

**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

Diplom-Biology Olga Bogatyrova

from Cherkassy, Ukraine

Oral-examination: .....



# **Mutations in regulators of the epigenome and their effects on the DNA methylome**

**Referees:**

**Prof. Dr. Stefan Wiemann**

**Prof. Dr. Christoph Plass**





**Division of Epigenomics and Cancer Risk Factors**

**Head of Division: Prof. Dr. Christoph Plass**

**German Cancer Research Center (DKFZ)**

**Heidelberg, Germany**



*To all the people that helped me on the way...*



## Contributions

Parts of the introduction section of this thesis was prepared, based on the published review “Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer” by Plass et al. (2013) in the journal *Nature Genetic Reviews*, on which I am a co-author [1]. The text sections may therefore contain input from co-authors and figures in the respective sections are partially modified from the review, which is indicated in the caption and in accordance with the copyright policy of the publishers.

The **chapter 7** of this thesis was prepared for manuscript under the title “***Epigenetic deregulation of miRNAs activates oncogenic pathways in prostate cancer***” by Bogatyrova O\*, Wuttig D.\* et al. For the Results **chapter 7**, text phrases, selected tables, figures and the respective captions are based on the prepared manuscript and contain input from co-authors.

Parts of the material and methods section concerning the Illumina HumanMethylation450 BeadChip assay and the Human Sentrix expression array analysis were conducted by the Genomics & Proteomics Core facility of the DKFZ. Data analysis and cluster performance of HumanMethylation450 BeadChip data was conducted with big help and supervision of Dr. Yassen Assenov.



## **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.

Heidelberg,

(Olga Bogatyrova)



**TABLE OF CONTENTS**

TABLE OF CONTENTS .....	14
<b>CHAPTER 1. SUMMARY .....</b>	<b>18</b>
Zusammenfassung .....	20
<b>CHAPTER 2. LIST OF ABBREVIATIONS.....</b>	<b>22</b>
<b>CHAPTER 3. INTRODUCTION .....</b>	<b>25</b>
3.1. EPIGENETICS .....	25
3.1.1. Epigenetics and epigenetic alterations.....	25
3.1.2. DNA modifications .....	25
3.1.3. Histone modifications.....	27
3.2. Major regulators of the epigenome .....	28
3.2.1 Complexes of epigenetic regulators in cancer .....	29
3.3. Non-coding RNAs.....	29
3.4. Epigenetic alterations in cancer .....	33
3.4.1. Global DNA methylation alterations in cancer .....	34
3.4.2. Cancer drivers vs. passengers .....	37
3.4.3. Cancer driver mutations in the epigenome.....	37
3.4.4. Integrative analysis of cancer genomes and epigenomes.....	37
3.4.5. Disruption of ncRNAs in cancer .....	38
3.5. PanCancer studies .....	41
3.6. Prostate cancer .....	42
3.6.1. Deregulation of ncRNA in prostate cancer .....	42
<b>CHAPTER 4. AIMS OF THE THESIS.....</b>	<b>46</b>
<b>CHAPTER 5. MATERIAL AND METHODS .....</b>	<b>48</b>
5.1. General equipment, disposables and chemical substances .....	48
5.2. Primers .....	48
5.3. Tissue samples.....	48
5.4. Data sets .....	51
5.5. DNA methylation analysis .....	51
5.6. miRNA expression analysis .....	54
5.7. Data analysis .....	55

RESULTS.....59

**CHAPTER 6. Mutations in regulators of the epigenome and their effects on the DNA methylome**  
.....59

6.1. ALTERATIONS IN EPIGENETIC REGULATORS IN CANCER .....61

6.1.1. Putative driver epigenetic regulators.....63

6.1.2. Alterations in complexes of epigenetic regulators in cancer .....69

6.2. Identification patterns/clusters of DNA methylation in cancer .....70

6.3. Correlative analysis of methylome data with defects in epigenetic regulators .....75

6.4. Differently methylated regions in cancer .....75

6.5. Discussion of results and deliveribles .....75

**CHAPTER 7. Global integrative analyses of deregulated miRNAs in prostate cancer** .....81

7.1. Global deregulation of miRNA expression in early-onset prostate cancer (EO-PCA) .....82

7.2. Genetic and epigenetic alterations associated with miRNA downregulation in EO-PCA .... 89

7.3. Genetic and epigenetic aberrations associated with miRNA upregulation in EO-PCA..... 89

7.4. Translocations associated with miRNAs deregulation in EO-PCA .....89

7.5. Tumor-suppressive and oncogenic miRNAs in EO-PCA.....89

7.6. Expression profile of miRNAs in different groups of EO-PCA.....94

7.7. Promoter DNA-methylation associated with miRNA deregulation in LO-PCA .....95

7.7.1. Validation of promoter DNA-methylation associated with miRNAs deregulation in PCA...95

7.7.2. Epigenetically regulated miRNAs activate key oncogenic pathways in PCA.....97

7.8. PTEN inactivation by miRNA in PCA .....100

7.9. miRNAs promoter DNA-methylation in cancer .....100

7.10. Discussion of results.....101

**CHAPTER 8. Mechanisms leading to deregulation of lncRNA in prostate cancer – an integrative analysis** .....104

8.1. Expression profile of lncRNAs in EO-PCA .....105

8.2. Genetic aberrations associated with lncRNA deregulation in EO-PCA.....108

8.3. Promoter DNA-methylation associated with lncRNA deregulation in EO-PCA .....108

8.4. Correlation of DNA methylation and expression deregulation of lncRNAs in EO-PCA .....111

8.5. Promoter DNA- methylation associated with lncRNAs deregualtion in different stages of PCA .....114

8.6. DNA methylation of promoter/regulatory regions of lncRNAs in PANCAN .....	114
8.7. Discussion of results.....	114
<b>REFERENCES.....</b>	<b>116</b>
<b>APPENDIX.....</b>	<b>121</b>
Supplemental Tables.....	121
Supplemental Figures.....	130
<b>PUBLICATIONS DURING THE THESIS.....</b>	<b>134</b>
<b>ACKNOWLEDGMENTS .....</b>	<b>136</b>



## CHAPTER 1. SUMMARY

---

### Abstract

Genome-wide profiling for genetic alterations in cancer has identified mutations in genes that are associated with epigenetic programming of genomes for DNA methylation patterns, histone modifications patterns and the positioning of nucleosomes. Here a systematic evaluation of the available cancer genome profiling data established by large international consortia, in order to identify recurrently mutated genes or pathways was described. Using curated list of approximately 700 epigenetic regulators and currently available genome-wide datasets on genetic and epigenetic alterations in cancers, the distribution of alterations in epigenetic regulators was described. Epigenetic genes were classified as potential oncogenic or those with tumor-suppressor function based on the location of mutations relative to functional domains and their frequencies. A panel of 50 epigenetic genes, including: DNMTs, histones (*H3F3A*, *HIST1H3B*), histone editors (*KDM5C*, *KDM6A*) and writers (*MLLs*, *SETD2*, *EZH2*, *ATM*) that can promote epigenetic changes in cancer was identified. Using correlative analysis of publicly available methylation data with information on deregulated epigenetic driver genes, many identified subtype-specific methylation clusters were correlated with groups of up to 3 epigenetic regulators. This analysis provides a source for the identification and link between methylation groups and deregulated epigenetic genes.

Major cancer specific methylation changes have been observed in promoters and gene bodies. Tissue-specific cancer methylation differences have been located in enhancers and regulatory regions of non-coding RNAs. Based on identified results, the major mechanism of non-coding RNA deregulation in cancer has been investigated on independent data cohort. Using integrative analysis of non-coding RNA in early-onset prostate cancer, non-coding RNAs were classified as tumor-suppressive and oncogenic. About 120 novel prostate cancer specific non-coding RNAs that have been epigenetically deregulated have been identified.

Our study on the defects in regulators of the epigenome will help to understand mechanisms leading to distinct epigenetic patterns and will allow the molecular validation of defined correlations in experimental settings.



## Zusammenfassung

Genom-weite Studien nach genetischen Veränderungen in Krebs haben Mutationen in Genen gefunden, die für die epigenetische Programmierung von DNA-Methylierungsmustern, Histonmodifikationen und die Positionierung von Nukleosomen verantwortlich sind. Hier wird systematisch auf die existierenden Daten der großen, internationalen Konsortien zurückgeblickt, um rekurrent-mutierte Gene oder Signalwege zu identifizieren. Durch die Nutzung einer kuratierten Liste von etwa 700 epigenetischen Regulatoren und den zurzeit verfügbaren genomweiten Datensätze der genetischen und epigenetischen Veränderungen in Krebs, wurde die Verteilung der mutierten epigenetischen Regulatoren beschrieben. Epigenetische Gene wurden als potenzielle Onkogene oder Tumorsuppressoren klassifiziert, je nach Frequenz und Lokalisation der Mutationen relativ zu funktionellen Proteindomänen. Eine Palette von 50 epigenetischen Genen wurde identifiziert, die epigenetische Veränderungen in Krebs hervorrufen können, inklusive DNMTs, Histonen (*H3F3A*, *HIST1H3B*), Histonmodifikatoren (*KDM5C*, *KDM6A*) und Histonbeschrifter (*MLLs*, *SETD2*, *EZH2*, *ATM*). Mittels Korrelationsanalysen zwischen öffentlich zugänglichen Methylierungsdaten und der Information über deregulierte epigenetische Krebsgene, konnten subtyp-spezifische Methylierungsmuster mit Gruppen von bis zu 3 epigenetischen Regulatoren assoziiert werden. Diese Analyse bietet eine

Quelle für die Identifizierung zwischen Methylierungsgruppen und deregulierten, epigenetischen Genen. Wesentliche Krebs-spezifische Methylierungsveränderungen wurden in Promotoren und Genkörpern beobachtet. Gewebsspezifische Krebsmethylierungsunterschiede wurden in Enhancern und regulatorischen Regionen für nicht-kodierende RNAs lokalisiert. Basierend auf den identifizierten Ergebnissen wurden die wesentlichen Mechanismen der nicht-kodierenden RNA-deregulierung in Krebs in einem unabhängigen Datensatz untersucht. Durch eine integrative Analyse der nicht-kodierenden RNA in früh-auftretendem Prostatakrebs wurden nicht-kodierende RNAs als Tumorsuppressoren und Onkogene klassifiziert. Hierdurch wurden etwa 120 neue, prostata-spezifische nicht-kodierende RNAs, welche epigenetisch dereguliert sind, beschrieben.

Diese Studie über die Defekte in epigenetischen Regulatoren wird dabei helfen, die Mechanismen, die zu veränderten epigenetischen Mustern führen, zu verstehen und wird die molekulare Validierung von definierten Korrelationen erlauben.



## CHAPTER 2. LIST OF ABBREVIATIONS

Abbreviation	Full text
<b>3'UTR</b>	3'untranslated regions
<b>5caC</b>	5-carboxycytosine
<b>5fC</b>	5-formylcytosine
<b>5hmC</b>	5-hydroxymethylcytosine
<b>5mC</b>	5-methylcytosine
<b>ACC</b>	Adrenocortical carcinoma
<b>Ago</b>	Argonaute protein
<b>AKT1</b>	<i>V-Akt Murine Thymoma Viral Oncogene Homolog 1</i>
<b>ALT</b>	alternative lengthening of telomeres
<b>ANRIL</b>	antisense non-coding RNA in the INK4 locus
<b>BCL2</b>	B-cell CLL/lymphoma 2
<b>Bp</b>	base pair
<b>BRCA</b>	Breast invasive carcinoma
<b>CBR3-AS1</b>	carbonyl reductase 3 antisense RNA 1
<b>CDH1</b>	Cadherin-1
<b>CDH1</b>	Cadherin-1
<b>CDKN2A</b>	cyclin-dependent kinase inhibitor 2A
<b>CGIs</b>	CpG islands
<b>CIMP</b>	CpG island methylator phenotype
<b>CNA</b>	copy-number alterations
<b>CNV</b>	copy-number variation
<b>COAD</b>	Colon adenocarcinoma
<b>CpGs</b>	cytosine-guanine dinucleotides
<b>crasiRNAs</b>	Centrosome-associated RNAs
<b>CTBP1-AS</b>	C-terminal binding protein 1 antisense RNA
<b>DAPK1</b>	death-associated protein kinase 1
<b>DICER1</b>	dicer 1, ribonuclease type III
<b>DICER1</b>	dicer 1, ribonuclease type III
<b>DMP</b>	Differently methylated probe
<b>DMR</b>	Differently methylated region
<b>DMRs</b>	differently methylated regions
<b>DNMT1</b>	DNA methyltransferase 1
<b>DNMT3a</b>	DNA methyltransferase 3A
<b>DNMT3b</b>	DNA methyltransferase 3B
<b>DNMTs</b>	DNA methyltransferases
<b>EMT</b>	epithelial mesenchymal transition
<b>ENCODE</b>	Encyclopedia of DNA Elements
<b>EO-PCA</b>	early onset prostate cancer
<b>ETS</b>	erythroblast transformation-specific gene family
<b>ETS</b>	E26 transformation-specific
<b>EZH2</b>	Enhancer of zeste homolog 2
<b>EZH2</b>	Enhancer of zeste homolog 2
<b>GBM</b>	Glioblastoma multiforme
<b>GEM</b>	genetically engineered mouse model
<b>GEO</b>	Gene Expression Omnibus
<b>GSEA</b>	Gene-set enrichment analysis
<b>H3K27ac</b>	acetylation of lysine 27 of histone 3

<b>H3K4me3</b>	trimethylation of lysine 4 of histone 3
<b>H3K4me3</b>	histone 3 lysine 4 trimethylation
<b>H3K9me3</b>	trimethylation of lysine 9 of histone 3
<b>HCA</b>	Hierarchical cluster analysis
<b>HMM</b>	Hidden Markov Model
<b>HRG</b>	high rearrangement group
<b>HSCs</b>	hematopoietic stem cells
<b>HUZAR</b>	Heidelberg Unix Sequence Analysis Resources
<b>ICGC</b>	International Cancer Genome Consortium
<b>IDH1</b>	isocitrate dehydrogenase 1
<b>IntOGene</b>	Integrative Onco Genomics
<b>IPA</b>	Ingenuity Pathway Analysis
<b>Jak1</b>	janus kinase 1
<b>Kdm6a</b>	lysine-specific demethylase 6A
<b>KICH</b>	Kidney Chromophobe
<b>LAML</b>	Acute Myeloid Leukemia
<b>Linc00963</b>	long intergenic non-protein-coding RNA 963
<b>LINES</b>	long interspersed elements
<b>lncRNA</b>	long non-coding RNAs
<b>LRG</b>	low rearrangement group
<b>LTR</b>	long terminal repeat
<b>LUAD</b>	Lung adenocarcinoma
<b>LUSC</b>	Lung squamous cell carcinoma
<b>MAGE-1</b>	melanoma antigen family A1
<b>MeCP</b>	Methyl-CpG immunoprecipitation
<b>miRNA</b>	micro RNAs
<b>MITF</b>	microphthalmia-associated transcriptional factor
<b>MLH1</b>	MutL Homolog 1, Colon Cancer, Nonpolyposis Type 2 (E. Coli)
<b>moRNAs</b>	microRNA-offset RNAs
<b>mRNAs</b>	messenger RNAs
<b>MSY-RNAs</b>	MSY2-associated RNAs
<b>ncRNAs</b>	non-coding RNAs
<b>O/E</b>	observed over expected CpG ratio
<b>p15AS</b>	p15 antisense lncRNA
<b>PAAD</b>	Pancreatic adenocarcinoma
<b>PAM</b>	Partitioning Around Medoids
<b>PARs</b>	Promoter-associated RNAs
<b>PASRs</b>	promoter-associated small RNAs
<b>PCA</b>	Prostate cancer
<b>PCA</b>	principal component analysis
<b>PCA3</b>	prostate cancer antigen 3
<b>PCAT1</b>	prostate cancer-associated transcript-1
<b>PCAT29</b>	prostate cancer-associated transcript-29
<b>PCGEM1</b>	prostate cancer gene expression marker 1
<b>PIN</b>	intraepithelial neoplasia
<b>piRNAs</b>	piwi-interacting RNAs
<b>Pol II</b>	RNA-polymerase II
<b>PRAD</b>	Prostate adenocarcinoma
<b>PRC2</b>	Polycomb Repressive Complex 2
<b>PRC2</b>	polycomb repressor complex 2

<b>pre-miRNA</b>	precursor miRNA
<b>pri-miRNAs</b>	primary miRNA transcript
<b><i>PRNCR1</i></b>	prostate cancer non-coding RNA 1
<b>PROMPTs</b>	promoter upstream transcripts
<b><i>PTEN</i></b>	<i>phosphatase and tensin homolog</i>
<b><i>PTEN</i></b>	<i>Phosphatase and tensin homolog</i>
<b>qPCR</b>	Quantitative PCR
<b><i>RAS</i></b>	Rat sarcoma
<b>READ</b>	Rectum adenocarcinoma
<b>RISC</b>	RNA-induced silencing complex
<b>rRNA</b>	ribosomal RNA
<b>SAM</b>	S-adenosyl methionine
<b>SChLAP1</b>	second chromosome locus associated with prostate
<b>SCNA</b>	Somatic copy number alteration
<b>SCNA</b>	somatic copy number alteration
<b>sdRNAs</b>	Sno-derived RNAs
<b>SFEs</b>	selected functional events
<b>SFEs</b>	selected functional events
<b>SINEs</b>	short interspersed elements
<b>siRNAs</b>	Small interfering RNAs
<b>SKCM</b>	Skin Cutaneous Melanoma
<b>snoRNAs</b>	Small nucleolar RNAs
<b>SNV</b>	single-nucleotide variation
<b>sRNAs</b>	small RNAs
<b>SRs</b>	structural rearrangement
<b>STAD</b>	Stomach adenocarcinoma
<b><i>TARBP2</i></b>	TAR (HIV-1) RNA Binding Protein 2
<b><i>TARBP2</i></b>	TAR (HIV-1) RNA Binding Protein 2
<b>TCGA</b>	The Cancer Genome Atlas
<b>tel-sRNAs</b>	Telomere small RNAs
<b>TERRAs</b>	telomeric repeat-containing RNAs
<b>THCA</b>	Thyroid carcinoma
<b>TSGs</b>	tumor suppressor genes
<b>TSSa-RNAs</b>	TSS-associated RNAs
<b>TSSs</b>	transcriptional start sites
<b>T-UCRs</b>	Tissue-specific ultraconserved regions
<b>UCRs</b>	ultraconserved regions
<b>UCSC</b>	University of California Santa Cruz
<b><i>VHL</i></b>	von Hippel-Lindau tumor suppressor
<b>WGS</b>	whole-genome sequencing
<b>xiRNAs</b>	X-inactivation RNAs
<b><i>XPO5</i></b>	exportin 5

## CHAPTER 3. INTRODUCTION

---

### 3.1. Epigenetics

#### 3.1.1. Epigenetics and epigenetic alterations

The term epigenetics is derived from the Greek “epi”, meaning “above, over”, and defines its function in executing and regulating information encoded in the DNA sequence. The term “epigenotype” was initially coined by Conrad Waddington in 1942, when he described the “epigenetic landscape” as the gene expression control that results in differentiation of cells leading to a variety of different tissues [2]. In 1996, Riggs and colleagues referred to epigenetics as “mitotically and/or mitotically heritable changes in gene function that cannot be explained by changes in the DNA sequence” [3]. Nowadays, the term “epigenetics” is defined as heritable changes in gene regulation that occur without alterations in the DNA sequence [4] and covers DNA and histone modifications, chromatin remodeling changes, transcription factor activity as well as non-coding RNAs: micro RNAs(miRNA) and long non-coding RNAs (lncRNA) regulating gene activities [5],[6]. In combination with genetic alterations, the patterns of epigenetic modifications serve as epigenetic markers to represent gene expression and chromatin states. A tight regulation of epigenetic patterns is required during normal developmental processes and epigenetic modifications are crucial for packaging and interpreting the genome under the influence of physiological factors [7],[8]. Through aberrant promoter DNA methylation, opening of the chromatin, and non-coding RNA regulation, epigenetics is involved in altered gene regulation and chromosomal breaks leading to deletions, translocations and other rearrangements of the chromosome during disease development like cancer.

#### 3.1.2. DNA modifications

DNA modifications have been recognized as key epigenetic marks for the maintenance of cellular states. Such modifications include canonical 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxycytosine (5caC). DNA modifications are established and maintained by enzymes, involved in adding modifications (“writers”), modifying these marks (“editors”), or translating a mark for other molecules (“readers”). Canonical DNA-modification, 5mC, is established by “writers”, DNA

methyltransferases (DNMTs), using S-adenosyl methionine (SAM) as the methyl group donor. DNMT3a and DNMT3b are *de novo* methyltransferases, preferentially targeting unmethylated cytosines in the context of a cytosine-guanine dinucleotides (CpGs) to initiate methylation which can occur in early embryonic stem cells and developmental processes or in abnormal processes in cancer cells [9]. In contrast, DNMT1 acts as a maintenance methyltransferase predominantly recognizing and methylating hemimethylated DNA by copying methylation patterns from the parental strand onto the newly synthesized strand [10]. DNA methylation is important for early development of mammals as confirmed by embryonically or postnatal lethality in mouse models with knockout of single DNMTs [11].

Methylated cytosines are generally present in a CpG dinucleotide context. The human genome consists of about  $2.7 \times 10^7$  CpGs, mostly occurring within repeat sequences (>90 %), or intronic regions (<10 %) and 1-2 % in CpG dense regions like CpG islands (CGIs). About 80-90% of all CpGs are usually methylated in the mammalian genome [12]. CGIs were defined as regions of DNA of at least 200 base pair (bp) length with the proportion of Gs or Cs, referred to as “GC content,” greater than 50%, and observed to expected CpG ratio (O/E) greater than 0.6 [12]. The observed to expected ratio is calculated by dividing the proportion of CpG dinucleotides in the region by what is expected by chance when bases are assumed to be independent outcomes of

a multinomial distribution. The formula to calculate the O/E ratio is  $\frac{O}{E} = \frac{\frac{\#CpG}{N}}{\frac{\#C \cdot \#G}{N * N}}$  where **N** is the

number of bp in the segment under consideration. Various computer algorithms have been developed to create CGI lists satisfying the definition, however, this definition is somewhat arbitrary because the choice of the cutoffs has a great influence on what is considered an island. Alternatively, a Hidden Markov Model (HMM) based method was used to create a list of CGIs, detected by jointly thresholding the result posterior probabilities and is available in UCSC (genome.ucsc.edu) [13]. CGIs are enriched around transcriptional start sites and serve as regulatory regions for genes often overlapping with gene promoters [14] and often found in constitutively active and highly expressed genes [15]. The majority of CGIs are unmethylated in normal adult cells, while repetitive sequences are highly methylated. [16]

In healthy development, DNA methylation is associated with tissue specific gene regulation [17], silencing of repetitive elements [18] and genomic stability [19]. It plays an important role in developmental processes like X chromosome inactivation by silencing genes on the inactive X, or genomic imprinting in repressing either paternally or maternally derived genes [20].

Impairment of DNA methylation during early embryonic development results in early developmental death, which demonstrates the essential role of DNA methylation in embryogenesis and development [21]. The important role of DNMT activity in development was demonstrated in hematopoietic stem cells (HSCs) in which the DNA methylation machinery was disrupted. These cells are defective in self-renewal and lose their ability to give rise to multilineage hematopoiesis [22].

Also about one-quarter of all methylated Cs in embryonic stem cells is found in non-CG context, suggesting that embryonic stem cells may use different methylation mechanisms to affect gene regulation[23]. Such methylation showed enrichment in gene bodies and depletion in protein binding sites and enhancers, disappeared upon induced differentiation of the embryonic stem cells, and was restored in induced pluripotent stem cells.

During the last years, the field gained interest in the role of aberrant methylation in regions outside of promoters, like in CGI “shores” (regions up to 2 kb adjacent to CGIs), enhancers, intragenic CGIs and gene bodies. For CGI shores and enhancers, high methylation variability in different cell types was observed [24]. Nevertheless the main reason why these regions are so hypervariable remains unclear. There are proposed hypothesis either explaining that these regions are functionally unimportant or alternatively suggesting that these regions are needed for regulation and differentiation of distinct cell types and therefore are hypervariable. Intragenic CGIs can be methylated in a tissue-specific fashion and regulate expression of alternative transcripts [25] and they are highly affected by aberrant methylation in cancer [26]. Methylation in gene bodies can be in a tissue-specific manner [27], is normally associated with active transcription and influences the regulation of RNA splicing [28].

### **3.1.3. Histone modifications**

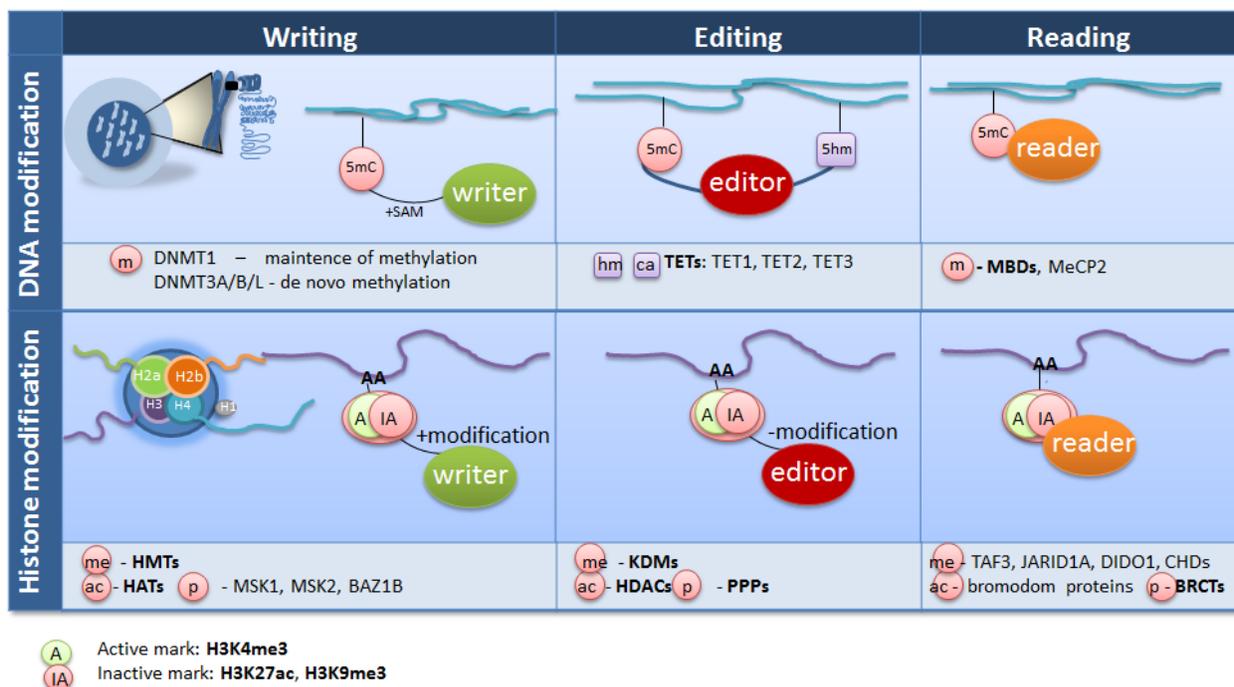
In eukaryotes the genomic DNA is packaged into higher chromatin structures in order to reduce the size, to create transcriptionally active and silent regions, to support DNA replication and to coordinate proper separation of genetic material to daughter cells. Chromatin is a complex of DNA and proteins structured in nucleosomes that consists of a 147bp DNA strand wrapped around a histone octamer composed by two histones of variant H2A, H2B, H3 and H4. . Histone octamers can be modified by a variety of modifications that preferentially are found at the histone tails, which extend out of the octamer. Modifications include mono-, di-, trimethylation, acetylation, phosphorylation or ubiquitination at defined positions. Histone modifications lead to

differences in the chromatin structure and, as a consequence, changes in gene expression. The best studied modifications are located on histones 3 and 4 and have been linked to certain functional states exerted by the genes nearby. Modifications determine if a chromosomal region is accessible for the binding of transcription factors or other regulatory molecules or if the gene *loci* are active or silent. For example active chromatin is characterized by trimethylation of lysine 4 of histone 3 (H3K4me3). Inactive promoter regions are marked by acetylation of lysine 27 of histone 3 (H3K27ac) or trimethylation of lysine 9 of histone 3 (H3K9me3) [29]. Histone acetylation represents usually an activating mark, whereas histone methylation can be activating or repressive depending on the modified amino acid.

### 3.2. Major regulators of the epigenome

Epigenetic modifications of DNA and histones, and/or alterations in chromatin-remodeling processes, determine active and repressive chromatin states of genes and of chromosomal regions. Such modifications can regulate the fine tuning of gene expression via promoter or enhancer methylation.

Enzymes that establish a mark on either DNA or the histone tail are termed 'writers'. These modifications can be removed or modified by 'editing' enzymes. The third class of enzymes includes the 'readers' of an epigenetic mark, which mediate the interaction of the mark with a transcriptional protein complex (**Figure 3-1**).



**Figure 3-1: Enzymes involved in DNA and histone modification pathways.** The top panel depicts DNA modifications, such as DNA methylation and demethylation, and the enzymes involved; the bottom panel shows histone modifications and the enzymes involved. Examples for each class of enzyme are given. 5hmC, 5-hydroxymethylcytosine; 5mC, 5-methylcytosine. A – active histone marks, IA- inactive marks.

### 3.2.1 Complexes of epigenetic regulators in cancer

Epigenetic gene regulation depends on the interplay of DNA methylation, histone marks and nucleosome positioning. Epigenetic marks are catalyzed by different epigenetic complexes. Such complexes consist of major core factors and axillary components. Polycomb group (PcG) complexes are epigenetic regulatory complexes that conduct transcriptional repression of target genes via modifying the chromatin. The two best characterized forms of PcG complexes, polycomb repressive complexes 1 and 2 (PRC1 and PRC2), are required for maintaining the stemness of embryonic stem cells and many types of adult stem cells. Target genes for PRCs are numerous and are changing with cell differentiation and during developmental processes. The major core component of PRC2 is EZH2 which is responsible for catalyzing di- and tri-methylation of Lys27 on histone H3 (H3K27me<sub>2/3</sub>) [30]. The other core PRC2 components are necessary for complex assembly and for proper enzymatic activity [31] [32].

### 3.3. Non-coding RNAs

Increasing evidence is showing that most epigenetic mechanisms of gene expression control include regulation by non-coding RNAs (ncRNAs), such as microRNAs (miRNAs), small RNAs (sRNAs) and long or large RNAs (lncRNAs) and other classes (Table 3-1). ncRNAs with different regulatory functions are recognized as common features of mammalian transcriptomes, especially after the discovery that most of the eukaryotic genomes are transcribed into RNAs that have no protein-coding potential [33, 34]. ncRNAs are able to direct the cytosine methylation and histone modifications for gene expression regulation. These molecules are very important in various epigenetic modification mechanisms such as transposon activity and silencing, position effect variegation, X-chromosome inactivation and paramutation, and will be explained in more details by each class.

#### microRNAs (miRNAs)

As best described class of ncRNAs, miRNAs, are small non-coding RNAs (about ~22 nucleotides in length) that mediate post-transcriptional gene silencing by controlling the translation or transcript stability of messenger RNAs (mRNAs). The majority of miRNAs are

transcribed by RNA-polymerase II (Pol II) as a primary miRNA transcript (pri-miRNAs) and further processed by the RNase ‘Drosha’ into the precursor miRNA (pre-miRNA). After active export to the cytoplasm, Dicer cleaves the pre-miRNAs into short mature miRNA fragments. The mature miRNA is loaded into the RNA-induced silencing complex (RISC), containing the single-stranded miRNA and Argonaute (Ago) proteins. By base-pairing of the miRNA to the 3′ untranslated regions (3′UTR) of target mRNAs, RISC can inhibit the expression of the target mRNA through mRNA degradation and blockade of translation [35]. Computational prediction reveals several hundred mRNA targets for each miRNA [36]. miRNAs are estimated to regulate translation over 60% of protein-coding genes. They are involved in the regulation of many processes including proliferation, differentiation, apoptosis and development. Whereas some miRNAs regulate specific individual targets, others can function as “master” regulators of a process, so key miRNAs regulate expression levels of hundreds of genes simultaneously. Many types of miRNAs regulate their targets cooperatively [37], [38].

### PIWI-interacting RNAs (piRNAs)

The newly identified small RNAs class of piwi-interacting RNAs (piRNAs) are Dicer-independent ncRNAs of 26-30bp in length which bind the PIWI subfamily of Argonaute family proteins that are involved in maintaining genome stability in germline cells. They are mapped mostly to repetitive elements, are important for transposon control in flies [39] and in silencing mechanisms of repetitive elements in vertebrates [40]. piRNAs are mapped to repeats in the genome, with most of them matching short interspersed elements (SINEs), long interspersed elements (LINEs) and long terminal repeat (LTR) retrotransposons [41].

**Table 3-1. Major classes of non-coding RNAs in mammals**

ncRNA class	Description
<b>Long (regulatory) non-coding RNAs (lncRNAs)</b>	The broadest class, lncRNAs, encompasses all non-protein-coding RNA species > ~200 nt, including mRNA-like ncRNAs. Their functions include epigenetic regulation, acting as sequence-specific tethers for protein complexes and specifying subcellular compartments or localization
<b>Small interfering RNAs (siRNAs)</b>	Small RNAs ~21–22 nt long, produced by Dicer cleavage of complementary dsRNA duplexes. siRNAs form complexes with Argonaute proteins and are involved in gene regulation, transposon control and viral defense
<b>microRNAs (miRNAs)</b>	Small RNAs ~22 nt long, produced by Dicer cleavage of imperfect RNA hairpins encoded in long primary transcripts or short introns. They associate with Argonaute proteins and are primarily involved in post-transcriptional gene regulation

<b>PIWI-interacting RNAs (piRNAs)</b>	Dicer-independent small RNAs ~26–30 nt long, principally restricted to the germline and somatic cells bordering the germline. They associate with PIWI-clade Argonaute proteins and regulate transposon activity and chromatin state
<b>Promoter-associated RNAs (PARs)</b>	A general term encompassing a suite of long and short RNAs, including promoter-associated RNAs (PASRs) and transcription initiation RNAs (tiRNAs) that overlap promoters and TSSs. These transcripts may regulate gene expression
<b>Small nucleolar RNAs (snoRNAs)</b>	Small RNAs (60–300 bp), traditionally viewed as guides of rRNA methylation and pseudouridylation. However, there is emerging evidence that they also have gene-regulatory roles
<b>X-inactivation RNAs (xiRNAs)</b>	Dicer-dependent small RNAs processed from duplexes of two lncRNAs, <i>XIST</i> and <i>TSIX</i> , which are responsible for X-chromosome inactivation in placental mammals
<b>Sno-derived RNAs (sdRNAs)</b>	Small RNA), some of which are Dicer-dependent, which are processed from small nucleolar RNAs (snoRNAs). Some sdRNAs have been shown to function as miRNA-like regulators of translation
<b>microRNA-offset RNAs (moRNAs)</b>	Small RNAs ~20 nt long, derived from the regions adjacent to pre-miRNAs. Their function is unknown
<b>tRNA-derived RNAs</b>	tRNAs can be processed into small RNA species by a conserved RNase (angiogenin). They are able to induce translational repression
<b>MSY2-associated RNAs (MSY-RNAs)</b>	MSY-RNAs are associated with the germ cell-specific DNA/RNA binding protein MSY2. Like piRNAs, they are largely restricted to the germline and are ~26–30 nt long. Their function is unknown
<b>Telomere small RNAs (tel-sRNAs)</b>	Dicer-independent ~24 nt RNAs principally derived from the G-rich strand of telomeric repeats. It has been suggested a role in telomere maintenance
<b>Centrosome-associated RNAs (crasiRNAs)</b>	A class of ~34–42 nt small RNAs, derived from centrosomes that show evidence of guiding local chromatin modifications

### Sno-derived RNAs (snoRNAs)

snoRNAs are intermediate-sized ncRNAs (60–300 bp). They are components of small nucleolar ribonucleoproteins (snoRNPs), which are complexes that are responsible for sequence-specific post-transcriptional modification of ribosomal RNA (rRNA) to facilitate rRNA folding and stability [42]. The sequences of snoRNAs are responsible for targeting the assembled snoRNPs to a specific target.

### Long-noncoding RNAs (lncRNAs)

As biggest class of ncRNAs, lncRNAs are a heterogeneous group of non-coding transcripts more than 200 nt long that are involved in many biological processes. This class of ncRNA makes up the largest portion of the mammalian non-coding transcriptome. Various mechanisms of transcriptional regulation of gene expression by lncRNAs have been proposed. Among these,

lncRNAs are known to mediate epigenetic modifications of DNA that act by recruiting chromatin remodeling complexes to specific *loci* [43].

Moreover, growing evidence indicates that lncRNAs are always transcribed from an imprinted region of mammalian genomes, which is very similar to the mechanisms used by the large ncRNA *XIST* in the X-chromosome inactivation in female cells, suggesting that they are also key players in this type of silencing process in autosome chromosomes [44]. ncRNAs are also involved in silencing of repeats in the genome mediated by small RNAs. Another function of ncRNAs was described by identification of *XIST* ncRNA. *XIST* is a very long ncRNA, 17 kb in length, that is able to physically bind and form complexes with the chromatin that is surrounding one of the X-chromosomes in females (for review see [45]). It was also shown that *XIST* expression can be controlled by epigenetic mechanisms such as DNA methylation and normal patterns of methylation are disrupted in cancer cells [46]. X-chromosome inactivation is a mechanism used by mammals for dosage compensation in female genomes: one copy of their sexual chromosome is inactivated by DNA methylation. ncRNA *XIST* was reported to be a major regulator of this process in association with proteins of the chromatin. Another long non-coding transcript termed *TSIX* is expressed from the strand opposite to *XIST* and is able to control its levels in the inactivation process by epigenetic mechanisms [43]. Both transcripts can directly influence the chromatin modifications and thereby alter the levels of proteins that bind to the DNA in the X-chromosome [47].

ncRNAs were also reported to be involved in expression control of imprinted regions. DNA imprinting is an epigenetic mechanism of regulation of monoallelic expression in autosome chromosomes of mammals. Generally, imprinting involves DNA methylation of one allele and it was reported that such regions have ncRNAs mapped or nearby and they are implicated in the silencing process. Deletion or knockdown of ncRNA transcripts in imprinted regions can have deleterious effects [48]. Imprinting clusters, such as *Igf2r*, *Kcnq1*, *Gnas*, and *Pws*, contain ncRNA in the anti-sense orientation of one of the silenced genes [49].

Recently, a long anti-sense ncRNA named *HOTAIR* was identified illuminating new mechanisms whereby transcription of ncRNAs dictates transcriptional silencing of a distant chromosomal region [50]. *HOTAIR* was implicated in mediating epigenetic silencing of a chromosomal domain in *trans* by guiding specific histone modifications [50]. There are still many questions about the mechanism of acting of such large ncRNAs in affecting the expression of genes in other *loci*, but this study suggests that ncRNAs might be essential epigenetic regulators and the same type of regulation may be occurring in other regions of the human genome.

Another class of lncRNAs are lincRNAs, which are transcribed from intergenic regions. These transcripts are identified by searching for chromatin signatures that are associated with active transcription in the regions across which transcriptional elongation takes place [51]. lincRNAs are found to regulate both the expression of neighboring genes and distant genomic sequences [52], [53].

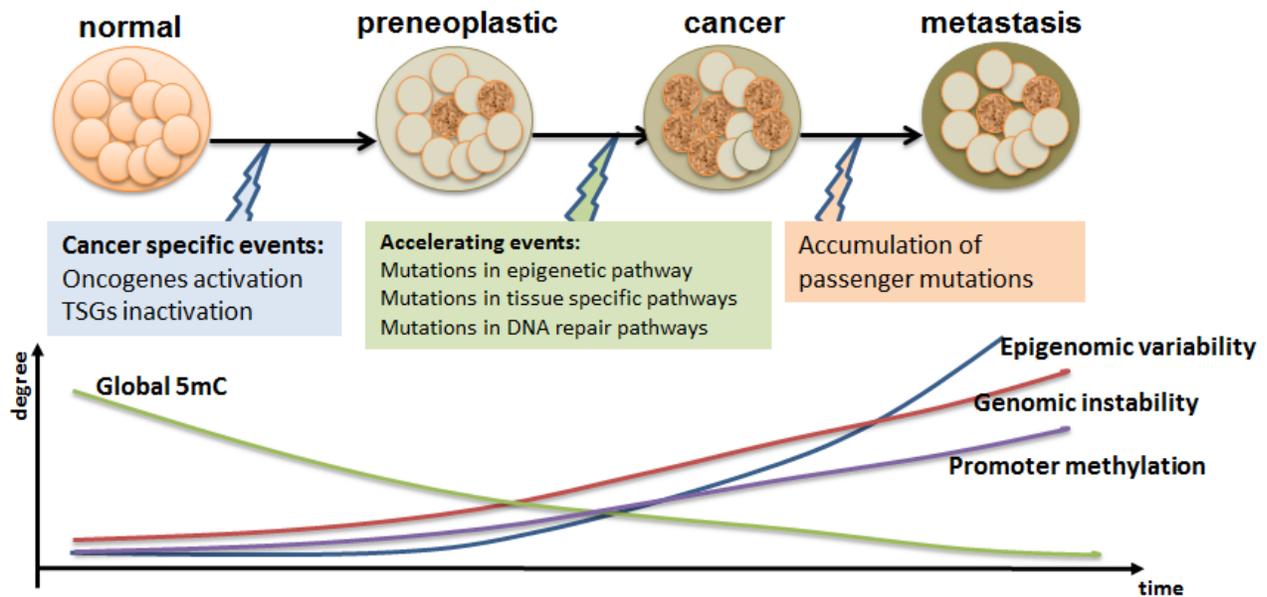
A final class of lncRNAs is those that are transcribed from ultraconserved regions (UCRs). UCRs are conserved DNA segments that are longer than 200 bp [54]. There are ~ 500 described UCRs, some of which overlap with coding exons, although it is believed that more than half of them do not encode any protein [55]. T-UCRs are a subclass of UCRs, that are transcribed and expressed in normal tissues either ubiquitously or in a specific pattern [56]. The functions of T-UCRs are remains unknown, but it has been shown that some of them bind to miRNAs [56].

### Other types of ncRNAs

Many classes of ncRNA have been described that are associated with the transcriptional start sites (TSSs) of genes: for example, promoter-associated small RNAs (PASRs), TSS-associated RNAs (TSSa-RNAs), promoter upstream transcripts (PROMPTs) and transcription initiation RNAs (tiRNAs) (**Table 3-1**). Their biological functions remain unknown, but it is hypothesized that they are involved in transcription regulation. Another type of lncRNA, known as telomeric repeat-containing RNAs (TERRAs), is transcribed from telomeres. TERRAs help to maintain the integrity of telomeric heterochromatin by regulating telomerase activity and are involved in the mechanism called alternative lengthening of telomeres (ALT) [57].

## 3.4. Epigenetic alterations in cancer

Epigenetic aberrations have been well established in cancer and occur in several other diseases, including diabetes, lupus, asthma and a variety of neurological disorders [58]. Feinberg and Vogelstein already in 1983 discovered that one major difference between cancer cells and healthy counterparts was the aberrant substantial DNA hypomethylation of genes in cancer patients with progressive hypomethylation in metastasis [59]. Epigenetic alterations in cancer have been investigated for more than 25 years, both on the single-gene level and on the genome-wide level. It was shown that both genetics and epigenetics cooperate during tumorigenesis and especially DNA methylation changes within CGIs have been described to occur frequently in cancer (**Figure 3-2**).



**Figure 3-2: Main events during cancerogenesis.** Schematic overview of steps during tumorigenesis from normal cell, passing cancer specific and tissue specific events (upper part). Cancerogenesis is characterized by loss of global DNA methylation and increase of genomic instability and epigenomic variability and promoter methylation (lower part).

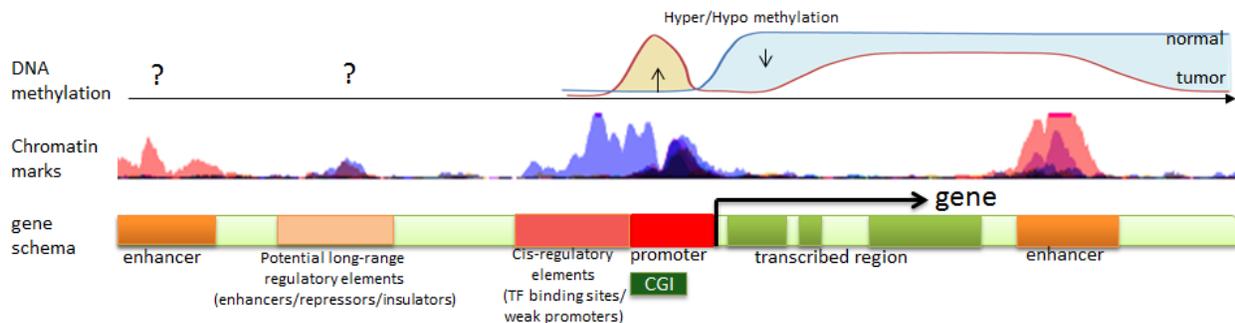
### 3.4.1. Global DNA methylation alterations in cancer

Hypermethylation of a GCI promoter is associated with gene silencing, which has been demonstrated for numerous tumor suppressor genes (TSGs). However, cancer genomes are also characterized by global loss of 5mC (hypomethylation), which mainly occurs at pericentromeric satellite DNA or other repetitive elements, including gene regulatory sequences. In normal cells, repetitive elements are highly methylated in the genome, which ensures integrity and stability of the genome. In addition, hypomethylation loosens the chromatin structure, leading to chromosomal instability such as translocations or deletions [1]

A **gain of methylation** is detected at the promoters of cancer related genes e.g. tumor suppressor genes and can lead to silencing of the respective gene (**Figure 3-3**). The first cancer-associated hypermethylation of a CGI of a TSG was reported in 1989 for the well-known retinoblastoma gene *RB1* [60, 61] and shortly after causally linked to TSG silencing in 1993 [62]. In the mid-90s, the importance of epigenetics in cancer development was further strengthened by the finding that known TSGs are preferentially or even exclusively silenced by DNA hypermethylation. Early examples of silenced TSGs were *CDKN2A* (cyclin-dependent kinase

inhibitor 2A) [63-65], *VHL* (von Hippel-Lindau Tumor Suppressor) [66], *MLH1* (MutL Homolog 1, Colon Cancer, Nonpolyposis Type 2 (E. Coli)) [67], and *DAPK1* (death-associated protein kinase 1) – a list of genes that has expanded ever since [68]. With the advent of genome profiling technologies, there are a lot more genes known to be aberrantly methylated in different tumor types.

**Loss of DNA methylation** was described already in 1982 and affects mainly regions that are highly methylated in normal tissue such as repetitive sequences or pericentromeric regions. Consequently this can lead to a more open chromatin structure, which might cause chromosomal breakage, aberrant rejoining or rearrangements [69]. Furthermore, loss of methylation in gene bodies could result as well in deregulation of gene expression, thereby contributing to tumorigenesis (**Figure 3-2-2**). Examples for hypomethylation of oncogenes include *RAS* (Rat sarcoma) in colonic adenocarcinomas and small cell lung cancer [70], *MAGE-1* (melanoma antigen family A1) [71] and *BCL2* (B-cell CLL/lymphoma 2) in B-cell chronic leukemia [72].



**Figure 3-3: Epigenetic regulation in cancer on gene-level.** A schematic view of epigenetic regulation of a gene in cancer vs. normal tissue represents exemplary “gene” with all regulatory elements like enhancers, promoters and transcribed regions, marked by different colors (lower part). Epigenetic changes in tumor vs. normal are plotted separately for DNA methylation as well as for chromatin marks (blue color denotes normal, red color – tumor). In DNA methylation plot depicts DNA hypermethylation of gene promoter and DNA hypomethylation of gene body in cancer.

Similar patterns of DNA methylation were observed across tumors from the same tissue types that indicated for specific mechanism, which causes such nonrandom pattern of methylation. These patterns can either be a reflection of the methylome seen in the cell of origin of a given tumor type or could be a result of evolving cancer cells further modifying the epigenome as a consequence of mutations in epigenetic enzymes [73]. Tumor methylome profiles have been used to subdivide tumors on tumor subgroups in order to better understand the tumor biology and mechanisms leading to tumor development. Medulloblastoma is a great example, where

four methylation patterns reflect differences in the cell of origin as well as in the mechanism of methylation deregulation [74]. DNA hypermethylation of a distinct set of genes was first described in colon cancer and designated as CpG island methylator phenotype (CIMP) [75]. This separation of CIMP positive and CIMP negative tumors is in the meantime also described in several other cancers including glioblastoma [76], gastric cancer [77], pancreatic adenocarcinoma [78], hepatocarcinoma [79], and in acute lymphocytic leukemia [80] and prostate [81]. The CIMP phenotype could reflect the differential activities of epigenetic enzymes in these subgroups shaping the epigenome in different ways.

### **3.4.2. Cancer drivers vs. passengers**

Recent genome sequencing activities by international consortia identified more somatic abnormalities in the cancer genome, but it is becoming clear that a fair amount have made no contribution in carcinogenesis. To address this phenomenon, the terminology that groups somatic mutations into two categories, **drivers** and **passengers** was developed. Stratton et al. described 'driver' mutations as those that confer growth advantage on the cells carrying them and have been positively selected during the evolution of the cancer. Remaining mutation were assigned as 'passengers' that do not confer growth advantage, but happened to be present in an ancestor of the cancer cell when it acquired one of its drivers [82]. Passenger mutations are found within the cancer genome since somatic mutations often occur during cell division without functional consequences. Hence, a cell that contains a passenger mutation will clonally expand carrying the mutation within all of the cells from that point forward.

The ability to discern between the two types of mutations can lead to a deeper understanding of cancer biology and empower the development of cancer therapeutics. But a complexity of cancer genomes makes it a key challenge nowadays. There are different tools/strategies used for investigation driver mutations. For example, one strategy exploits a number of structural signatures associated with mutations that are under positive selection. Such approach was fruitfully adopted in the past to identify most somatic cancer genes in studies of selected genome regions. Another way to find drivers is leveraging large cohorts of samples and using recurrence as an indicator of selection and to look for these signals of positive selection in the pattern of somatic mutations in genes across different tumors (relevant methods: MuSiC [83] or MutSig [84], InVex algorithm [85]). An alternative approach is to use functional annotation to infer driver status (relevant methods: IntOGen, OncodriveFM and OncodriveCLUST). There are around 100 genes that are known cancer drivers in literature, mostly assigned based on frequency, like

TP53, the gene encoding tumor protein p53 [86] or Kirsten rat sarcoma viral oncogene homolog (KRAS) [87].

### 3.4.3. Cancer driver mutations in the epigenome

Recent genome sequencing activities by international consortia identified defects in epigenetic enzymes that are responsible for the establishment of global epigenetic patterns, involving DNA methylation, histone modifications and chromatin remodeling [5, 29, 88]. Frequent and recurrent mutations have been described as the K27M and G34R/V mutations in the histone H3.3 variant H3.3A (encoded by *H3F3A*) in pediatric glioblastoma or chondrosarcomas [89, 90], isocitrate dehydrogenase 1 (*IDH1*) mutations in gliomas [91, 92], or *de novo* DNA methyltransferase 3A (*DNMT3a*) mutations in acute myeloid leukemia and glioblastoma [93, 94]. Recent studies revealed 125 pan-cancer driver genes recurrently mutated which are acting in epigenetic pathways [95]. Some of these mutations associate with distinct subgroups of tumors characterized by defined epigenetic patterns and clinical phenotypes. A different study identified 102 pan-cancer recurrent focal amplified/deleted regions, for which currently no known oncogenes/tumor suppressor genes have been identified. Interestingly these regions are enriched for genes involved in regulation of epigenetic patterns [96]. It was suggested that candidate driver mutations acting through global reprogramming of epigenetic patterns [1].

### 3.4.4. Integrative analysis of cancer genomes and epigenomes

In the past, a tumor genome study mainly focused on either genetic or epigenetic events. Nowadays this strategy was changed to a more comprehensive analysis of cancer genomes that includes different data types, like: gene mutations, copy-number aberrations, structural variations, epigenetic patterns and expression profiles of non-coding RNAs. But there is a huge lack of unified bioinformatic tools for such integrative analysis due to different approaches for performing the integration of data. Some tools are available to support visualization and pathway analysis of preselected cancer genes; however, a tool that encompasses all analyses is still needed. Table 3-2 provides a list of selected bioinformatic tools which allows the integration of at least two types of data sets.

Table 3-2. Selected web-based bioinformatic tools and web services for integrative cancer genome analysis

Bioinformatic tool or Webservices	Database used	Webservice or tool	Upload of data	Gene search	Chromosomal region search	mRNA expression	SNV	CNV	Methylation	miRNA expression	Protein	Pathways
cBioPortal for Cancer Genomics	TCGA	Webservice	-	✓	-	✓	✓	✓	-	-	✓	✓
PARADIGM, Broad GDAC Firehose	TCGA	Webservice	✓	✓	-	✓	✓	✓	✓	-	-	✓
WashU Epigenome Browser	ENCODE	Webservice	✓	✓	✓	✓	✓	✓	✓	-	-	✓
UCSC Cancer Genomics Browser	UCSC	Webservice	✓	✓	✓	✓	✓	✓	✓	✓	-	-
The Cancer Genome Workbench	TCGA	Webservice	-	✓	✓	✓	✓	✓	✓	-	-	-
EpiExplorer	ENCODE and ROADMAP	Webservice	✓	✓	✓	✓	✓	✓	✓	-	-	-
EpiGRAPH	ENCODE	Webservice	✓	✓	✓	✓	✓	✓	✓	-	-	-
Catalogue of Somatic Mutations in Cancer (COSMIC)	TCGA and ICGC	Webservice	-	✓	-	-	✓	✓	-	-	-	-
PCmtl, MAGIA, miRvar, CoMeTa etc	GEO and TCGA	Webservice	✓	✓	-	✓	-	-	-	✓	-	✓
ICGC	ICGC	Webservice	-	✓	-	✓	✓	✓	-	-	-	-
Genomatix	User defined	Tool	-	✓	-	✓	✓	✓	✓	-	-	✓
Caleydo	TCGA	Tool	-	✓	✓	✓	✓	✓	✓	✓	-	✓
Integrative Genomics Viewer (IGV)	ENCODE	Tool	-	✓	✓	✓	✓	✓	✓	-	-	-
iCluster and iClusterPlus	User defined	Tool	-	✓	-	✓	-	✓	-	-	-	-

CNV, copy-number variation; ENCODE, Encyclopedia of DNA Elements; ICGC, the International Cancer Genome Consortium; GDAC, Genomic Data Analysis Center; GEO, Gene Expression Omnibus; miRNA, microRNA; SNV, single-nucleotide variation; TCGA, The Cancer Genome Atlas; UCSC, University of California Santa Cruz.

Website with links for integrated analysis of microRNA and mRNA expression.

### 3.4.5. Disruption of ncRNAs in cancer

#### miRNAs in cancer

The roles of ncRNAs in tumorigenesis have most thoroughly been studied with respect to miRNAs [97]. Deregulated miRNAs multiply their influence on thousands of target genes many of them with cancer relevant functions. miRNA expression profiles differ between normal tissues and the tumors that are derived from them and also between tumor types [97],[98]. miRNAs can act as oncogenes or as tumor suppressors and can have key functions in tumorigenesis. Recently, by efforts of TCGA (The Cancer Genome Atlas) and ICGC (International Cancer

Genome Consortium) consortia, the miRNA profiles of different cancer types were revealed. There are few reports on the mechanisms that lead to miRNA deregulation [16],[99] e.g. genetic (such as somatic mutations or deletions/amplifications) or epigenetic events (changes in promoter methylation. Most of the previous studies were based on single (or a few) candidate miRNAs. Examples include promoter methylation for *miR-23b*, which results in activation of the proto-oncogenes [100] or hypermethylation of the *miR-26a* leading to activation of the histone modifier *EZH2* (Enhancer of zeste homolog 2) [101] in prostate cancer (PCA) or promoter hypermethylation of *miR-200* family, that leads to reduction of *CDH1* (Cadherin-1) in cancer (see full list of epigenetically disrupted miRNAs in Table 3-3) [102]. One of the first associations of miRNA deregulation due to a genetic event was reduced miR-15 and miR-16 expression in more than 50% of B cell chronic lymphocytic leukemia as a result of chromosome 13q14 deletion [103]. Interestingly, miRNAs are frequently located in fragile sites (heritable chromosomal sites which tend to form a gap or constriction and may tend to break) that are involved in ovarian and breast carcinomas and melanomas [104].

Even though there are reports on mechanisms that can lead to specific miRNA deregulation, it remains unclear what the major mechanisms of miRNAs deregulation in cancer are. Global genome wide data sets from ICGC or TCGA now allow new integrative approaches to decipher the mechanisms for each miRNA in different tumor entities. Such integrative analysis will need to cover all different data level and allow clarifying all mechanisms of miRNAs deregulation and/or finding the unique one.

**Table 3-3. Examples of miRNAs disrupted by genetic or epigenetic means in cancer**

miRNA name	Disruption	Consequence	Cancer type
<b>miR-124a</b>	CGI hypermethylation	<i>CDK6</i> overexpression	Colon, gastric, haematological
<b>miR-34b and miR-34c</b>	CGI hypermethylation	Metastasis	Many different tumor types
<b>miR-148a</b>	CGI hypermethylation	Metastasis	Colon, melanoma, breast
<b>miR-9</b>	CGI hypermethylation	Metastasis	Colon, melanoma, head and neck
<b>miR-200c</b>	CGI hypermethylation	EMT	Colon, breast, lung
<b>miR-141</b>	CGI hypermethylation	EMT	Colon, breast, lung
<b>miR-205</b>	CGI hypermethylation	EMT	Bladder
<b>miR-196b</b>	CGI hypermethylation	Unknown	Gastric
<b>miR-129-2</b>	CGI hypermethylation	<i>SOX2</i> overexpression	Colon, endometrial,

			gastric
<b>miR-137</b>	CGI hypermethylation	<i>CDC42</i> overexpression	Colon, head and neck
<b>miR-151</b>	Genomic gain	Metastasis	Hepatocellular carcinoma
<b>miR-517c and miR-520g</b>	Genomic gain	<i>WNT</i> upregulation	Neuroectodermal brain tumors
<b>miR-106b-25</b>	Genomic gain	<i>p21</i> and <i>BIM</i> depletion	Oesophageal adenocarcinoma
<b>miR-15 and miR-16</b>	Genomic deletion	<i>BCL2</i> overexpression	Haematological

EMT, epithelial-to-mesenchymal transition, table was modified from [105].

The recent tumor-specific genetic defects in the miRNA-processing machinery, such as in the genes that encode *TARBP2* (TAR (HIV-1) RNA Binding Protein 2) [106], *DICER1* (dicer 1, ribonuclease type III) [107] and *XPO5* (exportin 5) [108] was highlighted the relevance of these pathways in cellular transformation, in which such defects contribute to explain miRNA dysregulation in cancer. The miRNA expression profile of human tumors was characterized by a general defect in miRNA production that results in global miRNA downregulation. Such study indicated an alternative mechanism of miRNA deregulation in cancer.

### lncRNAs in cancer

Among various examples of the involvement of lncRNAs in cancer, the role of *HOTAIR* in human cancer is the best studied [109]. In epithelial cancer cells, *HOTAIR* overexpression causes genome-wide redistribution of promoter occupancy by Polycomb Repressive Complex 2 (PRC2) and H3K27me3. *HOTAIR* binds to the PRC2 thus preventing the repression of polycomb-targeted genes leading to changes in their expression across the genome. Cancer invasiveness is decreased when *HOTAIR* expression is lost, showing opposite effect compare to cells with higher than usual levels of PRC2 activity. Potentially *HOTAIR* might have an active role in modulating the cancer epigenome and mediating cell transformation. A similar function has been postulated for some other lincRNAs, such as lincRNA-p21, which functions as a repressor in p53-dependent transcriptional responses [52]. The *p15* antisense lncRNA, *p15AS*, which was first identified in human leukemia, has also been shown to induce the silencing of the *p15* tumor suppressor gene *locus* by inducing the formation of heterochromatin [110]. Recently, 121 prostate cancer associated intergenic non-coding RNA transcripts (termed the PCAT family) was identified, by novel tools for de novo identification of ncRNA transcripts from RNA-seq data (will be discussed in more details in chapter 3.4.1) [111].

### 3.5. PanCancer studies

In the last 10 years, the high increase in cancer data collection was led by the efforts of two major cancer consortia: TCGA and ICGC. Their major goal is to obtain a **comprehensive** description of **genomic, transcriptomic and epigenomic changes** in more than **50 different tumor types and/or subtypes** which are of clinical and societal importance across the globe. Recently, both consortia have raised their interest to perform comparisons between different tissue types to identify similarities and differences. PanCancer or PANCAN initiatives seek to combine analysis across tumor types in order to identify both similarities and differences in genomic/epigenomic/transcriptomic alterations. By combined efforts and expertise of different groups PANCAN initiative will also develop new bioinformatics tools and platforms, providing a foundation that should prove useful in future large-scale analysis projects. There are different threads for PANCAN analysis, such as identification of mutational drivers, network models, exposures and pathogens, data discovery and future directions.

The TCGA PanCancer initiative examines the similarities and differences among the genomic and cellular alterations found in the first dozen tumor types to be profiled by TCGA. The hierarchical classification of ~500 selected functional events (SFEs) of 3,299 TCGA tumors from 12 cancer types had shown two classes of tumors: either characterized by mutations (M class) or by copy number changes (C class) [112]. This distinction was clearest at the extremes of genomic instability, indicating the presence of different oncogenic processes. In another study, somatic copy number alteration (SCNA) patterns in 4,934 cancers from TCGA were characterized. In 37% of cancer samples, whole-genome doubling, associated with genetic alterations (such as *TP53* mutations, *CCNE1* amplifications and alterations of the *PPP2R* complex) was observed [96]. SCNAs that were internal to chromosomes tended to be shorter than telomere-bounded SCNAs, suggesting different mechanisms underlying their generation. Also recurrent focal SCNAs were observed in 140 regions, including 102 without known oncogene or tumor suppressor gene targets and 50 with significantly mutated genes. Amplified regions without known oncogenes were enriched for genes involved in epigenetic regulation [96]. In a different PANCAN TCGA study, 127 significantly mutated genes from well-known (for example, mitogen-activated protein kinase, phosphatidylinositol-3-OH kinase, Wnt/ $\beta$ -catenin and receptor tyrosine kinase signaling pathways, and cell cycle control) and emerging (for example, histone, histone modification, splicing, metabolism and proteolysis) cellular processes in cancer were identified [113]. The average number of mutations in these significantly mutated genes varies across tumor types; most tumors have two to six, indicating that the number of driver mutations required during oncogenesis is relatively small. Mutations in transcriptional

factors/regulators show tissue specificity, whereas histone modifiers are often mutated across several cancer types [113]. The results were also confirmed by recently identified 291 high-confidence cancer driver genes [114]. Such driver list was generated based on 3,205 tumors from 12 different cancer types. Among those genes, some have not been previously identified as cancer drivers and 16 have clear preference to sustain mutations in one specific tumor type.

Using molecular profiles of >3,000 tumors from 11 human cancer types in TCGA, systematically analyzed expression of miRNAs and mRNAs across cancer types to infer recurrent cancer-associated miRNA-target relationships was performed. miRNAs with recurrent target relationships were frequently regulated by genetic and epigenetic alterations across the studied cancer types [115]. Novel cancer specific miRNAs and miRNA families were identified, including miR-29 family, which recurrently regulates active DNA demethylation pathway members *TET1* and *TDG*.

### **3.6. Prostate cancer**

Prostate cancer (PCA) is the most common type of cancer in males and the second leading cause of death from cancer among men [116]. PCA has a rather slow progression and is generally considered characteristic for elderly males. Nonetheless, approximately 2% of tumors are diagnosed in men of 50 years or less [117]. This is referred to as early onset prostate cancer (EO-PCA). EO-PCA is analyzed as a separate entity by the German ICGC and should help to better understand PCA biology. The comparison analysis of age-related differences of structural rearrangement (SRs) has reported that EO-PCAs harbored androgen-regulated ETS (erythroblast transformation-specific gene family) gene fusions including *TMPRSS2:ERG* in prevalence, whereas elderly PCA carried primarily non-androgen SRs [118].

PCA genome-wide methylation studies have identified large numbers of differently methylated regions (DMRs) [119] [120], but the mechanism that lead to aberrant DNA methylation patterns remains unknown.

The clinical spectrum of PCA ranges from indolent tumors requiring no therapy to highly aggressive and often metastatic disease [121]. This clinical heterogeneity is caused by a complex pattern of genetic and epigenetic alterations, which are currently being uncovered by the ICGC and other projects. These alterations include the disruption of the chromatin modifier *CHD1* and the tumor suppressor gene *PTEN* (phosphatase and tensin homolog) by complex

mutational processes including deletions, mutations, translocations, microRNAs [118, 119, 122], or the activation of ETS transcription factors through structural rearrangements [119, 123].

It is well accepted that epigenetic alterations including histone modifications and DNA methylation participate in creating altered gene expression patterns [124]. In PCA, epigenetic aberrations are widespread and contribute to the deregulation of cellular processes, e.g. hormonal response, cell cycle or DNA damage repair [125, 126]. Epigenetic silencing of target genes by promoter hypermethylation is detected in more than 90% of PCA early during tumorigenesis. These include *GSTP1*, which is involved in the detoxification of potential carcinogens, and the tumor suppressors *RASSF1A* and *APC* [125, 127]. In addition, histone-modifying enzymes like *HDAC1*, *EZH2* or *LSD1* are upregulated in PCA [126]. Genome-wide methylation studies have identified large numbers of aberrantly methylated regions in PCA, demonstrating the complexity of epigenetic alterations and supporting their important role in PCA. Börno *et al.* have reported more than 147,000 cancer-associated epigenetic alterations [101] and Kim *et al.* have identified 2,481 promoter regions differentially methylated in PCA [120].

### 3.6.1. Deregulation of ncRNA in prostate cancer

By large-scale RNA-sequencing efforts of TCGA and ICGC, hundreds of novel cancer associated ncRNAs were identified. A set of 121 PCA-associated intergenic non-coding RNA transcripts were found, and termed due to PCA relevance as PCAT family [111]. PCATs have different role in prostate cancer development and can be oncogenic (like PCAT1, PCAT3, PCAT8) or have tumor-suppressor role (like PCAT29) (**Table 3-4**). One of the first described lncRNA in PCA is PCA3, originally described in 1999 [128] as up-regulated in PCA as compared to normal and benign prostate hyperplasia. This lncRNAs was used to develop a biomarker for early detection of PCA. The best studied examples of lncRNAs with androgene receptor (AR) co-activator functions are *PCGEM1* (prostate cancer gene expression marker 1) and *PRNCR1* (prostate cancer non-coding RNA 1), both overexpressed in PCA [129]. *PRNCR1* binds to AR and recruits *DOT1L* methyltransferase, which methylates AR and allows the interaction of *PCGEM1* and methylated AR that leads to activation of AR-regulated genes.

Table 3-4. Summary of studies on aberrant ncRNAs expression in prostate cancer

ncRNA name	AR-regulated	Expression in PCA	Related functions	Roles in PCA
<b>miRNAs</b>				
miR-21	+	↑	Motility, invasion, apoptosis resistance	Oncomir
miR-125b	+	↑	Apoptosis proliferation	Oncomir
miR-220/-221		↑		Oncomir
miR-101		↓	Metastasis	Tumor suppressor
miR-34 family	+	↓	Apoptosis	Tumor suppressor
miR-126		↓	Motility, invasion	Tumor suppressor
miR-146a		↓	Motility	Tumor suppressor
miR-200c		↓	EMT	Tumor suppressor
miR-200b		↓	EMT	Tumor suppressor
miR-141		↑↓	EMT	Tumor suppressor
miR-330		↓	Apoptosis	Tumor suppressor
<b>lncRNAs</b>				
Linc00963 (long intergenic non-protein coding RNA 963)		↑	cell viability, migration, invasion, apoptosis	
PCGEM1 (LINC00071) (prostate cancer gene expression marker 1)	+	↑	cell proliferation, apoptosis	Oncogene
PRNCR1 (PCAT8) (prostate cancer noncoding RNA 1)		↑	cell variability	Oncogene
CBR3-AS1 (PlncRNA-1) (carbonyl reductase 3 antisense RNA 1)	+	↑	cell variability, apoptosis	
PCA3 (DD3) (prostate cancer antigen 3)	+	↑	cell survival, cell growth	
CTBP1-AS (PCAT10) (C-terminal binding protein 1 antisense RNA)	+	↑	cell cycle progression, proliferation	Oncogene
ANRIL (p15AS) (antisense non-coding RNA in the INK4 locus)		↑	cell proliferation, senescence	
PCAT1 (prostate cancer-associated transcript-1)		↑	cell proliferation	Oncogene
SChLAP1 (PCAT114) (second chromosome locus associated with prostate 1)		↑	cell invasiveness, metastasis	Oncogene
PCAT29 (prostate cancer-associated transcript-29)	+	↓	cell migration, proliferation	Tumor suppressor
↑, up-regulated; ↓, down-regulated; EMT, epithelial mesenchymal transition; PCA prostate cancer.				



## CHAPTER 4. AIMS OF THE THESIS

---

Global reprogramming of epigenetic patterns, including gains or losses in DNA methylation and changes to histone marks has been observed in cancer. Furthermore, novel sequencing technologies are now enabling the resequencing of thousands of cancer genomes and produce a huge amount of cancer data. Such data reveals a wealth of mutations in genes encoding epigenetic regulators that are involved in DNA methylation and/or chromatin states. But there is a lack of understanding how these mutations have the potential to deregulate hundreds of targeted genes genome wide. A genome-wide unbiased approach to investigate the role of candidate driver mutations leading to epigenetic changes involved in the regulation of cancer in a systematic manner is required. The genome-wide identification of candidate driver mutations as well as their genome-wide methylation analysis would provide insight in an underlying mechanism of carcinogenesis and will help to understand the methylation subgroups of tumors.

The aim of this thesis is to identify epigenetic driver mutations and the mechanism of their deregulation in cancer. In addition to correlate defects in them with differently methylated patterns in cancer epigenome and also with clinical features of the disease. Taken together identification of candidate driver mutations and link to epigenetic patterns will help to understand the interconnection between genetic and epigenetic alterations in cancer.

### Thesis goals:

1. Creation of a list of candidate cancer driver genes/pathways for epigenetic alterations
2. Classification of epigenetic regulators as potential oncogenic or those with tumor-suppressor function based on the location of mutations relative to functional domains and their frequency to define driver epigenetic regulators.
  - identified alterations in epigenetic regulators were analyzed to assign epigenetic regulators with oncogenic or tumor-suppressor functions
3. Identification of the major mechanisms of deregulation of epigenetic genes/pathways occurs in a non-random and tumor-type specific manner.
4. Identification of driver epigenetic regulators in cancer
  - To achieve this aim an integrative analysis was applied to identify putative driver epigenetic regulators and driver epigenetic pathways
5. Correlation of analysis of affected genes or groups of genes with tumor subtypes based on methylation profiling data (450K Illumina arrays) to defects in regulators of the epigenome to understand mechanisms leading to distinct epigenetic patterns.

- Different cluster algorithms were used to identify methylation patterns/cluster
  - Correlation analysis of up to 3 epigenetic regulators at the same type was performed to identify link between defined methylation patterns and putative driver epigenetic genes/pathways
6. Identification of PANCAN and tissue-specific events
  7. Identification of main mechanisms of deregulation of non-coding RNAs in cancer.
    - using independent data cohort from early-onset prostate cancer a systematic genome-wide evaluation of non-coding RNA deregulation was performed to identify a major mechanism of non-coding RNA deregulation in prostate cancer

## CHAPTER 5. MATERIAL AND METHODS

---

### 5.1. General equipment, disposables and chemical substances

Standard and interchangeable laboratory instruments, disposables and chemical substances were used for all experiments in this thesis. Special and not interchangeable equipment or reagents are mentioned directly in the corresponding methods sections. If not explicitly stated otherwise, described reagents and kits were used according to the manufacturers' protocols.

### 5.2. Primers

Primer sequences for the indicated methods are listed in the appendix. All primers were synthesized by Sigma-Aldrich (Taufkirchen, Germany) using DESALT purification.

### 5.3. Tissue samples

All prostate tissue samples used in this study were obtained by radical prostatectomy at the University Medical Center Hamburg Eppendorf as described previously [118]. Written informed consent was obtained from each patient. Clinical data for all patients are presented in Suppl. Table S1.

#### Nucleic acid extraction

DNA and RNA including miRNA were isolated using the AllPrep DNA/RNA Mini kit (Qiagen, Hilden, Germany) as described previously [130].

**Table 5-1:** Material used for DNA isolation

Material	Manufacturer
<b>QIAamp DNA Mini Kit</b>	Qiagen, Hilden, Germany
<b>AllPrep DNA/RNA Mini Kit</b>	Qiagen, Hilden, Germany
<b>Nanodrop ND-1000</b>	Thermo Scientific, Rockford, USA
<b>ZR-96 Quick-gDNA Kit</b>	Zymo Research, Irvine, USA

Briefly, DNA from human cell lines was isolated using the QIAamp DNA Mini Kit (**Table 5-1**) according to the manufacturer's protocol. If co-purification of RNA and DNA from the same sample was required, the Qiagen AllPrep DNA/RNA Mini Kit was utilized as the supplier suggests.

Table 5-2: Material used for RNA isolation

Material	Manufacturer
RNeasy Mini Kit	Qiagen, Hilden, Germany
AllPrep RNA/DNA Mini Kit	Qiagen, Hilden, Germany
RNase-free DNase Set	Qiagen, Hilden, Germany
Nanodrop ND-1000	Thermo Scientific, Rockford, USA

Isolation of RNA from cell culture samples was achieved using the Qiagen RNeasy columns or the Qiagen AllPrep RNA/DNA Mini Kit (**Table 5-2**) according to the manufacturer's protocol. Additionally, RNA was on-column digested with DNase for 15 min to remove contaminating residual DNA as suggested in the manual. Nucleic acid concentration and purity was quantified with the Nanodrop system

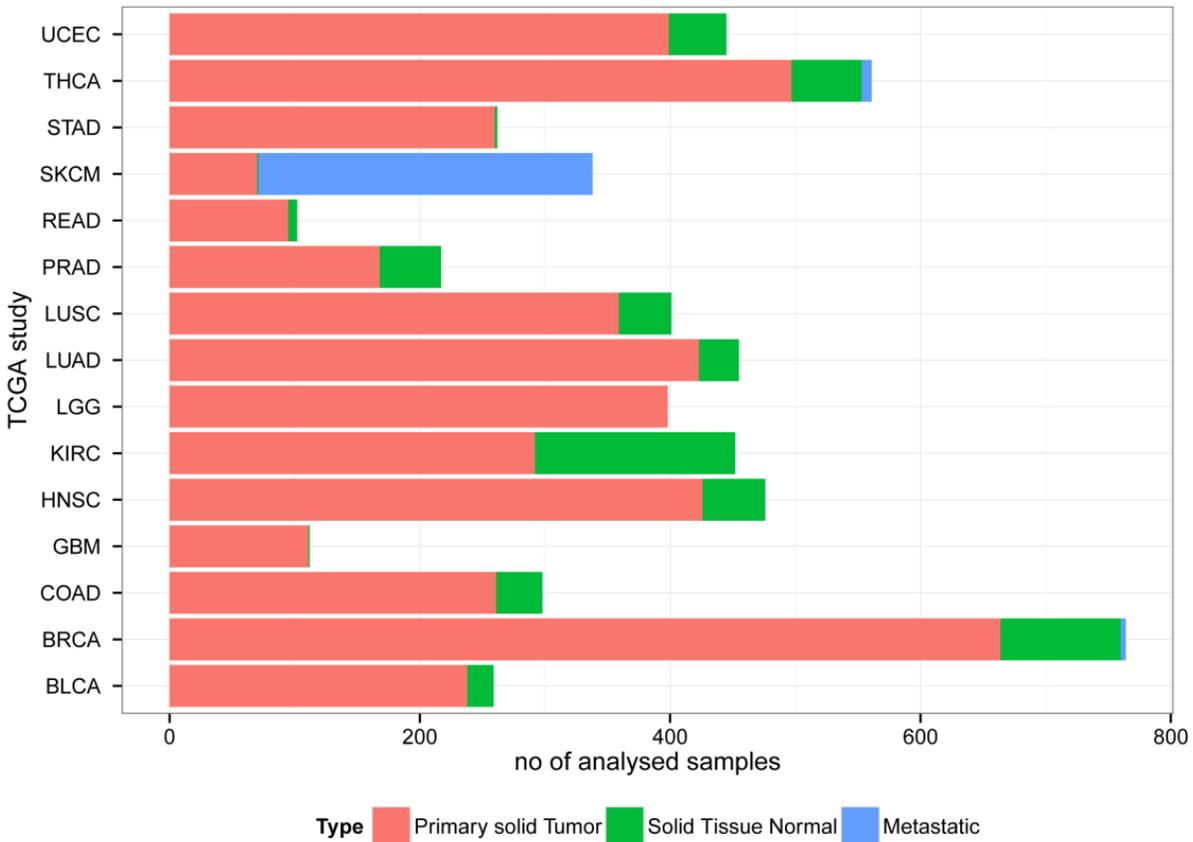
#### 5.4. Data sets

This thesis represents analysis of from several separate data sets. In order to skip the description of each data set in each chapter and for easy understanding, we assigned following abbreviations and used them during all results and discussion chapters.

1. **TCGA PANCAN data set** – currently available genome-wide datasets on genetic and epigenetic alterations in cancers from TCGA (<https://tcga-data.nci.nih.gov/tcga/>) from 14 cancer subtypes, covering 5738 samples (**Table 5-3, Figure 5-1**). TCGA PANCAN data set consist of methylation data (Illumina 450K arrays), mutations data (analyzed from whole genome-sequencing data), copy-number alterations (CNAs).

Table 5-3: TCGA PANCAN data set description

TCGA Cancer Types	TCGA study code	# Cases
Acute Myeloid Leukemia	LAML	200
Adrenocortical carcinoma	ACC	80
Breast invasive carcinoma	BRCA	1098
Colon adenocarcinoma	COAD	461
Glioblastoma multiforme	GBM	528
Kidney Chromophobe	KICH	66
Lung adenocarcinoma	LUAD	521
Pancreatic adenocarcinoma	PAAD	185
Prostate adenocarcinoma	PRAD	498
Rectum adenocarcinoma	READ	171
Skin Cutaneous Melanoma	SKCM	470
Stomach adenocarcinoma	STAD	443
Thyroid carcinoma	THCA	507
Lung squamous cell carcinoma	LUSC	510



**Figure 5-1: Histogram plot of total number of samples for each cancer data set used in this study.** Red color denotes primary tumors, blue – metastatic tumors, green – normal tissue. Cancer types were listed using cancer code form TCGA portal (see Table 5-3).

**2. EO-PCA data set** – data cohort from early-onset prostate cancer generated within the German ICGC project (ICGC EO-PCA dataset) [118] [Weischenfeldt et al., manuscript in preparation]. Data cohort consist of

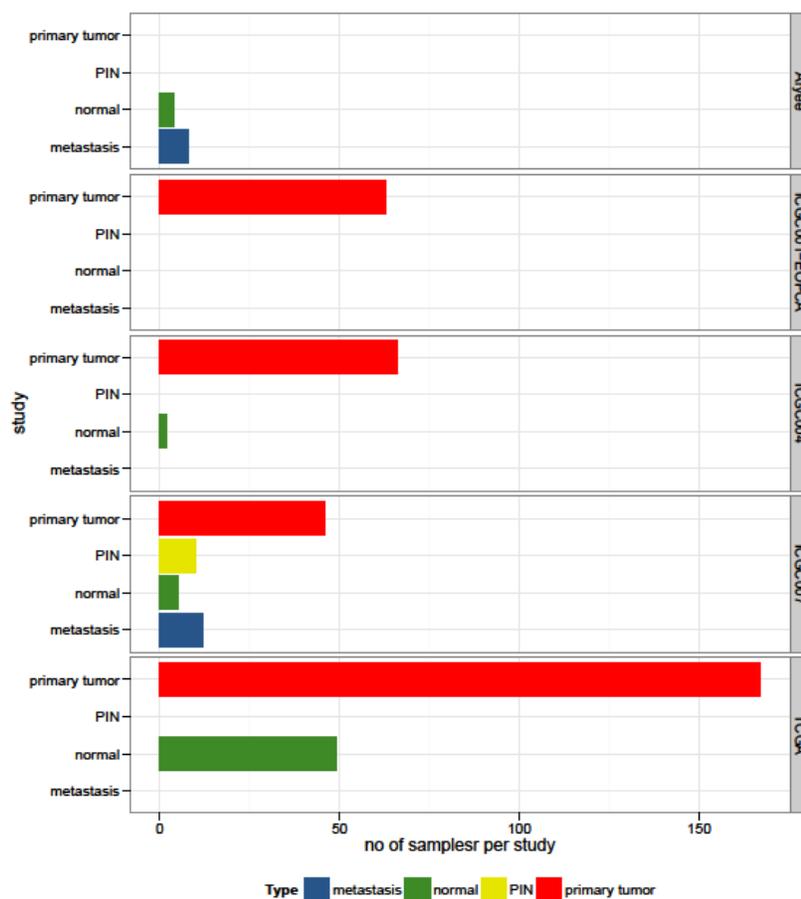
**miRNome** – miRNA expression profile from 66 EO-PCAs specimens and 8 normals, generated from small RNA sequencing

**transcriptome** – expression profile from 89 EO-PCAs specimens and 11 normals, generated from RNA sequencing data and consists coding and non-coding gene expression profiles.

**methylome** – methylation data for 155 EO-PCAs specimens and 15 normals, generated from MCIp-seq and Illumina 450K arrays.

**CNA** – copy-number alterations from 66 EO-PCAs specimens and 8 normals, generated from whole-genome sequencing (WGS).

**3. COMBINED PCA data set** – data cohort from prostate cancer with classical age distribution that includes 10 prostatic intraepithelial neoplasia samples (PINs), 342 primary tumors, and 19 metastatic samples, and 60 normal prostate epithelial samples. Data set covers methylation data only from Illumina 450K arrays. It was generated using publically available data from TCGA ([tcga-data.nci.nih.gov/tcga/](http://tcga-data.nci.nih.gov/tcga/)) and Aryee [131] as well as our own data from early onset prostate cancer (ICGC1-EOPCA) [Feuerbach et al., manuscript in preparation] and from late onset (ICGC004) [Bogatyrova et al., manuscript in preparation] and from heterogeneity study (ICGC007) [132].



**Figure 5-2:** Histogram depicting the sample numbers from COMBINED prostate cancer data set. Name of sample study is indicated on right axis. Color denotes sample type.

## 5.5. DNA methylation analysis

### EpiTYPER MassARRAY quantitative DNA methylation analysis

Table 5-3: Material used for EpiTYPER MassARRAY methylation analysis

Material	Manufacturer
<b>EZ DNA methylation kit</b>	Zymo Research, Irvine, USA
<b>HotStarTaq DNA Polymerase kit</b>	Qiagen, Hilden, Germany
<b>dATP, dCTP, dGTP, dTTP</b>	Fermentas, St. Leon-Rot, Germany
<b>“T” Cleavage MassCLEAVE reagent kit</b>	Sequenom, San Diego, USA
<b>RNase A</b>	Sequenom, San Diego, USA
<b>MassARRAY Nanodispenser</b>	Sequenom, San Diego, USA
<b>Sequenom MassARRAY MALDI-TOF Mass Spectrometer</b>	Sequenom, San Diego, USA
<b>QIAamp DNA Mini Kit</b>	Qiagen, Hilden, Germany
<b>Sequenome EpiTYPER software 1.2</b>	Sequenom, San Diego, USA
<b>Mastercycler 384</b>	Eppendorf, Hamburg, Germany
<b>Shrimp Alkaline Phosphatase (SAP)</b>	Sequenom, San Diego, USA

1µg genomic DNA was chemically modified with sodium bisulfite and subsequent quantitative DNA methylation analysis was performed by MassARRAY technique as described [133]. MassARRAY primers are listed in Suppl. Table S1. For each MassARRAY amplicon, the median of all methylation values of its single CpGs was used for further calculations. Methylation differences between sample groups are presented as  $\Delta\beta$  values. DMRs were identified by the Mann Whitney U test and subsequent Bonferroni correction of p values. Regions showing corrected p values < 0.05 were considered differentially methylated. Hierarchical clustering of MassARRAY methylation data was performed by the Multiple Experiment Viewer software v4.3. Correlation analyses were done by the Spearman method using IBM SPSS Statistics 19 (IBM Deutschland GmbH, Ehningen, Germany).

### Illumina HumanMethylation450 BeadChip

Table 5-4: Material used for Illumina 450k BeadChip

Material	Manufacturer
<b>Infinium HumanMethylation450 BeadChip Kit</b>	Illumina Inc., San Diego, CA, USA
<b>RnBeads R-Package [134]</b>	www.rnbeads.mpi-inf.mpg.de

For a representative, genome-wide analysis of DNA methylation in cancer samples, the Illumina HumanMethylation450 BeadChip (450k Chip) was employed by the DKFZ Genomics and Proteomics Core Facility according to the manufacturer’s instructions using up to 500 ng of DNA. Such array covers 483,854 probes addressing the methylation state of CpG sites located in CpG islands (CGIs), CGI shores, CGI shelves and open sea regions. Methylation of individual sites is calculated as beta-value ( $\beta$ -value) ranging from 0% (not methylated) to 100% (fully methylated).

For inter- and intra-sample data normalization, raw data was BMIQ-normalized using the RnBeads R-Package (<http://rnbeads.mpi-inf.mpg.de/>). For quality filtering, the single-nucleotide polymorphism (SNP)-calling probes (dbSNP132 Common, n=92428) and probes that had detection p-values value below 0.01 in at least one sample were excluded as well as probes missing information for a single sample per cancer study. Probes measuring methylation in a non CpG context (n=3156) were removed. No strong batch effects were identified.

### **DNA methylome deep sequencing and analysis**

Methyl-CpG immunoprecipitation (MCIp) for enrichment of highly methylated tumor and normal DNA and subsequent deep sequencing (MCIp-seq) was carried out as described previously [118]. Libraries were sequenced with single-end 50bp reads using the SOLiD 4 (Applied Biosystems, Life Technologies Corporation, Carlsbad, CA, USA) or the HiSeq2000 platform. Reads were mapped to the human genome and differential methylation was detected as described previously [118]. The seven tumor samples sequenced on the SOLiD 4 platform were compared to one normal control sequenced on the same platform [118], whereas the six tumors sequenced on the HiSeq2000 were compared to eight controls analyzed on the same platform.

### **DNA deep sequencing and analyses**

DNA library preparation and whole genome sequencing was performed on Illumina sequencers as described [118] with two complementary insert size libraries. Short insert size libraries were prepared with two 101bp reads spanning 187-301bp (paired-end), and large insert-size libraries with two 36bp reads spanning between 3,578-5,403bp (mate-pairs). Sequencing was performed on Illumina HiSeq to a minimum 30x whole-genome coverage for paired-end libraries.

Single nucleotide variants (SNVs) were called as described previously [135]. Structural rearrangements 200bp to Mb in size were detected as previously described [118]. Copy-number alterations were detected using BIC-seq [136]. Somatic structural rearrangements were filtered for lack of corresponding variant support in the germline control sample as described recently [118].

### **Detection of differentially methylated regions in the genome-wide validation dataset**

MeDIP seq data for all samples of the validation dataset were recently published [101]. Methylation data for all 500bp bins overlapping to miRNA regulatory regions [16] were extracted,

and tested for differential methylation by the Mann Whitney U test and FDR correction of p values using R Bioconductor. A regulatory region was considered differentially methylated if at least one included 500bp bin showed a FDR < 0.05.

### **5-aza-2'-deoxycytidine treatment of PCA cells**

Human PCA cell lines LNCaP and PC3 were obtained from ATCC (Manassas, VA) and authenticated by the Genomics and Proteomics Core Facility at the German Cancer Research Center in month year using a 24-plex SNP profiling assay. Cells were cultured in high level glucose RPMI 1640 medium (PAA, Coelbe, Germany), supplemental with 10% FCS. DNA demethylation treatment was performed with 0 (PBS), 1, 2.5 or 5  $\mu\text{mol/l}$  5-aza-2'-deoxycytidine (Sigma-Aldrich, St. Louis, MO) for 24, 48, 72, 96 and 120h hours by replacing the drug and medium every 24 hours. DNA demethylation efficiency was evaluated by quantitative DNA methylation analysis of repetitive elements (LINE-1) using MassARRAY. For each cell line, 5-aza-2'-deoxycytidine concentration and incubation time that showed the strongest demethylation effect was used for further analyses.

## **5.6. miRNA expression analysis**

### **miRNA deep sequencing and analysis**

Small RNA libraries were prepared with singleplex or custom multiplex adaptors and primers (kindly provided by NEB, Suppl. Table S12), and subsequently sequenced on an Illumina HiSeq2000 instrument (San Diego, CA) [118]. Raw sequencing reads were mapped to known human miRNAs (miRBase18.0) and normalized as described [118]. For the ICGC EO-PCA dataset, analyses were done on each tumor separately: all miRNAs with less than five reads in the relevant tumor or the mean of the four normal tissues were filtered out. MiRNAs showing a fold-change of at least 1.5 compared to the mean of the normal samples were considered as deregulated. In the validation dataset, differentially expressed miRNAs were identified using DESeq [137] after filtering out all miRNAs that had less than five reads in less than 50% of the samples. p values were corrected for multiple testing and miRNAs showing a FDR < 0.05 were considered as differentially expressed. Hierarchical clustering was performed using Genesis [138] according to Euclidean distance and average linkage algorithm, after data had been normalized by DESeq.

### **Quantitative PCR arrays and analysis**

Quantitative PCR (qPCR) screening of miRNAs was performed using the Array Human MicroRNA Set Cards v2.0 (Life Technologies, Darmstadt, Germany) as described previously [135, 139] with 100ng total RNA input. Raw data were obtained using automatic baseline calculation and a threshold of 0.15. Data was median-normalized plate-wise [135, 139]. Differentially expressed miRNAs were identified by LIMMA and subsequent FDR correction of p values, after miRNAs with CT values >35 in more than 70% of the samples had been filtered out [135, 139]. MiRNAs showing a FDR < 0.05 were considered differentially expressed. Normalized data were linearized using the Delta CT method [136]. Hierarchical clustering was performed on log<sub>2</sub> transformed data using Genesis according to Euclidean distance and average linkage algorithm.

### Quantification of pri-miRNA expression by qPCR

500ng total RNA was reverse transcribed using Superscript II (Invitrogen, Life Technologies) and random primers. qPCR was performed using TaqMan pri-miR assays (Applied Biosystems, Suppl. Table S13) and the Absolute QPCR Mastermix (Thermo Fisher Scientific, Schwerte, Germany) on the LightCycler 480 (Roche, Grenzach-Wyhlen, Germany). Expression levels were calculated with the DeltaDelta CT method [140] using the *TATA box binding protein (TBP)* for normalization.

## 5.7. Data analysis

For routine statistical analysis, GraphPad Prism 5 (GraphPad Software, La Jolla, USA) was applied. For group-wise comparison of two distributions from different samples/treatments, the two-tailed non-parametric Mann-Whitney-U test was used (not assuming Gaussian distributions). For experimental settings with replicates of paired treatments/samples, a two-tailed student's t test was applied. In all cases, significance levels were depicted as follows: \* = p<0.05; \*\* = p<0.01; \*\*\* = p<0.001.

### Prediction of functional domains

Program package HUZAR (Heidelberg Unix Sequence Analysis Resources) was used for domain prediction. The *in silico* functional domain prediction algorithms such as: pirsf, pfam, superfam, seg, smart, prints, ncoils, pfscan, signal, tigrfam were selected to identify domain in genes. Only domains, identified by at least 2 algorithms were selected as positive prediction.

### Identification of driver genes

List of tool, used for identification potential driver genes with short description:

[OncodriveFM](#): Identifies genes with a bias towards high functional mutations.

[OncodriveCLUST](#): Identifies genes with a significant regional clustering of mutations.

[MutSigCV](#): Identifies genes mutated more frequently than background mutation rate.

[OncodriveROLE](#): Classifies driver genes according to its mode of action in Activation or Loss of Function.

[IntOGen](#): Combined tool of all listed above methods ([www.intogen.org](http://www.intogen.org)).

### **Differently methylated analysis of sample groups**

Differential methylation analysis was conducted on site and region level according to the sample groups specified in the analysis. In the following analyses, p-values on the site level were computed using the limma method. I.e. hierarchical linear models from the R limma package were employed and fitted using an empirical Bayes approach on derived M-values. Differently methylated probes (DMPs) and regions (DMRs) were computed for tumor samples as in/decrease of methylation compare to normal per each cancer study individually, using differently methylated module from RnBead R package, as described in package vignette (<http://rnbeads.mpi-inf.mpg.de/data/RnBeads.pdf>). Differential methylation on the site level was computed based on a variety of metrics. Each site was assigned a rank based on three criteria:

- 1) the difference in mean methylation levels of the two groups being compared,
- 2) the quotient in mean methylation and 3) a statistical test (t-test or limma depending on the settings) assessing whether the methylation values in the two groups originate from distinct distributions. A combined rank is computed as the maximum (i.e. worst) rank among the three ranks. The smaller the combined rank for a site, the more evidence for differential methylation it exhibits. Only significant sites/regions were taken for further analysis.

### **Selection of probes for clustering algorithms**

We have used two strategies for probe selection:

**V50**: only probes that represent 50% of variability of the data were selected

**T10**: only 10 000 (10K) most variable probes were selected.

### **Clustering approaches for identification methylation patterns/subgroups**

Unsupervised clustering analysis using 3 major approaches: Partitioning Around Medoids (PAM), K-means and Hierarchical cluster analysis (HCA) were used.

**PAM**, most common realization of **k-medoids** algorithm: classical partitioning technique of clustering that clusters the data set of  $n$  objects into  $k$  clusters known *a priori*.

***k*-means:** algorithm of partition  $n$  observations into  $k$  clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Both the  $k$ -means and  $k$ -medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the  $k$ -means algorithm,  $k$ -medoids chooses datapoints as centers (medoids or exemplars) and works with an arbitrary matrix of distances between datapoints.

**HCA** is a method of cluster analysis which seeks to build a hierarchy of clusters.

Data was represented using principal component analysis (PCA). Comparison of the performance of different methods was done using average silhouette value.

### **Gene-set enrichment analysis (GSEA)**

GSEA was used to determine whether an a priori defined set of genes (protein coding or non-coding) shows statistically significant, concordant differences between two biological states (tumors vs. normals, different tumor groups). Analysis was done using [GSEA v2.1.0 Release](http://www.broadinstitute.org/gsea/) (<http://www.broadinstitute.org/gsea/>).



---

## RESULTS

---

### CHAPTER 6. MUTATIONS IN REGULATORS OF THE EPIGENOME AND THEIR EFFECTS ON THE DNA METHYLOME

Genetic mutations have been described in enzymes that regulate epigenetic patterns. For example, frequent and recurrent mutations have been found in H3F3A, the gene encoding for histone variant H3.3A in pediatric glioblastoma [89] [90], chondrosarcoma and giant cell tumors of the bone [141]. There were also found mutations in isocitrate dehydrogenase 1 (IDH1) in gliomas [91, 92], or the de novo DNA methyltransferase 3A (DNMT3a) in acute myeloid leukemia or glioblastoma [93] [94]. Based on the known functions of these genes one could predict that the observed mutations have the potential to deregulate, hundreds of direct and indirect targets genome-wide.

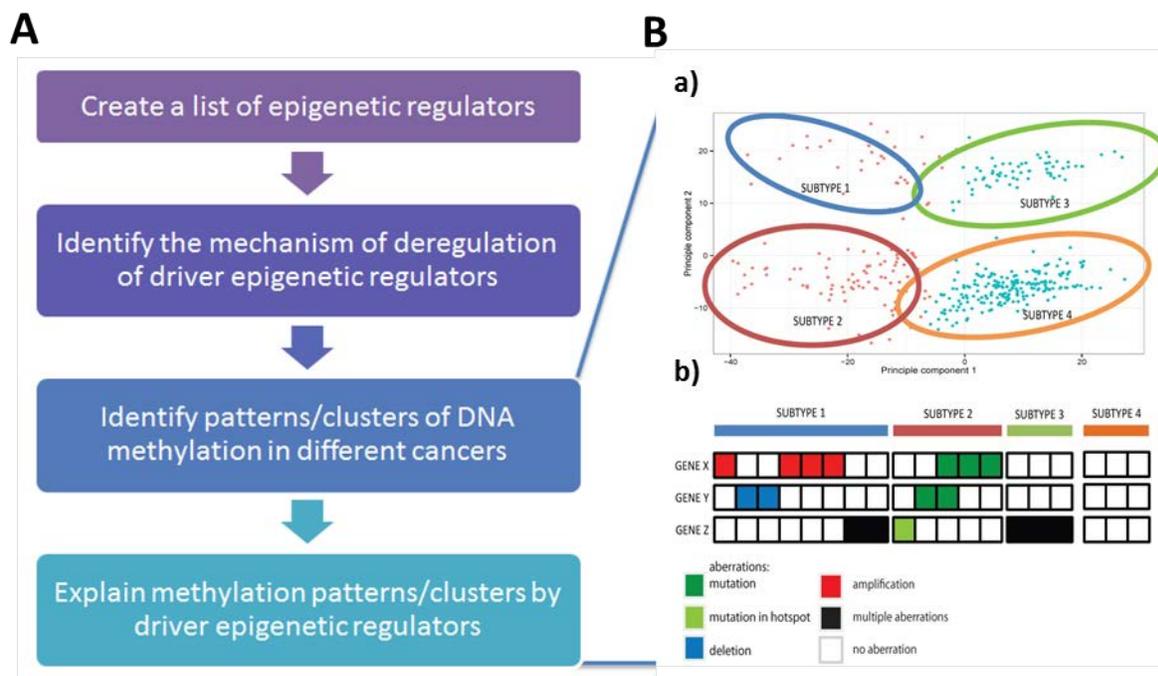
A genome-wide unbiased approach is required in order to further investigate, in a systematic manner, the role of candidate driver mutations for epigenetic changes in tumorigenesis. The genome-wide identification of candidate driver mutations as well as a global analysis of DNA methylation changes affecting expression of these genes would provide insights into an underlying mechanism of carcinogenesis and will help to identify the subgroups of tumors. In this chapter the role of specific genetic mutations will be correlated with associated DNA methylation patterns in the genome and also with clinical features of the disease subgroups. Functional characterization of the identified driver mutations will be carried out. Taken together identification of candidate driver mutations and the link to epigenetic patterns will help to understand the interplay between genetic and epigenetic alterations in cancer.

Utilizing the large amount of available cancer genome data obtained in international consortia such as TCGA and ICGC, we performed the studies on different cancer types in order to investigate similarities and differences in the mechanisms between different cancers. Such an approach, termed pan-cancer analysis (PANCAN), was initially introduced by the TCGA consortium. Recently, PANCAN mutation profiling revealed a list of 125 driver genes recurrently mutated which are acting in epigenetic pathways [95]. Some of these mutations associate with distinct subgroups of tumors characterized by defined epigenetic patterns and clinical phenotypes. A different study identified 102 PANCAN recurrent focal amplified/deleted regions, for which currently no known oncogenes/tumor suppressor genes were identified. Interestingly, these regions are enriched for genes involved in regulation of epigenetic patterns [96]. Based on

these data, we formulated and test the hypothesis that mutations in the three epigenetic pathways (DNA methylation, chromatin modification and chromatin remodeling) represent candidate driver mutations acting through global reprogramming of epigenetic patterns.

A systematic, integrative approach to the identification of candidate epigenetic driver genes and investigation of their link to epigenetic patterns with the goal to understand the interconnection between genetic and epigenetic alterations in cancer was created and is described in **Figure 6-1**. Briefly, the workflow consists of four major steps:

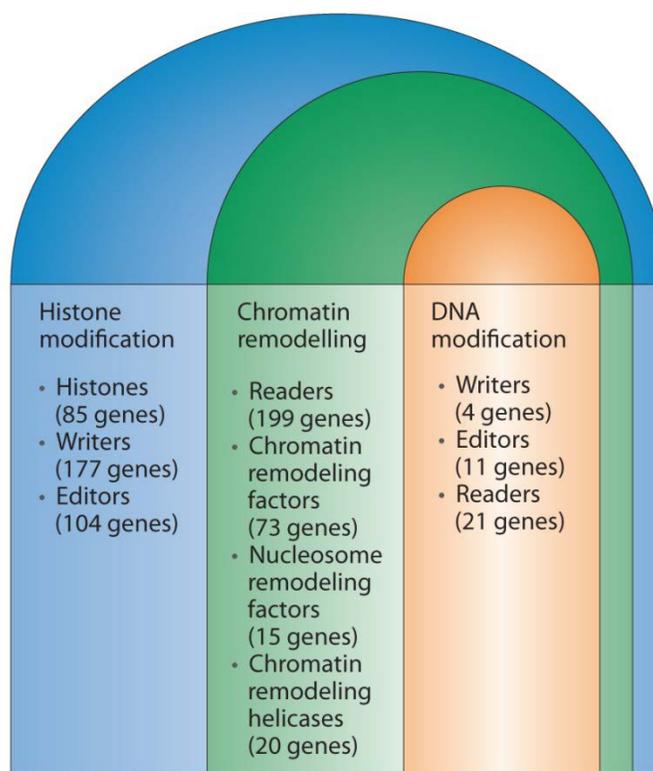
- 1) creation of a list of epigenetic regulators in order to identify potential driver genes;
- 2) identification of the major mechanism of deregulation of driver epigenetic regulators;
- 3) using methylation profiling (Illumina 450K arrays) for identification DNA methylation clusters/patterns in the analyzed cancers;
- 4) performing a correlation analysis between defined methylation patterns and altered driver epigenetic regulators or a group of them to identify the link between deregulation of epigenetic drivers and methylation profiles.



**Figure 6-1: Schematic workflow of integrative analysis of mutations in regulators of epigenome with DNA methylome.** (A) The workflow of integrative analysis step by step. (B) Zoom in step 3-4 for explanations: a) Principle component analysis (PCA) like example to present workflow from step 3. PCA of methylation data of thyroid cancer (THCA) was used to visualize 4 cancer subtypes. b) Exemplary alteration profile of each cancer subtype identified by methylation profiles.

## 6.1. Alterations of epigenetic regulators in cancer

As a first step into the investigations a list of about 700 candidate driver mutations for genes involved in establishing, maintaining and reading epigenetic patterns was created using available publications and web-browsers. This included the following web sites: Amigo (<http://amigo.geneontology.org/>), Cosmic (<http://www.sanger.ac.uk/>), PubMed, Pubmeth (<http://www.pubmeth.org/>), Mutation list (<http://www.cbioportal.org/>) and the UCSC Genome Browser (<https://genome.ucsc.edu/>). Genes were categorized into three major groups: histone modification, DNA methylation and chromatin remodeling. Each group was further divided into subclasses such as writers, editors and readers (**Figure 6-2, Supplemental Table S1**).

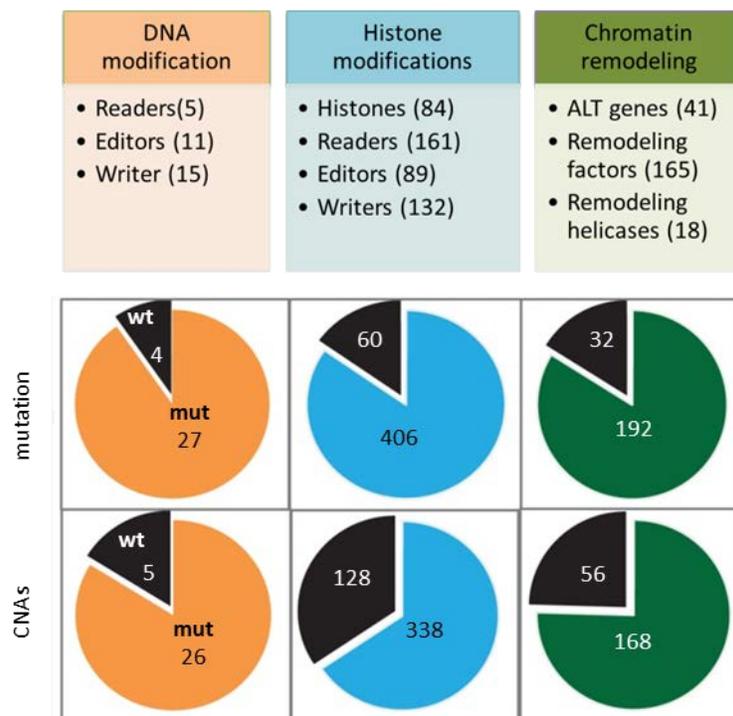


**Figure 6-2:** List of groups and subgroups of epigenetic regulators with number of genes in brackets. Gene names characteristic information is in Supplemental Table 6-2.

This list consists of Ensembl gene names and gene symbols, name of epigenetic group and subgroup, bioinformatical prediction of functional domains and a brief description of the functional role as inferred from the literature. For each epigenetic regulator bioinformatical prediction of functional domains was done as described in **chapter 5** to identify the most functionally relevant locations of mutations. Epigenetic regulators with predicted domains have been ranked based on known domain functions, as more prone to be potentially drivers in

cancer. The complete list of epigenetic drivers with function domain prediction is available in **Supplemental Table S2**.

Using currently available **TCGA PANCAN data set** (description in **chapter 5**) the different alterations in regulators of the epigenome by genetic mutations, copy number alterations (CNA), promoter methylation were analyzed. Significantly high enrichment of epigenetic regulators among the mutated genes (Fisher's exact test p-value  $p < 0.001$ ) and the genes with copy number alterations ( $p < 0.01$ ) indicate potential functional impact of defects in epigenetic regulators on a cancer genome. Up to 85% of the epigenetic regulators were found affected in cancer by either mutation and/or CNA (Figure 6-4). This finding suggests that there is not a unique mechanism of deregulation of epigenetic regulators in cancer and a systematic integrative approach is needed in order to determine driver regulators and major mechanism of their alterations



**Figure 6-4 Comparison of different identified alterations per group of epigenetic regulators.** The upper part of the figure indicates the total number of 709 epigenetic regulators. Each pie chart lists the frequency of genes that were found to be mutated or altered by CNA or by promoter methylation in at least two tumor samples and the numbers of unmutated genes (marked as wild type - wt) for each group of epigenetic enzymes. Black color denotes unmutated genes, orange for DNA modification, blue for histone modification, dark green for chromatin remodeling.

We analyzed the mutation distribution per sample in order to find out how many epigenetic regulators are affected per sample (**Figure 6-5**). A range from 2 (PRAD) up to 30 (LUAD)

average mutations per sample was identified in epigenetic regulators in the investigated tumor samples. Lung adenocarcinoma was showing the highest number of mutation that reflects high total mutation frequency. THCA and PRAD are showing the lowest frequency of mutations in epigenetic enzymes. The number of mutations is not significantly correlated to overall frequency of mutations in epigenetic key players.

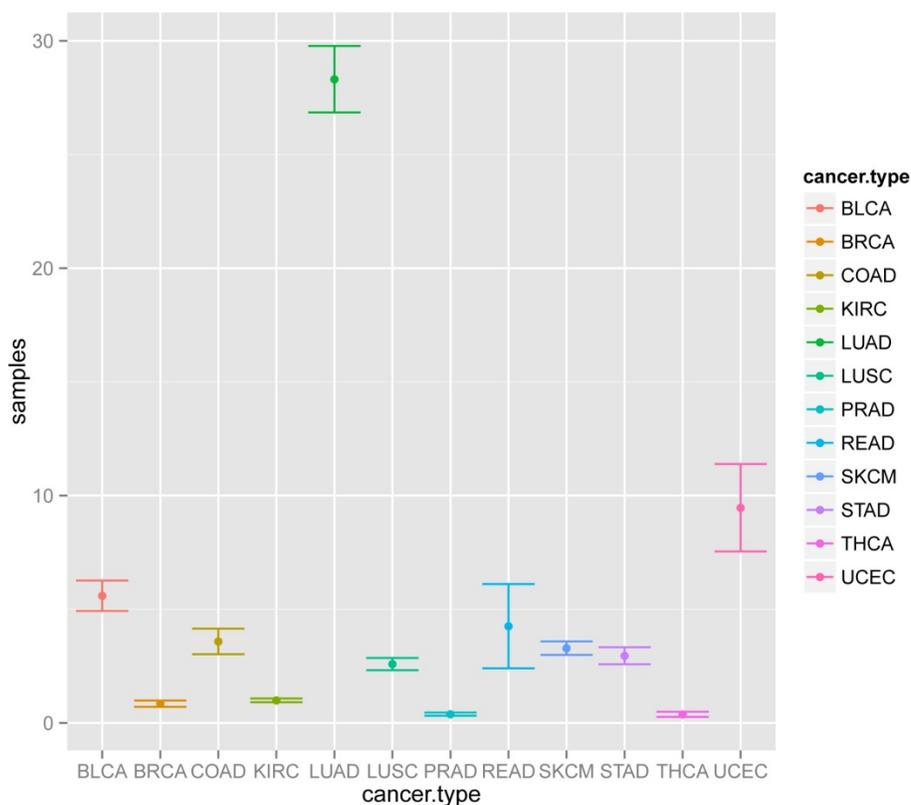


Figure 6-5: Distributions of number of non-synonymous somatic mutations in epigenetic regulators per sample in investigated PANCAN data cohort.

### 6.1.1. Putative driver epigenetic regulators

To select putative driver epigenetic regulators and identify major mechanism of their regulation, we performed a systematic integrative approach as outlined in **Figure 6-5**. Briefly, the workflow consists of five major steps. The first two steps are the identification PANCAN and tissue specific epigenetic regulators based on mutation frequency per cancer. Initially, the mutation profile of epigenetic regulators was established to determine recurrent mutations. Frequently mutated regulators were divided in two groups: either cancer-type specific events (PANCAN epigenetic regulator group), as those affected in >50% of investigated cancer types or tissue-

specific epigenetic regulator group, as those affected frequently in one cancer type but not in others. Step 3 was used to determine oncogenic and tumor-suppressor genes, based on alteration types. Genes that are commonly mutated in the same amino acid position and located in gained genomic locations were assigned as putative oncogenic. Alternatively, genes exhibiting a mutation spectrum throughout the gene body and located in deleted genomic locations were assigned as tumor-suppressor epigenetic regulators.

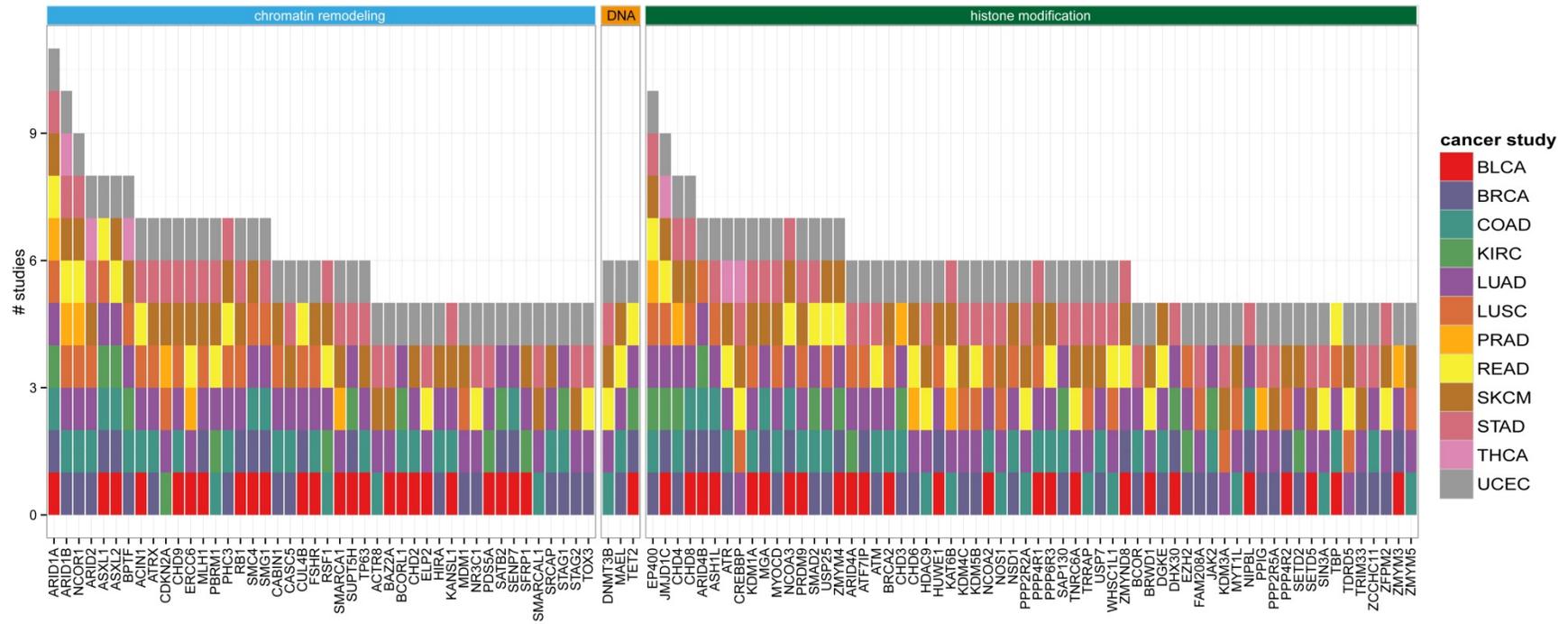


**Figure 6-5: Workflow of integrative approach of identification driver epigenetic regulators.**

### **PANCAN and tissue-specific epigenetic regulators**

Next, we classified epigenetic regulators into two groups: PANCAN and tissue-specific. To assign PANCAN mutated epigenetic regulators, 13 subtype specific frequently mutated genes were merged to select the ones mutated in more than of 50% of the tumor subtypes. In total 78 PANCAN epigenetic regulators were identified to be mutated across tumor types, including previously published *ARID1A/B* [142], *TETs* [143], *DNMT3A* [144] as top hits in the list (**Figure 6-5**).

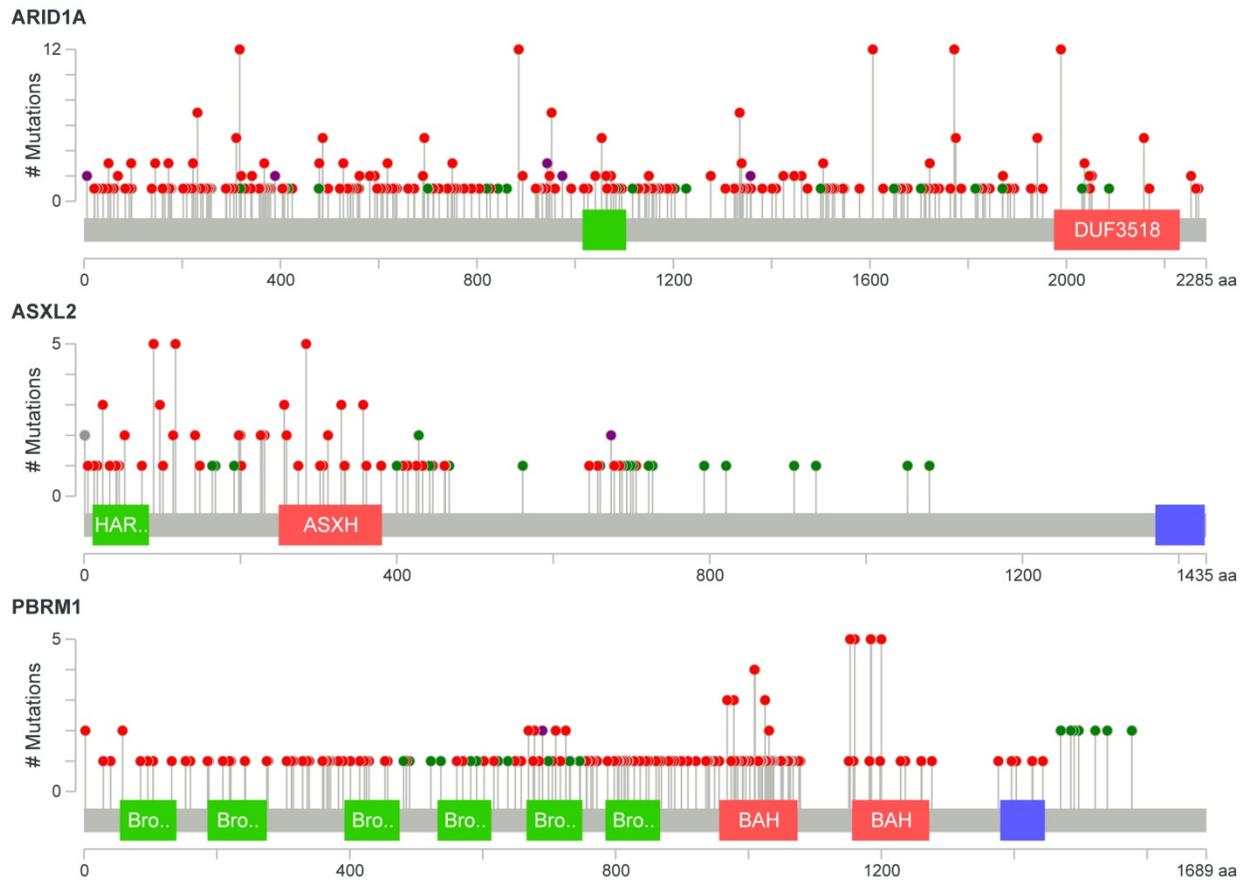
Frequently mutated epigenetic regulators, which were not assigned to be PANCAN regulators, were examined for tissue-specific ones. In total, 156 epigenetic regulators were assigned to be tissue-specific, as frequently mutated in a unique cancer type (**Supplemental Table S4**).



**Figure 6-5 Mutation distribution of PANCAN epigenetic regulators** (frequently affected in >50% cancer subtypes). Each bar chart represents cancer subtypes with mutation of epigenetic regulator.

### Oncogenic and tumor-suppressor epigenetic regulators

Next, mutations in epigenetic regulators have been evaluated for either gain-of-function or loss-of-function. For this, the location distribution throughout the genes and frequency of mutations has been evaluated. Mutations located in predicted functional domains were classified as being of high functional impact. Genes with up to 5 HotSpots (frequently mutated region that causes the same amino acid change) were evaluated as potential oncogenic. Regulators with more than 5 mutations spread throughout the coding region of a gene were classified as potential tumor suppressors. The lists of all identified HotSpots are provided in **Supplemental table S5**.



**Figure 6-6: Diagrams showing the location of mutations in protein coding regions of *ARID1A*, *ASXL2* and *PBRM1* genes.** Mutation needle plot represents all synonymous mutations found in the analyzed TCGA data set and plotted on protein annotation using MutationMapper (cbiportal.org). Protein domains –DUF3518-, HAR-, Bro-, BAH- and others – are marked by different colors and highlighted with box-shape according to the location in the protein. Mutation diagram circles are colored with respect to the corresponding mutation types (missense - green, truncating- red, in frame mutations – grey, residues affected by different types of mutations are colored by purple). The X-axis represents amino acids of the protein and schematic protein structure with colored domain prediction. The Y-axis represents number of mutations per amino acid.

In order to improve the assignment of genes to potential oncogenic or tumor-suppressive, additional alterations like CNA and promoter methylation of epigenetic regulators were added to the analysis. The simplified definitions used for our analysis, where:

**oncogene** is a gene that is frequently altered in different cancer types either by the same mutations or by allelic gains;

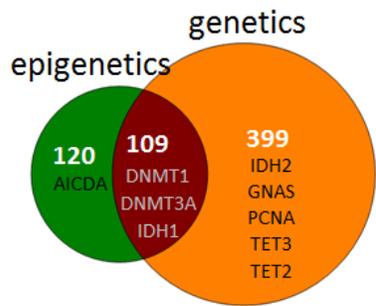
**tumor-suppressor** is a gene frequently altered in different cancer types either by different mutations spread over the entire gene or by allelic loss coupled with promoter hypermethylation.

We generated CNA profiles using the **TCGA PANCAN data set** to interrogate for frequently deleted and amplified epigenetic regulators. Level 2 provides all identified deleted and amplified regions per cancer sample in comparison to a pool of normal tissues. These profiles were used to identify highly deleted and amplified epigenetic regulators per cancer study. Firstly, tissue-specific frequently deleted (87 genes) and amplified (56 genes) regulators were identified. Using the same cut-off as in the mutation screen described above (>50% of the tumor subtypes), 59 PANCAN epigenetic regulators deregulated by CNA were identified in TCGA cancer cohort (top listed in **Supplemental Table S6**).

The mutation and CNA profiles were used to classify some of the epigenetic regulators as potential oncogenes or tumor-suppressors. More precisely, genes with HOTSPOT mutations or with mutations in predominantly functional domains and also located in commonly gained regions in cancer were defined as oncogenic regulators. Genes with mutations equally distributed over the gene sequence and in highly deleted regions were assigned to the group of tumor-suppressor regulators (both conditions should be positive). In total, 76 genes with potentially oncogenic function and 45 genes with potential tumor-suppressor function among the PANCAN epigenetic regulators were identified.

### **Epigenetic regulators, deregulated by promoter methylation**

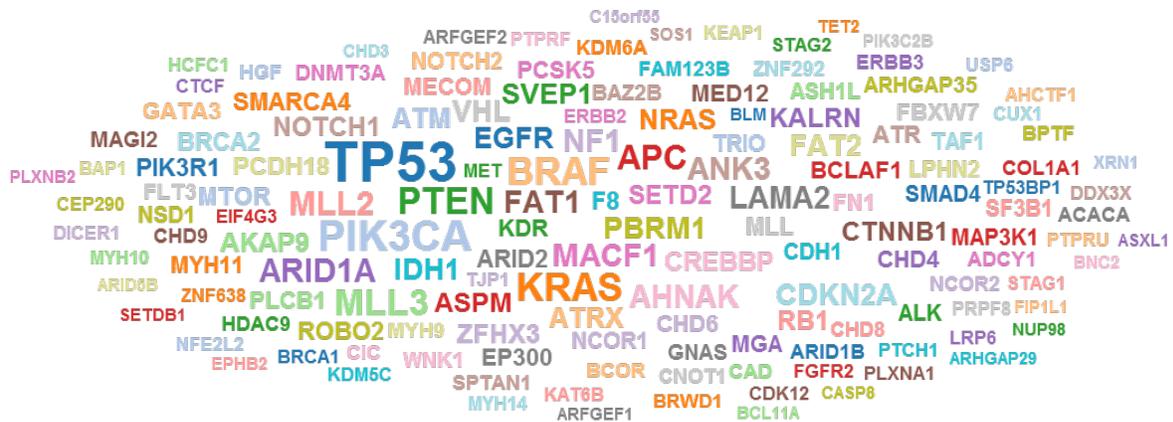
In order to unravel if promoter methylation may play a role in deregulation of epigenetic players, we performed an integrative analysis using methylome data. 58% of the epigenetic genes possess a DMR in the promoter region in at least 1 investigated cancer subtype. Moreover, 32% of the epigenetic regulators have hypermethylation of promoter in more than 50% of the cancers. These are potential tumor-suppressors and include many genes that have been reported before (e.g. *DNMT1*, *DNMT3A*, *IDH1*), as well as newly identified ones. In conclusion the genetic and epigenetic data gave strong evidence that genetic alterations, such as mutations and CNA is a major mechanism of deregulation for epigenetic genes (**Figure 6-8**). This observation could be further strengthened by examining additional lines of evidence, such as data on deregulation by non-coding RNAs.



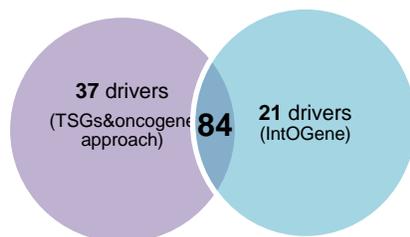
**Figure 6-8: Genetic and epigenetic events in the deregulation of epigenetic regulators in cancer.** Venn diagram showing the number of epigenetic regulators deregulated by a genetic mechanism (orange circle: mutations and CNAs), an epigenetic mechanism (green circle: promoter methylation), or both (red circle, not necessarily in the same cancer type).

### Potential driver epigenetic regulators in cancer

Since there are different approaches to identify driver cancer genes, we used a combination of multiple tools to obtain a list of driver genes that is both comprehensive and reliable. Briefly, the sets of 76 genes with potentially oncogenic function and 45 potential tumor-suppressors discussed above were combined with drivers defined by IntOGene (Integrative Onco Genomics, [www.intogen.org](http://www.intogen.org)). IntOgene was used to select driver genes based on protein affecting mutations. In total, 459 genes were defined as drivers in our sample cohort (**Figure 6-9**). Only ~23% (105 genes) of them were epigenetic regulators. The intersection of the 2 approaches produces a list of 84 reliable driver genes that were used for further analysis (**Figure 6-10**).



**Figure 6-9: Driver cloud plot represents the most recurrently mutated cancer driver genes generated by IntOGene.** The size of the gene symbol is relative to the number of samples with protein affecting mutations.



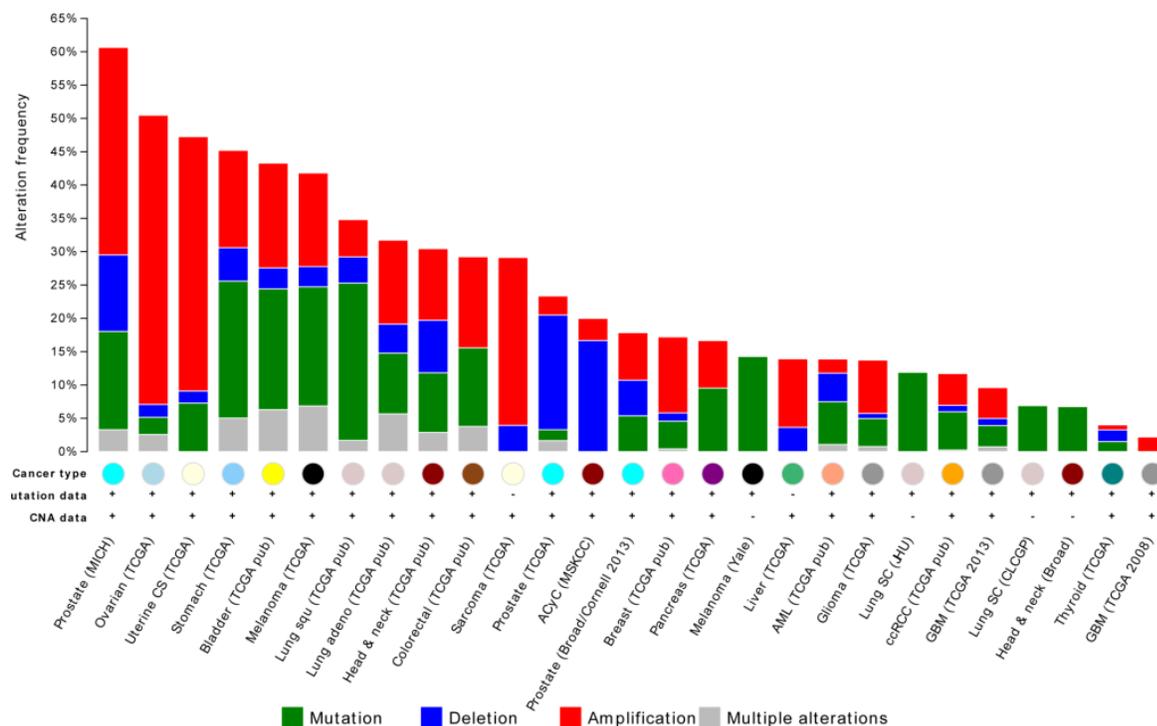
**Figure 6-10:** Venn diagram represents overlap of results between two approaches. Violet circus represents 121 drivers, identified by TSG&oncogene approach, described below and blue circus - 105 drivers, defined by IntOGene).

### 6.1.2. Alterations in complexes of epigenetic regulators in cancer

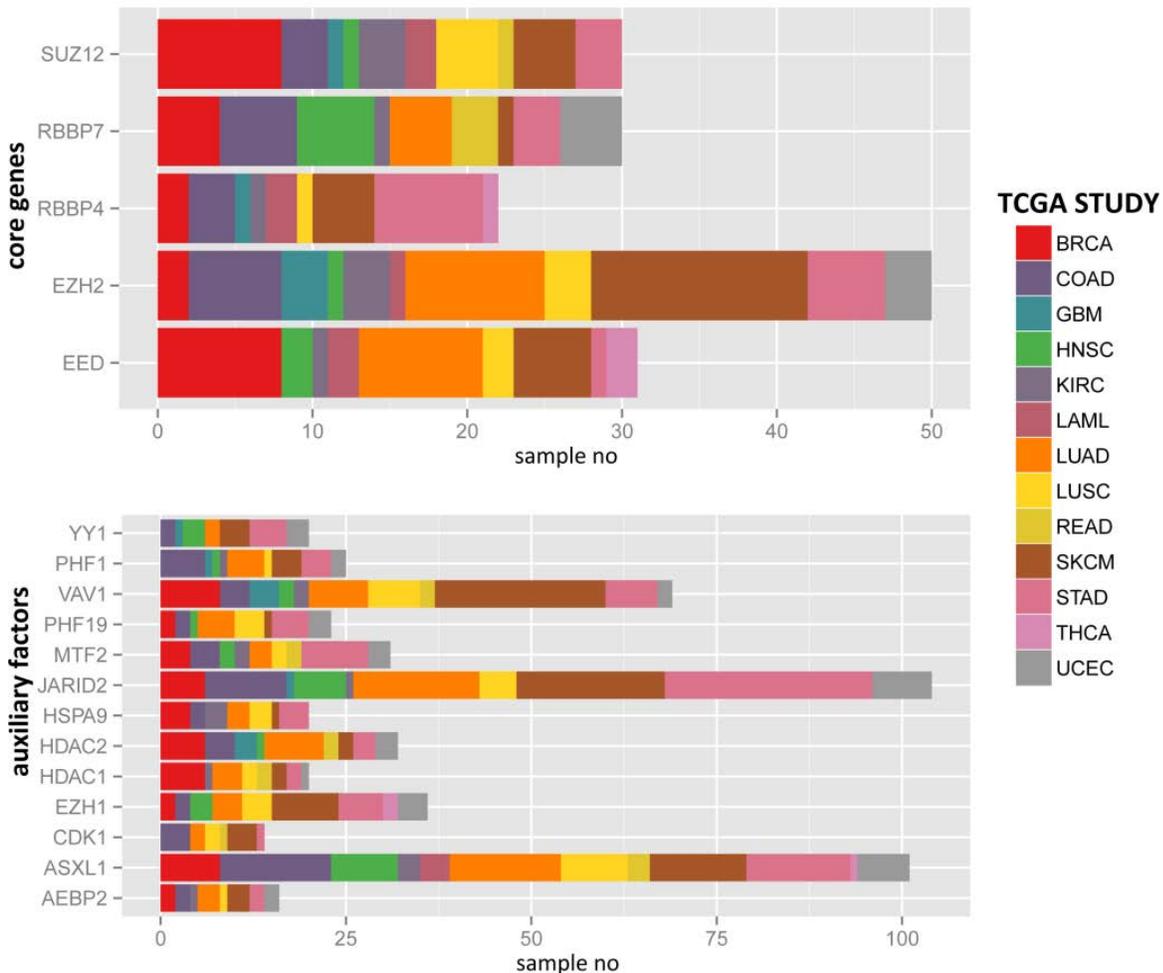
Alterations in different components have been identified by an integrative approach, indicating the complexity of regulation of epigenetic machinery in carcinogenesis. A novel systematic approach should involve the investigation of alterations in epigenetic regulators not alone but also consider other genes that are acting in the same epigenetic pathway or execute similar functions. To systematically screen alterations in different components of epigenetic complexes, cross-cancer analysis of defect of PRC2 genes (n=18) was performed using the cbio platform (cbio.org), which consists of genome-wide alterations data from different cancer types.

Up to 55% of the samples have defects in at least one component of the PRC2 complex in prostate cancer. Ovarian and lung cancers also show high percentage of affected samples, indicating that PRC2 is frequently altered and important for cancerogenesis (**Figure 6-11**).

In order to understand which alteration of component of the PRC2 complex might be a driver event, mutation distribution of each gene was further investigated using the TCGA data. Previously published driver genes were confirmed, including *EZH2* [145], *EED* [145] and *SUZ12* [145] in addition to novel putative drivers *JARID2* and *VAV1* (**Figure 6-12**).



**Figure 6-11: Cross-cancer alteration summary for PRC2 genes in different cancer types.** Histogram represents proportion of samples with alterations in any component of PRC2 genes. The vertical measures alteration frequency in percentage and the horizontal axis lists TCGA cancer studies. Bar color denote type of alterations.



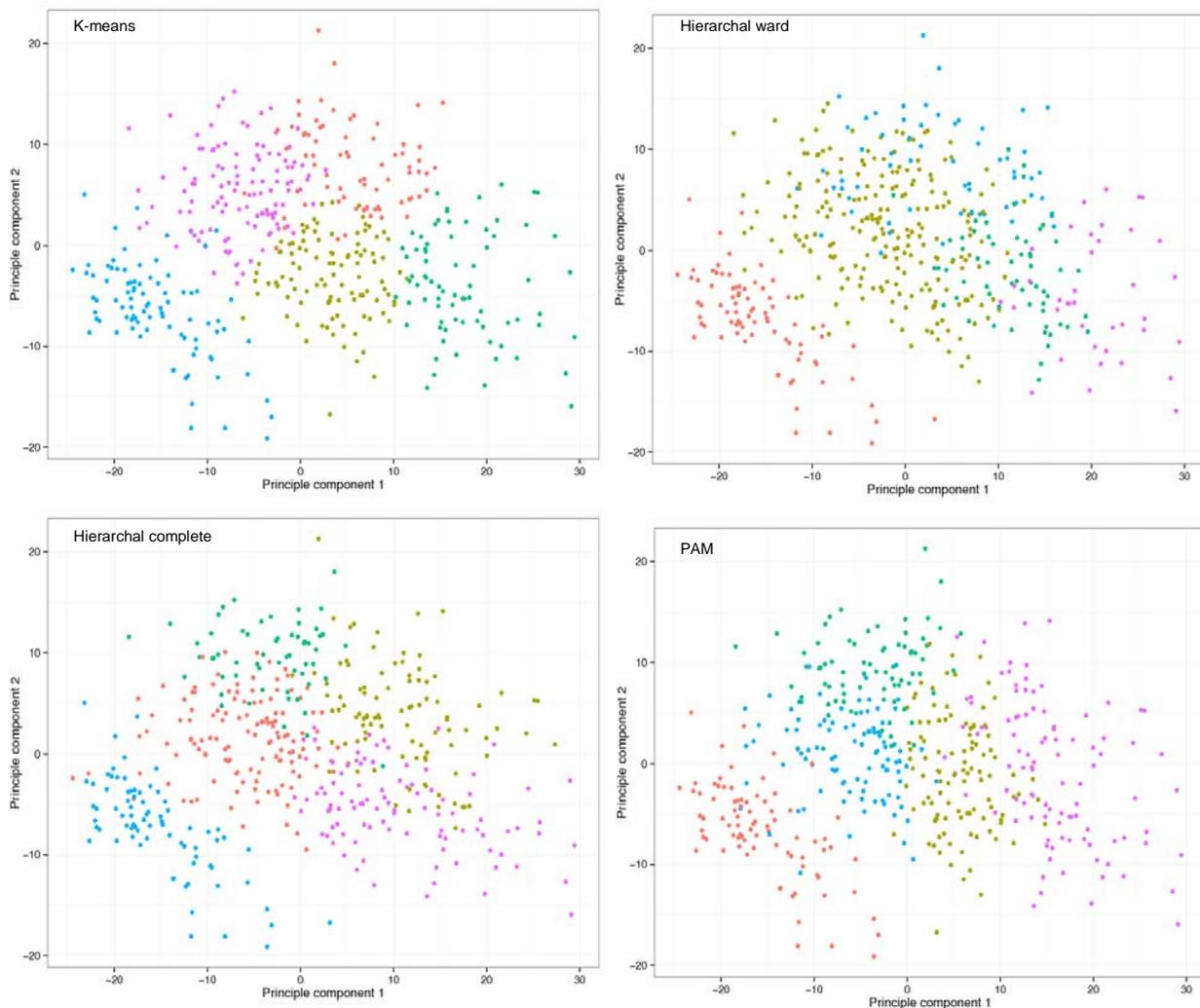
**Figure 6-12: Mutation distribution of PRC2 genes in different cancer types.** Histogram represents number of samples with mutations in each cancer study per PRC2 genes. Genes are listed on vertical axis; sample number is plotted on horizontal axis. Colors denote TCGA cancer studies.

## 6.2. Identification of patterns/clusters of DNA methylation in cancer

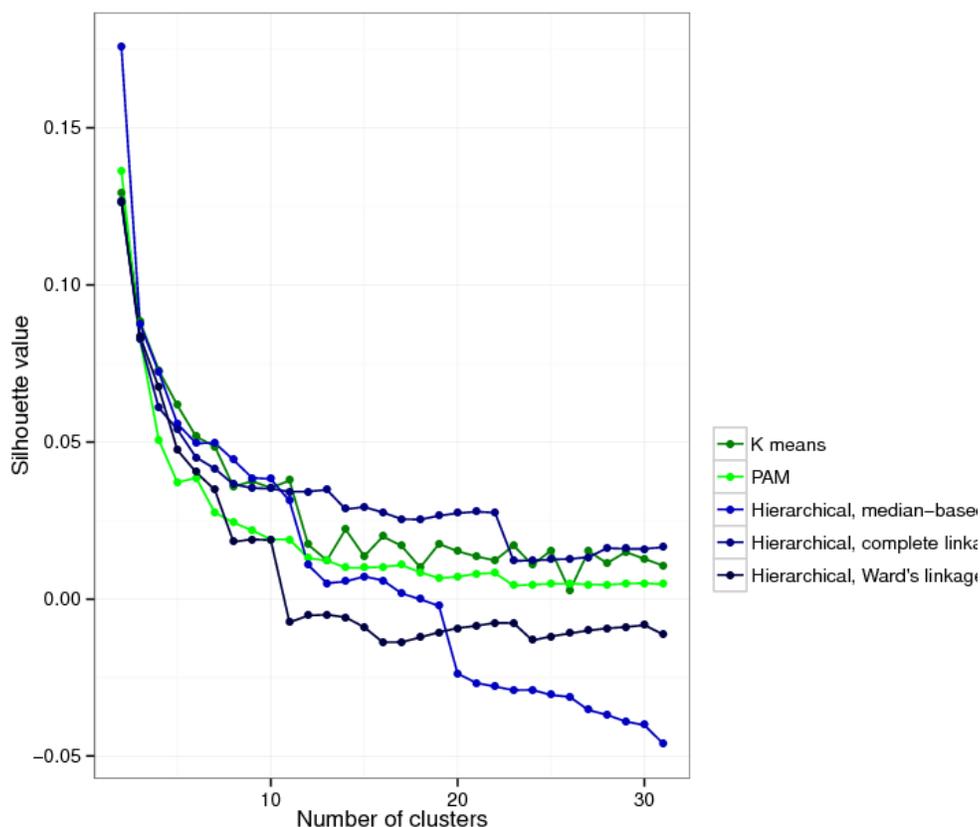
To determine the subgroups of DNA methylation clusters associating with mutation spectra, data from [TCGA PANCAN data set](#) of 450K Illumina arrays was used. All available raw (IDAT) files were downloaded from the TCGA data matrix and only samples for which genetic data is also available were used for further analysis. Raw data was preprocessed using the RnBeads analysis pipeline as described in the methods section. Briefly, raw files were loaded, low-quality probes as well as probes overlapping with known SNPs were filtered out, and then samples were normalized using the BMIQ algorithm.



The second strategy unifies the number of probes per cancer study. It selects the top 10K most variable probes for further analysis. Each cancer study was analyzed using 3 different cluster algorithms: PAM, K-means and Hierarchical (using both Ward and complete linkage) as described in detail in the **chapter 5.7**. Data was represented using principal component analysis (PCA). Comparison of performance of different clustering algorithm for breast cancer data set is plotted in **Figure 6-14**. As an example, the BRCA study contains 457 primary tumor samples and five different methylation patterns/subtypes have been identified in this dataset. Performance of the different clustering strategies was evaluated using average silhouette value, as shown in **Figure 6-15**.



**Figure 6-14: Comparison of outcome of different cluster procedures of selected probes (10K) for breast cancer data set using principal component analysis (PCA).** This plot represents comparison of 4 main cluster algorithms used for analysis on the same data set (top 10K most variable probes). For each algorithm PCA1 vs PCA2 are represented. Colors denote the 5 different methylation patterns/subtype identified in breast cancer patients.

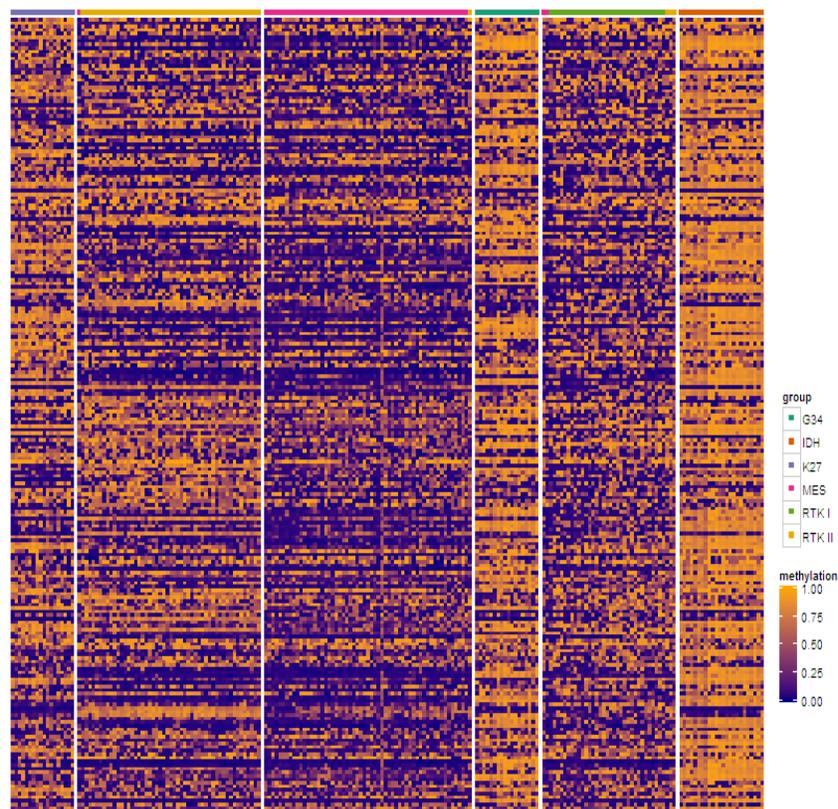


**Figure 6-15: Comparison of performance of studied clustering algorithms of selected probes (10K) for breast cancer data set using the average silhouette value as a grading scheme.** Every clustering method is denoted by a different color. The X axis represents number of clusters, and the Y axis – silhouette value.

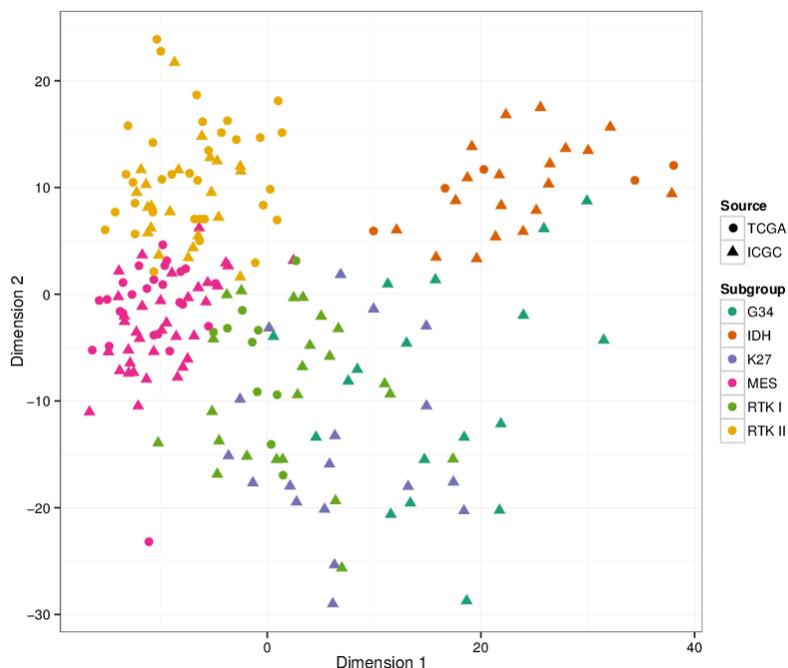
We decided to use a “gold standard” for the validation of the described approach to clustering and the selection of a single strategy. We focused on the glioblastoma data set, for which there are published well defined methylation clusters and identified epigenetic regulators that can explain methylation patterns [88]. In the glioblastoma data set consisting of 98 tumors, 6 major clusters have been previously identified. Knowing the number of identified clusters and using different clustering procedures, each of our clustering strategies was examined. In order to test which cluster algorithm fits best the published methylation patterns, the Rand index was computed. Adjusted Rand index values (calculated between the six subgroups of glioblastoma and identified clusters by our approach) rank K-means as the most appropriate clustering algorithm (**Figure 6-16**) and it was selected for further analysis. **Figures 6-17** and **6-18** confirm the identified clustering approach as well suited to the glioblastoma data set.

K means	0.455	0.686	0.798	0.905
PAM	0.378	0.488	0.574	0.634
Hierarchical, median-based linkage	0.015	0.014	0.014	0.014
Hierarchical, complete linkage	0.351	0.481	0.670	0.584
Hierarchical, Ward's linkage	0.326	0.342	0.442	0.521
	3	4	5	6
	Clusters			

**Figure 6-16: Performance of all applied clustering algorithms of selected probes (10K) for glioblastoma published cancer data set using adjusted Rand index values as a grading scheme.** Each row represents different clustering method. Different values for number of clusters are denoted by columns.



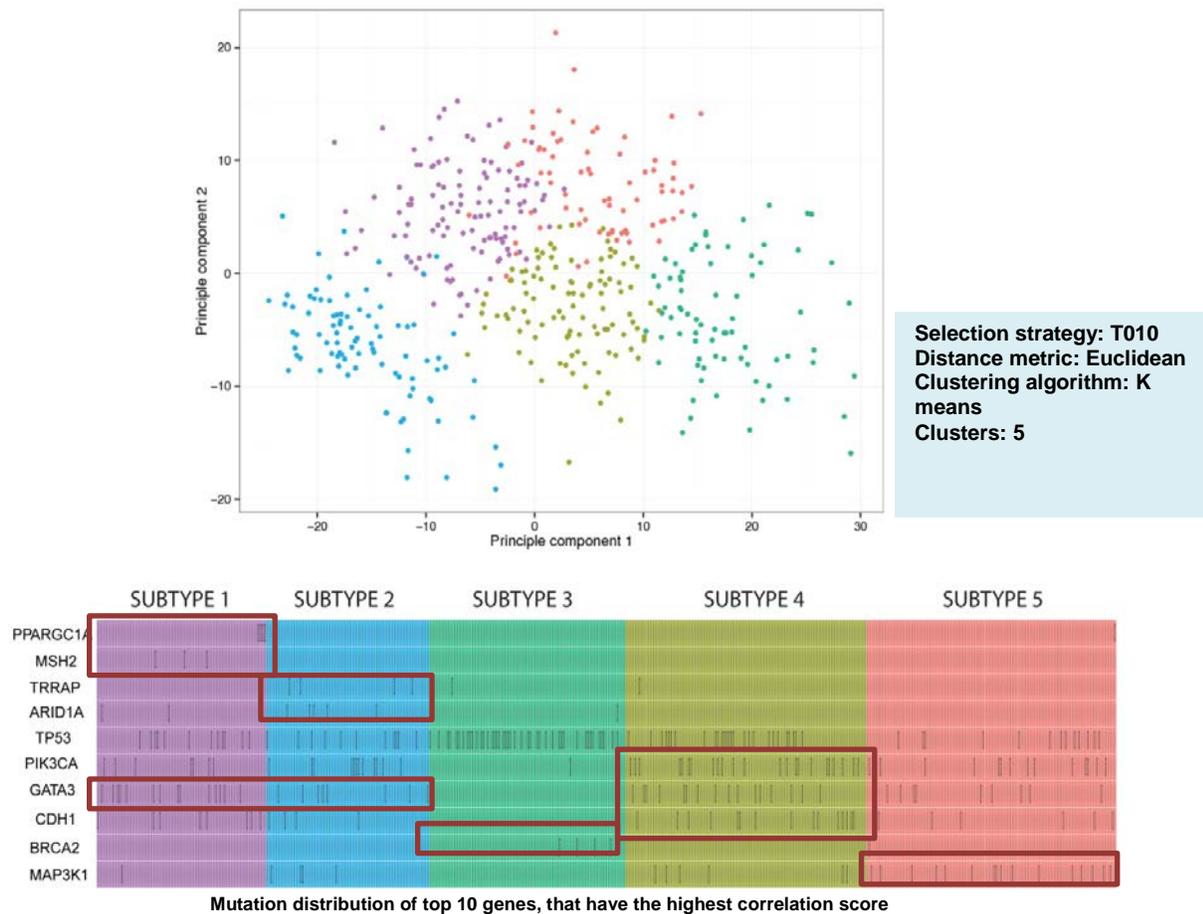
**Figure 6-17: Heatmap of selected probes (10K) for glioblastoma published cancer data set.** K-means clustering algorithm was used for analysis. Different colors in the column header denote previously identified methylation groups. Methylation degree is presented by grading schema: dark blue (no methylation) to orange (highly methylated). Each row denotes an Infinium 450k probe, and each column – a sample.



**Figure 6-18: Principal component analysis of selected probes (10K) for glioblastoma published cancer data set.** Different colors denote for 6 major methylation patterns/clusters. Shapes indicate the different sources that were used for obtaining the data.

### 6.3. Correlative analysis of methylome data with defects in epigenetic regulators

Since, different pattern/clusters of methylation data have been obtained (described in **chapter 6**), as well as putative driver epigenetic regulators (described in **chapter 6**) have been identified, correlative analysis of two data levels: methylation and defects in driver genes, was performed. For each cancer study independently, the samples have been grouped according to their clusters and for each group a list of alterations of epigenetic regulators were analyzed. Association between sample cluster and defects in the epigenetic regulators was quantified using Fisher's exact test. Since not all identified methylation patterns had any significant epigenetic regulator, this analysis was also conducted using all altered genes in genome. Including other genes could potentially lead to novel epigenetic regulators or can give a better understanding how previously published cancer drivers could be involved in deregulation of cancer methylomes. Up to 30 % of the identified methylation subgroups were found to be strongly correlated an alteration in the cancer genome, and 9% of the subgroups are associated with alterations in epigenetic regulators. The breast cancer data set is used to present an example outcome of this analysis (**Figure 6-19**).

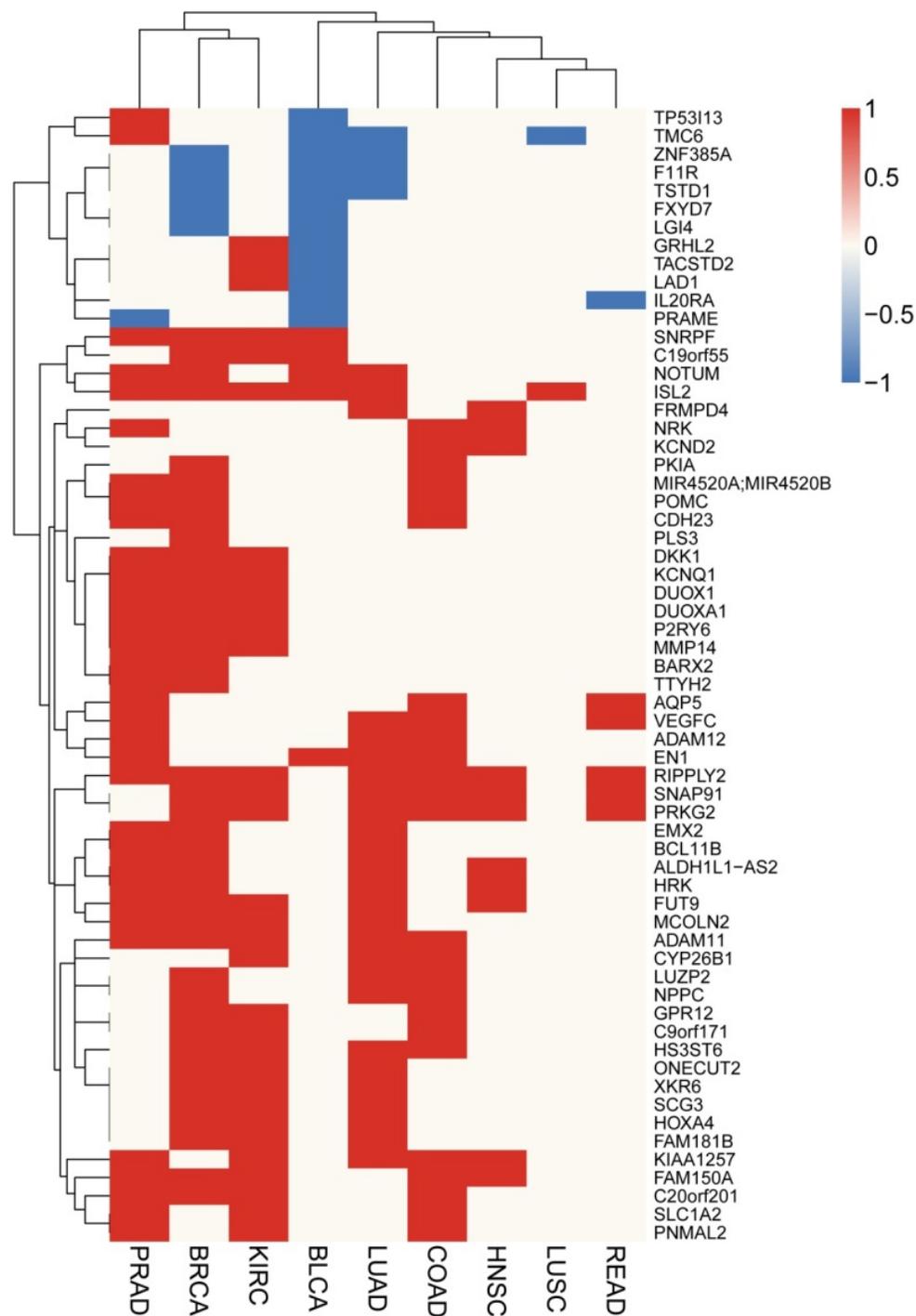


**Figure 6-19: Correlative analysis of selected probes (10K) for breast cancer data set.** Top panel presents PCA plot for K-means clustering of methylation data of breast cancer samples. Low panel presents mutation distribution of top 10 genes that have the highest association.

#### 6.4. Differently methylated regions in cancer

Differently methylated CpG sites and annotated in RnBeads regions (promoters, gene bodies, tiling 5kb regions) were identified using tumor vs. normal comparison per cancer study. Briefly, each cancer study was separated analyzed using the RnBeads pipeline for differently methylation sites and regions as described in methods in details. Only positive (significantly differently methylated in tumor vs. normal samples) sites/regions were taken for further analysis. Using the approach for identification PANCAN driver regulators, described in **chapter 6-3**, the positive sites/regions were combined and analyzed for identification PANCAN and tissue-specific events. In total 105 PANCAN promoter DMRs were identified in the investigated cancer

cohort. 72% of DMRs were consistently hypermethylated and 28% were hypomethylated in gene promoters (**Figure 6-20**). Such hypermethylated promoter DMRs represent putative tumor-suppressor genes. 87 PANCAN DMRs in gene bodies were identified in the data with almost equal distribution of hypo- and hypermethylated DMRs (**Figure 6-21**).



**Figure 6-20: Promoter differentially methylated regions (DMRs) in PANCAN.** Heatmap representing PANCAN DMRs in promoter regions of genes (>50% tumors) per TCGA cancer study. Color schema denotes methylation state of DMR. The vertical axis lists genes, and the horizontal one – cancer studies.



## 6.5. Discussion of results and deliverables

### Driver epigenetic regulators

Due to a big spectrum of tools for the identification of driver mutations, there is not a consensus in the field regarding which tool shows the most comprehensive and the most reliable results. One way is to use only one approach during analysis, another way is to combine multiple tools as the best approach to obtain a more “comprehensive and reliable” list of driver genes. This thesis develops a strategy to combine different tools with the goal of identifying putative epigenetic driver genes. The obtained list of genes could be further refined using novel tools and larger sample cohorts. The development of a unified method aims at avoiding false signals and allows comparing results obtained in different laboratories. An upcoming tool, The Cancer Genome Interpreter (developed by the Lopez-Bigas group in Barcelona), will also include drug information to already developed driver searcher that will help to interpret which mutations in a given tumor are important in cancer development and whether options for target therapies exist that are directed toward those driver events. Such new tool will provide a platform for clinicians to start a proper treatment just based on patient mutation profiling.

### Validation of identified drivers

Although using different approaches can provide quite reliable list of drivers, the validation of finding is required using additional data cohort, but more important in the laboratory. Validation of putative driver events requires equal consideration in the selection and interpretation of biological assays used to determine their relevance to cancer is required. Crucial steps for validation are the selection of the proper biological assay, since it is not entirely clear in what stage of the tumorigenic process the driver has a role; and relevant cell lines are [146]. The example of the *NOTCH1* gene, which acts as oncogene in leukemia and as tumor suppressor gene in head and neck and lung cancers presents a challenge in the functional validation of driver genes [146]. In some cases, cooperation with additional cancer-relevant genes to affect tumorigenesis or bypass oncogene-induced senescence is required. For example, the melanoma-associated oncogene *MITF* (microphthalmia-associated transcriptional factor) affects the proliferation of immortalized melanocytes only when co-expressed with a *BRAF*<sub>V600E</sub> mutant [147]. To confirm a driver mutation for a given gene genetically engineered mouse models (GEM) is currently used approach. Sequencing a mouse acute promyelocytic leukemia genome reveals a deletion in *Kdm6a* (lysine-specific demethylase 6A) and recurrent *Jak1* (janus kinase 1) <sub>V657F</sub> mutation. Human *JAK1*<sub>B658F</sub> mutation has been reported in acute promyelocytic

leukaemia, demonstrating relevance to use GEM models for the validation of driver mutations [148].

### **Integrative analysis of cancer genomes and epigenomes**

One of the biggest challenges in interpreting data from TCGA and ICGC is the lack of comprehensive unified integrative analyses. Several research groups have created their own pipelines and tools for integrating 2 or more data types. In this thesis own pipeline was used for data integration simply by adding extra data level: mutation -> CNA -> promoter methylation -> identification of drivers. There are other important genomic data types which are not addressed in this Chapter. Examples include non-coding mutations and regulation of genes by ncRNAs. These mechanisms were largely ignored due to very limited knowledge of their action on coding genes and will be discussed in depth in **chapter 9**.

### **Further validation and usage of obtained results**

This work provides a big spectrum of deliverables for follow-up investigations and in-depth studies. Such type of data can help to validate our findings on independent data cohort or compare findings in one cancer to other cancer types.

#### **Deliverables:**

- List of driver epigenetic regulators
- Methylome clusters per each cancer study
- Correlation matrix of epigenetic regulators and their complexes with methylation clusters
- Tissue-specific and cancer-specific differently methylation regions (DMRs)

### **Inclusion of data into Epigenome browser**

One of the benefits of obtained data is the possibility to include methylomes for cancers to available cancer genome browsers, like the Epigenome browser (<http://epigenomegateway.wustl.edu/>), the UCSC Genome Browser (<https://genome.ucsc.edu/>) or to IGV (<https://www.broadinstitute.org/igv/>). Such data can be used for visualization of the regions of interest per sample as well as for sample cohort.

---

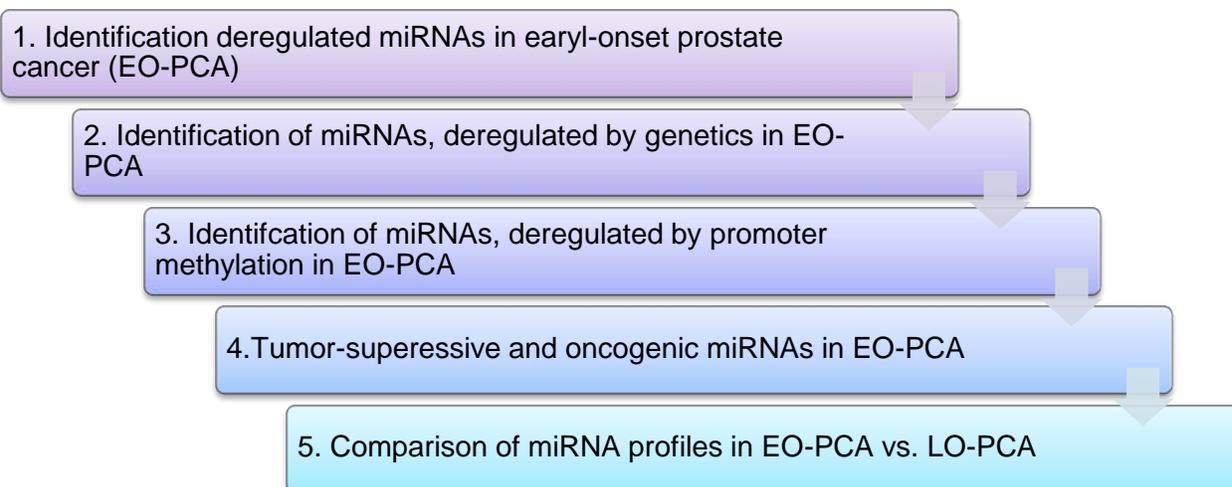
## CHAPTER 7.

---

### GLOBAL INTEGRATIVE ANALYSES OF DEREGULATED MIRNAS IN PROSTATE CANCER

In the previous chapter, 80% of tissue-specific DMRs were identified in regulated regions of miRNAs and lncRNAs. To assign functional relevance to identified DMRs in regulatory regions of miRNAs, a global integrative analysis was conducted in comprehensive data cohort of early-onset prostate cancer released by Germany ICGC consortium. Such systematic genome-wide evaluation of miRNA deregulation will allow identifying the mechanism of miRNA deregulation in PCA and will open a question of miRNAs role in prostate cancerogenesis.

In this chapter we work on the hypothesis that in addition to genetic events, epigenetic alterations lead to deregulation of miRNA transcription. In order to unravel the molecular alterations leading to miRNA deregulation in PCA we performed the first genome-wide integrative analysis of expression, DNA methylation and genetic alterations of miRNAs of 66 EO-PCAs specimens generated within the German ICGC project (ICGC EO-PCA dataset) [118] [Weischenfeldt et al., manuscript in preparation]. The workflow of the integrative analysis consists of five steps of data integration, starting with the analysis of expression profiles of miRNAs in EO-PCA, followed by the analysis of copy number alterations, integration with DNA methylation profiles (**Figure 7-1**).



**Figure 7-1:** Schematic workflow for the integrative analysis of mechanisms leading to miRNAs deregulation in EO-PCA.

**Step 1:** identify differently expressed miRNAs in EO-PCA using miRNA-seq data of 66 EO-PCAs samples and 8 normals as controls.

**Step 2:** identify commonly genetically altered miRNAs in EO-PCAs using copy number alteration (CNA) profiles of 66 EO-PCAs samples.

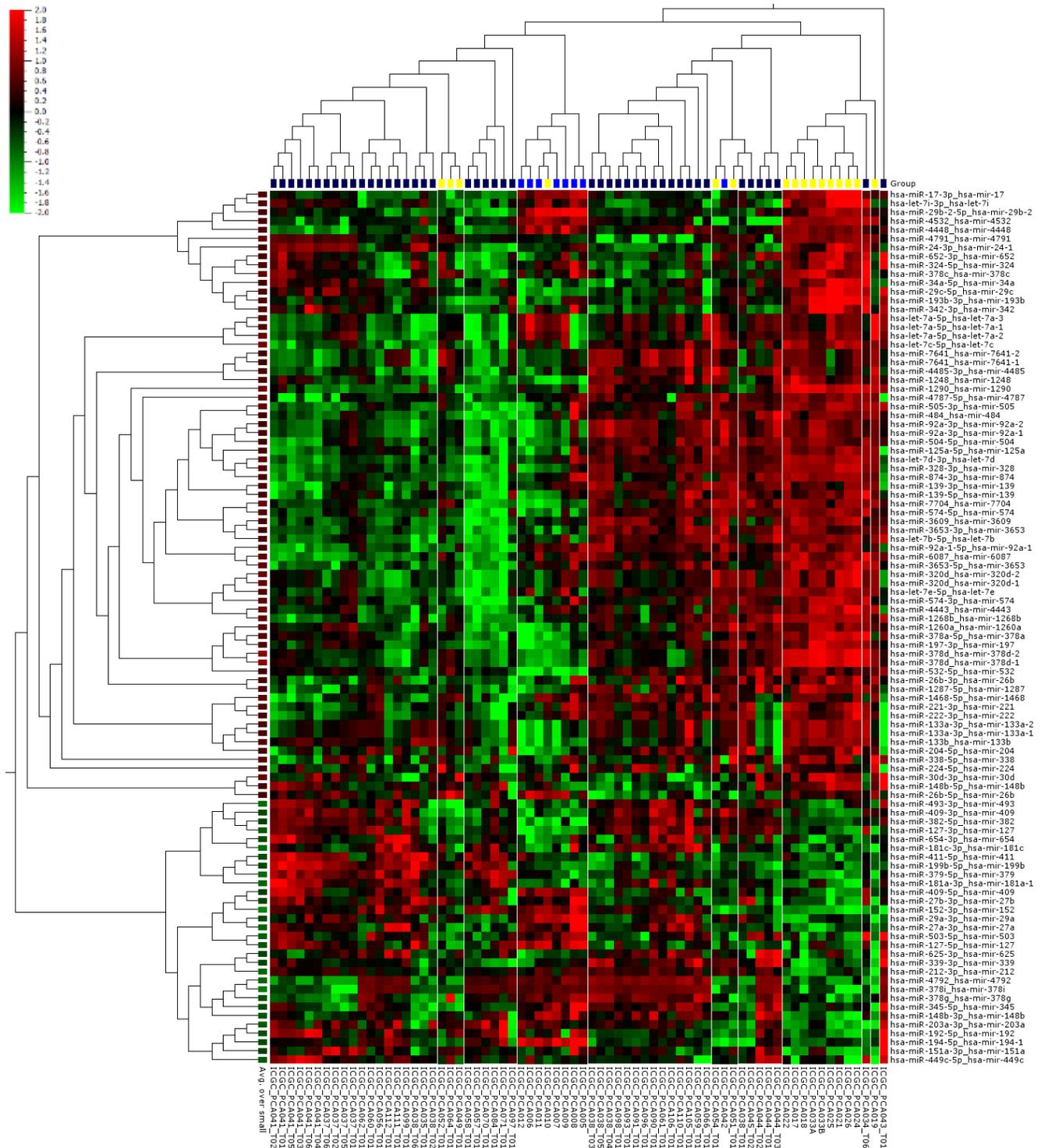
**Step 3:** Identify miRNAs with promoter/regulatory regions DNA methylation using MCip-seq data from 33 EO-PCA and 1 normal and extend analysis using Illumina 450K arrays from 155 EO-PCAs and 44 normals.

**Step 4:** Identify miRNAs, which are deregulated in EO-PCAs by two hits: promoter methylation and CNA. Based on data assign miRNAs to tumor-suppressive and oncogenic miRNAs in EO-PCAs.

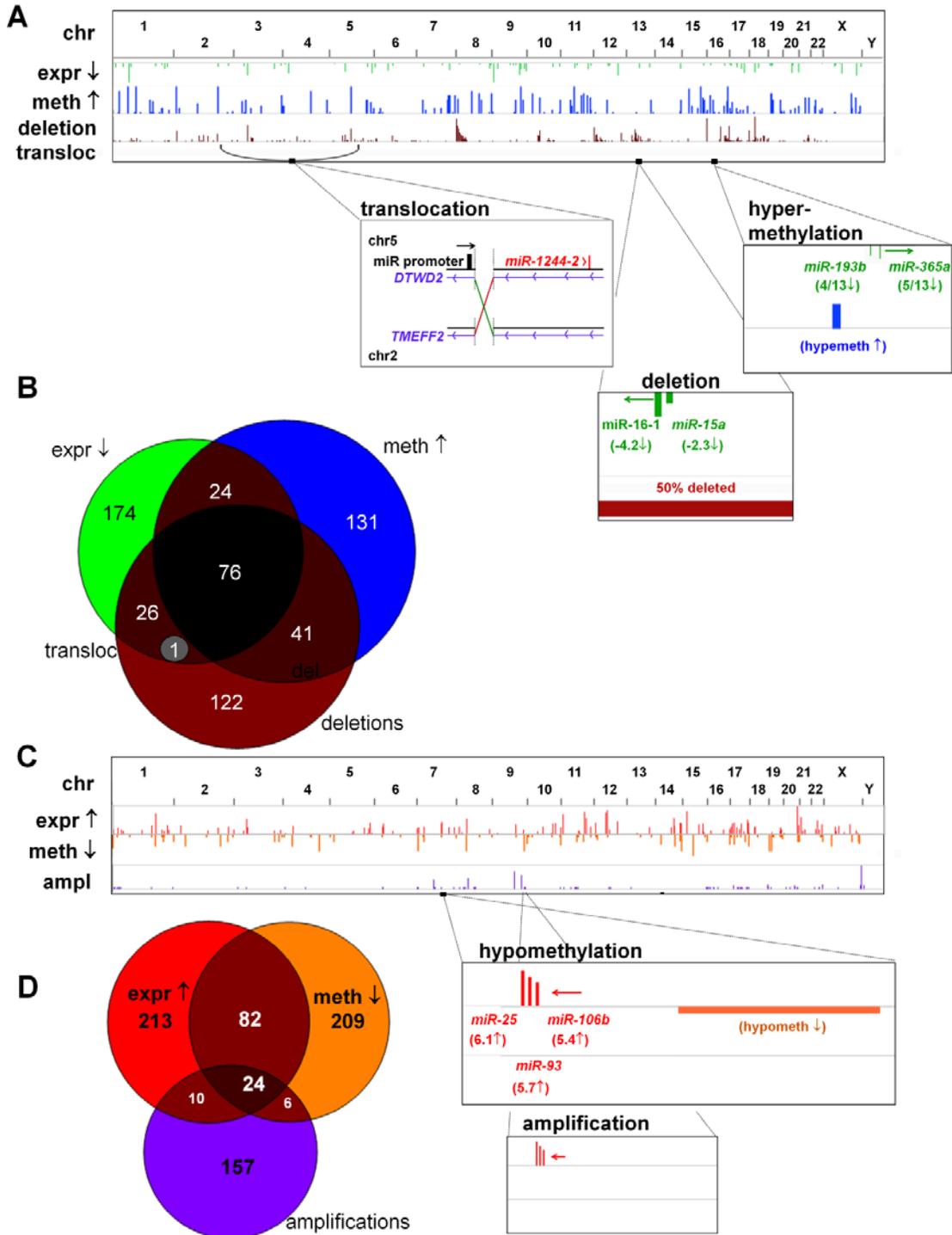
**Step 5:** Compare miRNAs profile of EO-PCA vs. LO-PCA, identify major mechanism of deregulation of miRNAs in PCA.

### **7.1. Global deregulation of miRNA expression in early-onset prostate cancer (EO-PCA)**

In order to comprehensively define molecular alterations leading to miRNA deregulation in prostate cancer, deep sequencing data on genome-wide miRNA expression, global DNA methylation, structural variations and single nucleotide variants (SNVs) of 66 EO-PCA specimens generated within the German ICGC project (ICGC EOPC dataset) [118], Feuerbach *et al.*, manuscript in preparation] were utilized. Using a read-counts cutoff of >100 reads, a total of 911 miRNAs were expressed in at least one sample. Using DeSeq algorithm to identify differently expressed miRNAs, pool of 66 tumors and 8 normal controls were used for comparison and in total 174 miRNAs were significantly downregulated and 213 miRNAs were upregulated sample cohort (top deregulated miRNA are shown in **Figure 7-1**). These deregulated miRNAs were distributed over the entire genome, and no apparent clustering at certain chromosomal regions was detectable (**Figure 7-1**). Notably only 57% of identified miRNAs were reported to be differently expressed in PCA before [145], providing 43% novel miRNAs, important for EO-PCA cancerogenesis and can be potentially early events in prostate cancer development.



**Figure 7-1: Heatmap of expression profile of miRNAs in 66 early-onset PCA cases (ICGC EO-PCA dataset).** Unsupervised clustering analysis of top 100 most variable miRNAs. Columns represent samples and rows represent corresponding miRNAs. Columns heading denotes for sample type: yellow- normal, blue –small tumors (d <= 3 sm), dark blue – large tumors (d >3sm). Expression level is presented as log<sub>10</sub> (normalized read counts). Bright green denotes 2 (low expression level), bright red denotes 2 (high expression level). Hierarchical clustering was performed using Euclidian distance and complete linkage using Qlucore software 3.1.

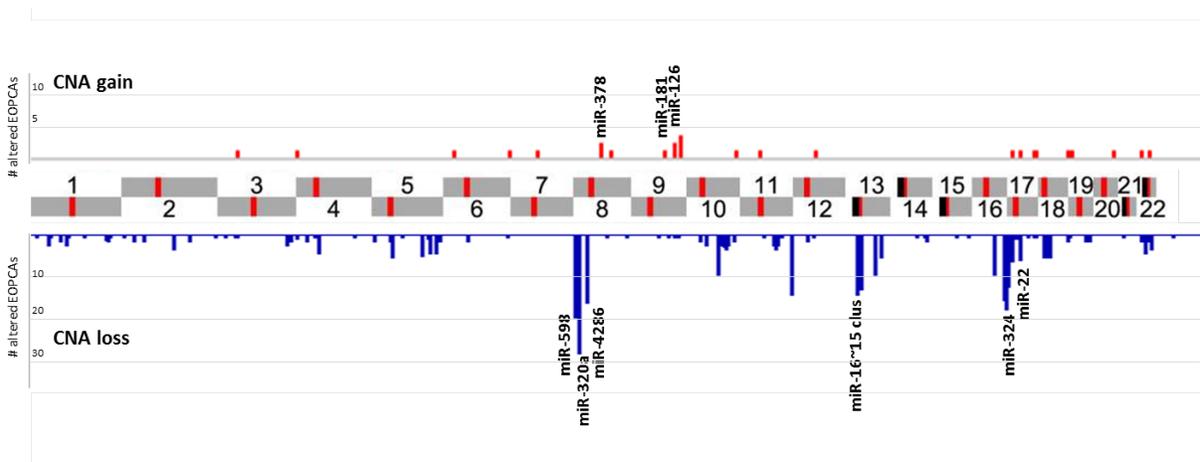


**Figure 7-2: Genome-wide genetic and epigenetic miRNA deregulation in 66 early-onset PCA cases (ICGC EO-PCA dataset).** (A) Genome-wide representation of miRNA downregulation (green bars), hypermethylation of miRNA promoters (blue bars), deleted miRNA precursors (dark red bars) and one translocation directly affecting a miRNA precursor. Bar height indicates the number of affected PCA samples. The detailed views show examples of miRNAs/miRNA clusters likely downregulated by translocation, deletion or hypermethylation (arrows indicate direction of transcription). (B) Venn diagram showing the overlap of mature miRNAs that are downregulated and affected by promoter hypermethylation, deletion and translocation in any of the 66 patients. Only

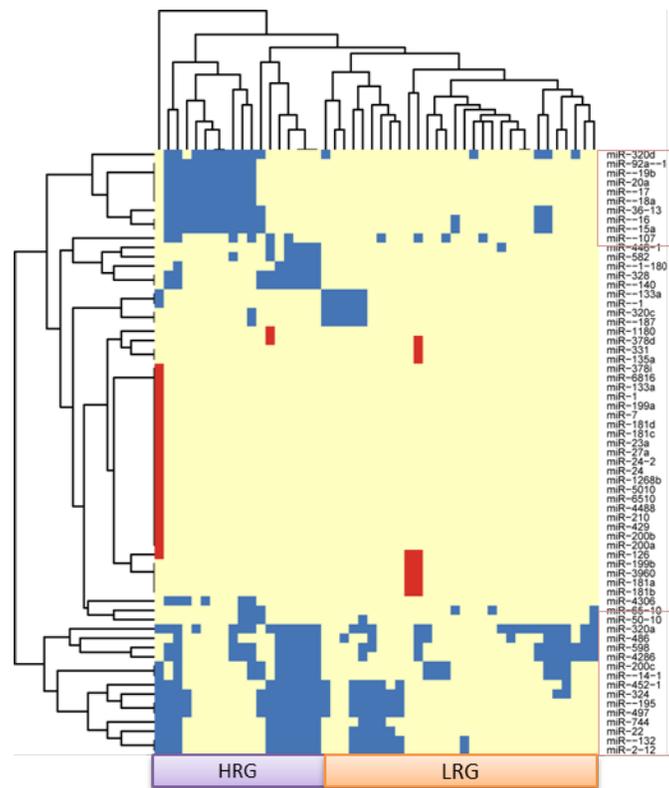
expressed miRNAs with mapped regulatory regions were considered. (C) Genome-wide representation of miRNA upregulation (red bars), promoter hypomethylation (orange bars) and amplified miRNA precursors. Bar height indicates the number of affected tumor samples. The detailed view shows a miRNA cluster likely upregulated by promoter hypomethylation (arrow indicates direction of transcription). (D) Venn diagram showing the overlap of mature miRNAs that are downregulated and affected by promoter hypermethylation, deletion and translocation in any of the 66 patients. Only expressed miRNAs with mapped regulatory regions were considered.

## 7.2. Genetic and epigenetic alterations associated with miRNA downregulation in EO-PCA

In order to unravel the molecular mechanisms leading to miRNA downregulation miRNA precursors which either carried single nucleotide variants (SNVs) or which were affected by deletions or promoter hypermethylation were searched. No SNVs were detected, suggesting that SNVs play only a minor role in miRNA deregulation. However, a total of 605 miRNA precursors were deleted in at least one tumor sample (**Figure 7-3**), and 100 of those overlapped with the 174 downregulated miRNAs (58%; see previous chapter, **Figure 7-2**), suggesting a causal link between deletion and downregulation in these cases. Notably, the most frequently deleted miRNA precursors were located in regions commonly deleted in PCA such as 8q, 13q and 17q [145, 149-152], also including *miR-15a~16-1* (13q14; **Figure 7-3**), which is frequently deleted in PCA [145]. In total, two times more miRNAs were observed to be affected by CNA losses compare to CNA gains. EO-PCA samples were separated in two distinguish classes using miRNA profile of genetically altered miRNAs. Identified clusters reflect the number of total genetic rearrangements and were named: high and low rearrangement groups (HRG and LRG respectively) (see **Figure 7-4**).

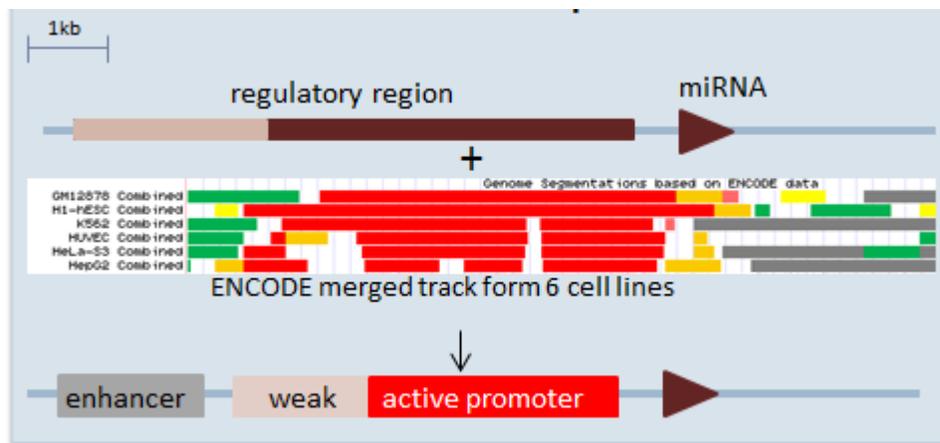


**Figure 7-3: Copy number genome-wide profile of miRNAs in 66 early-onset PCA cases (ICGC EO-PCA dataset).** Red/blue bars indicate the location on chromosome maps of miRNAs with CNA gains/loss respectively in PCAs. High of bar represents number of altered EO-PCA cases. The most frequently genetically altered miRNAs are named.

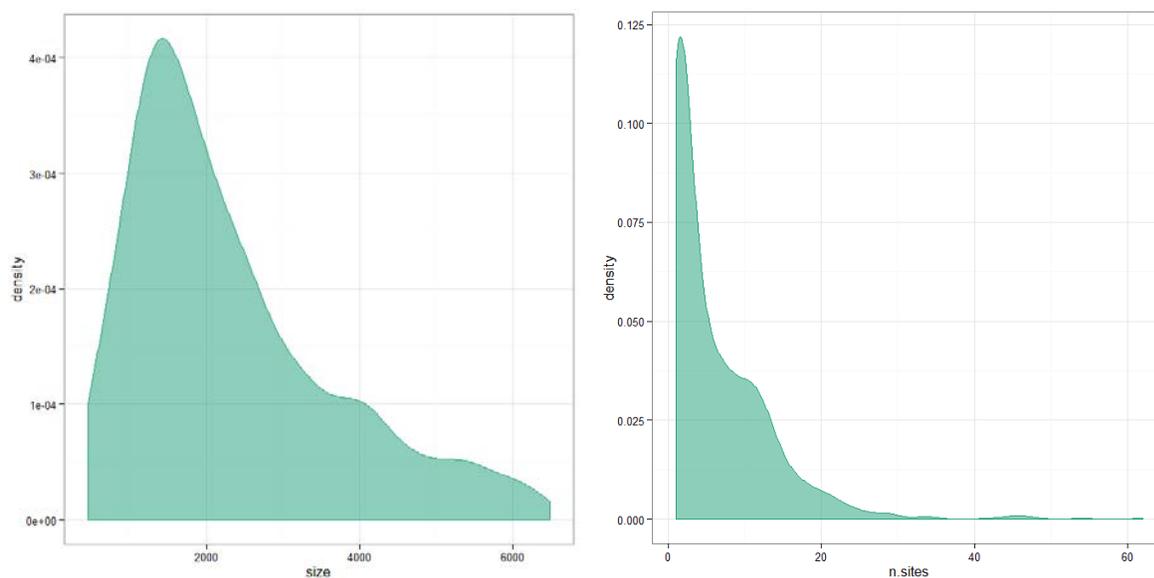


**Figure 7-4: Heatmap of genetically gained/loss miRNAs in 66 early-onset PCA cases (ICGC EO-PCA dataset).** Columns represent samples, rows represent frequently altered miRNAs. With blue/red color marked miRNAs, altered by CNA losses/gains respectively.

Since, not all downregulated miRNAs could be explained by genetic alterations, miRNAs associated with promoter methylation were screened using ICGC EO-PCA methylomes (MCIp-seq data). Using previously defined 1374 regulatory regions from Baer et al. [16] covering 329 annotated miRNAs and newly annotated miRNAs by miRBase (mirbase.org), a list of putative promoter regions was generated by overlapping with active promoter marks as annotated by ENCODE (Figure 7-4). ENCODE tracks were obtained by merging active histone marks (such as H3K4me3) of six published cell lines (<http://genome.ucsc.edu/ENCODE/>). MiRNAs putative promoter regions have been annotated by regions length and number of CG sites that are present in 450K arrays. In average, about 30% of all miRNAs predicted promoter regions are covered by 450K arrays with 10 probes per region (range 1 to 65 CpGs). Regulatory regions and coverage by 450K Illumina arrays are described in **Figure 7-5**.



**Figure 7-5: Workflow of identification of putative promoter regions of miRNAs.** 1374 regulatory regions from C.Baer (Baer C et al. Cancer Res. 2012) and newly annotated miRNAs with 35kb upstream transcription start sites were overlapped with promoter active tracks annotated by ENCODE.

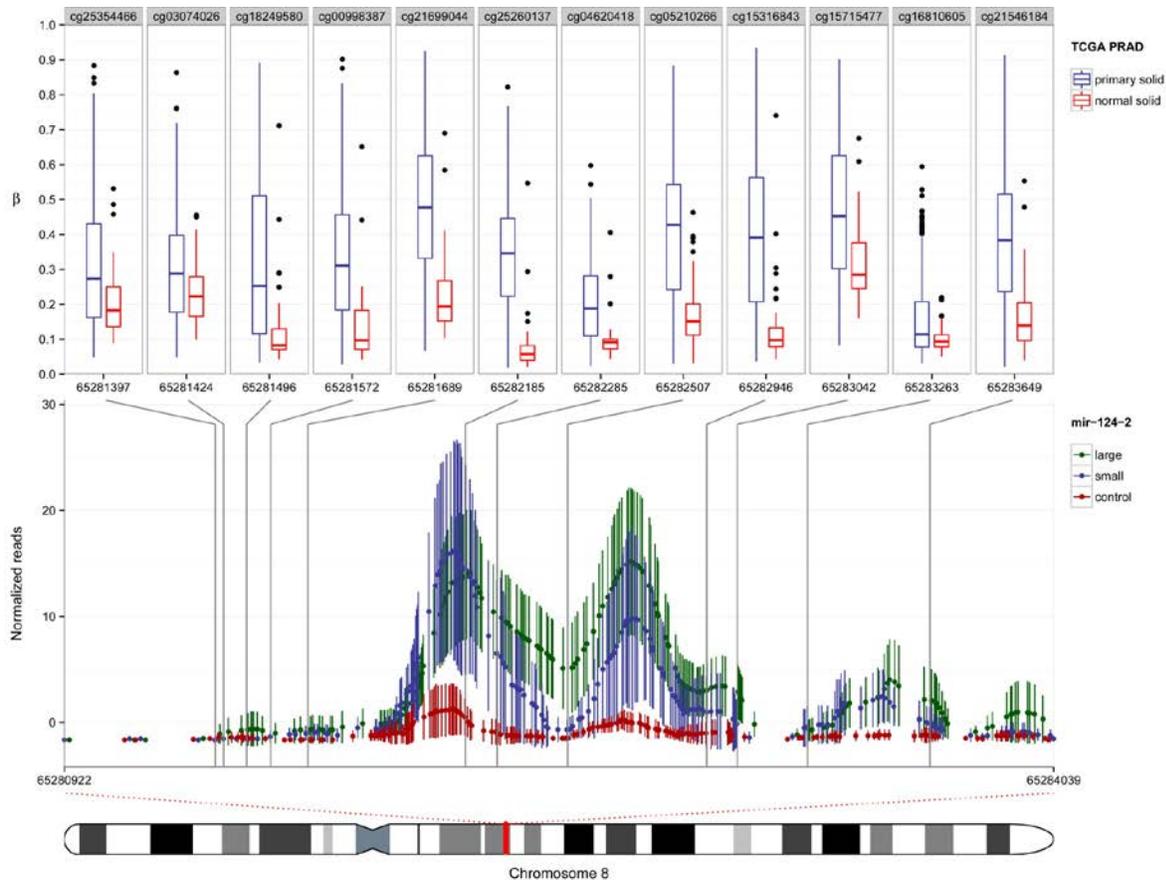


**Figure 7-5: Annotation of putative promoter regions of miRNAs.** Left density plot represents promoter region length distributions. Right density plot represents distribution of number of sites per region of miRNA promoters.

Using ICGC EO-PCA MCIP-seq data to analyze methylation of miRNA promoters, 237 miRNA precursors with hypermethylated promoter regions were identified (**Figure 7-2 A**). Of these, 109 miRNA precursors were recurrently in PCA cases (13%) and 163 of the 676 downregulated miRNAs (24%) were hypermethylated in at least one tumor. Downregulation of these 163 miRNAs, including cluster *miR-193b~365a* previously described to be epigenetically regulated in PCA cells [145] (**Figure 7-2 A**), might have been caused by promoter hypermethylation.

In order to compare the results of MCIP-seq data, we used independent TCGA published data cohort of prostate cancer samples (450K Illumina arrays). Since 450K arrays were not designed

to cover all promoter regions for miRNAs, only 34% of identified DMRs have at least 1 CpG probe on 450K and were taken for comparison. 67% of regions, covered by both methods have shown correlation of methylation data (example in **Figure 7-6**).



**Figure 7-6: Comparison of methylation degree of promoter of miRNA-124-2 obtained by MCIP-seq and using polished TCGA PRAD data set.** Upper part represents data obtained by TCGA PRAD data set: each plot shows methylation degree of CpG unit from 450K arrays. Blue color denotes primary tumors, red color – normal. Lower part presents data obtained by MCIP-seq arrays from ICGC EO-PCA data cohort. Each dot represents normalized read counts with a link to corresponding 450K array prob. Green/blue color denote large/small tumors, red color – normal.

The respective contribution of deletions and promoter hypermethylation to the downregulation of those miRNAs with annotated regulatory regions ( $n = 492$ ) was assessed. Both types of aberrations appeared to be considerably involved in miRNA downregulation (**Figure 7-2 C**): 126/492 (26%) miRNAs were deleted and 161/492 (33%) showed promoter hypermethylation in at least one PCA sample. This indicates that, besides miRNA deletion, promoter hypermethylation may considerably contribute to miRNA downregulation on a genome-wide scale.

### 7.3. Genetic and epigenetic aberrations associated with miRNA upregulation in EO-PCA

Upregulation of miRNA expression could be the result of copy number gains or hypomethylation of promoter regions. Only a single copy number gain (**Figure 7-1 A, C**) was detected in screened ICGC EO-PCA cohort, yet the affected *miR-4307* was not expressed, precluding upregulation by this amplification. In contrast, 180 miRNA precursors displayed promoter hypomethylation, 50 of which were frequently affected PCA cases. Of the 897 upregulated miRNAs, 156 (17%) showed promoter hypomethylation in at least one tumor including the recently described *miR-106b~93~25* cluster [145] (**Figure 7-1 A, C**) suggesting epigenetic activation as the most likely cause for upregulation. Among the 578 upregulated miRNAs with annotated regulatory regions, 133 (23%) had a hypomethylated promoter in at least one tumor sample (**Figure 7-1 A, C**). Taken all observations together, epigenetic alterations seem to be a major mechanism for miRNA activation in EO-PCA.

### 7.4. Translocations associated with miRNAs deregulation in EO-PCA

Using genomic breakpoints in **EO-PCA data cohort**, we searched for miRNAs, located in regions +/-10kb. In total 25 translocation events leading to the disruption miRNAs were observed. However no frequent events were identified. In order to examine the relevance of translocations on deregulation of miRNAs, bigger data cohort should be examined.

### 7.5. Tumor-suppressive and oncogenic miRNAs in EO-PCA

Next, deregulated in EO-PCA miRNAs have been evaluated for either oncogenic or tumor-suppressive function. For this, the mechanism of deregulation has been evaluated.

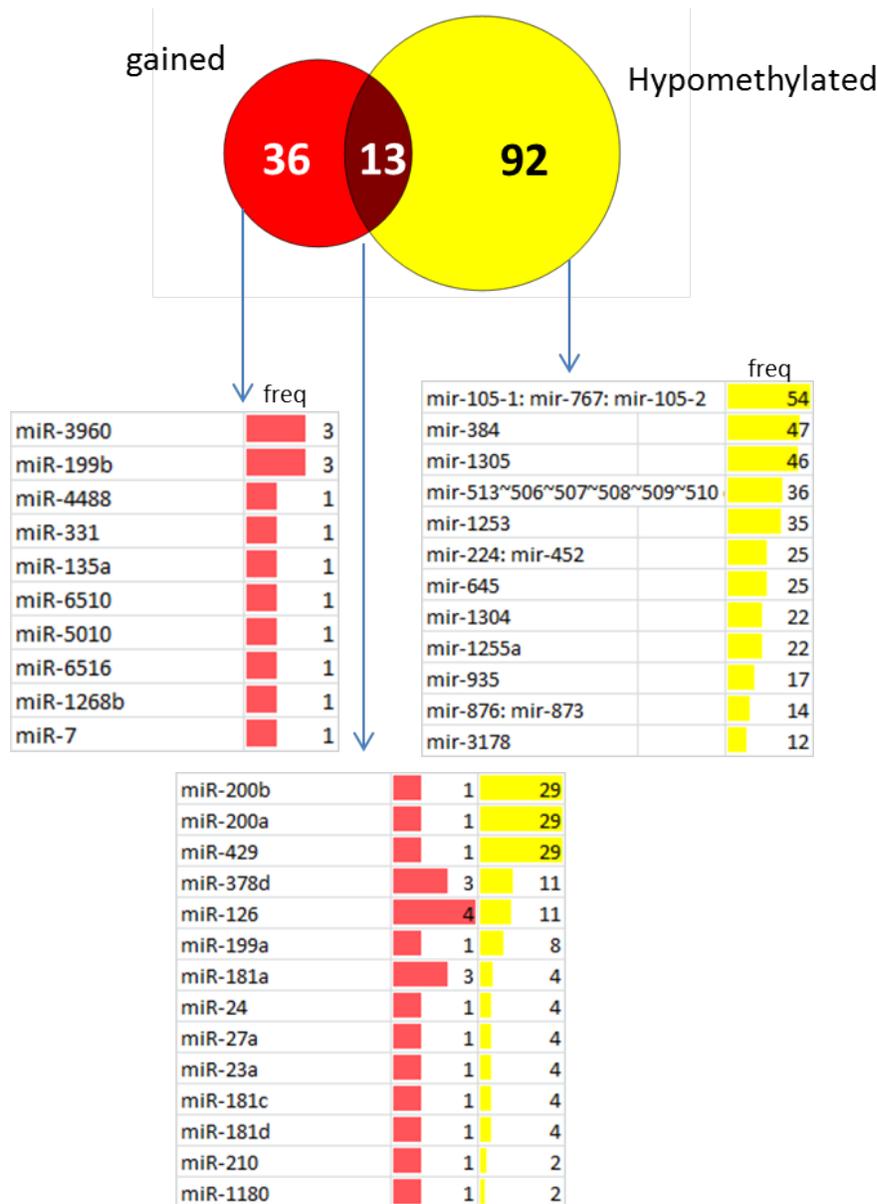
The simplified definitions for two groups used for our analysis, where:

**Oncogenic miRNA** is a miRNA that is frequently altered in EO-PCA either by promoter hypomethylation and/or by allelic gains;

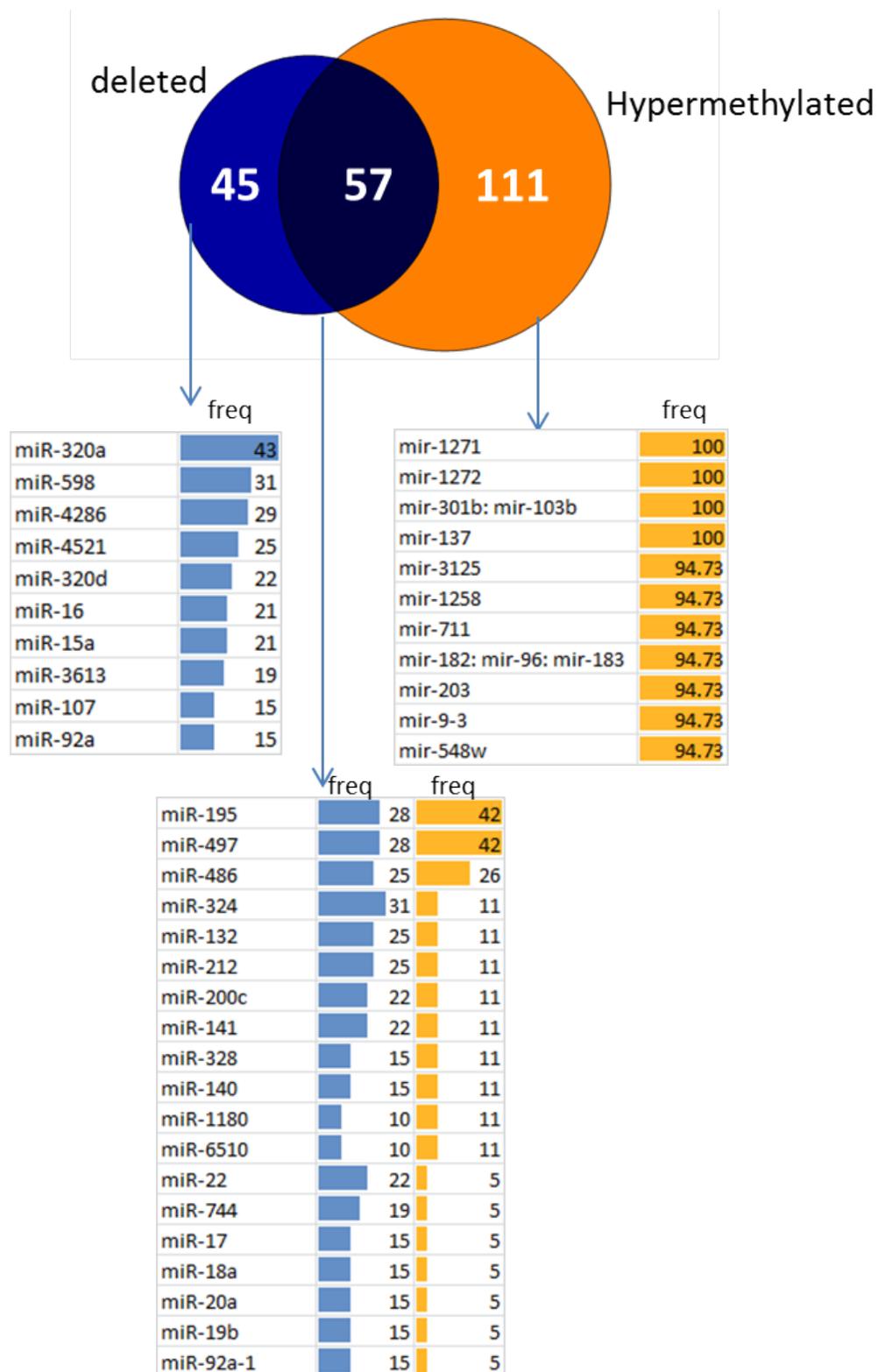
**Tumor-suppressive miRNA** is a miRNA frequently altered in EO-PCA either by promoter hypermethylation and/or by allelic loss.

In total, a list of 213 miRNAs were assigned to be tumor-suppressive (**Figure 7-8**) and 141 miRNAs were assigned to be oncogenic (**Figure 7-7**) for EO-PCAs. 57 (27%) of tumor-suppressive and 13 (9%) of oncogenic miRNAs were altered by two mechanisms at the same time. In order to examine, the role of alteration on expression of miRNAs, we compared the expression level of miRNAs in samples with no alterations to samples with one or two

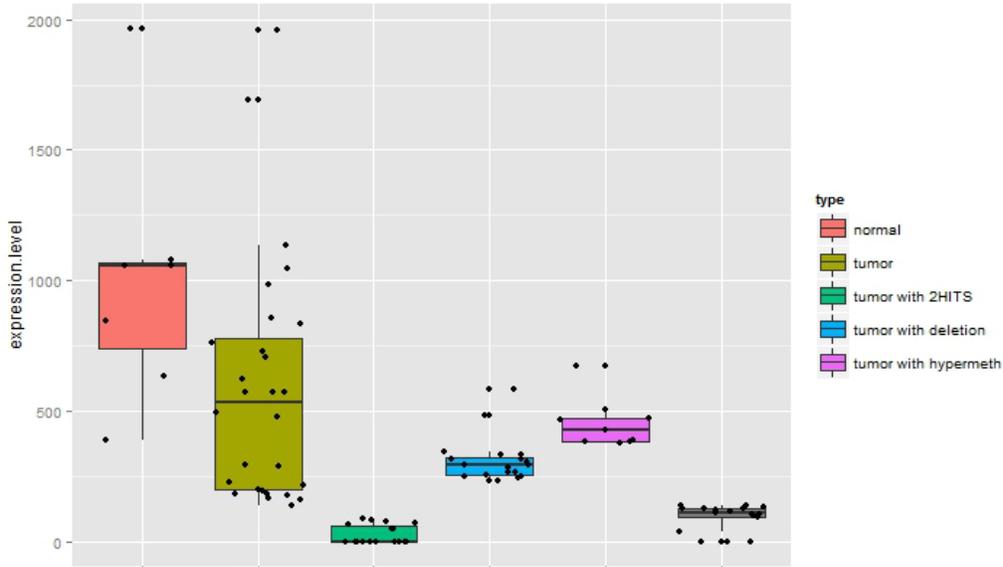
alterations. For tumor-suppressive/oncogenic miRNAs the decrease/increase of expression of miRNAs with one or two alterations suggested the functional impact of alterations on miRNAs transcription. The highest difference of expression was observed in samples with two alterations (miRNA-195 and 200b were used like examples in **Figure 7-9,10**).



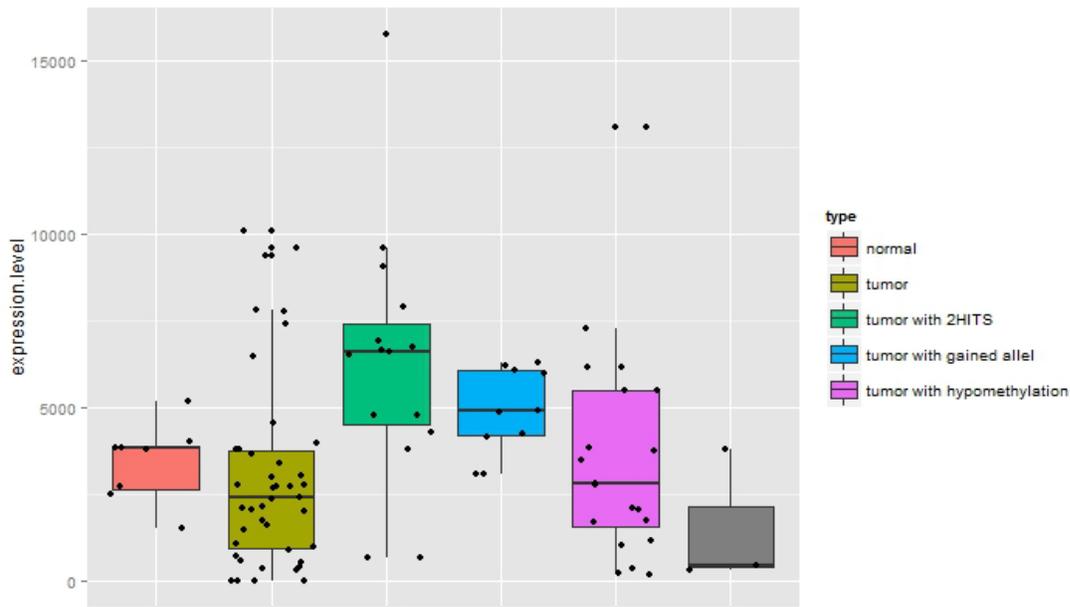
**Figure 7-7: Oncogenic miRNAs in EO-PCA cohort.** Venn diagram represents number of miRNAs, deregulated by genetic gains or hypomethylation of promoter. Top frequently altered miRNA per each of the group presented below diagram in table manner. Such tables describe name of miRNA and frequency of alterations in ICGC EO-PCA sample cohort. Red color denotes genetically gained miRNAs. Yellow color denotes hypomethylated miRNAs.



**Figure 7-8: Tumor-suppressive miRNAs in EO-PCA cohort.** Venn diagram represents number of miRNAs, deregulated by deletions or hypermethylated of promoter. Top frequently altered miRNA per each of the group presented below diagram in table manner. Such tables describe name of miRNA and frequency of alterations in ICGC EO-PCA sample cohort. Blue color denotes deleted miRNAs. Orange color denotes hypermethylated miRNAs.



**Figure 7-8: Comparison of expression level of tumor-suppressive miRNA-195 in EO-PCA cohort.** Expression level of genes, using normalized read counts is visualized by box plots. Comparison between expression level of miRNA-195 in normals, tumors with no alteration, tumors with two HITs (alterations), tumors with one alteration (deletion/hypermethylation of promoter) in miRNA-195 is plotted.

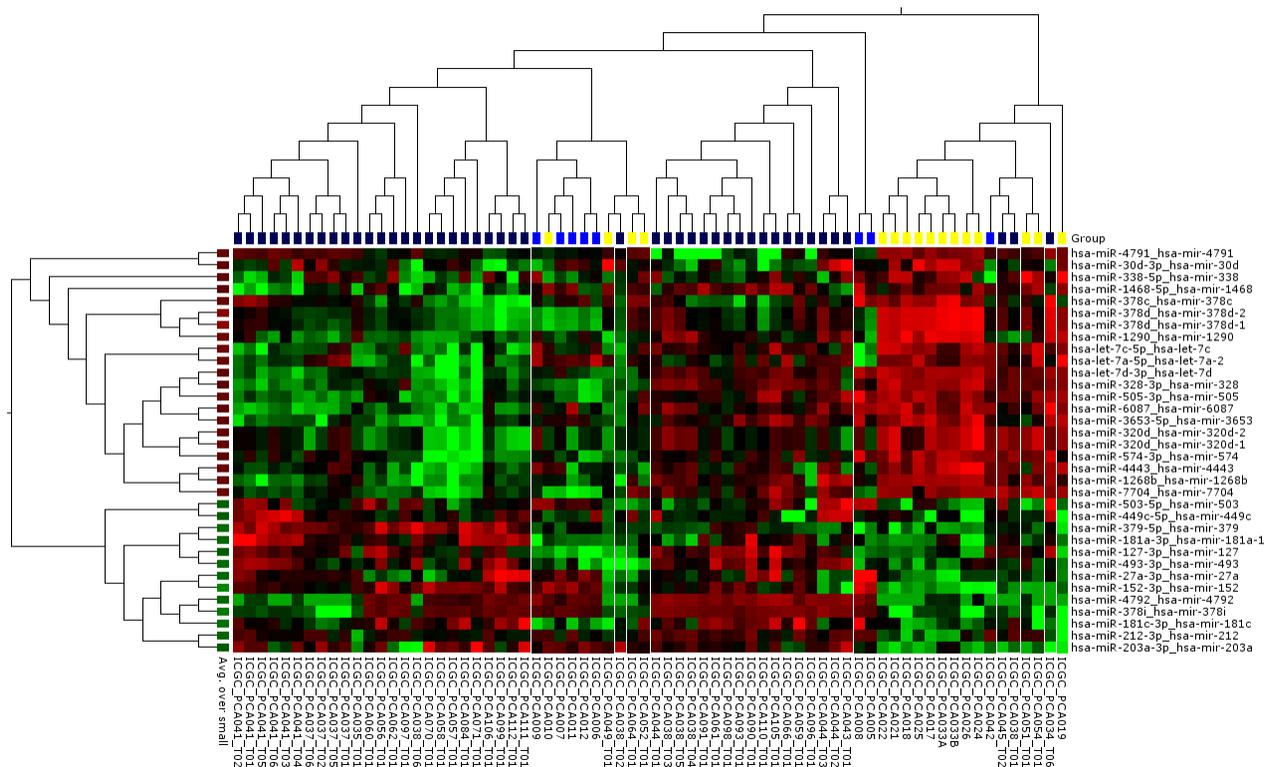


**Figure 7-10: Comparison of expression level of oncogenic miRNA-200b in EO-PCA cohort.** Expression level of genes, using normalized read counts is visualized by box plots. Comparison between expression level of miRNA-200b in normals, tumors with no alteration, tumors with two HITs (alterations), tumors with one alteration (deletion/hypomethylation of promoter) in miRNA-200b is plotted.

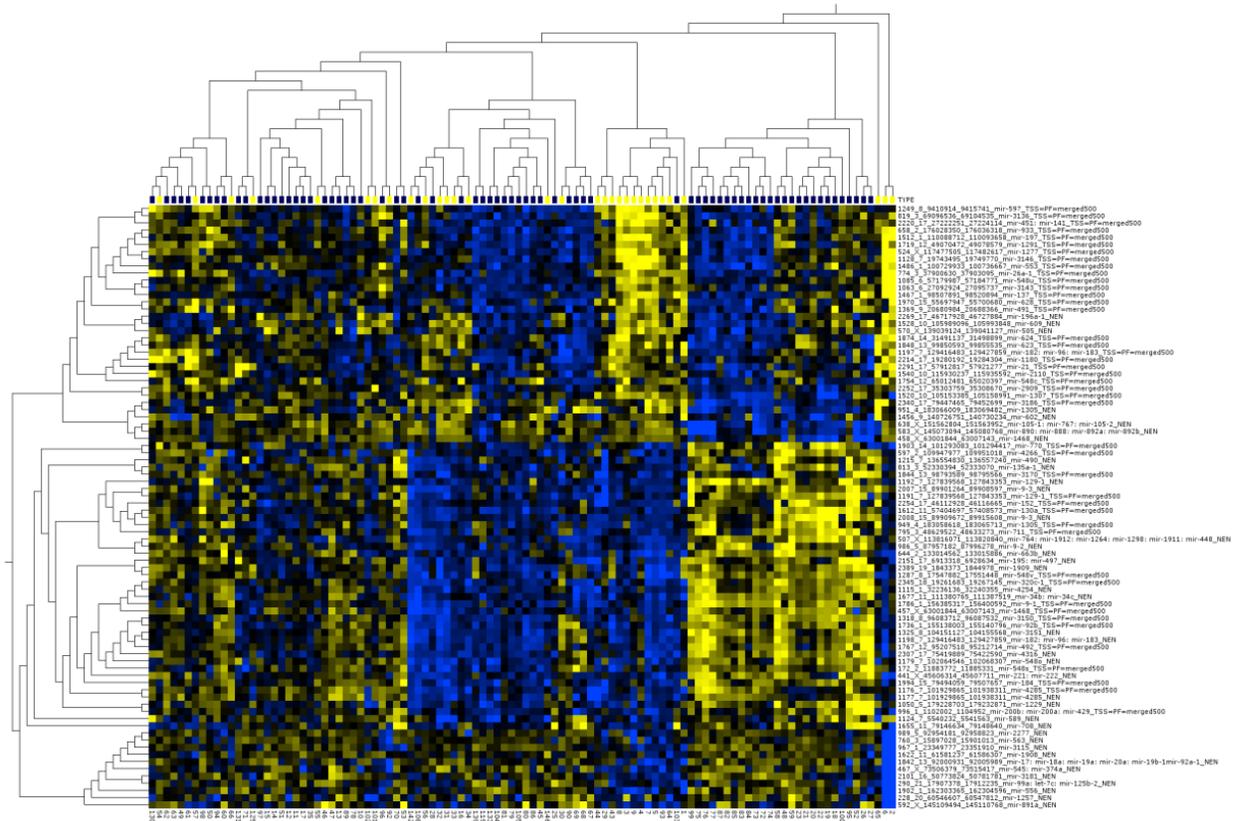
## 7.6. Expression profile of miRNAs in different groups of EO-PCA

Our sample cohort has two major groups of samples: 23 small tumors (diameter < 3 sm) and 35 large tumors (diameter  $\geq$  3sm). We asked a question, if the expression profile of miRNAs in EO-PCA between analyzed groups is different. Using DeSeq R package we performed large vs. small group comparison of miRNA expression profile. 32 miRNAs were significantly differently expressed between selected groups (**Figure 7-11**).

Next we address the question, whether promoter methylation is a major mechanism of deregulation of differently expressed miRNAs between two groups. Using RnBeads R package, we compared methylation degree of all annotated miRNA promoters, covered by 450K arrays. In total 84 differently methylated miRNAs between large vs. small tumors in EO-PCA with at least 1 DMR in promoter/regulatory region were identified (**Figure 7-12**).



**Figure 7-11: Comparison of expression level of miRNAs in large vs. small tumors in EO-PCA data set.** Heatmap represents unsupervised clustering of sample of identified 31 miRNAs. Columns represent sample names, rows – miRNA names.

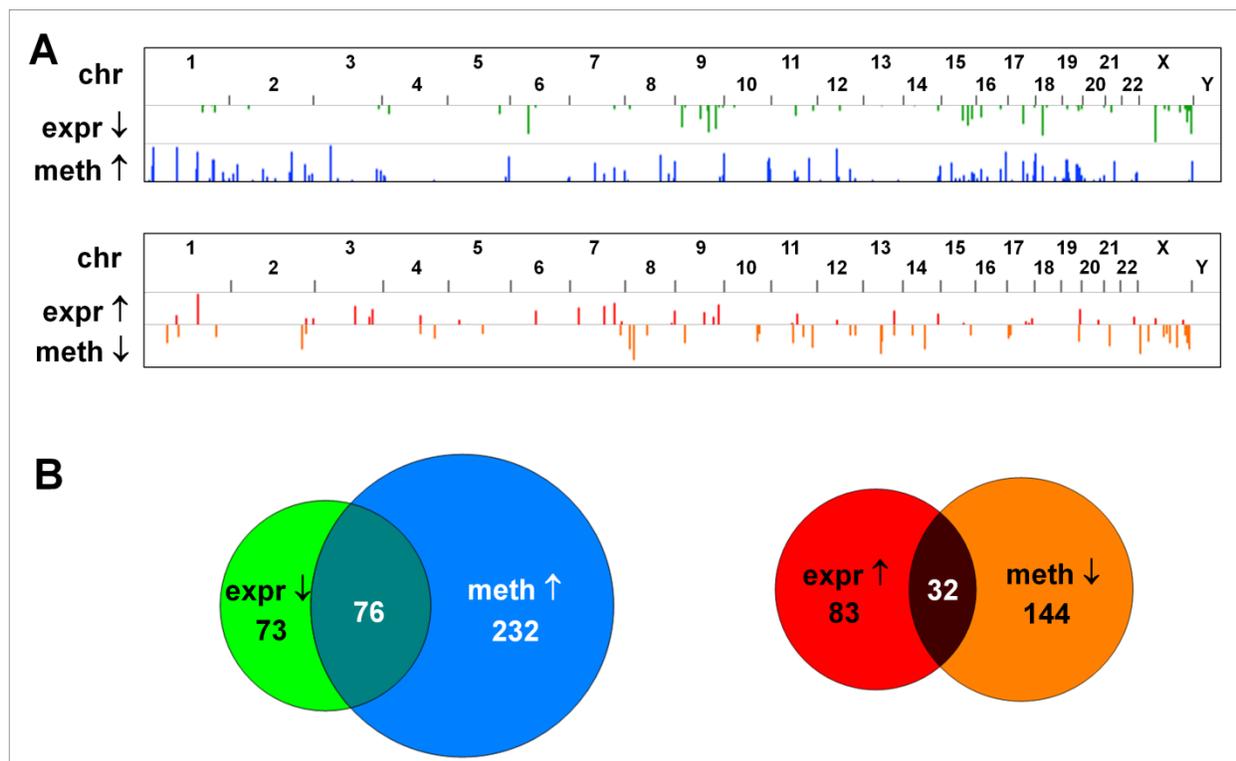


**Figure 7-12: Comparison of promoter methylation of miRNAs in large vs. small tumors in EO-PCA data set.** Heatmap represents unsupervised clustering of sample of identified 84 miRNAs. Columns represent sample names, rows – miRNA names.

### 7.7. Promoter DNA-methylation associated with miRNA deregulation in LO-PCA

We used an independent validation dataset of 51 PCA samples with a classical age distribution and 48 normal prostate tissues to validate our findings on miRNA deregulation and its frequent association to aberrant promoter methylation in the EO-PCA samples. Genome-wide miRNA expression data were generated by quantitative PCR (qPCR) arrays by Ruprecht Kuner from Prof.Dr.Holger Sültmann group (<http://www.dkfz.de/en/krebsgenomforschung/>). In addition, miRNA deep sequencing was performed by them on a sample subset in order to extend the panel of miRNAs. In total 311 miRNAs (142 up- and 169 downregulated) were differentially expressed. To identify differentially methylated regions (DMRs) in miRNA promoters we used the recently published genome-wide methylation profiles of all 51 PCA and 48 normal samples [145]. Significant deregulation of miRNA expression accompanied by differential promoter methylation throughout the genome was observed in concordant to previously observed results (see **chapter 7.1**) (**Figure 7-6 A**). 76/149 (51%) significantly downregulated miRNAs with annotated regulatory regions revealed promoter hypermethylation (**Figure 7-13 B**). Promoter

hypomethylation was observed for 32/115 (28%) upregulated miRNAs with annotated regulatory regions (Figure 7-13 B).



**Figure 7-13: Genome-wide miRNA deregulation and aberrant promoter methylation in a PCA validation dataset. (A) Upper panel,** genome-wide view on significantly downregulated miRNAs (green bars) and significantly hypermethylated regulatory regions (blue bars). **Lower panel,** significantly upregulated miRNAs (red bars) and significantly hypomethylated promoter regions (orange bars). Bar height indicates statistical significance ( $-\log_{10}p$ -value) **(B)** Overlap between significantly deregulated miRNAs and differentially methylated promoter regions. Only miRNAs with annotated regulatory regions were considered.

Hypomethylation of promoter regions of downregulated miRNAs and hypermethylation of upregulated miRNAs were also observed in investigated data cohort. This points into the lack of knowledge of annotation of miRNA promoter regions. Such identified regions could be potentially enhancers or insulators and further functional investigation is required.

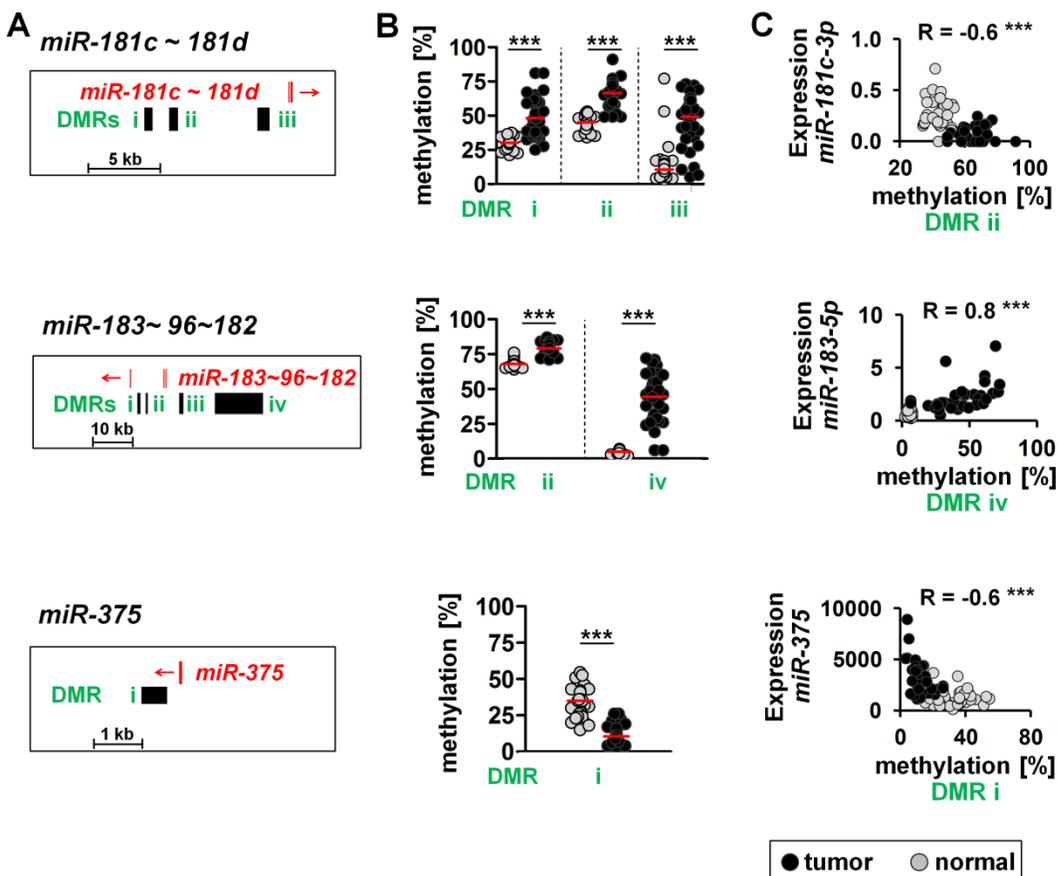
In summary, genome-wide analysis of an independent validation dataset of classical PCA cases supported that hyper- and hypomethylation of regulatory regions frequently contribute to miRNA deregulation in PCA as well as EO-PCA.

### 7.7.1. Validation of promoter DNA-methylation associated with miRNAs deregulation in PCA

Methylation levels of 15 selected miRNAs/miRNA clusters were used for technical validation by MassARRAY. Four of them were commonly upregulated and eleven downregulated in several

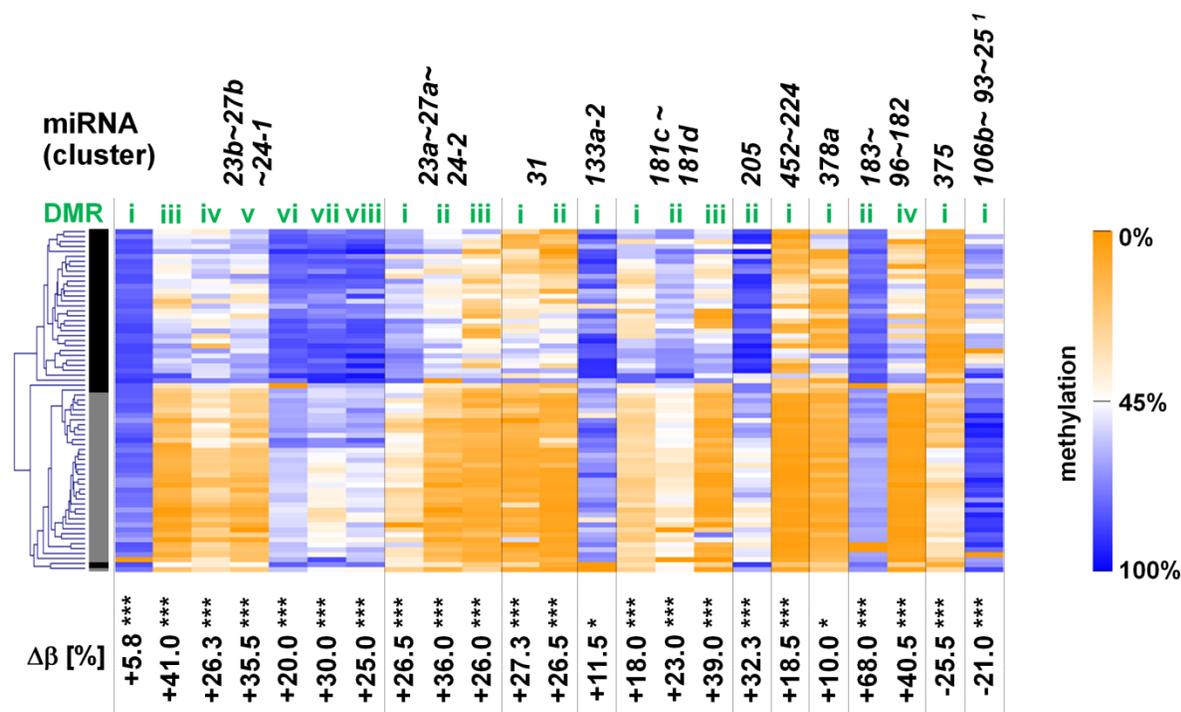
independent studies on PCA and hence, may act as key regulators in PCA. Eleven miRNA clusters became upregulated upon DNA demethylation by 5-Aza-2'-deoxycytidine treatment of PCA cells and the *miR-183~96~182* cluster, which is upregulated and hypermethylated in PCA, was downregulated by the treatment (**Supplemental Figure S2**). Thus, these 12 miRNA clusters are indeed epigenetically regulated. Their regulatory sequences carried 33 DMRs, 32 hyper- and one hypomethylated, in both the ICGC EO-PCA and the validation dataset.

MassARRAY analysis of 34 tumors and 35 normal samples of the validation dataset confirmed 27/32 (84%) hypermethylated and 1/1 (100%) hypomethylated DMRs. Methylation levels of 22 of these 28 validated DMRs (79%), corresponding to ten miRNA clusters, were correlated with the expression of the corresponding miRNAs including *miR-181c~181d* (**Figure 7-14**).



**Figure 7-14: Confirmation of differentially methylated regions (DMRs) in PCA tissues and their correlation with expression.** Representative results for the downregulated miRNA cluster *miR-181c~181d* with hypermethylated DMRs, and the upregulated *miR-183~96~182* and *miRNA-375* with hypermethylated and hypomethylated DMRs, respectively. **(A)** Schematic representation of the chromosomal location of miRNAs and DMRs analyzed by MassARRAY. Red arrows indicate direction of transcription. **(B)** Methylation levels of DMRs quantified by MassARRAY in normal prostate tissues (n=35) and PCA (n=34). Red lines indicate medians. (\* p < 0.05, \*\*\* p < 0.001).

These 22 DMRs and the hypomethylated *miR-106b-93-25* promoter described recently [145] separated tumor from normal samples in a hierarchical cluster analysis (**Figure 7-15**). Only three tumor samples were assigned incorrectly, a fact that might be explained by their low *GSTP1* methylation levels indicating a high content of stromal cells.



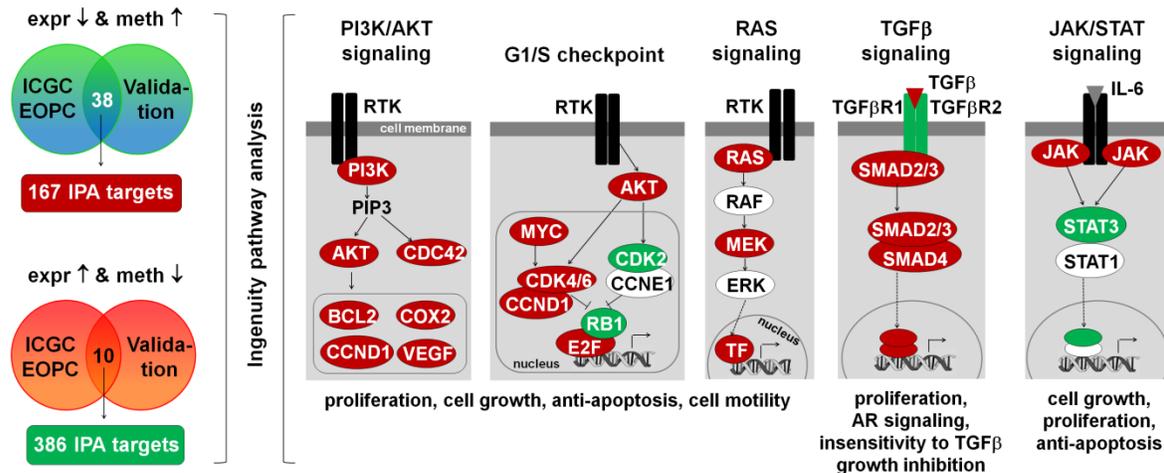
**Figure 7-15: Heatmap of differentially methylated regions (DMRs) in PCA validation cohort.** Methylation levels of all validated DMRs that are inversely (*miR-183-96-182*: directly) correlated to miRNA expression, including one hypomethylated DMR regulating *miR-106b-93-25* and recently described as in [118]. Color denotes methylation degree (blue denotes 100% methylation and yellow – no methylation of investigated regions) Below the heatmap, the methylation differences between PCA and normal prostate tissues ( $\Delta\beta$  [%]) and their statistical significances are shown (\*  $p < 0.05$ , \*\*\*  $p < 0.001$ ).

Methylation levels of validated DMRs were able to separate PCA from normal prostate tissues in hierarchical clustering. Taken together, 10 of 15 selected miRNA clusters (67%) commonly deregulated in PCA were validated by independent method and sample cohort. This finding supports global observation that aberrant promoter DNA-methylation of miRNAs represents an important mechanism of deregulation of miRNA expression in PCA.

### 7.7.2. Epigenetically regulated miRNAs activate key oncogenic pathways in PCA

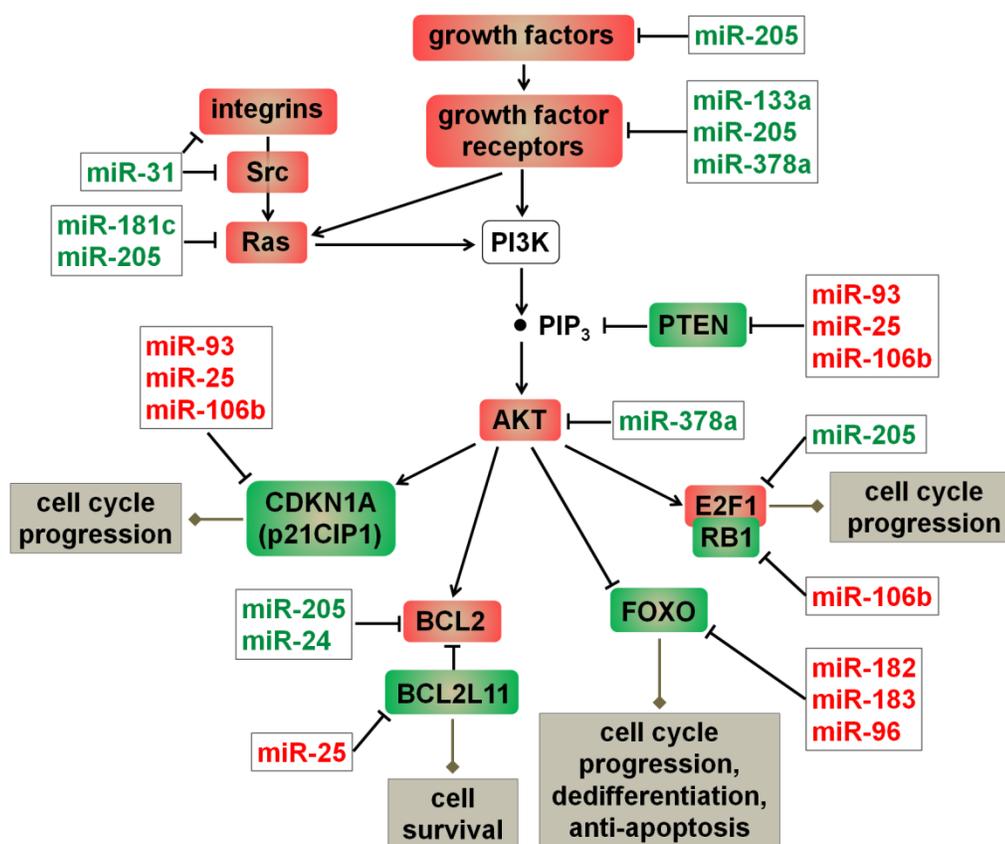
Screening both the ICGC EO-PCA and the validation dataset, we identified 37 miRNAs consistently downregulated and ten miRNAs consistently upregulated by promoter hyper- or hypomethylation, respectively. To identify gene regulatory networks possibly impaired in PCA by

these epigenetically deregulated miRNAs on a global level, all experimentally observed target genes of these 48 miRNAs were assigned using Ingenuity Pathway Analysis (IPA) assuming gene upregulation upon miRNA downregulation and *vice versa*. Mapping these target genes to molecular pathways by IPA the top five oncogenic pathways were: PI3K/AKT signaling, G1/S checkpoint, RAS, TGF $\beta$  and JAK/STAT signaling (**Figure 7-16**).



**Figure 7-16: Epigenetically regulated miRNAs activate key oncogenic pathways.** Left panel, 37 and 10 miRNAs displayed an inverse association between expression and promoter methylation in both the ICGC EO-PCA and the validation dataset (green/blue = down regulation + hypermethylation; red/orange = upregulation + hypomethylation). Their experimentally observed target genes were extracted from Ingenuity Pathway Analysis (IPA) and assumed to be upregulated (denotes by red) by downregulated miRNAs and downregulated (denotes by green) by upregulated miRNAs. Right panel, the top five pathways to which these target genes were assigned to by IPA are tumor-promoting. Four of these five pathways are clearly activated by the observed miRNA deregulation (TF: transcription factors).

Based on the miRNA deregulation in our datasets, the majority of the genes activating these pathways were upregulated, thereby leading to the activation of four pathways. Focusing on the experimentally verified target genes extracted from the literature of eleven miRNA clusters: the ten being validated to be epigenetically deregulated as described in the previous chapter and the *miR-106b~93~25* cluster which epigenetic upregulation was recently reported in PCA [145]. IPA revealed that 9/11 (82%) miRNA clusters regulate genes of PI3K/AKT/PEN signaling pathway (**Figure 7-17**), eight of which could lead to activation of this pathway. As examples, the pathway-activation of *AKT1* gene (*V-Akt Murine Thymoma Viral Oncogene Homolog 1*) is activated by downregulated miRNAs, whereas the pathway-inhibiting *PTEN* gene (*Phosphatase and tensin homolog*) is downregulated by upregulated miRNAs.

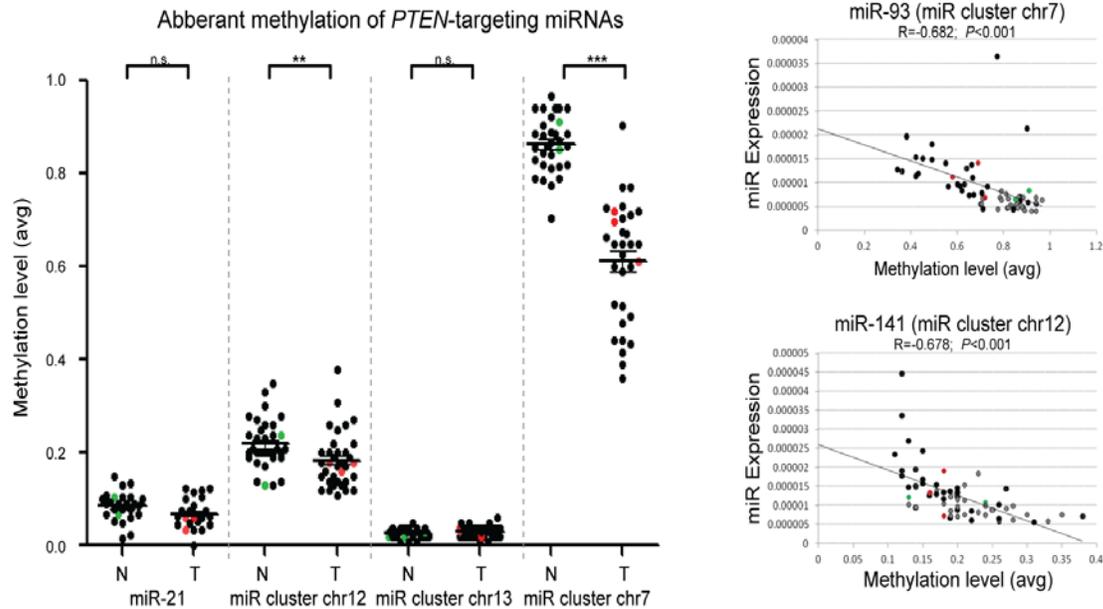


**Figure 7-17: Activation of PI3K/AKT/PTEN signaling by eleven confirmed epigenetically regulated miRNA clusters (see text for description).** Experimentally verified target genes were extracted from literature. miRNAs printed in green are consistently downregulated in PCA resulting in upregulation of their target genes (red). miRNAs printed in red are consistently upregulated in PCA resulting in downregulation of their target genes (green).

### 7.8. *PTEN* inactivation by miRNAs in PCA

Famous tumor-suppressor gene *PTEN* has different mechanisms of inactivation in prostate cancer such as complex mutational processes including deletions, mutations, translocations or the activation of *ETS* (E26 transformation-specific) transcription factors through structural rearrangements [145]. But regulation of *PTEN* expression by miRNAs remains unclear in prostate cancer. In order to understand alternative mechanisms of *PTEN* deregulation, 16 functional proved *PTEN*-targeting miRNA profiling was examined. Multiple upregulating of *PTEN*-targeting miRNAs indicated the alternative mechanism of deregulation of *PTEN* in PCA. 2 of investigated clusters of miRNAs had hypomethylated promoter regions correlating with miRNA expression (**Figure 7-18, Supplemental Figure S3**). *PTEN*-targeting miR-106b-5p that displays oncogenic activity in a PCA model [145] was upregulated 2.4x-4.6x in all seven EO-PCA samples. Obtained data indicates miRNAs contribution to *PTEN* inactivation in prostate

cancer. Due to the lack of a sample cohort with *PTEN* genetic data, mutually exclusivity of each inactivation mechanisms was not analyzed.



**Figure 7-18: Hypomethylation and upregulation of two *PTEN*-targeting miRNA cluster.** MassARRAY results represented by dot-plots (left panel) revealed promoter hypomethylation of 2 *PTEN*-targeting miRNA clusters (cluster on chr7 and chr12). Right panel displays correlations between miRNA expression and average methylation levels across the promoter regions ( $R$  corresponds to Spearman rank correlation values). Horizontal lines depict median values. Black dots denotes elderly-onset prostate cancer, grey and green dots denotes for normal prostate epithelium (>50years old and <=50 years), red dots denotes for EO-PCAs.

## 7.9. miRNAs promoter DNA-methylation in cancer

In order to further investigate the role of promoter methylation of miRNAs in cancer in a systematic manner, a PANCAN approach is required. PANCAN is an approach analyzing differences between cancer types that allows identifying cancer- as well as tissue- specific features. Such approach allows unraveling whether observed promoter methylation of miRNAs in PCA is also present in other cancer subtypes. In order to perform comprehensive screen of DMRs in promoters of miRNAs we used TCGA available resources. Using [TCGA PANCAN data set](#) (described in [chapter 5.4](#)), DMR search in promoter/regulatory regions of miRNAs was performed. In total 127 DMRs (~15%) in promoter/regulatory regions of miRNAs were present in >50% of investigated studies and assigned as PANCAN DMRs. Where 875 DMRs (~85%) DMRs were unique per study and were assigned as tissue-specific DMRs. Obtained results supported the hypothesis that DNA-methylation of promoter regions is one of major mechanisms of deregulation of miRNAs in cancer.

## 7.10. Discussion of results

Next generation profiling data for cancer genomes allows an integrated analysis of genetic and epigenetic alterations. In the majority of the studies the power of an integrative analysis is not used to its full potential and focus is either on genetic or epigenetic alterations. In this study we investigated the potential of integrative genome analysis on the regulation of miRNAs in EO-PCA. We utilized a comprehensive dataset of 66 EO-PCA and demonstrated on the genome-wide level that - besides miRNA deletion - epigenetic alterations, namely differential methylation of regulatory regions, represent a major mechanism for miRNA deregulation in PCA.

Strikingly, we did not only uncover promoter hypermethylation for up to half (33% and 51%) of all downregulated miRNAs in the 66 EO-PCA and an independent validation dataset of 51 PCA, but as a novel mechanism also promoter hypomethylation was detected for one fourth (23% and 28%) of the upregulated miRNAs. The former included known epigenetically regulated miRNAs like *miR-205* [153, 154], *miR-34c* [155] and *miR-193b* [156]. The frequent co-occurrence of hypomethylation and miRNA activation detected in our genome-wide analyses is in agreement with recent data in chronic lymphocytic leukemia [16]. The lack of genetic aberrations that could cause miRNA upregulation in the 66 EO-PCA further indicates that epigenetic mechanisms like DNA hypomethylation are essential for miRNA activation in (early-onset) PCA. We confirmed the epigenetic regulation for ten of 15 selected miRNA clusters that are commonly deregulated in PCA using the MassARRAY technology. Furthermore, we demonstrated that epigenetically deregulated miRNAs contribute to the activation of key oncogenic signaling pathways in PCA. A large fraction of the validated epigenetically regulated miRNAs activates AKT/PTEN signaling by targeting diverse genes within this pathway. Thus, providing a novel mechanism of deregulation besides known genetic alterations of *PTEN* [118, 119, 122, 151]. This also supports the view that individual tumors, despite bearing different miRNA expression patterns, are able to activate the same pathway by targeting different genes.

In several cancer types, including colorectal [157] and breast cancer [155, 158], melanoma [159] and leukemia [16] aberrant promoter methylation has been extensively described to be an important source for deregulation of miRNA expression. In PCA, large-scale studies on epigenetically regulated miRNAs were mainly performed by 5-aza-2'-deoxycytidine-mediated inhibition of DNA methyltransferases in PCA cells leading to global DNA demethylation and reactivated expression of epigenetically regulated miRNAs [153, 156]. In PCA tissues, the global picture was lacking so far, and aberrant methylation of miRNA promoters associated with deregulated expression was demonstrated only for single miRNA candidates including the

tumor-suppressive *miR-205* [154] and *miR-31* [160]. The hypermethylated DMRs for both miRNAs identified in our study overlap to the promoter regions described in other studies on PCA [154, 160], providing independent support for our results. Several other epigenetically regulated miRNAs validated in this study have also been described to be epigenetically controlled in PCA or other human tumor cells: these include *miR-133a* [161], *miR-181c* [162], *miR-23b* [100], *miR-27a* [156], *miR-27b* [163], *miR-375* [156] and *miR-378a* [163, 164]. We add here the exact localizations of the differentially methylated regulatory regions for these miRNAs in PCA tissues. The activation of *miR-375* by hypomethylation has previously been described in breast cancer for a DMR upstream [165] of the DMR identified by us, which is almost unmethylated in PCA. This cell type-specific shift of regulatory regions has previously been described for *miR-21* [166].

In addition to the classical view of an inverse correlation between aberrant promoter methylation and differential expression, we observed a number of downregulated miRNAs with hypomethylated regulatory regions and upregulated miRNAs with hypermethylated regions. This could be explained by the fact that the investigated regulatory regions carrying the chromatin mark histone 3 lysine 4 trimethylation (H3K4me3) [16] does not necessarily represent promoter regions but could also function as enhancers [167]. Furthermore, other suppressive (H3K9me3, H3K27me3) or activating chromatin marks (H3K36me3, H3K27 acetylation) may overrule the regulatory effect of promoter methylation in certain genomic contexts [124]. For example, the regulatory region of the oncogenic *miR-183-96~182* cluster [168, 169], which is commonly upregulated in PCA, was significantly hypermethylated. This regulatory region might function as a silencer that becomes inhibited by hypermethylation. Interestingly, the region is a target of polycomb-repressive complexes based on the ENCODE chromatin data and thus, might be hypermethylated by the polycomb repressor complex 2 (PRC2) member EZH2, whose expression is highly correlated to the methylation level of this regulatory region. Therefore, our data indicates that the prevailing hypothesis of “promoter hypermethylation causes inactivation” is not sufficient to explain all alterations of miRNA expression in PCA tissues.

The mechanisms leading to epigenetic reprogramming in miRNA promoter sequences are unknown. The currently uncovered mutation spectrum in PCA [119, 170] might shed some light on this question. In these data sets we found several genes involved in epigenetic pathways including DNA methylation, histone modifications and chromatin remodeling (e.g. *CHD1*, *HDAC9* and *MLL2*) [119, 170]. It is intriguing to speculate that these defects lead to global alterations of

epigenetic patterns including the regulatory regions of miRNAs. Alternative mechanism may include environmental or tumor micro-environmental stimuli that influence epigenetic patterns.

In conclusion, our study shows that aberrant DNA hyper- and hypomethylation of miRNA regulatory regions is a frequent event and represents a major process directed towards transcriptional deregulation of miRNAs in PCA, thereby effectively activating key oncogenic pathways. However, a large fraction of deregulated miRNAs could not be explained by genetic alterations or aberrant DNA methylation indicating that additional mechanisms including other epigenetic mechanisms like histone modifications, aberrant transcription factor activity or posttranscriptional regulation are involved in miRNA deregulation.

In the chapter 7-8 of the thesis, we focused on the identification major mechanisms of deregulation of non-coding RNAs (miRNAs and lncRNAs correspondently), by using independent comprehensive data cohort of early-onset prostate cancer. We observed 35 translocations that could potentially lead to deregulation of miRNAs, but due to a small sample size, we were not able to estimate the functional impact of them. Testing bigger sample cohorts as well as verifying identified novel rearrangements in another cancer types would provide novel mechanism of deregulation of ncRNAs.

In this thesis we followed two major approaches: identification differences between tumor and normal or the comparison within tumor samples to identify novel subtypes. Both approaches have benefits but also some limitations. The tumor vs. normal approach was used to identify major differences in tumor cells compare to normal controls. The biggest advantage is to be able to statistically distinguish major differences. However, one of the limitations of this approach is the assumption that all tumors have the same cell composition and tumor content. This is not problematic for the genetic datasets, however due to tissue-specific DNA methylation and expression patterns this represents a major problem since the cell or cell-type of tumor origin is often unknown. There are different methods described in the field already to evaluate tumor purity, based on genetic events (ABSOLUTE) or expression profiles (ESTIMATE) or methylation events (*GSTP1* for prostate cancer). All methods were showing different results and can be implemented only to specific data level. But universal tool for simple reliable estimation tumor purity is still missing.

We also used to approach to compare tumors only. This approach operates under the assumption that all tumors have similar tumor content and were raised from the same cell of origin cell. There is, however, a lack of knowledge in the field for proper normal control or for methods for the identification from already existing ICGC and TCGA data cell of origin of tumor. Identification of the cellular origins of cancer is also crucial in enhancing our understanding of the mechanisms regulating the different steps of tumor initiation and progression.

## CHAPTER 8.

---

### MECHANISMS LEADING TO DEREGLATION OF LNCRNA IN PROSTATE CANCER

---

Long non-coding RNAs (lncRNAs) represent a class of molecules that exert regulatory functions to their target genes by recruitment of chromatin modifying protein complexes [171]. Deregulation of lncRNAs results in altered expression of target genes. An additional level of epigenetic deregulation in cancer genomes comes from deregulation of lncRNAs, which are reported to act as regulators of tumorigenic pathways [172]. Deregulated lncRNAs multiply their influence on thousands of target genes, many of them with cancer relevant functions as for example *HOTAIR*, *XIST* etc. [173]. Recent studies provided lists of cancer-specific and tumor type-specific deregulated lncRNAs. Although the contribution of deregulated lncRNA expression to cancer pathogenesis is evident, it is still not clear which mechanisms lead to aberrant expression of lncRNAs.

For a long time, the systematic genome-wide evaluation of mechanisms leading to lncRNA deregulation was not performed due to the lack of data and comprehensive promoter annotation for lncRNAs. This has recently changed due to the utilization of RNA-seq data for cancer patients by efforts of big cancer genome consortia, better annotation of lncRNA genes in entire genome [16] and large scale profiling activities in the ENCODE projects for annotation of regulatory regions [174]. Large cancer genome characterization consortia (e.g. TCGA and ICGC) have released comprehensive datasets on genetic, transcriptomic and epigenetic levels that will now allow an integrative analysis to identify the mechanism of lncRNA deregulation. In PCA datasets are available that foster a comparison of early-onset (**EO-PCA data set**) with late onset prostate cancer (ICGC, TCGA and others).

In this chapter we work on the hypothesis that in addition to genetic events, epigenetic alterations lead to deregulation of lncRNA transcription. In order to unravel the molecular alterations leading to lncRNA deregulation in PCA we performed the first genome-wide integrative analysis of expression, DNA methylation and genetic alterations of lncRNAs of 89 EO-PCAs specimens generated within the German ICGC project (ICGC EO-PCA dataset) [118] [Feuerbach et al., manuscript in preparation]. The workflow of the integrative analysis consists of six steps of data integration starting with the analysis of expression profiles of lncRNAs in EO-PCA, followed by the analysis of copy number alterations, integration with DNA methylation profiles and subsequent addition of non-coding mutations (**Figure 8-1**).

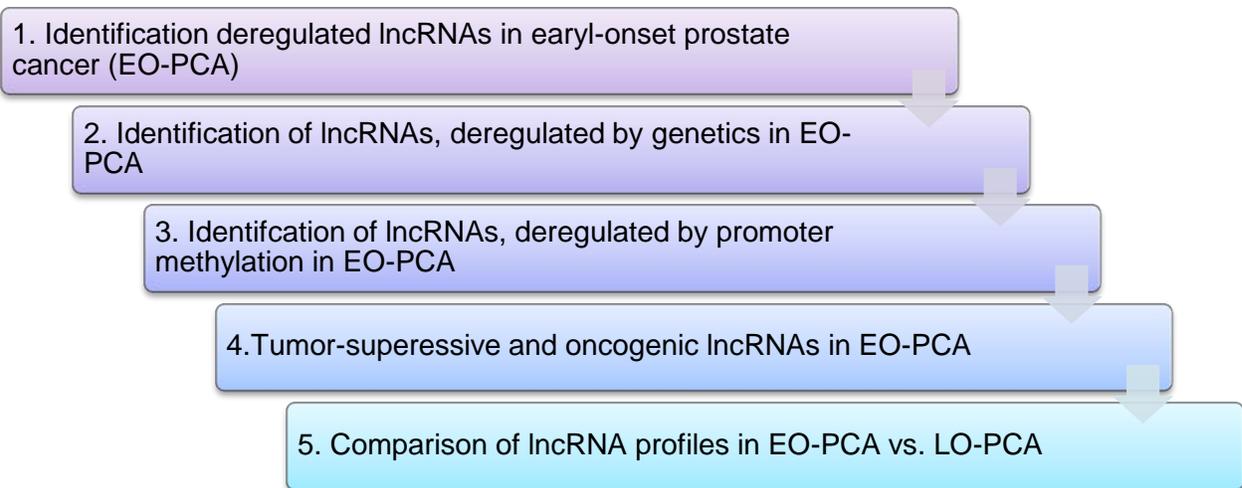


Figure 8-1: Schematic workflow for the integrative analysis of mechanisms leading to lncRNAs deregulation in EO-PCA.

**Step 1:** Identify differently expressed lncRNAs in EO-PCA using RNA-seq data of 78 EO-PCAs samples and 11 normals as controls

**Step 2:** Identify commonly genetically altered lncRNAs in EO-PCAs using copy number alteration (CNA) profiles of 66 EO-PCAs samples

**Step 3:** Identify lncRNAs with promoter/regulatory regions DNA methylation using MCip-seq data from 33 EO-PCA and 1 normal and extend analysis using Illumina 450K arrays from 155 EO-PCAs and 44 normals.

**Step 4:** Identify lncRNAs, which are deregulated in EO-PCAs by two hits: promoter methylation and CNA. Based on data assign miRNAs to tumor-suppressive and oncogenic lncRNAs in EO-PCAs.

**Step 5:** Compare lncRNAs profile of EO-PCA vs. LO-PCA, identify major mechanism of deregulation of lncRNAs in PCA.

### 8.1. Expression profile of lncRNAs in EO-PCA

We started our analysis with the evaluation of differently expressed lncRNAs in **EO-PCA transcriptom data set** (described in details in **chapter 5.4**). Using RNA-seq data of 78 EO-PCAs samples and 11 normals as controls, 12045 lncRNAs were quantified with more than 10 reads in samples cohort. Using DeSeq R package to identify differently expressed miRNAs, pool of 78 tumors and 11 normal controls were used for comparison and in total 76 lncRNAs were significantly downregulated and 57 lncRNAs were upregulated in EO-PCA samples (top deregulated lncRNAs are shown in **Figure 8-2**).

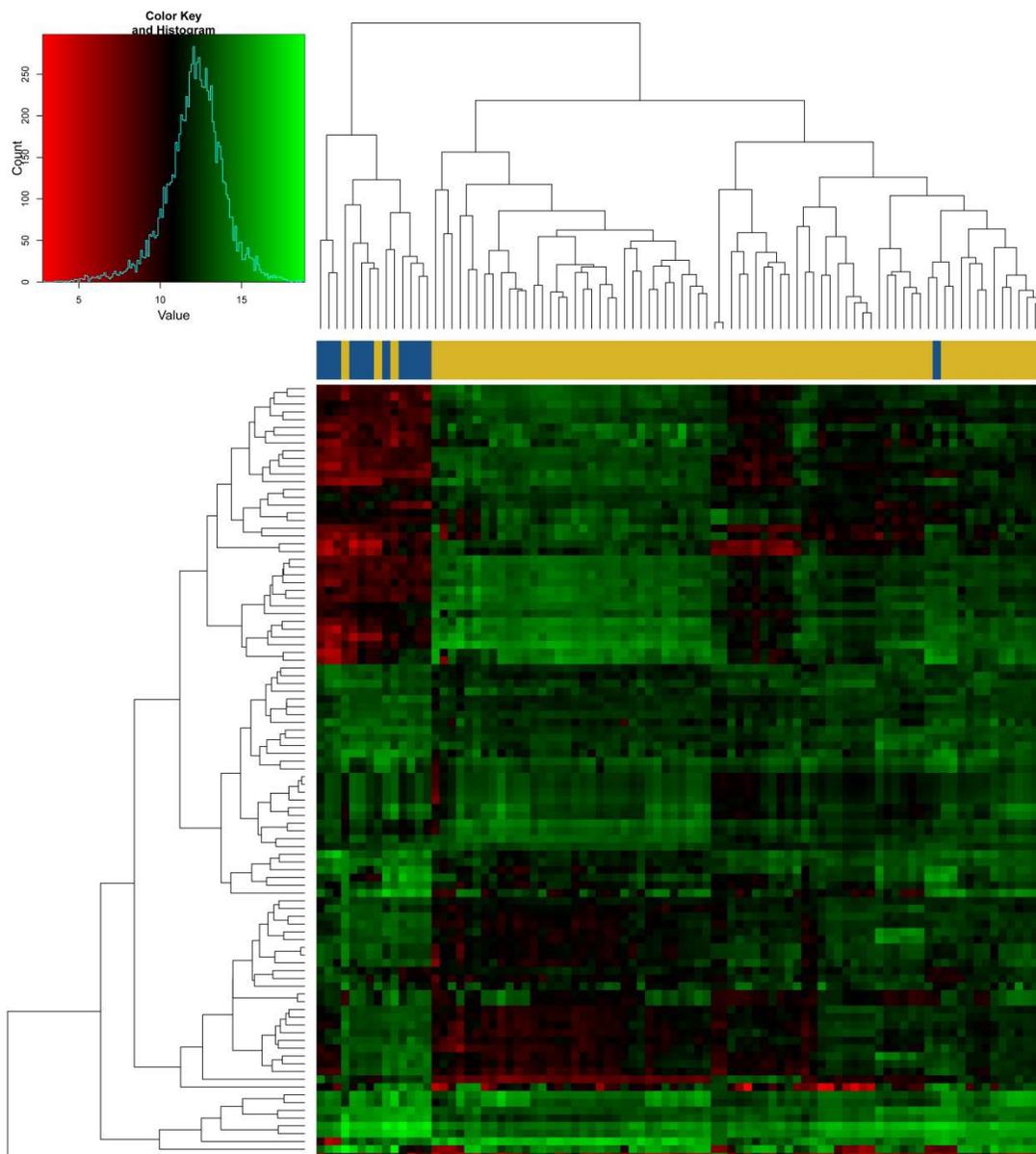
Within the list of deregulated lncRNAs of EO-PCA (see **Table 8-1**) are published lncRNAs *PCAT1* [111], *PCA3* [128] known from previous publications on late onset prostate cancer. However, they were not within the top 10 lncRNAs, suggesting that in EO-PCA additional lncRNAs participates in tumorigenesis. Similar to miRNAs (**chapter 7**) the deregulated lncRNAs were distributed over the entire genome and no apparent clustering at certain chromosomal regions was detectable which would suggest a chromosome *loci* specific mechanism. This indicates that deregulation of lncRNAs underlies individual events for each gene.

**Table 8-1. Top 20 up-regulated lncRNAs in EO-PCA**

lncRNA name	group	log2FC	p-value	adjusted p-value
<i>RP5-1092A3.4</i>	antisense	2.274551	0.001715	0.072588
<i>RP11-457M11.5</i>	lincRNA	2.380836	0.001501	0.067332
<i>RP1-80N2.3</i>	lincRNA	2.610246	0.000626	0.039055
<i>CTD-3060P21.1</i>	antisense	2.613297	0.001153	0.057306
<i>RP11-299G20.2</i>	antisense	2.884136	0.002739	0.097687
<i>RP11-368I7.2</i>	antisense	2.885027	0.000451	0.032183
<i>DLX6-AS1</i>	antisense	3.125954	0.000312	0.02469
<i>CTD-2265M8.2</i>	antisense	3.020961	0.000223	0.019667
<i>AC116614.1</i>	antisense	3.075482	0.001166	0.057548
<i>CTD-3064H18.4</i>	antisense	3.406407	0.000118	0.013134
<i>AP006748.1</i>	lincRNA	4.040963	0.000583	0.03786
<i>AC002511.2</i>	antisense	3.986033	1.33E-05	0.002741
<i>AC073133.2</i>	lincRNA	3.94981	4.70E-05	0.006875
<i>CTC-523E23.5</i>	lincRNA	3.940505	0.000106	0.012331
<i>AC002511.3</i>	lincRNA	3.967112	1.06E-05	0.002416
<b><i>PCAT1</i></b>	<b>lincRNA</b>	<b>4.042015</b>	<b>0.000174</b>	<b>0.016953</b>
<i>CTD-2126E3.3</i>	antisense	4.021008	2.26E-05	0.004057
<i>AC144450.2</i>	lincRNA	5.606544	6.36E-06	0.001692
<i>ERVH48-1</i>	antisense	5.834426	0.000225	0.01971
<b><i>PCA3</i></b>	<b>antisense</b>	<b>6.47371</b>	<b>9.05E-11</b>	<b>3.44E-07</b>
<i>AC144450.1</i>	Antisense	6.534477	3.36E-10	8.52E-07
<i>AC074389.9</i>	lincRNA	6.861704	9.12E-06	0.002174

Unsupervised clustering of the top 100 differently expressed lncRNAs identified 3 tumor subgroups but they did not associate with any clinical feature (such as stage, age, Gleason score) or genetic alterations (*PTEN*, *ERG* status) of the samples (**Figure 8-2**). Suggesting prostate cancer progression is different between samples rather than cancer subgroup specific lncRNAs expression. Another explanation can be a different tumor content or cell of origin of the

samples. These different hypotheses will be addressed in detail in the discussion section (chapter 9).



**Figure 8-2: Heatmap of top 100 differently expressed lncRNAs in EO-PCA samples and normal.** Every row represent lncRNA, columns indicate samples (yellow – tumors, blue - normals). In the color palette used to represent expression normalized read counts from RNA-seq data, bright green denotes high expression (high number of reads) of lncRNAs, dark red denotes lowly expressed lncRNAs. Hierarchical clustering was performed using Manhattan distance and complete linkage using DESeq R package.

## 8.2. Genetic aberrations associated with lncRNA deregulation in EO-PCA

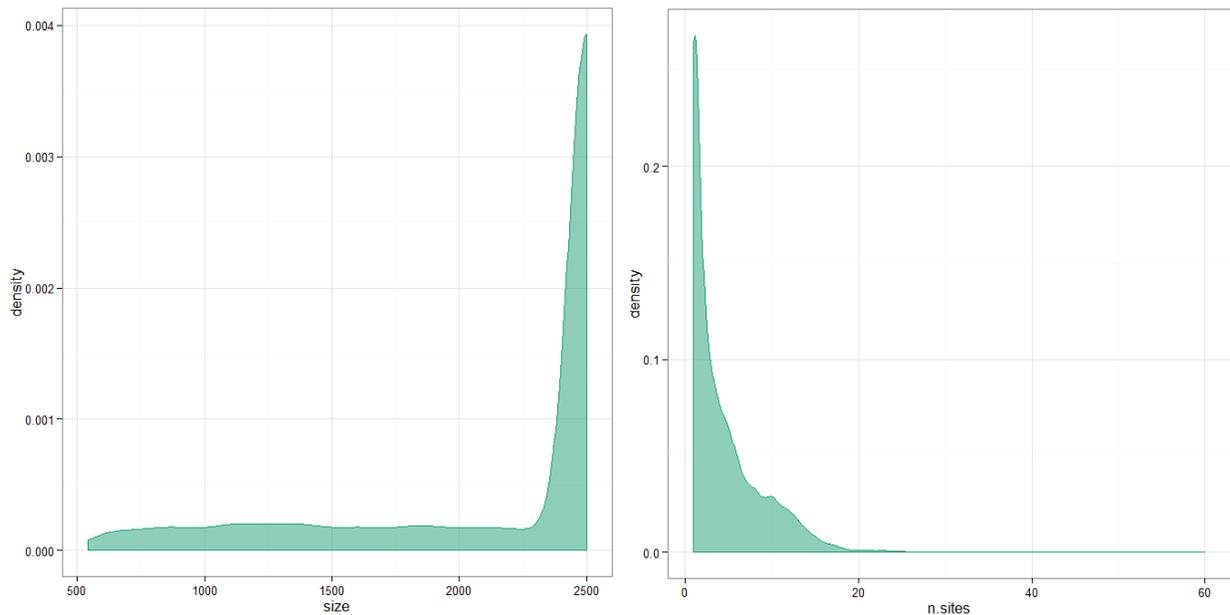
In order to identify the molecular mechanisms leading to lncRNA downregulation we searched for lncRNAs which either carried single nucleotide variants (SNVs) or which were affected by deletions. No SNVs were detected in the lncRNA sequences suggesting that SNVs do not play a role in the deregulation of lncRNAs. Next we addressed if mutations in regulatory sequences (promoter sequences) of lncRNAs could be identified. Using available ENCODE data (ROADMAP, histone marks from prostate cancer cell lines) we created a list of regulatory regions/potential promoters for lncRNAs and used this list to search for SNVs in the EO-PCA data set. We found that 10% of the deregulated lncRNAs have SNVs in potential promoter regions, suggesting the novel mechanism of lncRNAs deregulation. The effects of these SNVs should be further investigated in experimental settings and their importance further validated in larger sample cohorts. Next we assessed if larger genetic alterations such as deletions could have an effect on lncRNAs expression. Interestingly, a total of 102 lncRNAs were deleted in tumor samples and 56 of those overlapped with the 76 downregulated lncRNAs (27%) suggesting a causal link between deletion and downregulation.

Notably, the most frequently deleted lncRNAs were located in regions commonly deleted in PCA such as 3p14 [145]. Upregulation of lncRNAs expression could be the result of copy number gains or SNVs in their sequence or in their promoter region. 76 lncRNAs were found amplified in tumor samples and 45 of those overlap with upregulated lncRNAs. Only 3% of upregulated lncRNAs have SNV in potential promoter regions with no significant correlation to expression. This can be due to the lack of promoter regions description or the small number of samples analyzed for mutations. Such novel mechanism of lncRNAs deregulation by promoter SNV should be deeply investigated using bigger sample cohorts, what will be possible within an ICGC PanCancer project.

## 8.3. Promoter DNA-methylation associated with lncRNA deregulation in EO-PCA

Since, not all downregulated lncRNAs could be explained by genetic alterations, we searched for lncRNAs associated with promoter methylation using ICGC EO-PCA methylomes (450K Illumina data). Using GENOCODE17 ([www.genencodegenes.org](http://www.genencodegenes.org)), 12060 newly annotated lncRNAs were extracted. A list of promoters for these lncRNAs was generated by overlapping regions expanding from 2kb upstream to 500bp downstream of the transcription start sites (TSS) with active promoter marks as annotated by ENCODE. ENCODE tracks were obtained by merging active histone marks (such as H3K4me3) of six published cell lines ([www.genome.ucsc.edu/ENCODE/](http://www.genome.ucsc.edu/ENCODE/)). lncRNAs promoter/regulatory regions have been

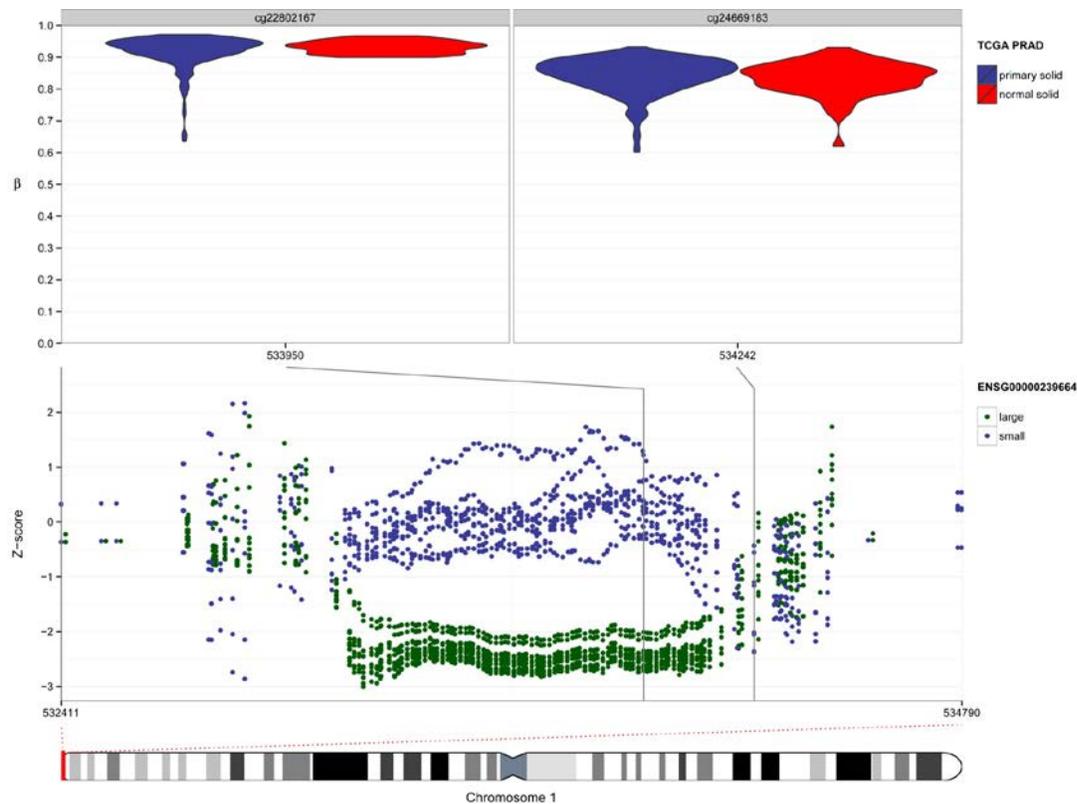
annotated by regions length and number of CG sites that are present in 450K arrays. In average, about 60% of all lncRNA predicted promoter regions are covered by 450K arrays with a mean 10 probes per region (range 1 to 60 CpGs). Description of regulatory regions and coverage by 450K Illumina arrays are described in **Figure 8-3**.



**Figure 8-3: Annotation of putative promoter regions of lncRNAs.** Left density plot represents promoter region length distributions. Right density plot represents distribution of number of sites per region of miRNA promoters.

Using ICGC EO-PCA MChp-seq data to analyze methylation of lncRNA promoters, 137 lncRNA with DMR in promoter regions were identified.

In order to compare the results of MChp-seq data, we used independent TCGA published data cohort of prostate cancer samples (450K Illumina arrays). Since 450K arrays were not designed to cover all promoter regions for miRNAs, only 42% of identified DMRs have at least 1 CpG probe on 450K and were taken for comparison. 67% of regions, covered by both methods have shown correlation of methylation data (example in **Figure 8-4**).

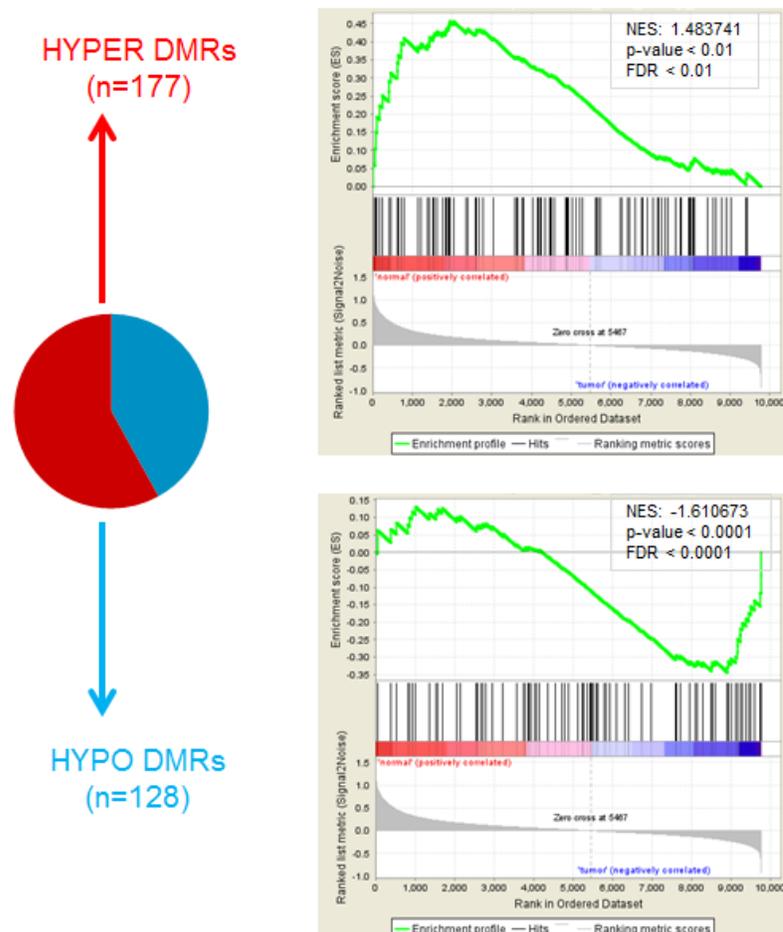


**Figure 8-4: Comparison of methylation degree of promoter of lncRNA obtained by MCIP-seq and using polished TCGA PRAD data set.** Upper part represents data obtained by TCGA PRAD data set: each plot shows methylation degree of CpG unit from 450K arrays. Blue color denotes primary tumors, red color – normal. Lower part presents data obtained by MCIP-seq arrays from ICGC EO-PCA data cohort. Each dot represents normalized read counts with a link to corresponding 450K array prob. Green/blue color denote large/small tumors, red color – normal.

Using methylation data from 66 EO-PCA samples and 120 normals, 177 hyper- and 128 hypomethylated regions (DMRs) in promoter/regulatory regions covering 254 lncRNAs were detected. No specific methylated samples subgroup was identified by unsupervised clustering of 1000 most differently methylated promoter/regulatory regions of lncRNAs. The respective contribution of deletions and promoter hypermethylations to the downregulation of those lncRNAs with annotated regulatory regions was assessed.

#### 8.4. Correlation of DNA methylation and expression deregulation of lncRNAs in EO-PCA

To confirm a functional impact of promoter DMRs on expression of lncRNAs, gene-set enrichment analysis (GSEA) was performed. Significant hyper and hypo methylated DMRs in promoter/regulatory regions of lncRNAs and expression of all expressed lncRNAs were used for unbiased correlation. GSEA has shown a significant enrichment for hyper and hypo DMRs with deregulated lncRNAs (**Figure 8-5**), suggesting that promoter methylation of lncRNAs is one of mechanisms of lncRNA regulation.

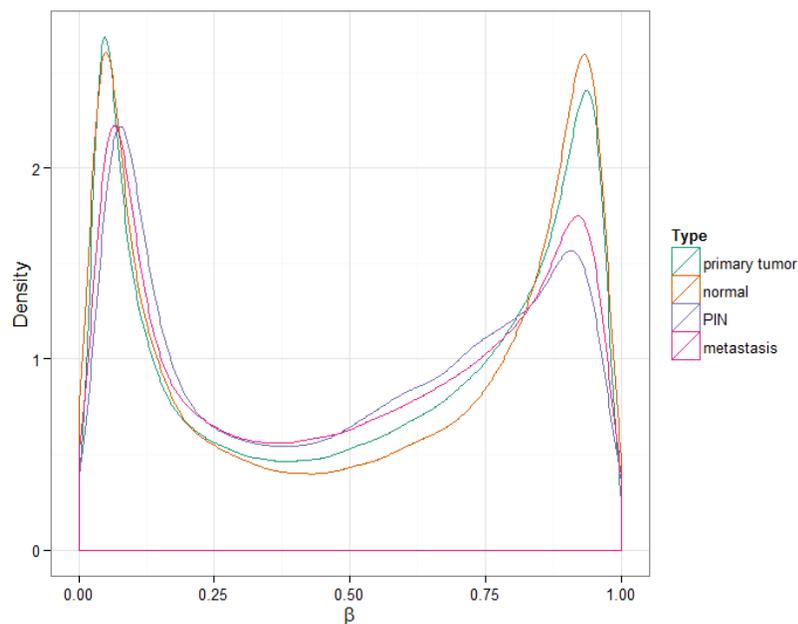


**Figure 8-5: Schematic representation of GSEA of hyper/hypo DMRs in promoter/regulatory regions of deregulated lncRNAs and expression of lncRNAs in EO-PCA data cohort.** Segmented cycle plot on left side represents total no of methylated DMRs. Color denotes for gain (red) and loss (blue) of methylation in EO-PCA vs. normal control samples. Right panel represents results of GSEA analysis.

### 8.5. Promoter DNA-methylation associated with lncRNAs deregulation in different stages PCA

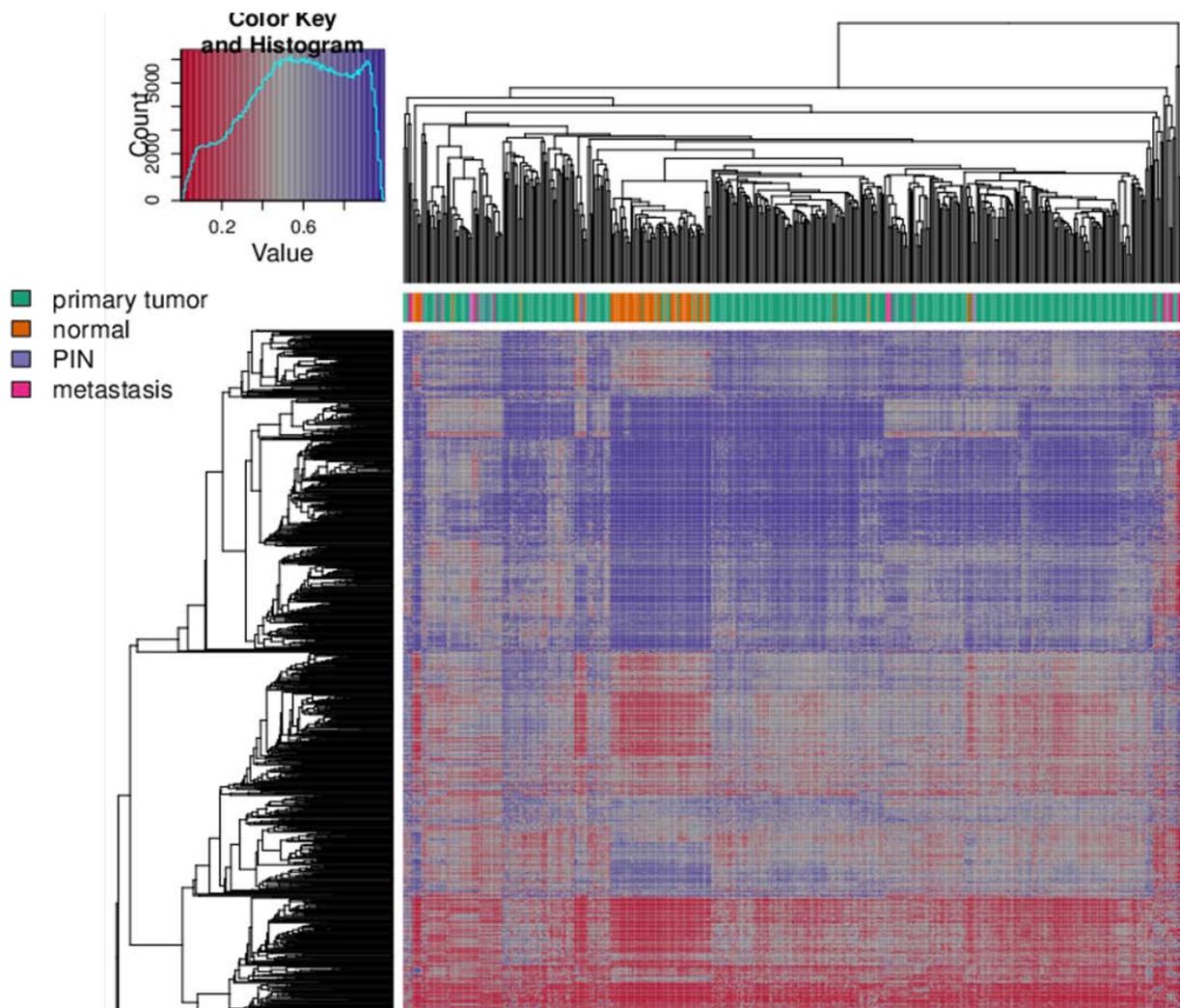
Dysregulation of lncRNAs has been associated with the multistep progression of PCA from prostatic intraepithelial neoplasia (PIN), localized adenocarcinoma to metastatic castration-resistance PCA (CRPC). There are few studies with identified unique broad spectrum of both tumor suppressive and oncogenic miRNAs and non-coding RNA deregulated in different stages and contributing to prostate carcinogenesis. In this chapter, we asked the question, whether promoter methylation associated with lncRNA deregulation is associated with different stages of prostate cancer development.

Using extended sample cohort **COMBINED PCA data set** (description in **chapter 5.4**), classical age distribution and includes PINs and metastasis from published data, firstly we asked the question, whether there are differences in DNA methylation of lncRNA promoters between different stages of prostate cancer. We checked the distribution of methylation in all annotated promoter/regulatory regions of lncRNAs and observed a shift towards 50% methylation peak in tumor, PIN and metastasis samples compare to normal samples. Such redistribution of methylation values indicates global methylated differences in these regulator regions between different sample types of prostate cancer (**Figure 8-6**).



**Figure 8-6:** Density plots depicting frequency (X-axis) of methylation  $\beta$ -values (Y-axis) for each of sample group (primary tumor-green, normal – yellow, PIN - blue, and metastasis - red).

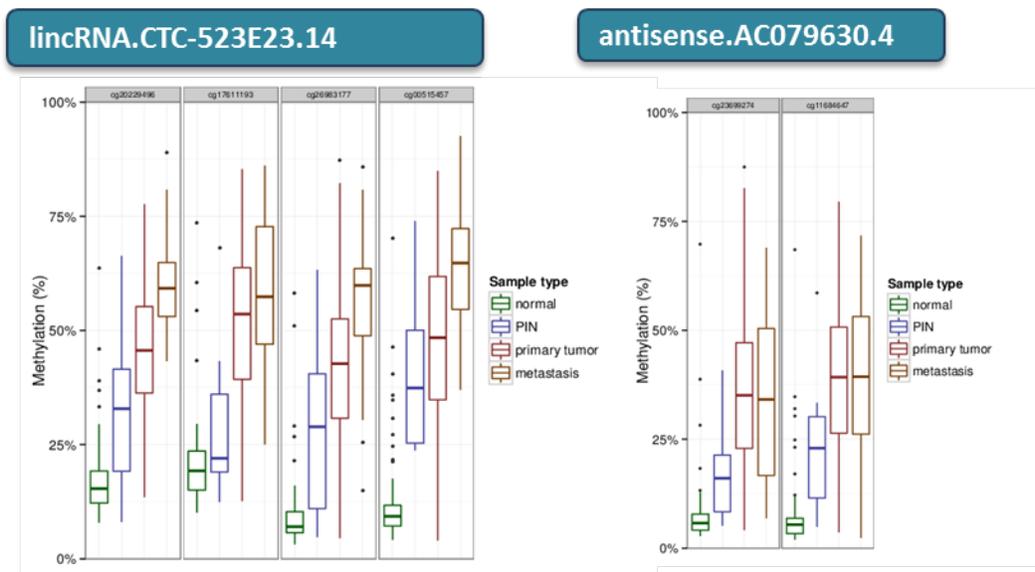
Next we performed unsupervised cluster analysis of all annotated promoter regions for lncRNAs using data cohort. We were able to separate normal samples to separate cluster, but have not observed a complete separation of different tumor stages (**Figure 8-10**). This indicates that there are more heterogeneity in samples as well as highlight a possible influence of cell of origin.



**Figure 8-7: Heatmap of 1000 differently methylated DMRs, covering promoter/regulatory regions of lncRNAs in EO-PCA samples and normal.** Every row denotes one of 1000 selected lncRNAs, and columns are samples (color denotes sample type as explained in color legend). In the color palette used to represent methylated degree, blue denotes 1 (high methylation level), dark red denotes 0 (no methylation). Hierarchical clustering is performed using Euclidian distance and complete linkage using RnBeads R package.

Although we were not able based on methylation profile of lncRNAs separate different stages of prostate cancer, we have identified a list of 56 lncRNAs with methylation increase from low methylated in normal to high methylated in metastasis samples. Examples of progressive DMRs

in promoter/regulatory region of lincRNA *CTC-523E23.14* and antisense *AC079630.4* are plotted in **Figure 8-8**. Such promoter DMRs after further validation can be used for diagnosis or survival of prostate cancer. Performing functional assays to characterize lncRNAs with progressive DMRs can help to find new mechanisms of prostate cancer progression and explain the role of lncRNAs in prostate tumorigenesis.



**Figure 8-8: Methylation data on promoters of lincRNA *CTC-523E23.14* and antisense *AC079630.4*.** Box-plot of methylation values of selected CpG, covering promoter/regulatory regions of lncRNA. Methylation degree (X-axis), different CpGs with CG id from 450K array (Y-axis). Different colors denotes for different sample type: green-normal, blue – PIN, red – primary tumor and orange – metastasis.

## 8.6. DNA methylation of promoter/regulatory regions of lncRNAs in PANCAN

The same approach as for miRNA (described in **chapter 7.9**) was used to in order to further investigate the role of promoter methylation of lncRNAs for changes in the regulation of lncRNAs expression in cancer. Using promoter/regulatory regions, DMR search was performed in **TCGA PANCAN data set (description in chapter 5.4)**. 88 DMRs (~10%) in promoter/regulatory regions of lncRNAs were pan-cancer specific, compare to 987 DMRs (~90%) DMRs were tissue specific. This may reflect recently published tissue-specific expression of lncRNAs in cancer [175].

## 8.7. Discussion of results

Next generation profiling data for cancer genomes allows an integrated analysis of genetic and epigenetic alterations, such as DNA methylation. In the majority of studies the power of an integrative analysis is not used to its full potential and focus is either on genetic or epigenetic alterations. The molecular mechanism of lncRNA deregulation in prostate cancer as well as in PANCAN to some extent was investigated. Based on our data we were able to identify 3 major mechanisms of lncRNA deregulation: non-coding mutation in promoter/regulatory region of lncRNA, promoter methylation of lncRNAs and loss/gain of alleles, carrying lncRNA gene (**Figure 8-9**). Further investigations are needed to unravel it with increased sample cohort and improved technique, that can allow distinguish allele specific events.

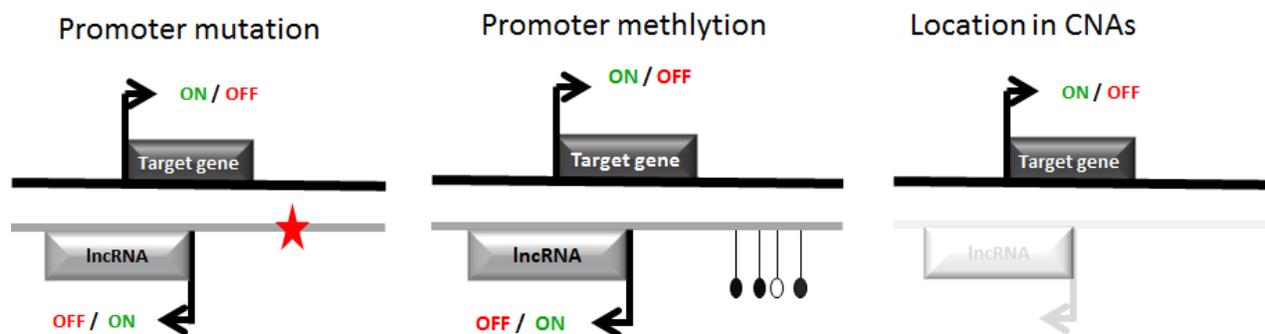


Figure 8-9. Schematic representation of major mechanisms of lncRNA deregulation in cancer.

Obtained data is a great resource and can be used in further studies for validation differently-regulated lncRNAs. Such resource will allow pick up interesting candidates for prostate cancer, as well as for other cancer types for functional studies to understand the mechanism and role of specific lncRNA in cancerogenesis.

Summarizing all results and discussion, the future directions of the thesis are quit broad and can be continued by:

1. Improving of integrative analysis by adding comprehensive non-coding RNA data
2. Adding non-coding mutations that can lead to deregulation of epigenetic regulators
3. Searching for novel cancer drivers and validating of defined epigenetic drivers
4. Creating a comprehensive 'validation of drivers' approach
5. Developing new drugs targeting specific driver mutation in epigenetic regulators
6. Investigation of functional relevance of translocation on deregulation of non-coding RNAs.
7. Identification the cell of origin of all tumor type

## REFERENCES

1. 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.
2. Waddington, C., *The epigenotype*. Endeavour, 1942. **1**(18-20).
3. Russo, V.E.A., R.A. Martienssen, and A.D. Riggs, *Epigenetic mechanisms of gene regulation*. 1996: Cold Spring Harbor Laboratory Press.
4. Bird, A., *Perceptions of epigenetics*. Nature, 2007. **447**(7143): p. 396-8.
5. Shen, H. and P.W. Laird, *Interplay between the cancer genome and epigenome*. Cell, 2013. **153**(1): p. 38-55.
6. Costa, F.F., *Non-coding RNAs, epigenetics and complexity*. Gene, 2008. **410**(1): p. 9-17.
7. Jenuwein, T. and C.D. Allis, *Translating the histone code*. Science, 2001. **293**(5532): p. 1074-80.
8. Berger, S.L., *The complex language of chromatin regulation during transcription*. Nature, 2007. **447**(7143): p. 407-12.
9. 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.
10. Groth, A., et al., *Chromatin challenges during DNA replication and repair*. Cell, 2007. **128**(4): p. 721-33.
11. Meissner, A., et al., *Genome-scale DNA methylation maps of pluripotent and differentiated cells*. Nature, 2008. **454**(7205): p. 766-70.
12. Gardiner-Garden, M. and M. Frommer, *CpG islands in vertebrate genomes*. J Mol Biol, 1987. **196**(2): p. 261-82.
13. Wu, H., et al., *Redefining CpG islands using hidden Markov models*. Biostatistics, 2010. **11**(3): p. 499-514.
14. Bird, A., et al., *A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA*. Cell, 1985. **40**(1): p. 91-9.
15. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription*. Genes Dev, 2011. **25**(10): p. 1010-22.
16. Baer, C., et al., *Extensive promoter DNA hypermethylation and hypomethylation is associated with aberrant microRNA expression in chronic lymphocytic leukemia*. Cancer Res, 2012. **72**(15): p. 3775-85.
17. Mohn, F., et al., *Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors*. Mol Cell, 2008. **30**(6): p. 755-66.
18. Yoder, J.A., C.P. Walsh, and T.H. Bestor, *Cytosine methylation and the ecology of intragenomic parasites*. Trends Genet, 1997. **13**(8): p. 335-40.
19. Chen, R.Z., et al., *DNA hypomethylation leads to elevated mutation rates*. Nature, 1998. **395**(6697): p. 89-93.
20. Bird, A., *DNA methylation patterns and epigenetic memory*. Genes Dev, 2002. **16**(1): p. 6-21.
21. Li, E., T.H. Bestor, and R. Jaenisch, *Targeted mutation of the DNA methyltransferase gene results in embryonic lethality*. Cell, 1992. **69**(6): p. 915-26.
22. Trowbridge, J.J., et al., *DNA methyltransferase 1 is essential for and uniquely regulates hematopoietic stem and progenitor cells*. Cell Stem Cell, 2009. **5**(4): p. 442-9.
23. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences*. Nature, 2009. **462**(7271): p. 315-22.
24. Schmidl, C., et al., *Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity*. Genome Res, 2009. **19**(7): p. 1165-74.
25. Maunakea, A.K., et al., *Conserved role of intragenic DNA methylation in regulating alternative promoters*. Nature, 2010. **466**(7303): p. 253-7.
26. Nguyen, C., et al., *Susceptibility of nonpromoter CpG islands to de novo methylation in normal and neoplastic cells*. J Natl Cancer Inst, 2001. **93**(19): p. 1465-72.
27. Hellman, A. and A. Chess, *Gene body-specific methylation on the active X chromosome*. Science, 2007. **315**(5815): p. 1141-3.
28. Laurent, L., et al., *Dynamic changes in the human methylome during differentiation*. Genome Res, 2010. **20**(3): p. 320-31.
29. Dawson, M.A. and T. Kouzarides, *Cancer epigenetics: from mechanism to therapy*. Cell, 2012. **150**(1): p. 12-27.
30. Margueron, R., et al., *Ezh1 and Ezh2 Maintain Repressive Chromatin through Different Mechanisms*. Molecular Cell, 2008. **32**(4): p. 503-518.
31. Pasini, D., et al., *Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity*. Embo Journal, 2004. **23**(20): p. 4061-4071.
32. Margueron, R., et al., *Role of the polycomb protein EED in the propagation of repressive histone marks*. Nature, 2009. **461**(7265): p. 762-U11.
33. Costa, F.F., *Non-coding RNAs: new players in eukaryotic biology*. Gene, 2005. **357**(2): p. 83-94.
34. Mattick, J.S. and I.V. Makunin, *Non-coding RNA*. Hum Mol Genet, 2006. **15 Spec No 1**: p. R17-29.

35. Inui, M., G. Martello, and S. Piccolo, *MicroRNA control of signal transduction*. Nat Rev Mol Cell Biol, 2010. **11**(4): p. 252-63.
36. Farh, K.K., et al., *The widespread impact of mammalian MicroRNAs on mRNA repression and evolution*. Science, 2005. **310**(5755): p. 1817-21.
37. He, L. and G.J. Hannon, *MicroRNAs: small RNAs with a big role in gene regulation*. Nat Rev Genet, 2004. **5**(7): p. 522-31.
38. Mendell, J.T., *MicroRNAs: critical regulators of development, cellular physiology and malignancy*. Cell Cycle, 2005. **4**(9): p. 1179-84.
39. Brennecke, J., et al., *Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila*. Cell, 2007. **128**(6): p. 1089-103.
40. Houwing, S., et al., *A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish*. Cell, 2007. **129**(1): p. 69-82.
41. Aravin, A.A., et al., *Developmentally regulated piRNA clusters implicate MILI in transposon control*. Science, 2007. **316**(5825): p. 744-7.
42. King, T.H., et al., *Ribosome structure and activity are altered in cells lacking snoRNPs that form pseudouridines in the peptidyl transferase center*. Mol Cell, 2003. **11**(2): p. 425-35.
43. Navarro, P., et al., *Tsix-mediated epigenetic switch of a CTCF-flanked region of the Xist promoter determines the Xist transcription program*. Genes Dev, 2006. **20**(20): p. 2787-92.
44. Pauler, F.M., M.V. Koerner, and D.P. Barlow, *Silencing by imprinted noncoding RNAs: is transcription the answer?* Trends Genet, 2007. **23**(6): p. 284-92.
45. Ng, K., et al., *Xist and the order of silencing*. EMBO Rep, 2007. **8**(1): p. 34-9.
46. Kawakami, T., et al., *XIST unmethylated DNA fragments in male-derived plasma as a tumour marker for testicular cancer*. Lancet, 2004. **363**(9402): p. 40-2.
47. Sado, T., Y. Hoki, and H. Sasaki, *Tsix silences Xist through modification of chromatin structure*. Dev Cell, 2005. **9**(1): p. 159-65.
48. Sleutels, F., R. Zwart, and D.P. Barlow, *The non-coding Air RNA is required for silencing autosomal imprinted genes*. Nature, 2002. **415**(6873): p. 810-3.
49. Yang, P.K. and M.I. Kuroda, *Noncoding RNAs and intranuclear positioning in monoallelic gene expression*. Cell, 2007. **128**(4): p. 777-86.
50. 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.
51. 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.
52. Huarte, M., et al., *A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response*. Cell, 2010. **142**(3): p. 409-19.
53. Guttman, M., et al., *lincRNAs act in the circuitry controlling pluripotency and differentiation*. Nature, 2011. **477**(7364): p. 295-300.
54. Orom, U.A., et al., *Long noncoding RNAs as enhancers of gene expression*. Cold Spring Harb Symp Quant Biol, 2010. **75**: p. 325-31.
55. Bejerano, G., et al., *A distal enhancer and an ultraconserved exon are derived from a novel retroposon*. Nature, 2006. **441**(7089): p. 87-90.
56. Calin, G.A., et al., *Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas*. Cancer Cell, 2007. **12**(3): p. 215-29.
57. Feuerhahn, S., et al., *TERRA biogenesis, turnover and implications for function*. FEBS Lett, 2010. **584**(17): p. 3812-8.
58. Kelly, T.K., D.D. De Carvalho, and P.A. Jones, *Epigenetic modifications as therapeutic targets*. Nat Biotechnol, 2010. **28**(10): p. 1069-1078.
59. Feinberg, A.P. and B. Vogelstein, *Hypomethylation distinguishes genes of some human cancers from their normal counterparts*. Nature, 1983. **301**(5895): p. 89-92.
60. Greger, V., et al., *Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma*. Hum Genet, 1989. **83**(2): p. 155-8.
61. Sakai, T., et al., *Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene*. Am J Hum Genet, 1991. **48**(5): p. 880-8.
62. Ohtani-Fujita, N., et al., *CpG methylation inactivates the promoter activity of the human retinoblastoma tumor-suppressor gene*. Oncogene, 1993. **8**(4): p. 1063-7.
63. Merlo, A., et al., *5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers*. Nat Med, 1995. **1**(7): p. 686-92.
64. Herman, J.G., et al., *Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers*. Cancer Res, 1995. **55**(20): p. 4525-30.
65. Gonzalez-Zulueta, M., et al., *Methylation of the 5' CpG island of the p16/CDKN2 tumor suppressor gene in normal and transformed human tissues correlates with gene silencing*. Cancer Res, 1995. **55**(20): p. 4531-5.
66. Herman, J.G., et al., *Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma*. Proc Natl Acad Sci U S A, 1994. **91**(21): p. 9700-4.

67. Veigl, M.L., et al., *Biallelic inactivation of hMLH1 by epigenetic gene silencing, a novel mechanism causing human MSI cancers*. Proc Natl Acad Sci U S A, 1998. **95**(15): p. 8698-702.
68. Jones, P.A. and S.B. Baylin, *The fundamental role of epigenetic events in cancer*. Nat Rev Genet, 2002. **3**(6): p. 415-28.
69. Yamada, Y., et al., *Opposing effects of DNA hypomethylation on intestinal and liver carcinogenesis*. Proc Natl Acad Sci U S A, 2005. **102**(38): p. 13580-5.
70. Feinberg, A.P. and B. Vogelstein, *Hypomethylation of ras oncogenes in primary human cancers*. Biochem Biophys Res Commun, 1983. **111**(1): p. 47-54.
71. De Smet, C., et al., *The activation of human gene MAGE-1 in tumor cells is correlated with genome-wide demethylation*. Proc Natl Acad Sci U S A, 1996. **93**(14): p. 7149-53.
72. Hanada, M., et al., *bcl-2 gene hypomethylation and high-level expression in B-cell chronic lymphocytic leukemia*. Blood, 1993. **82**(6): p. 1820-8.
73. Bender, S., et al., *Reduced H3K27me3 and DNA Hypomethylation Are Major Drivers of Gene Expression in K27M Mutant Pediatric High-Grade Gliomas*. Cancer Cell, 2013. **24**(5): p. 660-72.
74. Hovestadt, V., et al., *Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing*. Nature, 2014. **510**(7506): p. 537-41.
75. Toyota, M., et al., *CpG island methylator phenotype in colorectal cancer*. Proc Natl Acad Sci U S A, 1999. **96**(15): p. 8681-6.
76. Li, Q., et al., *Concordant methylation of the ER and N33 genes in glioblastoma multiforme*. Oncogene, 1998. **16**(24): p. 3197-202.
77. Toyota, M., et al., *Aberrant methylation in gastric cancer associated with the CpG island methylator phenotype*. Cancer Res, 1999. **59**(21): p. 5438-42.
78. Ueki, T., et al., *Hypermethylation of multiple genes in pancreatic adenocarcinoma*. Cancer Res, 2000. **60**(7): p. 1835-9.
79. Shen, L., et al., *DNA methylation and environmental exposures in human hepatocellular carcinoma*. J Natl Cancer Inst, 2002. **94**(10): p. 755-61.
80. Toyota, M., et al., *Methylation profiling in acute myeloid leukemia*. Blood, 2001. **97**(9): p. 2823-9.
81. Gu, L., et al., *BAZ2A (TIP5) is involved in epigenetic alterations in prostate cancer and its overexpression predicts disease recurrence*. Nat Genet, 2015. **47**(1): p. 22-30.
82. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome*. Nature, 2009. **458**(7239): p. 719-724.
83. Dees, N.D., et al., *MuSiC: Identifying mutational significance in cancer genomes*. Genome Research, 2012. **22**(8): p. 1589-1598.
84. Chin, L., et al., *Comprehensive genomic characterization defines human glioblastoma genes and core pathways*. Nature, 2008. **455**(7216): p. 1061-1068.
85. Hodis, E., et al., *A Landscape of Driver Mutations in Melanoma*. Cell, 2012. **150**(2): p. 251-263.
86. Defromental, C.C. and T. Soussi, *Tp53 Tumor Suppressor Gene - a Model for Investigating Human Mutagenesis*. Genes Chromosomes & Cancer, 1992. **4**(1): p. 1-15.
87. Suda, K., K. Tomizawa, and T. Mitsudomi, *Biological and clinical significance of KRAS mutations in lung cancer: an oncogenic driver that contrasts with EGFR mutation*. Cancer and Metastasis Reviews, 2010. **29**(1): p. 49-60.
88. You, J.S. and P.A. Jones, *Cancer genetics and epigenetics: two sides of the same coin?* Cancer Cell, 2012. **22**(1): p. 9-20.
89. Sturm, D., et al., *Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma*. Cancer Cell, 2012. **22**(4): p. 425-37.
90. Schwartzentruber, J., et al., *Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma*. Nature, 2012. **482**(7384): p. 226-31.
91. Turcan, S., et al., *IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype*. Nature, 2012. **483**(7390): p. 479-83.
92. Figueroa, M.E., et al., *Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation*. Cancer Cell, 2010. **18**(6): p. 553-67.
93. Shih, A.H., et al., *The role of mutations in epigenetic regulators in myeloid malignancies*. Nat Rev Cancer, 2012. **12**(9): p. 599-612.
94. Yan, X.J., et al., *Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia*. Nat Genet, 2011. **43**(4): p. 309-15.
95. Vogelstein, B., et al., *Cancer genome landscapes*. Science, 2013. **339**(6127): p. 1546-58.
96. Zack, T.I., et al., *Pan-cancer patterns of somatic copy number alteration*. Nat Genet, 2013. **45**(10): p. 1134-1140.
97. Esquela-Kerscher, A. and F.J. Slack, *Oncomirs - microRNAs with a role in cancer*. Nat Rev Cancer, 2006. **6**(4): p. 259-69.
98. Hammond, S.M., *MicroRNAs as tumor suppressors*. Nat Genet, 2007. **39**(5): p. 582-3.
99. Balatti, V., et al., *miR Deregulation in CLL*. Adv Exp Med Biol, 2013. **792**: p. 309-25.
100. Majid, S., et al., *miR-23b represses proto-oncogene Src kinase and functions as methylation-silenced tumor suppressor with diagnostic and prognostic significance in prostate cancer*. Cancer Res, 2012. **72**(24): p. 6435-46.

101. Börno, 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.
102. Davalos, V., et al., *Dynamic epigenetic regulation of the microRNA-200 family mediates epithelial and mesenchymal transitions in human tumorigenesis*. *Oncogene*, 2012. **31**(16): p. 2062-74.
103. Calin, G.A., et al., *Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia*. *Proc Natl Acad Sci U S A*, 2002. **99**(24): p. 15524-9.
104. Calin, G.A., et al., *Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers*. *Proc Natl Acad Sci U S A*, 2004. **101**(9): p. 2999-3004.
105. Esteller, M., *Non-coding RNAs in human disease*. *Nat Rev Genet*, 2011. **12**(12): p. 861-74.
106. Melo, S.A., et al., *A TARBP2 mutation in human cancer impairs microRNA processing and DICER1 function*. *Nat Genet*, 2009. **41**(3): p. 365-70.
107. Hill, D.A., et al., *DICER1 mutations in familial pleuropulmonary blastoma*. *Science*, 2009. **325**(5943): p. 965.
108. Melo, S.A., et al., *A genetic defect in exportin-5 traps precursor microRNAs in the nucleus of cancer cells*. *Cancer Cell*, 2010. **18**(4): p. 303-15.
109. Plath, K., et al., *Role of histone H3 lysine 27 methylation in X inactivation*. *Science*, 2003. **300**(5616): p. 131-5.
110. Yu, W., et al., *Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA*. *Nature*, 2008. **451**(7175): p. 202-6.
111. 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.
112. Ciriello, G., et al., *Emerging landscape of oncogenic signatures across human cancers*. *Nat Genet*, 2013. **45**(10): p. 1127-1133.
113. Kandoth, C., et al., *Mutational landscape and significance across 12 major cancer types*. *Nature*, 2013. **502**(7471): p. 333-9.
114. Tamborero, D., et al., *Comprehensive identification of mutational cancer driver genes across 12 tumor types*. *Sci Rep*, 2013. **3**: p. 2650.
115. Jacobsen, A., et al., *Analysis of microRNA-target interactions across diverse cancer types*. *Nat Struct Mol Biol*, 2013. **20**(11): p. 1325-32.
116. Ahmed, H., *Promoter Methylation in Prostate Cancer and its Application for the Early Detection of Prostate Cancer Using Serum and Urine Samples*. *Biomark Cancer*, 2010. **2010**(2): p. 17-33.
117. Heidenreich, A., et al., *EAU guidelines on prostate cancer*. *Eur Urol*, 2008. **53**(1): p. 68-80.
118. 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.
119. Berger, M.F., et al., *The genomic complexity of primary human prostate cancer*. *Nature*, 2011. **470**(7333): p. 214-20.
120. Kim, J.H., et al., *Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer*. *Genome Res*, 2011. **21**(7): p. 1028-41.
121. Bill-Axelsson, A., et al., *Radical prostatectomy versus watchful waiting in localized prostate cancer: the Scandinavian prostate cancer group-4 randomized trial*. *J Natl Cancer Inst*, 2008. **100**(16): p. 1144-54.
122. Reid, A.H., et al., *Novel, gross chromosomal alterations involving PTEN cooperate with allelic loss in prostate cancer*. *Mod Pathol*, 2012. **25**(6): p. 902-10.
123. Tomlins, S.A., et al., *Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer*. *Nature*, 2007. **448**(7153): p. 595-9.
124. Suva, M.L., N. Riggi, and B.E. Bernstein, *Epigenetic reprogramming in cancer*. *Science*, 2013. **339**(6127): p. 1567-70.
125. Perry, A.S., et al., *The epigenome as a therapeutic target in prostate cancer*. *Nat Rev Urol*, 2010. **7**(12): p. 668-80.
126. Jeronimo, C., et al., *Epigenetics in prostate cancer: biologic and clinical relevance*. *Eur Urol*, 2011. **60**(4): p. 753-66.
127. Jeronimo, C., et al., *A quantitative promoter methylation profile of prostate cancer*. *Clin Cancer Res*, 2004. **10**(24): p. 8472-8.
128. Isaacs, J.T., *The biology of hormone refractory prostate cancer. Why does it develop?* *Urol Clin North Am*, 1999. **26**(2): p. 263-73.
129. 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.
130. Brase, J.C., et al., *TMPRSS2-ERG -specific transcriptional modulation is associated with prostate cancer biomarkers and TGF-beta signaling*. *BMC Cancer*, 2011. **11**: p. 507.
131. Aryee, M.J., et al., *DNA methylation alterations exhibit intraindividual stability and interindividual heterogeneity in prostate cancer metastases*. *Sci Transl Med*, 2013. **5**(169): p. 169ra10.
132. Brocks, D., et al., *Intratumor DNA Methylation Heterogeneity Reflects Clonal Evolution in Aggressive Prostate Cancer*. *Cell Reports*, 2014. **8**(3): p. 798-806.
133. Claus, R., et al., *Quantitative analyses of DAPK1 methylation in AML and MDS*. *International Journal of Cancer*, 2012. **131**(2): p. E138-E142.

134. Assenov, Y., et al., *Comprehensive analysis of DNA methylation data with RnBeads*. *Nature Methods*, 2014. **11**(11): p. 1138-1140.
135. Jones, D.T., et al., *Dissecting the genomic complexity underlying medulloblastoma*. *Nature*, 2012. **488**(7409): p. 100-5.
136. Xi, R., et al., *Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion*. *Proc Natl Acad Sci U S A*, 2011. **108**(46): p. E1128-36.
137. Anders, S. and W. Huber, *Differential expression analysis for sequence count data*. *Genome Biol*, 2010. **11**(10): p. R106.
138. Sturn, A., J. Quackenbush, and Z. Trajanoski, *Genesis: cluster analysis of microarray data*. *Bioinformatics*, 2002. **18**(1): p. 207-8.
139. Brase, J.C., et al., *Circulating miRNAs are correlated with tumor progression in prostate cancer*. *Int J Cancer*, 2011. **128**(3): p. 608-16.
140. 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.
141. Behjati, S., et al., *Distinct H3F3A and H3F3B driver mutations define chondroblastoma and giant cell tumor of bone*. *Nature Genetics*, 2013. **45**(12): p. 1479-U105.
142. Wu, J.N. and C.W.M. Roberts, *ARID1A Mutations in Cancer: Another Epigenetic Tumor Suppressor?* *Cancer Discovery*, 2013. **3**(1): p. 35-43.
143. Abdel-Wahab, O., et al., *Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies*. *Blood*, 2009. **114**(1): p. 144-147.
144. Yang, L.B., R. Rau, and M.A. Goodell, *DNMT3A in haematological malignancies*. *Nature Reviews Cancer*, 2015. **15**(3): p. 152-165.
145. de Andres, M.C., et al., *Epigenetic regulation during fetal femur development: DNA methylation matters*. *PLoS One*, 2013. **8**(1): p. e54957.
146. Watson, I.R., et al., *Emerging patterns of somatic mutations in cancer*. *Nature Reviews Genetics*, 2013. **14**(10): p. 703-718.
147. Garraway, L.A., et al., *Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma*. *Nature*, 2005. **436**(7047): p. 117-122.
148. Wartman, L.D., et al., *Sequencing a mouse acute promyelocytic leukemia genome reveals genetic events relevant for disease progression*. *Journal of Clinical Investigation*, 2011. **121**(4): p. 1445-1455.
149. Brookman-Amisshah, N., et al., *Allelic imbalance at 13q14.2 approximately q14.3 in localized prostate cancer is associated with early biochemical relapse*. *Cancer Genet Cytogenet*, 2007. **179**(2): p. 118-26.
150. Lu, W., et al., *Allelotyping analysis at chromosome 13q of high-grade prostatic intraepithelial neoplasia and clinically insignificant and significant prostate cancers*. *Prostate*, 2006. **66**(4): p. 405-12.
151. Taylor, B.S., et al., *Integrative genomic profiling of human prostate cancer*. *Cancer Cell*, 2010. **18**(1): p. 11-22.
152. Tomlins, S.A., et al., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer*. *Science*, 2005. **310**(5748): p. 644-8.
153. Hulf, T., et al., *Discovery pipeline for epigenetically deregulated miRNAs in cancer: integration of primary miRNA transcription*. *BMC Genomics*, 2011. **12**: p. 54.
154. Hulf, T., et al., *Epigenetic-induced repression of microRNA-205 is associated with MED1 activation and a poorer prognosis in localized prostate cancer*. *Oncogene*, 2012.
155. Yu, F., et al., *MicroRNA 34c gene down-regulation via DNA methylation promotes self-renewal and epithelial-mesenchymal transition in breast tumor-initiating cells*. *J Biol Chem*, 2012. **287**(1): p. 465-73.
156. Rauhala, H.E., et al., *miR-193b is an epigenetically regulated putative tumor suppressor in prostate cancer*. *Int J Cancer*, 2010. **127**(6): p. 1363-72.
157. Suzuki, H., et al., *Genome-wide profiling of chromatin signatures reveals epigenetic regulation of MicroRNA genes in colorectal cancer*. *Cancer Res*, 2011. **71**(17): p. 5646-58.
158. Wee, E.J., et al., *Mapping the regulatory sequences controlling 93 breast cancer-associated miRNA genes leads to the identification of two functional promoters of the Hsa-mir-200b cluster, methylation of which is associated with metastasis or hormone receptor status in advanced breast cancer*. *Oncogene*, 2012. **31**(38): p. 4182-95.
159. Asangani, I.A., et al., *Genetic and epigenetic loss of microRNA-31 leads to feed-forward expression of EZH2 in melanoma*. *Oncotarget*, 2012. **3**(9): p. 1011-25.

## APPENDIX

## Supplemental Tables

**Supplemental Table S1:** Primers used for MassARRAY analyses (capital letters represent sequences that are complementary to DNA, lowercase letters represent tags for MassARRAY analyses). Y = C/T, R = G/A

miRNA cluster_DMR	Forward primer	Reverse primer
miR-23b~27b~24-1_i	aggaagagagTAGAAGAAAGGATTGAAG ATGGG	cagtaatcagactcactatagggagaaggctCCTCCCCAC CCTTCAATATTC
miR-23b~27b~24-1_ii	aggaagagagGGGATGGTYGTTAGGATT TTG	cagtaatcagactcactatagggagaaggctTACTCTATTA TCATAACCAAATCC
miR-23b~27b~24-1_iii	aggaagagagGTATAATGTGTTGTAGAA AGGAAG	cagtaatcagactcactatagggagaaggctTAAACAAAA ACTTCATTCTCCTAC
miR-23b~27b~24-1_iv	aggaagagagGATAGATTGGAGAGGGG GAG	cagtaatcagactcactatagggagaaggctAATAAACCTA AAACTAATAACAACC
miR-23b~27b~24-1_v	aggaagagagGGAAGTGGAGTGGTTTAT ATAG	cagtaatcagactcactatagggagaaggctATCTTTATAA AATAAACATTTACTCC
miR-23b~27b~24-1_vi	aggaagagagTTAGGTATTGAGGGAATT AGTTAG	cagtaatcagactcactatagggagaaggctATTACCCCT CACCTAAAAACC
miR-23b~27b~24-1_vii	aggaagagagGATGAGTTTAGATTAGTT ATTTTATAG	cagtaatcagactcactatagggagaaggctCTCCAATT CTCTACACAAACTC
miR-23b~27b~24-1_viii	aggaagagagGAGTATATGTGGTATATA TAGTG	cagtaatcagactcactatagggagaaggctAAATAATTTA AATAATCTAAAACCTCC
miR-23a~27a~24-2_i	aggaagagagGTTAAGTTTTGTTTTTATAG GTTAGG	cagtaatcagactcactatagggagaaggctTCCAAATAC CAACCTCTAACCC
miR-23a~27a~24-2_ii	aggaagagagGGTGGGTAYGATAYGGG GG	cagtaatcagactcactatagggagaaggctAACAAACAA ACCTTACCTATAAATC
miR-23a~27a~24-2_iii	aggaagagagGGTAGTAAGTTTGGGATA TTTAG	cagtaatcagactcactatagggagaaggctCTCCTACAT CCCRCTCCCTC
miR-29b-1~29a_i	aggaagagagAGTAGTTTTTAYGGAAGA TATAG	cagtaatcagactcactatagggagaaggctAATTCAATAA CAAACACTTTAATTC
miR-29b-1~29a_iii	aggaagagagTGAATAAATAGTTTTAGTT TGTTTG	cagtaatcagactcactatagggagaaggctCCTCAAATTA TAATCTTCATTATTC
miR-31_i	aggaagagagTAAAGTGATAGTAATTTTA GGTGG	cagtaatcagactcactatagggagaaggctCCTCCTCTTA TCCAAATCTCAA
miR-31_ii	aggaagagagGAAAGGAGGGGGAGGGA AG	cagtaatcagactcactatagggagaaggctTCCTCTCC CTTAACTCTAAC
miR-133a-2_i	aggaagagagGTTTTATTAGATGTYGATT TGGG	cagtaatcagactcactatagggagaaggctAAACCCCTC CTACCTTTAATAAA
miR-149_ii	aggaagagagGTAATAAGTGATAGGTT TGAGTG	cagtaatcagactcactatagggagaaggctATAACAATA ACCRCTAAATAACC

miR-149_iii	aggaagagagAGGGATAGTGTGAGTTTG TGG	cagtaatcgactcactatagggagaaggctCAAACACCC ACTAAACTCTCC
miR-149_iv	aggaagagagAGGAGGAATAGAATTTAT AGATAG	cagtaatcgactcactatagggagaaggctCCTTACAAAA ACCTCAAAC TAATA
miR-149_v	aggaagagagTGGTGATAGTGTGAGGGA GTG	cagtaatcgactcactatagggagaaggctCCTAAATAAC CCRACCCTACC
miR-149_vi	aggaagagagGGTAYGTAGGTGTGTATA TATTG	cagtaatcgactcactatagggagaaggctCAACAAATTA ACTTTAACTTCTCTC
miR-149_vii	aggaagagagGGGTGTGGGGAGGGGAG G	cagtaatcgactcactatagggagaaggctCAATCCCTAT CCRCC TAAACTC
miR-181c~181d_i	aggaagagagTTTAGGTTTGTGATATTT TAGTG	cagtaatcgactcactatagggagaaggctAAATACACAA ACACACCACATTC
miR-181c~181d_ii	aggaagagagGATTGTTTATAGATGTATT AAGTTAG	cagtaatcgactcactatagggagaaggctCAATAAACA CTATCCCTAAATATC
miR-181c~181d_iii	aggaagagagGTGGAGGTGGYGGATAA AGAG	cagtaatcgactcactatagggagaaggctCACCTTCTAA AACCCTTTCCCC
miR-205_ii	aggaagagagGGTATGGAGTTGATAATT ATGAG	cagtaatcgactcactatagggagaaggctTCTAACTATC TCTATTCCTAAATC
miR-452~224_i	aggaagagagAGGATTTYGGGAGTTTAT TTTATAG	cagtaatcgactcactatagggagaaggctCTCCAATCR AAATAAATCTCCATC
miR-1250~338~657_i	aggaagagagTGTGTTGGGGTTTAGGGT TTG	cagtaatcgactcactatagggagaaggctAACCCCAAC TTCRCCTTCAAC
miR-378a_i	aggaagagagGAGGGGTAGAGTTATTTT TGGG	cagtaatcgactcactatagggagaaggctATAATTAATT ATTACTAAACCAACAC
miR-183~96~182_i	aggaagagagTAGTATGGGTTTATTATGA GTAGG	cagtaatcgactcactatagggagaaggctCATCCCATC TCACTCCACCC
miR-183~96~182_ii	aggaagagagTATTTAGATAGGTTAAGTA AGGTG	cagtaatcgactcactatagggagaaggctTATCCAATAA TATCCAATATCAAAC
miR-183~96~182_iii	aggaagagagTTAGGGATTTATAGGGTG TGTG	cagtaatcgactcactatagggagaaggctCCCTTTACAT TAAACACCAATAC
miR-183~96~182_iv	aggaagagagAGGGGGTGAGGTAGTGG AAG	cagtaatcgactcactatagggagaaggctCTAAAAC TC RTCCAATTATCCAC
miR-100~let-7a-2_i	aggaagagagAAGGGGAAGAGAAGGAG AATG	cagtaatcgactcactatagggagaaggctTATTTAAACC TATTACCACAAACC
miR-100~let-7a-2_ii	aggaagagagGTTAAATGATTTAGGAGT TATTATT	cagtaatcgactcactatagggagaaggctCTACACTATA AATTCRCCTTCC
miR-7-1_i	aggaagagagAGATAATGGAATTGTGGA GTAG	cagtaatcgactcactatagggagaaggctTTTCAATCAA CCTCATATATTCC
miR-375_i	aggaagagagATTAAGGGAGTGTAGGG AGTG	cagtaatcgactcactatagggagaaggctACCCTCCCC AACCTCCTATC
LINE-1	aggaagagagTTTATATTTTGGTATGATT TTGTAG	cagtaatcgactcactatagggagaaggctACCCAAAAT CAACTATCCACCCTT

Supplemental Table S2. List of epigenetic regulators

Group	Subgroup	Gene names
Histone modification	Histones	H3F3A,HIST2H2AA3,HIST2H2AB,HIST2H2AC,HIST2H2BE,HIST2H2BF,HIST2H3C,HIST2H3D,HIST3H2A,HIST3H2BB,HIST3H3,H1FOO,H1FX,H2AFZ,H2AFY,HIST1H1A,HIST1H1B,HIST1H1C,HIST1H1D,HIST1H1E,HIST1H1T,HIST1H2AA,HIST1H2AB,HIST1H2AC,HIST1H2AD,HIST1H2AE,HIST1H2AG,HIST1H2AH,HIST1H2AI,HIST1H2AJ,HIST1H2AK,HIST1H2AL,HIST1H2AM,HIST1H2BA,HIST1H2BB,HIST1H2BC,HIST1H2BD,HIST1H2BE,HIST1H2BF,HIST1H2BG,HIST1H2BH,HIST1H2BI,HIST1H2BJ,HIST1H2BK,HIST1H2BL,HIST1H2BM,HIST1H2BN,HIST1H2BO,HIST1H3A,HIST1H3B,HIST1H3C,HIST1H3D,HIST1H3E,HIST1H3F,HIST1H3G,HIST1H3H,HIST1H3I,HIST1H3J,HIST1H4A,HIST1H4B,HIST1H4C,HIST1H4D,HIST1H4E,HIST1H4F,HIST1H4G,HIST1H4H,HIST1H4I,HIST1H4J,HIST1H4K,HIST1H4L,H2AFV,H2AFY2,H2AFX,H1FNT,H2AFJ,H3F3C,HIST4H4,H3F3B,H1F0,H2AFB1,H2AFB2,H2AFB3,H2BFM,H2BFWT
	Editors	EYA3,HDAC1,KDM1A,KDM4A,MYSM1,PPP2R5A,USP21,ANKRD44,HDAC4,KDM3A,PPP1CB,SMEK2,ANKRD28,BAP1,HDAC11,PPP2R3A,PPP4R2,PPP2R2C,PPP3CA,DUSP1,HDAC3,KDM3B,PPP2CA,PPP2R2B,EYA4,HDAC2,KDM1B,PPP2R5D,SIRT5,HDAC9,EYA1,PPP2CB,PPP2R2A,PPP3CC,KDM4C,PPP2R4,PPP6C,JMJD1C,PARG,PPP2R2D,PPP3CB,SIRT1,KDM4D,PPP1CA,PPP2R1B,PPP2R5B,PPP6R3,SIRT3,ANKRD52,HDA C7,KDM2B,KDM5A,PPP1CC,SIRT4,PPP2R3C,PPP2R5C,PPP2R5E,PPP4R4,SMEK1,U SP3,KDM8,PPP4C,USP7,HDAC5,JMJD6,PPM1D,SIRT7,USP36,PPP4R1,KDM4B,PPP2 R1A,PPP5C,PPP6R1,SIRT2,SIRT6,EYA2,USP16,USP25,HDAC10,PPP6R2,BRCC3,HD AC6,HDAC8,KDM5C,KDM6A,PHF8,PPP2R3B,KDM5D,UTY
	Readers	AHDC1,ARID4B,ATXN7L2,CHD5,DMAP1,GATAD2B,KDM5B,LRIF1,MEAF6,MIER1,POG Z,RBBP4,RBBP5,RLF,RNF2,TADA1,TAF12,TAF13,TAF5L,TRIM33,VPS72,ZMYM4,ZMY M6,ZNF687,ZZZ3,EPC2,MTA3,MYT1L,ORC2,PIIG,SAP130,SUPT7L,ATXN7,DHX30,FA M208A,RUVBL1,TADA3,UBXN7,YEATS2,GABRG1,ING2,SMARCA5,TADA2B,ZNF518B, CHD1,NIPBL,PWWP2A,SAP30L,TAF7,TAF9,BRPF3,PHF1,RPS10,SUPT3H,TAF11,TAF8 ,TBP,ATXN7L1,CRCP,GATAD1,ING3,JAZF1,LRWD1,MGAM,PHF14,TAF6,TRIM24,TRR AP,ASH2L,TAF2,ZFPM2,ZBTB43,ZNF462,EPC1,PCGF6,SMC3,TAF3,TAF5,C11orf30,CF L1,EED,KDM2A,MEN1,MTA2,PRDM10,TAF10,TAF6L,TRIM66,BCL7A,CDK2AP1,CHD4, EP400,HCFC2,PPHLN1,SETD1B,SUDS3,YEATS4,KBTBD7,KPNA3,ZMYM2,ZMYM5,ARI D4A,BRMS1L,CHD8,GTF2A1,HOMEZ,MAX,MBIP,MTA1,UBR7,MGA,MORF4L1,SIN3A,Z NF592,FLYWCH1,KLHL36,PRR14,SETD1A,SF3B3,ATXN7L3,CBX4,CHD3,DGKE,KAT2A ,KDM6B,MBTD1,PHF12,RAI1,SRSF2,SUZ12,TADA2A,USP22,ADNP,CXXC1,MBD2,SM CHD1,ZNF532,GATAD2A,MIER2,RUVBL2,SIN3B,TRIM28,ZNF428,ADNP,CHD6,CSR P2BP,MRGBP,TAF4,ZMYND8,BRWD1,WRB,BRD1,L3MBTL2,TCF20,HCFC1,MORF4L2,RB BP7,SMC1A,TAF1,ZMYM3
Chromatin remodeling	Writers	ASH1L,PADI4,PARP1,PRDM2,PRKAA2,PRKAB2,PRMT6,SETDB1,SMYD3,ANTXR1,BA RD1,BUB1,HAT1,NCOA1,PRKAG3,SMYD1,SMYD5,TLK1,ATR,BTD,DTX3L,DZIP3,GSK3 B,KAT2B,PAK2,PARP3,RNF168,SETD2,SETD5,SETMAR,UBE2E1,CLOCK,PRMT10,SE TD7,NSD1,PRDM6,PRDM9,PRKAA1,STK10,UBE2B,CDYL,EHMT2,FYN,HSD17B8,RING 1,RNF8,UBR2,BAZ1B,CDK5,EZH2,PRKAG2,UBE2H,ELP3,KAT6A,NCOA2,PBK,PRKDC, EHMT1,GTF3C4,JAK2,RNF20,CHUK,KAT6B,MAP3K8,MASTL,ATM,CHEK1,DDB1,DDB2 ,KAT5,NAT10,PAK1,PRMT3,RAG1,RPS6KA4,SUV420H1,ASCL1,CDK17,CDK2,MAP3K1 2,PRKAB1,PRKAG1,PRMT8,SETD8,BRCA2,CDK8,SETDB2,PARP2,PRMT5,RPS6KA5,S ETD3,VRK1,CREBBP,KAT8,PRDM7,PRKCB,PRMT7,RNF40,UBE2I,AURKB,CDK3,EZH1 ,GSG2,KAT7,PRKCA,SMYD4,MALT1,AURKC,CARM1,DAPK3,DOT1L,GSK3A,PKN1,PR MT1,SAE1,SUV420H2,UBA2,NCOA3,STK4,HLCS,PRMT2,EP300,LIMK2,RBX1,HUWE1, OGT,OPN1LW,RPS6KA3,UBA1,UBE2A,CDY1,CDY2A
	Chromatin remodeling factors	BRDT,HIST2H4A,ITGB3BP,NASP,RERE,CENPA,CENPO,FSHR,HJURP,KANSL3,SATB2 ,TNP1,PBRM1,SMARCC1,TP63,WDR82,CENPH,CENPK,NPM1,NR3C1,CENPQ,HMGA1 ,MYB,PHF10,ACTL6B,CBX3,SMARCD3,CHRAC1,MYC,NPM2,TOP1MT,CDKN2A,CENP P,TTF1,BNIP3,ELP4,RSF1,BAZ2A,HNF1A,KANSL2,SMARCC2,SMARCD1,SYCP3,RB1, FOXA1,MIS18BP1,NKX2-1,BAHD1,CASC5,OIP5,CENPN,SALL1,CENPV,KANSL1,SMARCD2,SMARCE1,SOX9,S UPT4H1,SUPT6H,KLF1,SUPT5H,PHF20,TOP1,MIS18A,CENPI

DNA modification	Chromatin remodeling helicase	CHD1L,RAD54L,TTF2,SMARCAL1,SMARCA1,CHD7,RAD54B,SMARCA2,BTAF1,ERC6,HELLS,CHD2,INO80,CHD9,SRCAP,SMARCA4,ATRX,SMARCA1
	nucleosome remodeling factors	UCHL5,ACTL6A,HLTF,DPF2,NAP1L4,NFRKB,MCRS1,YY1,BPTF,DPF1,NAP1L3
	Editors	GADD45A,TET3,TET2,TET1,ALKBH3,AICDA,APOBEC1,TDG,ALKBH1,FTO, GADD45B
	Readers	MBD4,MBD1,MBD3,UHRF1,PCNA
	Writers	MAEL,DNMT3A,IDH1,ZFP57,ASZ1,MGMT,TRDMT1,HINFP,SPI1,FOS,IDH2,DNMT1,DNMT3B,GNAS,DNMT3L

Supplemental Table S3. Predicted functional domains in epigenetic regulators

group	Gene names
<b>Bromodomain-containing proteins</b>	ASH1L, ATAD2, ATAD2B, BAZ1A, BAZ1B, BAZ2A, BAZ2B, BPTF, BRD1, BRD2.1, BRD2.2, BRD3.1, BRD3.2, BRD4.1, BRD4.2, BRD7, BRD8.1, BRD8.2, BRD9, BRDT.1, BRDT.2, BRPF1, BRPF1, BRPF3, BRWD1.1, BRWD1.2, BRWD3.1, BRWD3.2, CECR2, CREBBP, EP300, KAT2A, KAT2B, KIAA2026, MLL, PBRM1.1, PBRM1.2, PBRM1.3, PBRM1.4, PBRM1.5, PBRM1.6, PHIP.1, PHIP.2, SMARCA2, SMARCA2, SMARCA4, SP100, SP110, SP110, SP140, SP140L, TAF1.1, TAF1.2, TAF1L.1, TAF1L.2, TRIM24, TRIM28, TRIM33, TRIM33, TRIM66, ZMYND11, ZMYND8
<b>CHROMOdomain-containing proteins</b>	ARID4A, ARID4B, CBX1, CBX2, CBX3, CBX4, CBX5, CBX6, CBX7, CBX8, CDY1, CDY1B, CDY2A, CDY2B, CDYL, CDYL2, CHD1.1, CHD1.2, CHD2.1, CHD2.2, CHD3.1, CHD3.2, CHD4.1, CHD4.2, CHD5.1, CHD5.2, CHD6.1, CHD6.2, CHD7.1, CHD7.2, CHD8.1, CHD8.2, CHD9.1, CHD9.2, KAT5, MORF4L1, MPHOSPH8, MSL3L1, MYST1, SMARCC1, SMARCC2, SUV39H1, SUV39H2
<b>HATdomain-containing proteins</b>	ATAT1, CLOCK, CREBBP, ELP3, EP300, GTF3C4, HAT1, KAT2A/GCN5L2, KAT2B/PCAF, KAT5/TIP60, MYST1, MYST2, MYST3, MYST4, NCOA1, NCOA3, TAF1, TAF1L
<b>HDAC,SIRTdomain-containing proteins</b>	HDAC1, HDAC10.1, HDAC10.2, HDAC10.1, HDAC10, HDAC11, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6.1, HDAC6.2, HDAC7,

	HDAC8, HDAC9, SIRT1, SIRT2, SIRT3, SIRT4, SIRT5, SIRT6, SIRT6, SIRT6, SIRT7
<b>KDMdomain-containing proteins</b>	JARID2, JHDM1D, JMJD1C, JMJD5, KDM1A, KDM1B, KDM1B, KDM2A, KDM2B, KDM3A, KDM3B, KDM4A, KDM4B, KDM4C, KDM4D, KDM4DL, KDM5A, KDM5B, KDM5C, KDM5D, KDM6A, KDM6B, MINA, NO66, PHF2, PHF8, UTY
<b>MBTdomain-containing proteins</b>	L3MBTL.1, L3MBTL.2, L3MBTL.3, L3MBTL2.1, L3MBTL2.2, L3MBTL2.3, L3MBTL2.4, L3MBTL3.1, L3MBTL3.2, L3MBTL3.3, L3MBTL4.1, L3MBTL4.2, L3MBTL4.3, MBTD1.1, MBTD1.2, MBTD1.3, MBTD1.4, SCMH1.1, SCMH1.2, SCML2.1, SCML2.2, SFMBT1.1, SFMBT1.2, SFMBT1.3, SFMBT1.4, SFMBT2.1, SFMBT2.2, SFMBT2.3, SFMBT2.4
<b>PHDdomain-containing proteins</b>	AIRE.1, AIRE.2, ASH1L, ASXL1, ASXL2, ASXL3, BAZ1A, BAZ1B, BAZ2A, BAZ2B, BPTF.1, BPTF.2, BRD1.1, BRD1.2, BRPF1.1, BRPF1.2, BRPF3.1, BRPF3.2, CHD3.1, CHD3.2, CHD4.1, CHD4.2, CHD5.1, CHD5.2, CXXC1, DIDO1, DNMT3A, DPF1.1, DPF1.2, DPF1.2, DPF2.1, DPF2.2, DPF3.1, DPF3.2, G2E3.1, G2E3.2, ING1, ING2, ING3, ING4, ING5, INTS12, JHDM1D, KDM2A, KDM2B, KDM4A.1, KDM4A.2, KDM4B.1, KDM4B.2, KDM4C.1, KDM4C.2, KDM5A.1, KDM5A.2, KDM5A.3, KDM5B.1, KDM5B.2, KDM5B.3, KDM5C.1, KDM5C.2, KDM5D.1, KDM5D.2, MLL.1, MLL.2, MLL.3, MLL.4, MLL2.1, MLL2.2, MLL2.3, MLL2.4, MLL2.5, MLL2.6, MLL2.7, MLL3.1, MLL3.2, MLL3.3, MLL3.4, MLL3.5, MLL3.6, MLL3.7, MLL3.8, MLL4.1, MLL4.2, MLL4.3, MLL4.4, MLL5, MLLT10.1, MLLT10.2, MLLT6.1, MLLT6.2, MTF2.1, MTF2.2, MYST3.1, MYST3.2, MYST4.1, MYST4.2, NSD1.1, NSD1.2, NSD1.3, NSD1.4, NSD1.5, PHF1.1, PHF1.2, PHF10.1, PHF10.2, PHF11, PHF12.1, PHF12.2, PHF12.1, PHF13, PHF14.1, PHF14.2, PHF14.3, PHF14.4, PHF15.1, PHF15.2, PHF16.1, PHF16.2, PHF17.1, PHF17.2, PHF19.1, PHF19.2, PHF2, PHF20, PHF20L1, PHF21A, PHF21B, PHF23, PHF3, PHF5A, PHF6.1, PHF6.2, PHF7.1, PHF7.2, PHF8, PHRF1, PYGO1, PYGO2, RAG2, RAI1, RPH3A, RSF1, SHPRH, SP100, SP110, SP140, SP140L, TAF3, TCF19, TCF20, TRIM24, TRIM28.1, TRIM28.2, TRIM33.2, TRIM66.1, TRIM66.2, UBR7, UHRF1, UHRF2, WHSC1.1, WHSC1.2, WHSC1.3, WHSC1.4, WHSC1L1.1, WHSC1L1.2, WHSC1L1.3, WHSC1L1.4, WHSC1L1.5, ZMYND11, ZMYND8
<b>PMTdomain-containing proteins</b>	ASH1L, CARM1, DOT1L, EHMT1, EHMT2, EZH1, EZH2, EZH2, MDS1, MLL, MLL2, MLL3, MLL4, MLL5, NSD1, PRDM1, PRDM10, PRDM11, PRDM12, PRDM13, PRDM14, PRDM15, PRDM16, PRDM2, PRDM4, PRDM5, PRDM6, PRDM7, PRDM8, PRDM9, PRMT1, PRMT2, PRMT3, PRMT5, PRMT6, PRMT7.1, PRMT7.2, PRMT8, SETD1A, SETD1B, SETD2, SETD3, SETD4, SETD5, SETD6, SETD6, SETD7, SETD8, SETDB1, SETDB2, SETMAR, SMYD1, SMYD2, SMYD3, SMYD4, SMYD5, SUV39H1, SUV39H2, SUV420H1, SUV420H2, WHSC1, WHSC1L1
<b>PWWPdomain-containing proteins</b>	BRD1, BRPF1, BRPF3, DNMT3A, DNMT3B, GLYR1, GLYR1, HDGF, HDGFL1, HDGFRP2, HDGFRP3, MBD5, MSH6, MUM1, NSD1.1, NSD1.2, PSIP1, PWWP2B, WHSC1.1, WHSC1.2, WHSC1L1.1, WHSC1L1.2, ZCWPW1, ZCWPW2, ZMYND11, ZMYND8
<b>TUDORdomain-containing proteins</b>	AKAP1, ARID4A, ARID4B, CCDC101.1, CCDC101.2, FMR1.1, FMR1.2, FXR1.1, FXR1.2, FXR2.1, FXR2.2, JMJD2A, JMJD2B,

JMJD2C, LBR, MTF2, PHF1, PHF19, PHF20.1, PHF20.2, PHF20L1.1, PHF20L1.2, RNF17.1, RNF17.2, RNF17.3, RNF17.4, RNF17.5, SETDB1.1, SETDB1.2, SETDB1.3, SMN1, SMN2, SMNDC1, SND1, STK31, TDRD1.1, TDRD1.2, TDRD1.3, TDRD1.4, TDRD10, TDRD12.1, TDRD12.2, TDRD3, TDRD5, TDRD6.1, TDRD6.2, TDRD6.3, TDRD6.4, TDRD6.5, TDRD6.6, TDRD6.7, TDRD7.1, TDRD7.2, TDRD7.3, TDRD9, TDRKH, TP53BP1.1, TP53BP1.2, UHRF1.1, UHRF1.2, UHRF2.1, UHRF2.2, ZGPAT, LOC100129278.1, LOC100129278.2, LOC100129278.3, LOC100129278.4, LOC100129278.5, LOC100129278.6, LOC100129278.7, LOC100129278.8



Supplemental Table S5. Epigenetic regulators with identified HOTPOST mutations

Group epigenetic regulators	Gene name	# mutations	# HOTSPOTS	max# mutations in HOTSPOT	HOTSPOT aa.coordinates
chromatin remodeling	SENP7	71	5	3	S540, S638, S639, S671, S704
	CDKN2A	67	9	6	E120, E69, Q50, R29, R58, R7, R80, W110, W59
	PBRM1	213	10	5	R1009, R1010, R1025, R1153, R1160, R1184, R1185, R1200, R968, R978
	TP63	57	7	5	M1, R415, R496, R500, R509, R590, R594
	RSF1	67	5	4	R592, R645, R813, R844, R97
	FSHR	56	1	3	V166
	KANSL2	15	4	3	E154, E406, E440, E623
	SMARCA1	41	4	3	R198, R233, R334, R54
	NAP1L4	11	2	3	R13, R44
	SFRP1	14	1	8	AA13
	SMC4	55	3	6	R1194, R1227, R1252
	ASXL2	106	8	5	R24, R256, R284, R329, R357, R97, SS117, SS89
	VHL	48	7	5	E148, E189, Q73, Q96, S65, S68, W88
	HMG20A	13	1	4	R98
	MGEA5	18	2	3	R533, R586
	NAP1L2	19	2	3	ED221, ED79
ARID1A	389	26	12	Q145, Q479, Q528, Q96, R1054, R1335, R1339, R1505, R1606, R1722, R1772, R1775, R1941, R1989, R2158, R231, R310, R317, R367, R486, R50, R618, R693, R750, R885, R952	
ARID1B	164	1	3	Q1307	
ARID2	174	3	3	Q571, R125, R274	
ASXL1	55	2	3	Q507, Q512	
DNA modifications	MAEL	21	2	3	R236, R267
histone modification	PARG	9	1	6	A99
	PPM1D	14	1	4	E525
	PPP2R5B	23	2	4	R241, R268
	ANKRD28	18	2	3	R25, R58
	BAP1	28	1	3	M1
	HDAC9	207	1	3	R365
	USP36	15	2	3	KKK563, KKK958
	HIST1H2AA	7	1	3	T102
	TBP	91	4	37	Q40H, Q43H, Q60H, Q63H
	EP400	190	5	8	Q2646H, Q2690H, Q2726H, Q2727H, Q2763H
	MGA	129	5	6	R1043, R1242, R2187, R2357, R2396
	BRWD1	18	1	5	R969
	SIN3A	22	1	4	R1104
	ADNP	9	1	3	R306
	ATXN7	10	1	3	Q155
	CHD1	15	1	3	M1
	CHD4	76	3	3	R340, R344, R347
	FAM208A	25	2	3	E287, E683
	PHF12	11	2	3	R218, R731
	PPIG	42	2	3	R561, R576
	ZMYM4	34	3	3	Q368, Q620, Q709
	PIK3CA	30	1	7	E545
	SMAD2	28	2	7	S434, S464
	CTCF	52	2	6	R120, R448
	TARBP2	30	5	4	E100, E161, E170, E191, E92
	BCOR	81	7	3	R1462, R1480, R1514, R187, R208, R357, R384
	NOS1	65	2	3	W1143, W1177
	TDRD5	36	1	3	R518
	ATM	57	2	4	R250, R2993
	PRDM9	86	1	4	H826
	ASH1L	43	1	3	E77
	BRCA2	47	1	3	Q2870
	CDYL	46	3	3	E164, E218, E32
	LIMK2	27	4	3	R118, R286, R343, R364
	SETD2	69	2	3	R1278, R1322
	SUV420H1	27	1	3	P583
	PHF2	24	2	9	S2165PPLP, S985SPPLP
MYOCD	85	5	4	R288, R384, R89, W238, W533	
WHSC1L1	34	3	3	R1261, R1299, R1310	

Supplemental Table S6. PANCAN deleted/amplified epigenetic regulators

Group	Subgroup	gene name	# affected cancers
DNA modifications	Writers	DNMT3A	8
		DNMT1	7
	editors	TET1	8
		TET2	7
Histone modification	Writers	MLL3	9
		OGT	9
		ATM	9
		CREBBP	8
		MLL2	8
		SETD2	8
		ATR	8
		BRCA2	8
		NCOA2	7
		NCOA3	7
		ASH1L	7
		CARM1	7
		MLL	7
		SETD5	7
		BAZ1B	7
		PRKCB	7
		RPS6KA3	7
		HUWE1	7
		MDM2	7
		UBA1	7
	MGA	9	
	Readers	CHD4	8
		CHD5	8
		TAF1	8
		ZNF462	8
		ADNP	7
		ATXN7	7
		CHD1	7
		CHD3	7
		CHD8	7
		DHX30	7
		EP400	7
		GABRG1	7
		HCFC1	7
		NIPBL	7
		POGZ	7
		RAI1	7
		SMC1A	7
		SMCHD1	7
		TAF6	7
		TRIM33	7
		TRRAP	7
		ZMYM3	7
		ZMYND8	7
	ZNF687	7	
	Editors	KDM6A	9
		HDAC6	8
		HDAC9	8
		KDM3B	8
		HDAC4	7
		KDM2A	7
		KDM5A	7
		PHF8	7
EYA4		7	
PPP6R3		7	
Chromatin remodeling	nucleosome remodeling factors	USP36	7
		BPTF	8
		HLTF	7
	Chromatin remodeling helicase	NFRKB	7
		ATRX	9
		SMARCA2	8
		CHD2	7
		CHD7	7
		SMARCA1	7
	chromatin remodeling factors	SRCAP	7
		ERCC6	7
		MYB	8
		SMARCC2	8
		SUPT5H	8
		PBRM1	7
RB1	7		

## Supplemental Figures

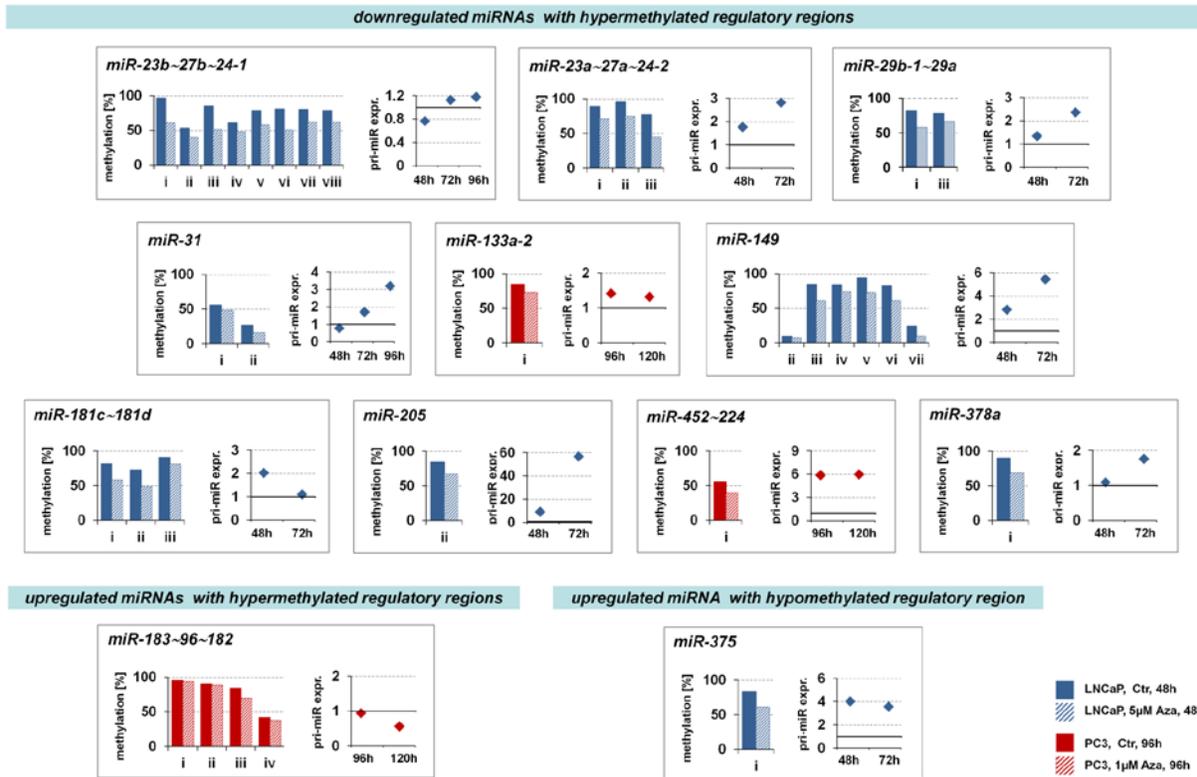
**A**

	Downregulated	Downregulated & Deleted	Downregulated & Hypermethylated	Downregulated & Deleted & Hypermethylated	All downregulated
EOPC-005	139	12	20	0	171
EOPC-006	189	15	30	2	236
EOPC-007	181	9	37	1	228
EOPC-008	122	3	20	2	147
EOPC-009	174	8	32	0	214
EOPC-010	185	4	41	0	230
EOPC-011	187	7	35	0	229
EOPC-017	148	0	28	0	176
EOPC-019	229	3	26	3	261
EOPC-021	138	11	18	3	170
EOPC-024	179	0	16	0	195
EOPC-025	107	5	26	1	139
EOPC-026	188	0	14	0	202
in any patient	246	85	120	41	492

**B**

	Upregulated	Upregulated & Hypomethylated	All upregulated
EOPC-005	277	6	283
EOPC-006	164	7	171
EOPC-007	202	4	206
EOPC-008	303	29	332
EOPC-009	203	2	205
EOPC-010	152	8	160
EOPC-011	194	8	202
EOPC-017	178	18	196
EOPC-019	135	7	142
EOPC-021	242	13	255
EOPC-024	210	20	230
EOPC-025	212	35	247
EOPC-026	207	15	222
in any patient	445	133	578

**Supplemental Figure S1: Genome-wide genetic and epigenetic miRNA deregulation in 13 early-onset PCA cases (ICGC EOPC dataset).** (A) For each single patient, the numbers of downregulated miRNAs without any detected genetic or epigenetic aberration and downregulated miRNAs whose precursors are deleted, whose promoters are significantly hypermethylated or that exhibit both aberrations in exactly the same tumor are shown. Only miRNAs with mapped regulatory regions were considered. (B) For each single patient, the numbers of upregulated miRNAs without any detected genetic or epigenetic aberration and upregulated miRNAs whose promoters are significantly hypomethylated in exactly the same tumor are shown. Only miRNAs with mapped regulatory regions were considered.



**Supplemental Figure S2: Pri-miR expression and methylation levels of the corresponding differentially methylated regions (DMRs) upon 5-aza-2'-deoxycytidine treatment of PCA cells.** Results are shown for the 12 miRNA clusters that were upregulated (*miR-183-96~182* downregulated) accompanied by demethylation of their DMRs upon treatment. Optimal treatment concentrations and time periods were determined by analysis of LINE-1 methylation levels. Pri-miR expression was determined by quantitative PCR, methylation levels were quantified by MassARRAY. Methylation levels are shown for each DMR analyzed (i, ii, ...) after 48h (LNCaP, blue) or 96h (PC3, red) of treatment. Expression levels are normalized to the PBS control (Ctr) and are shown for the respective treatment time (48 or 96h) and onwards. Bold solid lines represent no expression change compared to the PBS-treated control.

	mature miRNA	EO-PCA		Elderly-onset PCA	Methylation analysis (Mass ARRAY)		
		Upreg. (N=7)	fold-change	fold-change	promoter region	Methylation	correlation with miRNA expression (R, p)
chr.13 cluster	hsa-miR-17-5p	7	4.0	1.4	chr13:91995949-92000723	n.s.	n.s.
	hsa-miR-19a-3p	6	2.6	1.6			
	hsa-miR-19b-3p	6	2.7	1.5			
	hsa-miR-20a-5p	7	5.2	1.5			
	hsa-miR-92a-3p	5	2.3	n.s.			
chr.7 cluster	hsa-miR-106b-5p	7	3.3	1.5	chr7:99695817-99701753	hypo	-0.543, <0.001
	hsa-miR-93-5p	7	5.2	1.4			-0.682, <0.001
	hsa-miR-25-3p	7	3.3	1.6			-0.625, <0.001
chr.12 cluster	hsa-miR-141-3p	6	2.0	1.6	chr12:7057258-7074488	hypo	-0.678, <0.001
chr.1 cluster	hsa-miR-214-3p	3	1.5	-1.2	not analyzed	not analyzed	
chr.14 cluster	hsa-miR-494	4	1.5	1.3	not analyzed	not analyzed	
chr.X cluster	hsa-miR-221-3p	3	1.1	-2.2	chrX:45610594-45611811	not analyzed	
	hsa-miR-222-3p	1	-2.0	-2.2			
	hsa-miR-21-5p	5	1.9	n.s.	chr17:57912817-57921277	n.s.	n.s.
	hsa-miR-22-3p	1	-1.3	-1.3		not analyzed	
	hsa-miR-26b-5p	4	1.1	1.2		not analyzed	

**Supplemental Figure S3: Genome-wide genetic and epigenetic miRNA deregulation in 13 early-onset PCA cases (ICGC EO-PCA dataset).** (A) For each single patient, the numbers of downregulated miRNAs without any detected genetic or epigenetic aberration and downregulated miRNAs whose precursors are deleted, whose promoters are significantly hypermethylated or that exhibit both aberrations in exactly the same tumor are shown. Only miRNAs with mapped regulatory regions were considered. (B) For each single patient, the numbers of upregulated miRNAs without any detected genetic or epigenetic aberration and upregulated miRNAs whose promoters are significantly hypomethylated in exactly the same tumor are shown. Only miRNAs with mapped regulatory regions were considered.



## PUBLICATIONS DURING THE THESIS

### International Conferences

#### Peer reviewed journals

Brocks D., Assenov Y., Minner S., **Bogatyrova O.**, ....Plass. C, *Intratumor DNA Methylation Heterogeneity Reflects Clonal Evolution in Aggressive Prostate Cancer*, **Cell Reports**, 2014

Oakes C., .... **Bogatyrova O.**, ....Plass. C, *Evolution of DNA methylation is linked to genetic aberrations in chronic lymphocytic leukemia*, **Cancer Discovery**, 2013

Plass C.,... **Bogatyrova O.**, *Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer*  
**Nat Genet review**, 2013.

Kostareli E, Holzinger D, **Bogatyrova O.**, .., Hess J. *HPV-related methylation signature correlates with survival in oropharyngeal squamous cell carcinomas*,  
**J Clin Invest.** 2013 May 1. doi:pri: 67010. 10.1172/JCI67010

Goepfert B, Konermann C, Schmidt CR, **Bogatyrova O.**,..., Weichenhan D. *Global alterations of DNA methylation in cholangiocarcinoma targets the Wnt signaling pathway*  
**Hepatology**, 2013

Weischenfeldt, J.,...**Bogatyrova O.**, 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.

### International Conferences

International PhD cancer conference, Manchester, June, 2015

Gordon Conference, Toscana, Italy, Apr.2015

The 5<sup>th</sup> Wiessenburg Symposium Biriciana, 15-17 Sep 2014, Wiessenburg, Germany

Integrative and computational Biology Symposium, 27-28 Mar. 2014, Barcelona, Spain

2nd Heidelberg Forum for Young Life Scientists, DKFZ Heidelberg, 2/2012

38<sup>th</sup> FEBS Congress, 6-11 July 2013, St.Peterburg, Russia

EMBL 9th Transcription and Chromatin conference, 2013, EMBL, Heidelberg, Germany

7<sup>th</sup> ICGC Annual meeting in Heidelberg, Germany, Dec, 2013

5<sup>th</sup> Annual meeting of NGFN-Plus and NGFN-transfer, 11-13 Dec, 2012, Heidelberg, Germany

6<sup>th</sup> Scientific workshop of ICGC, 20-22 Mar. 2012, Cannes, France



## ACKNOWLEDGMENTS

I would like to thank the many people in the lab that made the completion of this thesis not only (scientifically possible), but also a great time of my life!

- Prof. Dr. Christoph Plass, for giving me the chance to pursue this PhD project in your group, the constant scientific and personal support you supplied, for organizing our regular “Altenbach Retreats” and parties, for making it possible for me to present my data on several international conferences, correcting this thesis, the external cooperation’s you set up, and, and, and.....
- PD Dr. Dieter Weichenhan, for helping me and teaching me all lab work, during my first year of PhD, for maintaining the lab organization in a way that efficient working was possible, and for giving an important technical feedback.
- Prof. Dr. Stefan Wiemann for very useful and important comments and feedback during all my work and especially during my TAC meetings
- Prof. Dr. Benedikt Brors for incredibly productive collaboration and lead of PANCAN project
- Prof. Dr. Peter Lichter for important help during my last years and giving feedback on my work
- Prof. Dr. Holger Sültmann for productive collaboration and work on miRNA project
- Dr. Yassen Assenov, for being my tutor in learning bioinformatics and being my friend in real life
- Dr. Daniella Wuttig for productive collaboration and work on miRNA project and great discussion out of lab
- Marion Bähr and Oliver Mücke, Jana Peterson for helping me out during all experiments, especially Marion for her support with samples.
- Dr. Manuela Zucknick for anytime biostatistical support during all my work
- Cholangiocarcinoma team: Lea and Christopher for learning how to be a productive team and how to deal with problems
- Melanie Weiss for great work on lncRNA project
- All the members of the C010 group that have not yet been mentioned, for great working team!!

In addition to the people who helped me on the scientific level, I also want to acknowledge important non-scientific contributors to this thesis:

- To Dr. Marina Laplana for all her help in work and real life...for all her support....for being my best ... friend!!!
- The H2.03.072 office members, David, Yassen and Michael, Dr. Schmidtko for great working atmosphere!!
- The Lab members that became personal friends and made the lab a place I always enjoyed coming to: Christian, Cornelia, Kathi, Miri, Maria, Melanie, Tania, Mridul, Chris....
- Last but not least, I want to thank to my family for all the past and future support and always believing in me!