 2014. **13**(8): p. 3854-5.

Feng, J., et al., *The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator.* Genes Dev, 2006. **20**(11): p. 1470-84.

Feng, Y., et al., *Methods for the study of long noncoding RNA in cancer cell signaling.* Methods Mol Biol, 2014. **1165**: p. 115-43.

Finn, R.D., et al., *The Pfam protein families database: towards a more sustainable future.* Nucleic Acids Res, 2016. **44**(D1): p. D279-85.

Flatau, E., et al., *DNA methylation in 5-aza-2'-deoxycytidine-resistant variants of C3H 10T1/2 C18 cells.* Mol Cell Biol, 1984. **4**(10): p. 2098-102.

Fu, X., et al., *Regulation of apoptosis by a prostate-specific and prostate cancer-associated noncoding gene, PCGEM1.* DNA Cell Biol, 2006. **25**(3): p. 135-41.

Fu, Y., et al., *Epigenetic regulation of proprotein convertase PACE4 gene expression in human ovarian cancer cells.* Mol Cancer Res, 2003. **1**(8): p. 569-76.

Fussner, E., R.W. Ching, and D.P. Bazett-Jones, *Living without 30nm chromatin fibers.* Trends Biochem Sci, 2011. **36**(1): p. 1-6.

Gagnon, K.T., et al., *Analysis of nuclear RNA interference in human cells by subcellular fractionation and Argonaute loading.* Nat Protoc, 2014. **9**(9): p. 2045-60.

Gilbert, L.A., et al., *CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes.* Cell, 2013. **154**(2): p. 442-51.

Goldberg, T., et al., *LocTree3 prediction of localization.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W350-5.

Goldstein, J., et al., *Does Choline PET/CT Change the Management of Prostate Cancer Patients With Biochemical Failure?* Am J Clin Oncol, 2014.

Gong, C. and L.E. Maquat, *lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements.* Nature, 2011. **470**(7333): p. 284-8.

Goyal, A., et al., *Challenges of CRISPR/Cas9 applications for long non-coding RNA genes.* Nucleic Acids Res, 2016.

Grasso, C.S., et al., *The mutational landscape of lethal castration-resistant prostate cancer.* Nature, 2012. **487**(7406): p. 239-43.

Greco, S.A., et al., *Thrombospondin-4 is a putative tumour-suppressor gene in colorectal cancer that exhibits age-related methylation.* BMC Cancer, 2010. **10**: p. 494.

Gregory, R.I. and R. Shiekhattar, *Chromatin modifiers and carcinogenesis.* Trends Cell Biol, 2004. **14**(12): p. 695-702.

Guttman, M., et al., *Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.* Nature, 2009. **458**(7235): p. 223-7.

Guttman, M., et al., *lincRNAs act in the circuitry controlling pluripotency and differentiation.* Nature, 2011. **477**(7364): p. 295-300.

Guttman, M., et al., *Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins.* Cell, 2013. **154**(1): p. 240-51.

Gyorffy, B., et al., *Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer.* PLoS One, 2013. **8**(12): p. e82241.

Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.

Hata, K., et al., *Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice.* Development, 2002. **129**(8): p. 1983-93.

Hayward, S.W., et al., *Establishment and characterization of an immortalized but non-transformed human prostate epithelial cell line: BPH-1.* In Vitro Cell Dev Biol Anim, 1995. **31**(1): p. 14-24.

He, S., et al., *LongTarget: a tool to predict lncRNA DNA-binding motifs and binding sites via Hoogsteen base-pairing analysis.* Bioinformatics, 2015. **31**(2): p. 178-86.

He, Y., et al., *The antisense transcriptomes of human cells.* Science, 2008. **322**(5909): p. 1855-7.

He, Y.F., et al., *Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA.* Science, 2011. **333**(6047): p. 1303-7.

Hermann, A., R. Goyal, and A. Jeltsch, *The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites.* J Biol Chem, 2004. **279**(46): p. 48350-9.

Heyer, E.E. and M.J. Moore, *Redefining the Translational Status of 80S Monosomes.* Cell, 2016. **164**(4): p. 757-69.

Holliday, R., *The inheritance of epigenetic defects.* Science, 1987. **238**(4824): p. 163-70.

Holliday, R. and J.E. Pugh, *DNA modification mechanisms and gene activity during development.* Science, 1975. **187**(4173): p. 226-32.

Hu, S., X. Wang, and G. Shan, *Insertion of an Alu element in a lncRNA leads to primate-specific modulation of alternative splicing.* Nat Struct Mol Biol, 2016. **23**(11): p. 1011-1019.

Ingolia, N.T., L.F. Lareau, and J.S. Weissman, *Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes.* Cell, 2011. **147**(4): p. 789-802.

Irizarry, R.A., et al., *The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.* Nat Genet, 2009. **41**(2): p. 178-86.

Iyer, L.M., et al., *Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids.* Cell Cycle, 2009. **8**(11): p. 1698-710.

Iyer, M.K., et al., *The landscape of long noncoding RNAs in the human transcriptome.* Nat Genet, 2015. **47**(3): p. 199-208.

Jeltsch, A. and R.Z. Jurkowska, *New concepts in DNA methylation.* Trends Biochem Sci, 2014. **39**(7): p. 310-8.

Jin, S.G., C. Guo, and G.P. Pfeifer, *GADD45A does not promote DNA demethylation.* PLoS Genet, 2008. **4**(3): p. e1000013.

Jinek, M., et al., *A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity.* Science, 2012. **337**(6096): p. 816-21.

Johnsson, P., et al., *A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells.* Nat Struct Mol Biol, 2013. **20**(4): p. 440-6.

Jones, M.J., S.J. Goodman, and M.S. Kobor, *DNA methylation and healthy human aging.* Aging Cell, 2015. **14**(6): p. 924-32.

Jones, P.A. and S.B. Baylin, *The epigenomics of cancer.* Cell, 2007. **128**(4): p. 683-92.

Jones, P.A. and G. Liang, *Rethinking how DNA methylation patterns are maintained.* Nat Rev Genet, 2009. **10**(11): p. 805-11.

Jones, P.L., et al., *Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription.* Nat Genet, 1998. **19**(2): p. 187-91.

Kaczkowski, B., et al., *Transcriptome Analysis of Recurrently Deregulated Genes across Multiple Cancers Identifies New Pan-Cancer Biomarkers.* Cancer Res, 2016. **76**(2): p. 216-26.

Kan, Z., et al., *Diverse somatic mutation patterns and pathway alterations in human cancers.* Nature, 2010. **466**(7308): p. 869-73.

Kang, S., et al., *miR-124 exhibits antiproliferative and antiaggressive effects on prostate cancer cells through PACE4 pathway.* Prostate, 2014. **74**(11): p. 1095-106.

Kapranov, P., et al., *RNA maps reveal new RNA classes and a possible function for pervasive transcription.* Science, 2007. **316**(5830): p. 1484-8.

Kareta, M.S., et al., *Reconstitution and mechanism of the stimulation of de novo methylation by human DNMT3L.* J Biol Chem, 2006. **281**(36): p. 25893-902.

Khalil, A.M., et al., *Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression.* Proc Natl Acad Sci U S A, 2009. **106**(28): p. 11667-72.

Kim, J.H., et al., *Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer.* Genome Res, 2011. **21**(7): p. 1028-41.

King, J.C., et al., *Cooperativity of TMPRSS2-ERG with PI3-kinase pathway activation in prostate oncogenesis.* Nat Genet, 2009. **41**(5): p. 524-6.

Kobayashi, Y., et al., *DNA methylation profiling reveals novel biomarkers and important roles for DNA methyltransferases in prostate cancer.* Genome Res, 2011. **21**(7): p. 1017-27.

Kondo, Y., et al., *Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation.* Nat Genet, 2008. **40**(6): p. 741-50.

Kotake, Y., et al., *Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene.* Oncogene, 2011. **30**(16): p. 1956-62.

Kouzarides, T., *Chromatin modifications and their function.* Cell, 2007. **128**(4): p. 693-705.

Kozak, M., *Initiation of translation in prokaryotes and eukaryotes.* Gene, 1999. **234**(2): p. 187-208.

Krohn, A., et al., *Recurrent deletion of 3p13 targets multiple tumour suppressor genes and defines a distinct subgroup of aggressive ERG fusion-positive prostate cancers.* J Pathol, 2013. **231**(1): p. 130-41.

Kron, K., et al., *Discovery of novel hypermethylated genes in prostate cancer using genomic CpG island microarrays.* PLoS One, 2009. **4**(3): p. e4830.

Kumar-Sinha, C., S.A. Tomlins, and A.M. Chinnaiyan, *Recurrent gene fusions in prostate cancer.* Nat Rev Cancer, 2008. **8**(7): p. 497-511.

Lamprecht, B., et al., *Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma.* Nat Med, 2010. **16**(5): p. 571-9, 1p following 579.

Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

Lanz, R.B., et al., *A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex.* Cell, 1999. **97**(1): p. 17-27.

Li, J., et al., *Expression of long non-coding RNA DLX6-AS1 in lung adenocarcinoma.* Cancer Cell Int, 2015. **15**: p. 48.

Livak, K.J. and T.D. Schmittgen, *Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.* Methods, 2001. **25**(4): p. 402-8.

Lorincz, M.C., et al., *Dynamic analysis of proviral induction and De Novo methylation: implications for a histone deacetylase-independent, methylation density-dependent mechanism of transcriptional repression.* Mol Cell Biol, 2000. **20**(3): p. 842-50.

Luger, K., et al., *Crystal structure of the nucleosome core particle at 2.8 A resolution.* Nature, 1997. **389**(6648): p. 251-60.

Maiti, A. and A.C. Drohat, *Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites.* J Biol Chem, 2011. **286**(41): p. 35334-8.

Marchler-Bauer, A., et al., *CDD: NCBI's conserved domain database.* Nucleic Acids Res, 2015. **43**(Database issue): p. D222-6.

Maruyama, R. and H. Suzuki, *Long noncoding RNA involvement in cancer.* BMB Rep, 2012. **45**(11): p. 604-11.

Mayer, W., et al., *Demethylation of the zygotic paternal genome.* Nature, 2000. **403**(6769): p. 501-2.

Merry, C.R., et al., *DNMT1-associated long non-coding RNAs regulate global gene expression and DNA methylation in colon cancer.* Hum Mol Genet, 2015. **24**(21): p. 6240-53.

Metivier, R., et al., *Cyclical DNA methylation of a transcriptionally active promoter.* Nature, 2008. **452**(7183): p. 45-50.

Mili, S. and J.A. Steitz, *Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses.* RNA, 2004. **10**(11): p. 1692-4.

Mills, I.G., *Maintaining and reprogramming genomic androgen receptor activity in prostate cancer.* Nat Rev Cancer, 2014. **14**(3): p. 187-98.

Morini, M., et al., *Mutually exclusive expression of DLX2 and DLX5/6 is associated with the metastatic potential of the human breast cancer cell line MDA-MB-231.* BMC Cancer, 2010. **10**: p. 649.

N. Mottet, J.B., E. Briers, R.C.N. van den Bergh, M. Bolla, N.J. van Casteren, P. Cornford, S. Culine, S. Joniau, T. Lam, M.D. Mason, V. Matveev, H. van der Poel, T.H. van der Kwast, O. Rouvière, T. Wiegel, *Guidelines on Prostate Cancer.* European Association of Urology, 2015. **2015**.

Nan, X., et al., *Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex.* Nature, 1998. **393**(6683): p. 386-9.

Nelson, B.R., et al., *A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle.* Science, 2016. **351**(6270): p. 271-5.

Neri, F., et al., *Dnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs.* Cell, 2013. **155**(1): p. 121-34.

Nuytten, M., et al., *The transcriptional repressor NIPP1 is an essential player in EZH2-mediated gene silencing.* Oncogene, 2008. **27**(10): p. 1449-60.

Ohno, S., *So much "junk" DNA in our genome.* Brookhaven Symp Biol, 1972. **23**: p. 366-70.

Okano, M., et al., *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development.* Cell, 1999. **99**(3): p. 247-57.

Olsson, M., et al., *Genome-wide methylation profiling identifies novel methylated genes in neuroblastoma tumors.* Epigenetics, 2016. **11**(1): p. 74-84.

Orom, U.A., et al., *Long noncoding RNAs with enhancer-like function in human cells.* Cell, 2010. **143**(1): p. 46-58.

Oswald, J., et al., *Active demethylation of the paternal genome in the mouse zygote.* Curr Biol, 2000. **10**(8): p. 475-8.

Paul, J. and J.D. Duerksen, *Chromatin-associated RNA content of heterochromatin and euchromatin.* Mol Cell Biochem, 1975. **9**(1): p. 9-16.

Pedersen, A.G. and H. Nielsen, *Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis.* Proc Int Conf Intell Syst Mol Biol, 1997. **5**: p. 226-33.

Pellacani, D., et al., *DNA hypermethylation in prostate cancer is a consequence of aberrant epithelial differentiation and hyperproliferation.* Cell Death Differ, 2014. **21**(5): p. 761-73.

Petersen, T.N., et al., *SignalP 4.0: discriminating signal peptides from transmembrane regions.* Nat Methods, 2011. **8**(10): p. 785-6.

Peterson, C.L. and J.L. Workman, *Promoter targeting and chromatin remodeling by the SWI/SNF complex.* Curr Opin Genet Dev, 2000. **10**(2): p. 187-92.

Petrovics, G., et al., *Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients.* Oncogene, 2004. **23**(2): p. 605-11.

Pienta, K.J. and P.S. Esper, *Risk factors for prostate cancer.* Ann Intern Med, 1993. **118**(10): p. 793-803.

Plass, C., et al., *Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer.* Nat Rev Genet, 2013. **14**(11): p. 765-80.

Ponting, C.P., P.L. Oliver, and W. Reik, *Evolution and functions of long noncoding RNAs.* Cell, 2009. **136**(4): p. 629-41.

Potosky, A.L., et al., *The role of increasing detection in the rising incidence of prostate cancer.* JAMA, 1995. **273**(7): p. 548-52.

Prensner, J.R., et al., *Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression.* Nat Biotechnol, 2011. **29**(8): p. 742-9.

Prensner, J.R., et al., *The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex.* Nat Genet, 2013. **45**(11): p. 1392-8.

Prensner, J.R., et al., *The lncRNAs PCGEM1 and PRNCR1 are not implicated in castration resistant prostate cancer.* Oncotarget, 2014.

Quek, X.C., et al., *lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs.* Nucleic Acids Res, 2015. **43**(Database issue): p. D168-73.

Quinn, J.J. and H.Y. Chang, *Unique features of long non-coding RNA biogenesis and function.* Nat Rev Genet, 2016. **17**(1): p. 47-62.

Rashid, F., A. Shah, and G. Shan, *Long Non-coding RNAs in the Cytoplasm.* Genomics Proteomics Bioinformatics, 2016. **14**(2): p. 73-80.

Ren, S., et al., *RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings.* Cell Res, 2012. **22**(5): p. 806-21.

Riggs, A.D., *X inactivation, differentiation, and DNA methylation.* Cytogenet Cell Genet, 1975. **14**(1): p. 9-25.

Riggs, A.D., R.A. Martienssen, and V.E.A. Russo, *Introduction*. 1996. 1996.

Rinn, J.L. and H.Y. Chang, *Genome regulation by long noncoding RNAs.* Annu Rev Biochem, 2012. **81**: p. 145-66.

Rinn, J.L., et al., *Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs.* Cell, 2007. **129**(7): p. 1311-23.

Robledo, R.F., et al., *The Dlx5 and Dlx6 homeobox genes are essential for craniofacial, axial, and appendicular skeletal development.* Genes Dev, 2002. **16**(9): p. 1089-101.

Rougier, N., et al., *Chromosome methylation patterns during mammalian preimplantation development.* Genes Dev, 1998. **12**(14): p. 2108-13.

Salamov, A.A., T. Nishikawa, and M.B. Swindells, *Assessing protein coding region integrity in cDNA sequencing projects.* Bioinformatics, 1998. **14**(5): p. 384-90.

Salinas, C.A., et al., *Prostate cancer in young men: an important clinical entity.* Nat Rev Urol, 2014. **11**(6): p. 317-23.

Schein, A., et al., *Identification of antisense long noncoding RNAs that function as SINEUPs in human cells.* Sci Rep, 2016. **6**: p. 33605.

Schmitz, K.M., et al., *TAF12 recruits Gadd45a and the nucleotide excision repair complex to the promoter of rRNA genes leading to active DNA demethylation.* Mol Cell, 2009. **33**(3): p. 344-53.

Scotto-Lavino, E., G. Du, and M.A. Frohman, *5' end cDNA amplification using classic RACE.* Nat Protoc, 2006. **1**(6): p. 2555-62.

Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2016.* CA Cancer J Clin, 2016. **66**(1): p. 7-30.

Sigrist, C.J., et al., *New and continuing developments at PROSITE.* Nucleic Acids Res, 2013. **41**(Database issue): p. D344-7.

Slavoff, S.A., et al., *Peptidomic discovery of short open reading frame-encoded peptides in human cells.* Nat Chem Biol, 2013. **9**(1): p. 59-64.

Society, A.C., *Global Cancer Facts & Figures 3rd Edition* American Cancer Society 2015(2015).

Srikantan, V., et al., *PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer.* Proc Natl Acad Sci U S A, 2000. **97**(22): p. 12216-21.

Steijger, T., et al., *Assessment of transcript reconstruction methods for RNA-seq.* Nat Methods, 2013. **10**(12): p. 1177-84.

Steurer, S., et al., *TMPRSS2-ERG fusions are strongly linked to young patient age in low-grade prostate cancer.* Eur Urol, 2014. **66**(6): p. 978-81.

Strahl, B.D. and C.D. Allis, *The language of covalent histone modifications.* Nature, 2000. **403**(6765): p. 41-45.

Taberlay, P.C., et al., *Three-dimensional disorganisation of the cancer genome occurs coincident with long range genetic and epigenetic alterations.* Genome Res, 2016.

Tahiliani, M., et al., *Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1.* Science, 2009. **324**(5929): p. 930-5.

Takayama, K., et al., *Androgen-responsive long noncoding RNA CTBP1-AS promotes prostate cancer.* EMBO J, 2013. **32**(12): p. 1665-80.

Taylor, B.S., et al., *Integrative genomic profiling of human prostate cancer.* Cancer Cell, 2010. **18**(1): p. 11-22.

Thillainadesan, G., et al., *TGF-beta-dependent active demethylation and expression of the p15ink4b tumor suppressor are impaired by the ZNF217/CoREST complex.* Mol Cell, 2012. **46**(5): p. 636-49.

Thomson, J.P., et al., *CpG islands influence chromatin structure via the CpG-binding protein Cfp1.* Nature, 2010. **464**(7291): p. 1082-6.

Tomlins, S.A., et al., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.* Science, 2005. **310**(5748): p. 644-8.

Trinklein, N.D., et al., *An abundance of bidirectional promoters in the human genome.* Genome Res, 2004. **14**(1): p. 62-6.

Troshin, P.V., J.B. Procter, and G.J. Barton, *Java bioinformatics analysis web services for multiple sequence alignment--JABAWS:MSA.* Bioinformatics, 2011. **27**(14): p. 2001-2.

van Heesch, S., et al., *Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes.* Genome Biol, 2014. **15**(1): p. R6.

Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.

Waddington, C.H., *The epigenotype. 1942.* Int J Epidemiol, 2012. **41**(1): p. 10-3.

Wang, C., et al., *mCOPA: analysis of heterogeneous features in cancer expression data.* J Clin Bioinforma, 2012. **2**(1): p. 22.

Wang, Z.A., et al., *Luminal cells are favored as the cell of origin for prostate cancer.* Cell Rep, 2014. **8**(5): p. 1339-46.

Waterhouse, A.M., et al., *Jalview Version 2--a multiple sequence alignment editor and analysis workbench.* Bioinformatics, 2009. **25**(9): p. 1189-91.

Weischenfeldt, J., et al., *Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer.* Cancer Cell, 2013. **23**(2): p. 159-70.

Weiss, M., C. Plass, and C. Gerhauser, *Role of lncRNAs in prostate cancer development and progression.* Biol Chem, 2014. **395**(11): p. 1275-90.

Wilhelm, M., et al., *Mass-spectrometry-based draft of the human proteome.* Nature, 2014. **509**(7502): p. 582-7.

Wolff, E.M., et al., *Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer.* PLoS Genet, 2010. **6**(4): p. e1000917.

Xu, B., et al., *The prognostic role of ERG immunopositivity in prostatic acinar adenocarcinoma: a study including 454 cases and review of the literature.* Hum Pathol, 2014. **45**(3): p. 488-97.

Yan, X., et al., *Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers.* Cancer Cell, 2015. **28**(4): p. 529-40.

Yang, L., et al., *lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs.* Nature, 2013. **500**(7464): p. 598-602.

Yap, K.L., et al., *Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a.* Mol Cell, 2010. **38**(5): p. 662-74.

Yegnasubramanian, S., et al., *DNA hypomethylation arises later in prostate cancer progression than CpG island hypermethylation and contributes to metastatic tumor heterogeneity.* Cancer Res, 2008. **68**(21): p. 8954-67.

Yu, J., et al., *An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression.* Cancer Cell, 2010. **17**(5): p. 443-54.

Yuasa, K., et al., *Transcriptional regulation of subtilisin-like proprotein convertase PACE4 by E2F: possible role of E2F-mediated upregulation of PACE4 in tumor progression.* Gene, 2007. **402**(1-2): p. 103-10.

Zhao, J., et al., *Genome-wide identification of polycomb-associated RNAs by RIP-seq.* Mol Cell, 2010. **40**(6): p. 939-53.

Zhao, Y., et al., *NONCODE 2016: an informative and valuable data source of long non-coding RNAs.* Nucleic Acids Res, 2016. **44**(D1): p. D203-8.

Zucchelli, S., et al., *SINEUPs are modular antisense long non-coding RNAs that increase synthesis of target proteins in cells.* Front Cell Neurosci, 2015. **9**: p. 174.

9. PUBLICATIONS DURING THE THESIS

9.1 Publications

Weiss, M., Plass C. and Gerhäuser C. *Role of lncRNAs in prostate cancer development and progression* Biological. Chemistry, 2014: 395(11): p.1275–1290

Gu, L., Frommel S., Oakes C. C., Simon, R., Grupp, K., Gerig, Y. C., Bär, D., Robinson D. M., Baer C., Weiss, M., Gu, Z., Schapira, M., Kuner, R., Sültmann, H., Provenzano M., ICGC project on Early Onset Prostate Cancer, Yaspo, M-L., Brors, B., Korbel, J., Schlomm, T., Sauter, G., Eils, R., Plass, C. and Santoro, R., *BAZ2A (TIP5) is involved in epigenetic alterations in prostate cancer and its overexpression predicts disease recurrence* Nature Genetics, 2015. 47(1): p. 22-30.

9.2 Presentations

Weiss, M., Bogatyrova O., ICGC project on Early Onset Prostate Cancer, Gerhäuser, C. and Plass, C., *Epigenetic alterations of TERT promoter in pancreatic cancer*
Annual DKFZ PhD Student Poster Session, November 2015, Heidelberg, Germany. (poster presentation)

Weiss, M., Bogatyrova O., ICGC project on Early Onset Prostate Cancer, Gerhäuser, C. and Plass, C., *Mechanisms of lncRNAs as mediators of DNA (de)methylation in prostate cancer*
EMBL non-coding genomes October 2015, Heidelberg, Germany. (poster presentation)

Weiss, M.,Feuerbach, L., Plass, C., Brors, B. and Gerhäuser, C., *Integrative analysis of tumor suppressor-ncRNAs associated with recurrently deleted regions in prostate cancer*
RNA@DKFZ November 2015, Heidelberg, Germany. (oral presentation)

10. ACKNOWLEDGMENTS

4 years older, a Germanized French-English accent and “some” white hairs later, I have reached the final chapter of my thesis! This thesis covers only a brief history of my time spent in this Division. Many persons helped me on many different projects in many ways and I would like to thank them through the following part.

THANK YOU:

- **Prof. Dr. Christoph Plass** and **Dr. Clarissa Gerhäuser** for offering me the opportunity to pursue my thesis in your respective groups, as well as for your invaluable support and guidance throughout the last four years. I especially appreciated having the freedom to develop and pursue my own scientific ideas during my time working with you.
- Special thanks to **Clarissa** for having helped me improving my writing skills through correction of my TAC reports, the review and this thesis.
- **PD. Dr. Popanda**, **Prof. Dr. Wiemann** and **Dr. Tamás Fischer** my TAC members for their support, time, and insightful help during my TAC committees.
- **Olga Bogatyrova** for her help with the integration of RNA sequencing and DNA methylation data presented in this thesis.
- **Brigitta Dapkute**, **Mariana Pardo-Contreras** and **Valentin Swingle**, the young and bright working force I had the pleasure to train and work with. You helped me so much; I can’t thank you enough for your help.
- To the colleagues who also became my friends, **Schmanneke**, **Schmusti**, **Schmoni**, **Schmoli** and **Schtania**. I miss you all very much as well as our troubleshooting, helping, singing, witheling, laughing, and sometimes more emotional moments! I appreciated your ongoing support in good and bad times. It wouldn’t have been such pleasant working place without all of you, my dear Sch-company!
- **Annette**, **Karin**, **Oli**, **Marion**, **Moni**, **Peter** and **Reinhard I** for your availability, help and support.
- **Constance Baer**, **Oliver Mücke**, **Chris Oakes**, **Olga Bogatyrova**, **Soo-Zin** and **Christopher Schmidt** for guiding me through the lab and learning the scientific techniques on your side during my first year.
- To my dear **Annika** for translating my summary and being such a great (spreading) desk sharer. I will miss listening to your phone calls with Mani about brötchen.

- To the best office mates ever **Tania**, **Justyna** and **Annika** for letting me take control over the heater ;).
- All present and past members of the Plass and Gerhäuser group, **Daniel**, **Daniela**, **David**, **Dieter**, **"little" Clarissa**, **Constance**, **Katharina**, **Maria**, **Michael**, **Miriam**, **Mridul**, **Pavlo**, **Reinhard II**, **Reka**, **Simin**, **Sina**, **Suzanna**, **Yassen**,..

Et plus personnellement un immense MERCI :

- A mes **parents** et **Jean-Pierre** de me suivre et de me trimballer moi et mes affaires aux quatre coins de l´Europe et comme vous le savez, bientôt ça continue..
  Un merci tout spécial à mon petit frère de spécialement venir à Heidelberg pour décoincer une chaussette de ma machine à laver. Oui, je me devais d'écrire ça dans ma thèse!
  Plus sérieusement, merci d'être toujours là pour moi.

- A **Bonbon**, de m'encourager, me soutenir, me conseiller, me supporter, de m'écouter me plaindre (très, très rarement..) tout au long de mes thèses, mes changements de villes, mes changements de projets et tellement plus.. Un énormissime merci d'avoir corrigé cette thèse et surtout de toujours trouver le temps et des solutions.
  Bien que loin des yeux, merci d'avoir été la personne la plus présente pour moi ces dernières années.

Merci à vous tous, I thank all of you

Mélanie