Ieva Rauluševičiūtė

Computational Analysis of DNA Methylation and Gene Expression Patterns in Prostate Cancer

Master's thesis in Molecular Medicine

Trondheim, June 2018

Supervisors:
dr. Morten Beck Rye and prof. Finn Drabløs

Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

NTNU
Norwegian University of Science and Technology

ABSTRACT

DNA methylation is an important contributor for prostate cancer development and progression. It has been studied experimentally for years, but, lately, high-throughput technologies are able to produce genome-wide DNA methylation data that can be analyzed using various computational approaches. Thus, this study aims to bioinformatically investigate different DNA methylation and gene expression patterns in prostate cancer. DNA methylation data from three datasets (TCGA, *Absher* and *Kirby*) was correlated with gene expression data in order to distinguish different regulation patterns. Classically, increased DNA methylation in promoter regions is being associated with gene expression downregulation. Although, results of the present project demonstrate another robust pattern, where DNA hypermethylation in promoter regions of 1,058 common genes is accompanied by upregulated expression.

After analyzing expression and methylation values in the same samples from TCGA dataset, expression overcompensation in a dataset as an explanation for upregulation was excluded. Further reasons behind the pattern were investigated using TCGA DNA methylation data with extended list of probes and includes the presence of methylated positions in CpG islands, distance to transcription start sites and alternative TSSs. As compared with the downregulated genes in the classical pattern, upregulated genes were shown to have more positions in CpG islands and closer to TSSs. Moreover, the presence of alternative TSS in prostate was demonstrated, also disclosing the limitations of methylation detection systems.

In conclusion, indications from a present study suggest a possible DNA methylation role in gene expression upregulation in prostate cancer, which prompts to reconsider the classical believe that DNA methylation always leads to gene suppression.

ACKNOWLEDGEMENTS

This thesis is a part of MSc in Molecular Medicine study program in Faculty of Medicine and Health Sciences (MH) at Norwegian University of Science and Technology (NTNU) in Trondheim, Norway.

I would like to thank my supervisor dr. Morten Beck Rye for an opportunity to work with him, for his continuous advices and expertise, incredible patience and support throughout this project. Thank You very much for all the lessons and fruitful discussions.

I am also grateful to prof. Finn Drabløs for allowing me to join his research group in the Department of Clinical and Molecular Medicine. Thank You for the guidance and many suggestions.

Finally, my education would not be possible without the support of my family, which I am thankful for.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

27k	Infinium HumanMethylation27 BeadChip
3'UTR	3'-untranslated region
5'UTR	5'-untranslated region
5mC	5'-methylcytosine
AR	Androgen receptor
BCR	Biochemical recurrence
bp	Base pair
BPH	Benign prostatic hyperplasia
BS-seq	Bisulphite sequencing
c	Combined enrichment score
CGI	CpG island
ChIP-seq	Chromatin immunoprecipitation sequencing
CIN	Chromosomal instability
CRPC	Castration-resistant prostate cancer
DBTSS	DataBase of Transcriptional Start Sites
DNase	Deoxyribonuclease
DNMT	DNA methyltransferase
DOWNDOWN	'Loss of methylation — downregulated expression' regulation pattern
DOWNUP	'Loss of methylation — upregulated expression' regulation pattern
DRE	Digital rectal examination
DTCs	Disseminated tumor cells
ES	Enrichment score
FDR	False discovery rate
GEO	Gene Expression Omnibus
GO	Gene Ontology
GSEA	Gene set enrichment analysis
HERVs	Human endogenous retroviruses
HM450	Illumina Infinium HumanMethylation450 BeadChip
kb	Kilobase
LINE	Long interspersed nuclear element
miRNA	Micro RNA

NES	Normalized enrichment score
NTX	N-telopeptide of collagen
PCa	Prostate cancer
PcG	Polycomb Group
PIN	Prostatic intraepithelial neoplasia
PRC1	Polycomb repressive complex 1
PRC2	Polycomb repressive complex 2
PRC4	Polycomb repressive complex 4
PRCs	Polycomb repressive complexes
PSA	Prostate-specific antigen
RARP	Robot-assisted radical prostatectomy
RP	Radical prostatectomy
SAM	S-adenosyl-L-methionine
SINE	Short interspersed nuclear element
SNV	Single nucleotide variation
TCGA	The Cancer Genome Atlas
TRUS	Transrectal ultrasound
TSG	Tumor suppressor gene
TSS	Transcription start site
UPDOWN	‘Gain of methylation — downregulated expression’ regulation pattern
UPUP	‘Gain of methylation — upregulated expression’ regulation pattern

1. INTRODUCTION

Prostate cancer (PCa) is the second most common type of cancer worldwide and the most common in Norway [1, 2]. Survival rates of the local tumor are known to be high, but aggressive forms are especially dangerous and difficult to treat. PCa is very heterogenic and there is much to be discovered. Research of DNA methylation and gene expression is an upcoming field since it can provide new insights on a development and progression of the cancer.

DNA methylation is often described as an epigenetic switch for turning gene expression on or off — when expression is needed, methylation marks are removed, and when a gene needs to be suppressed, DNA is methylated. Computational methods allow to perform a genome-wide analysis of gene expression and many methylation positions, which gives an overview of DNA methylation and gene expression status, possible interactions between two processes in PCa.

Aim of the research project — using computational methods investigate the relationship between the gain of methylation and upregulated expression in prostate cancer and compare with other DNA methylation and gene expression patterns.

Objectives:

1. Analyze available literature on DNA methylation in prostate cancer and find available DNA methylation datasets in prostate cancer and normal prostate tissue.
2. Process and prepare the data, calculate statistical differences in prostate cancer, compared with normal tissue samples. Distinguish DNA methylation and gene expression patterns and analyze genes that consistently follow the patterns in all datasets.
3. Compare DNA methylation and gene expression values derived from the same samples for genes following ‘Gain of methylation — upregulated expression’ regulation pattern in order to confirm or deny overcompensation of expression due to DNA methylation.
4. Perform detailed structural and functional analysis of methylation in genes following only ‘Gain of methylation — upregulated expression’ regulation pattern and compare the findings with genes from ‘Gain of methylation — downregulated expression’ regulation pattern only.

1.1. Prostate cancer

1.1.1. Prostate cancer problem and epidemiology

According to the most recent worldwide statistics in 2012, there were 1.1 million new cases of prostate cancer (PCa) estimated (15% of all new cancer cases in men worldwide) and 307,000 deaths registered (6.6% of all deaths of cancer in men) [1]. Prostate cancer is the second most common cancer type among men worldwide after lung cancer, but PCa is detected more frequently in developed countries and is the most common type in men [1]. The occurrence rates are higher in Northern and Western Europe, as well as Australia, New Zealand and Northern America [1]. Worldwide, PCa is the 5th deadliest cancer among men [1].

Statistical data is more recent for the population of United States. Prostate cancer is the most common type of cancer among men here (Figure 1) [3-5]. It accounts for almost 1 in 5 new cancer cases diagnosed [5]. PCa rates increases with age — 1 in 12 men older than 70 will get prostate cancer, while 1 in only 403 men younger than 50 will get prostate cancer [5]. In average 1 in 9 men develop prostate cancer in their lifetime [5].

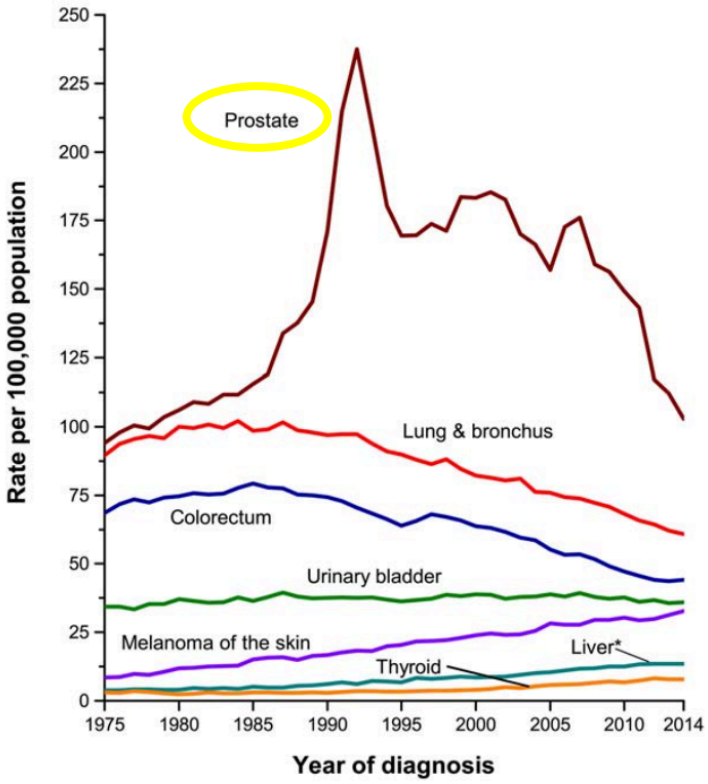


Figure 1 **Incidence rates for different cancer types in men in United States (US) from years 1975 to 2014.** The most common type of cancer in men is prostate cancer. The peak of prostate cancer rate in early 1990s was caused by the widespread testing of prostate-specific antigen (PSA). Adapted from [5].

In United States 180,890 new prostate cancer cases (21.5% of all cancers in men) were estimated for 2016, while for 2017 and 2018 161,360 (19.3% of all cancers in men) and 164,690 (19.2% of all PCa cases estimated) new cases respectively [3-5]. Despite the drop in a number of new estimated cases in 2017 and 2018 (compared with 2016) the death rate increases. In 2016, 26,120 deaths estimated (8.3% of all male deaths of cancer estimated), while in 2017 and 2018 — 26,730 (8.5% of all male deaths of cancer estimated) and 29,430 (9.1% of all male deaths of cancer estimated) deaths respectively [3, 4]. Prostate cancer remains the second (after lung and bronchus) deadliest cancer in men [3-5].

In Norway, 5,118 new cases of prostate cancer were registered in 2016. It is also the most common type of cancer (nearly 30% of all cancers in men in 2016) [2]. In 2016, 957 men died from PCa in Norway [2].

Fortunately, prostate cancer survival rate is highest of all cancer types, but sometimes survival statistics are difficult to interpret because of the differences in detection practice [5]. As mentioned before, prostate cancer is more common in developed countries mostly because of the PSA testing [1]. In the late 1980s and early 1990s prostate cancer detection rates increased because of the widespread prostate-specific antigen (PSA) testing [5]. PSA test became the main method for early diagnostics and cancer monitoring. The procedure is low invasiveness, low cost and can be repeated many times, but, unfortunately, PSA testing can also lead to overdiagnosis and overtreatment [6]. Because of that US Preventive Services Task Force recommended to reduce the testing in men older than 75 [5, 6]. General screening using PSA test is also not recommended in Norway, but it still remains the main test that is performed before scheduling the biopsy [2].

Problems of the PSA testing prompt scientists to look for other promising methods to diagnose and monitor prostate cancer. DNA methylation in cancer is one such potential area of research.

Other reason why prostate cancer research is important today is that even if survival rates of localized cancer are high, aggressive forms of prostate cancer are still highly dangerous, spread fast, are difficult to predict and thus difficult to treat. Furthermore, because of the high heterogeneity PCa, it complicates the studies of this type of cancer and unify the findings [7]. DNA methylation studies in prostate cancer could potentially reveal the important changes that would help to differentiate aggressive disease from localized cancer and shed the light on gene regulation patterns.

1.1.2. Prostate gland and prostate cancer development

Prostate is an exocrine gland, part of the male reproductive system, main function of which is to produce semen fluid [8, 9]. The gland can be divided into three zones: central, peripheral and transitional zones. The tissue of prostate consists of fibromuscular stroma and epithelial glands, that is a start point for the development of prostate adenocarcinoma [8, 9]. Glandular epithelium has three different cell types: neuroendocrine, luminal secretory and basal [8]. Luminal cells are responsible for androgen receptor (AR) and prostate-specific antigen (PSA) secretion. Expression of AR and PSA is regulated by androgens [8]. The fibromuscular stroma is made up of fibroblasts, endothelial cells, dendritic cells, nerves, smooth muscle cells as well as mast cells and lymphocytes [8]. Stroma cells are responsible for androgen-dependent production of growth factors [8]. Particularly interesting is a stromal-epithelial crosstalk in prostate. This event is an important factor for regulation of growth and development of the prostate and hormonal responses in a gland [8, 10]. It means that stroma also plays an important role in cancer development, despite the fact that cancer evolves from epithelial cells. Alterations of paracrine interactions in a tumor microenvironment — stroma — creates a microenvironment for tumor growth [9, 10].

Changes in prostate gland epithelium cells usually leads to a formation of pre-cancerous phase, known as prostatic intraepithelial neoplasia (PIN). In PIN benign prostatic acini and ducts are lined by cytological atypical cells [11]. PIN is being considered as a precursor for cancer and high-grade PIN is associated with prognosis of aggressive PCa [11]. Proliferative inflammatory atrophy (PIA) is also a contributor to high-grade PIN formation and/or prostate carcinogenesis (Figure 2) [11]. PIA contains highly proliferative, but failed-to-differentiate prostate epithelial cells in periphery of the prostate — hotspot for PCa [12].

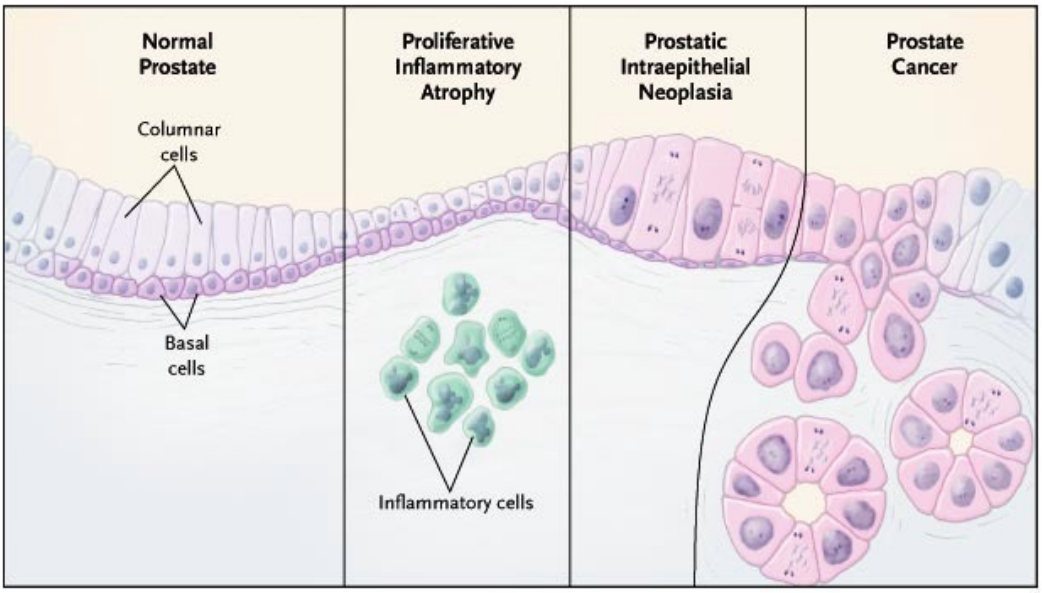


Figure 2 **From normal prostate to prostate cancer.** Proliferative inflammatory atrophy (PIA) and prostatic intraepithelial neoplasia (PIN) are precursory lesions to prostate cancer. Adapted from [12].

Prostate adenocarcinoma more frequently develops in the peripheral zone of the gland (80-85%). Transitional zone, together with central zone, are less common sites of origin for PCa (10-15% and 5-10%, respectively) [13].

Benign prostatic hyperplasia (BPH) is a very common non-malignant prostate lesion that may lead to increased chances to develop prostate cancer, but today this statement is debatable due to the differences in histology and localization [9, 14]. Prostate size increases from an approximate normal size of 30 g to larger than 150 g in older men [9]. BPH creates unpleasant symptoms, such as urinary incontinence and reduces the quality of life, but it is not lethal [14]. There are some overlapping features between PCa and BPH: hormone-independent growth and response to antiandrogen therapy (a type of drug-based androgen-deprivation therapy) [14]. Furthermore, shared risk factors for both diseases are chronic inflammation, genetic variation and metabolic disruption [14]. In conclusion, PCa and BPH share common features rather than BPH causes PCa (Figure 3).

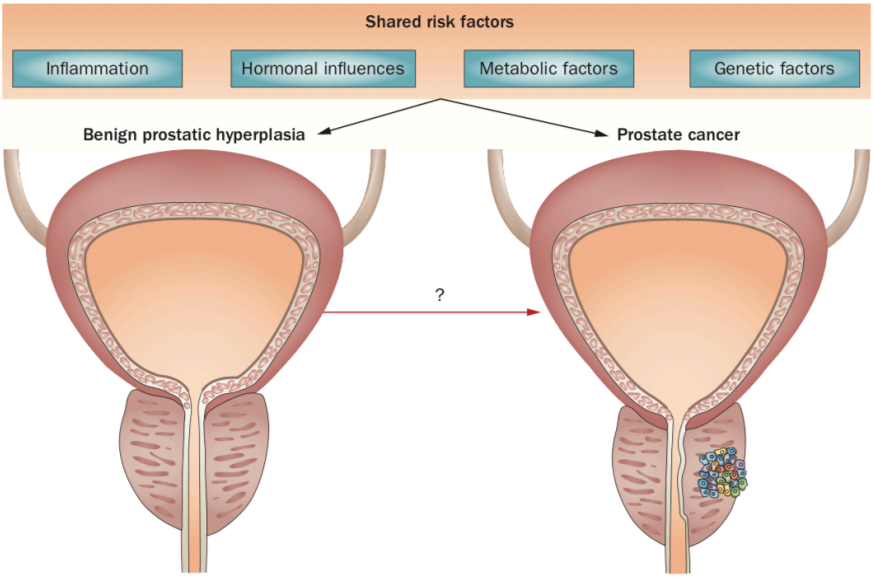


Figure 3 **Benign prostatic hyperplasia (BPH) and prostate cancer share risk factors, but whether BPH is a premalignant condition is still unclear.** Inflammation, hormonal influences, metabolic and genetic factors are shared features in BPH and prostate cancer. Adapted from [14].

Aggressive forms: metastatic and castration-resistant prostate cancer

Prostate cancer might be the most common type of cancer in men, as mentioned before, but, fortunately, not the deadliest. Localized and not progressing PCa does not require a treatment at many times, but aggressive forms, such as metastatic and castration-resistant prostate cancer (CRPC), are especially hard to treat. It is critical to develop procedures for distinguishing advanced and aggressive prostate cancer from indolent cancer to be able to prescribe the best treatment strategy [15]. Various genetic and epigenetic changes are responsible for development of aggressive PCa forms (Figure 4). Unfortunately, because of the wide variety of molecular changes, PCa differs from one individual to another and even in different lesions of the same case [12].

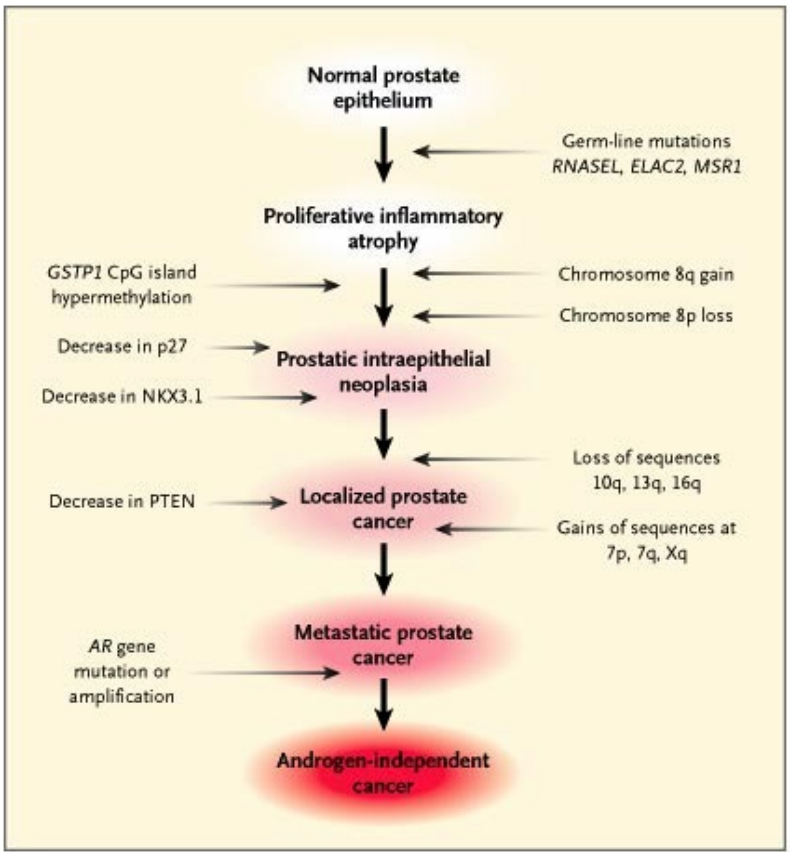


Figure 4 **Molecular pathogenesis of prostate cancer.** Somatic mutations, gene deletions and amplifications, chromosomal rearrangements, all together with aberrant DNA methylation and other epigenetic alterations are involved in cancer development and progression. Adapted from [12].

First site of spreading of prostate cancer is bones (osteoblastic metastases, causing the overproduction of bone cells) and it is often the only clinically detectable site of metastasis for PCa [13, 16]. PCa patients, as well as breast cancer patients, are the most severely affected by bone metastasis [17]. Reasons why PCa metastasizes to the bones and mechanisms behind osteoblastic metastases are not well known in comparison with another type of bone metastases, called osteolytic metastases. Higher Gleason score, a loss of *PTEN* gene, gain of mitogen-activated protein kinase (RAS/MAPK) signaling, increased levels of N-telopeptide of collagen (NTX) are believed to contribute to osteoblastic disease [17]. Process, such as reverse hematopoiesis, where red marrow tissue collects disseminated tumor cells (DTCs), occurs and makes bone metastases uniquely dangerous [17].

Metastases of any organ are dangerous, but bone metastases are exceptional. They are life-threatening and, unfortunately, in most cases incurable, frequently followed by hardly manageable pain [17, 18]. Challenging progressive prostate cancer forms encourage scientists to study early PCa and investigate molecular changes could lead to metastatic or castrate-resistant prostate cancer.

Androgen receptor (AR)

Androgen receptor (AR) is extremely important for a healthy prostate development and functioning, but alterations in a pathway of androgens and AR contribute to cancer initiation and progression, as well as CRPC [8, 19]. In a healthy cell, testosterone triggers the ligand-dependent transactivation of AR, leading to cell proliferation, differentiation and survival [20]. Expression activation pathway for genes, responsible for those functions, is described in detail in Figure 5.

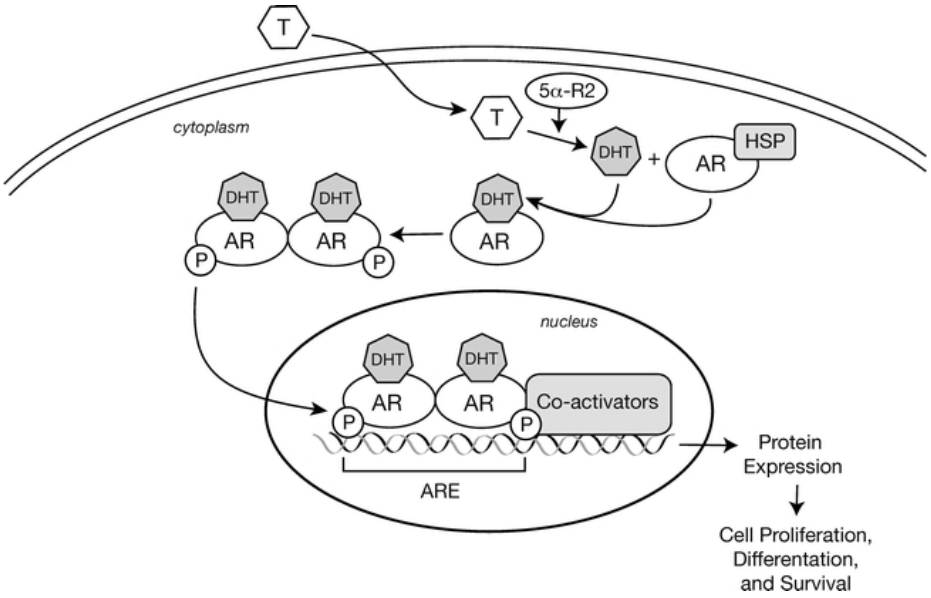


Figure 5 **Ligand-dependent gene expression activation by androgen receptor.** Testosterone (T) enters prostate gland epithelium cell and is converted to dihydrotestosterone (DHT) by 5α-reductase (5α-R2). Heat shock proteins dissociate from androgen receptor (AR) and DHT bind to AR, which leads to dimer formation and interaction with androgen response elements (ARE). Other co-activators are recruited, and gene expression is activated. Adapted from [20].

Continued transactivation of AR is a cause of CRPC, also known as androgen-independent prostate cancer (AIPC) [16, 19]. This phenotype of PCa is lethal and characterized by increasing tumor size, new metastatic spread and rising levels of prostate-specific antigen (PSA) [19]. Inappropriate activity of AR can be caused by AR amplification due to overexpression, gain-of-function AR mutations, overexpression of AR co-factors, ligand-independent AR activation by growth factors or cytokines, intracrine AR production or constitutively active AR splice variants [19, 20].

One of the ways to use defects in the AR pathway as an advantage is androgen-deprivation therapy (ADT), where for chemical castration luteinizing-hormone-releasing hormone (LRHR) agonist is used (drug-based ADT) [19]. Furthermore, ADT can also be achieved by medical or surgical castration (surgery-based ADT), which is a primary way to

manage progressive PCa as well as a therapy for androgen-dependent tumors [8, 16]. Unfortunately, many patients develop CRPC and fail the therapy, causing the death of the patient [8].

Genomic rearrangements

Chromosomal rearrangements are common events in prostate cancer development. In PCa, rearrangements result in gene fusions, amplifications and deletions [21]. The most well-defined fusion in PCa is *TMPRSS2-ERG*, which, according to Perner et al. occurs in almost 50% of PCa cases [21, 22]. During the fusion event 5'-UTR of androgen-driven *TMPRSS2* gene fuses with the oncogenic transcription factors from *ETS* TF family (*ERG*, *ETV1* or *ETV4*) [22]. Fusion leads to aberrant *ERG* gene expression and drives prostate tumorigenesis [22]. Amplification of oncogenes *MYC*, *AR* and *PIK3CA* is also a signature rearrangement for PCa, as well as *PTEN* tumor suppression gene deletion [21].

Genomic rearrangement can also occur in a much larger scale, causing chromosomal instability (CIN), where whole chromosomes or arms of chromosomes are being altered (Figure 4) [21].

Telomeres

As it is common in other types of cancer, prostate cancer also activates telomerase to maintain telomeres, but the difference in PCa is that telomeres are unusually short in cancer cells compared with normal prostate tissue (average length of 5.4 kb in cancer compared with 6.6 kb in normal tissue) [23]. Shortened telomeres can affect genome integrity, as well as drive aneuploidy and somatic copy number alterations [23]. Such alterations has been associated with deletions of 8p and amplification of 8q (Figure 4) [21]. Shortening of the telomeres can result in genomic alterations, such as chromothripsis — "multiple translocation events occurring in a single catastrophic event leading to imperfect rearrangement and repair of one or a few shattered chromosomes" [23]. Kovtun et al. observed chromotripsis in around 30% of different grade PCa samples with more incidents among the lower grade tumors, suggesting the role of chromotripsis in cancer initiation [24].

Diet

It was shown that a diet might also influence the development of prostate cancer. For example, lycopene and other carotenoids, vitamin E, selenium, marine omega-3 fatty acids, soy, polyphenols and isoflavones are considered as protective dietary elements, while milk, calcium, high doses of zinc, saturated fat, grilled meats and heterocyclic amines contribute to the increased risk of PCa [25]. Diet's role in the process of cancer development is still a controversial topic, partially because of the lack of research projects in this field.

1.1.3. Current methods for diagnosis of prostate cancer. Pros and cons of methods for early diagnosis

Prostate-specific antigen (PSA) test

PSA is a serine protease that helps to liquefy the seminal fluid by proteolyzing gel proteins of seminal fluid [26]. PSA is a part of a glandular kallikrein-related peptidases. Genes for all 15 kallikreins are clustered together in human genome (280 kb locus of chromosome 19q133-4) [26]. PSA is encoded by *KLK3* and transcription is regulated by androgens [26]. Most of the times, PSA concentration in blood increases significantly because of the prostate cancer, which makes the molecule a suitable biomarker for PCa detection [27]. But other conditions, such as BPH or prostatitis, as well as older age, can also lead to increased levels of PSA [6, 26]. It is important to know that PSA levels in blood escalate not due to a higher expression of PSA, but probably because of the disruption of the basement membrane and loss of basal cell layer, epithelial cell polarity and ductal lumen architecture [26].

PSA test is the most frequently used test for early diagnosis for PCa [28]. During testing, concentration of PSA in blood is measured, where it varies from <0.1 to 10^4 ng/ml with a median of around 0.6 ng/ml [26]. In men with advanced PCa, high concentrations (above 10^2 ng/ml) are observed [26]. 4 ng/ml is considered a threshold for prostate cancer detection [26, 28]. If PSA concentration increases after radical prostatectomy (RP) by 0.4 ng/ml and another increase follows, metastatic progression is suspected, and such condition is defined as a biochemical recurrence (BCR) of PCa [29].

PSA test is relatively non-invasive, cheap and easy to perform, so can be repeated many times. However, it lacks sensitivity to detect early carcinogenesis, causing false-negative and false-positive results, and PSA levels can, unfortunately, vary independently from prostate carcinogenesis — increase when there is no cancer and decrease or stay unchanged with the cancer in presence [6, 26, 28, 30].

Digital rectal examination (DRE)

Before PSA test, digital rectal examination (DRE) was the most used screening method for prostate cancer [31]. During DRE a specialist is feeling the peripheral zone of prostate and can detect abnormalities of the prostate. One of the main drawbacks of this technique is high variability in results and poor correlation with biopsy results— each test is performed by a different expert and conclusions might differ significantly [31]. However, DRE is still being used as an additional test in order to gain more information about the particular tumor that is being treated.

Biopsy

If after DRE, PSA (with the result ≥4 ng/ml) and other possible tests, prostate cancer is suspected, needle biopsies are recommended [13, 28]. Procedure is considered as *the gold standard* for PCa diagnosis and is usually necessary to confirm PCa diagnosis. Prostate biopsy can be performed transrectally or guided by transrectal ultrasound (TRUS) — transperineally [32, 33]. During a biopsy 10 samples are usually taken, based on various recommendations [32].

Biopsies in cancer diagnostics raise concerns, because procedure is invasive and can cause serious side effects, such as infections, erectile dysfunction, urinary incontinence [6]. A biopsy can also be false-negative, leading to a repetitive biopsy [34].

Radical prostatectomy (RP)

After the diagnosis of prostate cancer, radical prostatectomy (RP) is usually chosen as a treatment (together with radiation). RP is a procedure, during which prostate is completely removed [33]. Robot-assisted radical prostatectomy (RARP) is a more innovative, less invasive procedure, allowing shorter hospital stay and decreasing blood loss [33]. Radical prostatectomy, like biopsies, has post-operative side effects — possible erectile dysfunction and urinary incontinence [33].

Evaluation of the tumor

Prostate cancer can be classified clinically by assigning a stage T1-4, based on the amount of cancer cells found [13]. For example, if the cancer is found in less than 5% of tissue, T1a stage is assigned, while T4 means that the tumor is fixed or invades adjacent structures. Another parameter for PCa — pathological T (pT) category— can be determined from radical prostatectomy (RP) samples in combination with the clinical staging [13]. There are three categories: pT2, pT3 and pT4, with pT3 having two subcategories pT3a and pT3b. pT2 means that tumor is organ-confined, while pT4 tumor is fixed or invasive.

American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (UICC) have developed a system to determine the stage of the cancer that is called the tumor-node-metastasis (TNM) cancer staging system [13]. It is based on the extent of the tumor, which is described by the value T, that uses clinical and pathological categorization. The system also assigns values N, for involvement of nearby lymph nodes (N0 meaning ‘no positive regional nodes’ and N1 meaning ‘metastases in regional nodes’), and value M, for presence of metastases (M0 meaning ‘no’ and M1 meaning ‘yes’) [13].

Gleason score is a grading system specific to prostate carcinoma, based on the architectural pattern of the tumor [35]. It was created by Donald F. Gleason in 1966 and modified several times: in 1974-1977, 2005 and with slight updates in prognostic evaluation in 2013 [35, 36]. Gleason grading system intends to group tumors into four groups with some cases (for example, clear cells) falling into separate pattern (pattern 5) [35]. A pathologist identifies one of five patterns, gives a primary grade and later assigns a secondary grade to the observed pattern of the sample. The final Gleason score is a sum of primary and secondary grades, for instance, 4 + 3 = 7. Score 2 means that the tumor is well-differentiated or low-grade, while grade 10 tumor is poorly differentiated.

Gleason score has also an ability to predict the outcome of PCa, which made this grading system widely-used for decades, but today the accuracy is arguable. Usually, patterns 3, 4 and 5 (Gleason score 8-10) are considered as high grade cancers with poor outcome [35]. In 2013, prognostic grade groups were proposed after correlating data with biochemical recurrence: prognostic grade group I with Gleason score ≤6, group II with score 3 + 4 = 7, III with score 4 + 3 = 7, IV with score 4 + 4 = 8 and, finally, prognostic grade group V with Gleason scores 9 and 10 [36]. Particularly important are prognostic groups II and III, since the final score is 7 for both groups, but the architecture of tumor is, in fact, different. The separation of these two groups allows to reflect tumor behavior more accurately [36].

After summarizing the methods, currently used for the detection of prostate cancer, it is clear that the search for new approaches is urgent to avoid unnecessary surgical procedures caused by false positive diagnosis and to make PCa surveillance easier and more accurate. Scientists are looking for new ways to detect and monitor PCa with non-invasive procedures (for example, testing urine samples). DNA methylation biomarkers could be that new way. To find PCa-specific DNA methylation biomarkers, studies of DNA methylation patterns in prostate cancer are needed.

1.2. DNA methylation

1.2.1. General concepts of DNA methylation

DNA methylation in a cell is used as an epigenetic control mechanism of the information encoded by DNA [37]. It is one of the key mechanisms for the regulation of gene expression, genetic imprinting, X chromosome inactivation, embryogenesis and aging [37, 38]. Various cell functions can be affected by altered gene expression and DNA methylation [11].

DNA is methylated by transferring a methyl group from S-adenosyl-L-methionine (SAM) to 5' carbon atom of a cytosine, usually located in CpG dinucleotide, creating 5-methylcytosine (5mC) (Figure 6) [37, 38]. Around 70-80% of all cytosine residues in CpG dinucleotides are methylated and are found in repetitive, non-transcribed DNA regions [11]. Methylated cytosine not in the context of CpG dinucleotides is found in embryonic stem cells and plants [39].

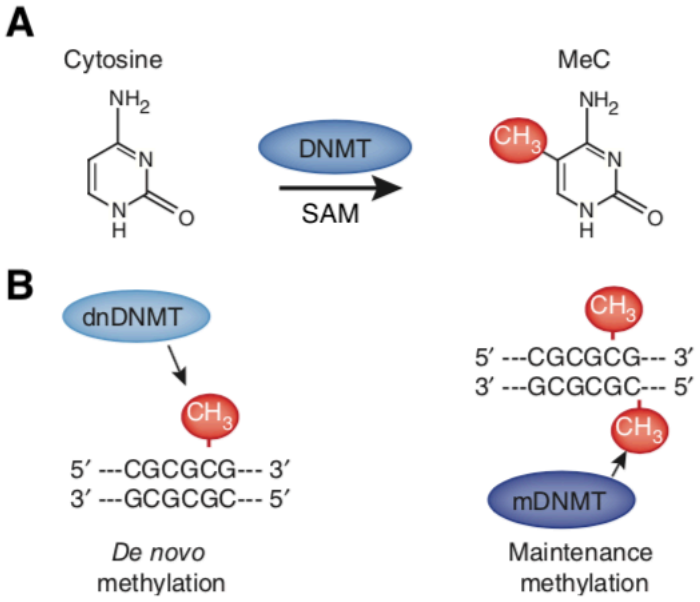


***Figure 6* DNA methylation.** (A) DNA is methylated at 5' position of cytosine residue by DNA methyltransferase (DNMT) transferring methyl group from a donor S-adenosyl-L-methionine (SAM). (B) DNA can be methylated either *de novo* or adding methyl groups at the complementary strand of hemi-methylated DNA. Adapted from [40].

Methyl groups are added by the DNA methyltransferases (DNMTs) [37]. The main DNMT is DNMT1, which is expressed at high levels and maintains cytosine methylation in the cell cycle [37]. DNMT1 maintains methylation by using hemi-methylated DNA as a template [38]. DNMT3a and DNMT3b are responsible for *de novo* methylation during early development (Figure 6) [37, 38]. Unfortunately, 5-methylcytosine (5mC) can mutate, because of the spontaneous deamination, but these mutations are recognized by the repair machinery and repaired [37].

CpG islands

CpG islands (CGIs) are defined as regions with high density of CpG dinucleotides, between 300 and 3000 bp long with greater than 50% GC content and observed/expected ratio of CpG to GpC greater than 0.6 [37, 41, 42]. Depending on the definition, there are around 50,000 CGIs in human genome [37]. Around 15% of the CpG sites are found in

islands, located in promoter regions (70% of them are promoters of protein-coding genes, usually housekeeping genes) [37, 43].

It is important not to forget the regions around CpG islands — the shores and the shelves — which can carry information no less important for gene regulation. For example, methylation of CpG shores is associated with reduced gene expression in highly conserved tissue-specific manner [44]. It is known that shore and shelve regions are more variable in different cancers, such as colon cancer [37, 44].

CpG islands in intergenic regions are usually methylated in normal cells, but the variation in methylation patterns is most significant in different tissues [43]. Intergenic regions contain transposable (such as LINE) and viral elements (such as HERVs) that are largely inactivated by bulk methylation in human genome [45]. This is done in order to repress potentially dangerous transcription of this genetic material.

DNA methylation and gene expression

DNA methylation may affect gene regulation by several different possible mechanisms. Firstly, 5mCs in CpG islands can physically block the access for transcription factors (TFs), which results in a suppressed initiation of gene expression. Secondly, methylated regions can be recognized by various proteins, containing methyl-CpG binding domains (MBDs), such as Methyl-CpG Binding Protein 2 (MeCP2), Methyl-CpG Binding Domain Protein 1 (MBD1), MBD2, MBD3, and MBD4 [37, 46, 47]. Although it is important to note that methylated sequences can be recognized and bound by proteins without MBDs, for instance, zinc-finger domain proteins. [47]. These proteins help to recruit other proteins to the site of methylation, which, for example, can lead to gene suppression, mediated by histone deacetylases (HDACs) or Polycomb proteins [37, 46, 47].

Particularly, gene promoter/TSS region methylation is being associated with gene expression regulation. The classical relationship between promoter DNA methylation and gene expression is gene silencing due to DNA methylation [48]. Unmethylated promoter regions can attract transcription factors and initiate gene transcription, while methylation of these regions silences the transcription (Figure 7) [46]. This DNA methylation-gene expression relationship is well-studied in promoters with CpG islands at their TSSs, but DNA methylation can also occur in non-CGI promoters [46]. Unfortunately, details on this methylation pattern are not known yet [46].

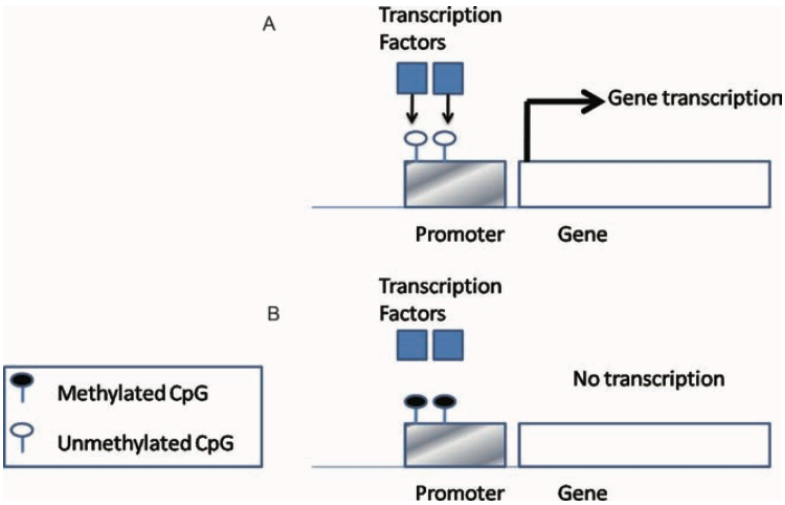


Figure 7 **DNA methylation is one of the mechanisms that regulates gene expression. (**A) Transcription factors can bind gene promoter region with unmethylated CpG and initiate gene expression. (B) When CpGs in promoter are methylated, transcription factors are unable to bind, which results in gene silencing [49].

CpG-rich promoters are typically unmethylated in normal cells, which means that they are active and marked with active histone modifications, such as trimethylation of lysine 4 on histone H3 (H3K4me3) [46, 50]. Relationship between DNA methylation, histone marks and micro RNA (miRNA) is known as epigenetic crosstalk, which is an important mechanism for gene regulation [43]. Genome-wide epigenetic profiling shows that highly methylated promoter regions have low levels of H3K4me3 active mark and H3K27ac in highly methylated distal regulatory regions [50]. It is known that H3K4me3 prevents DNMTs from binding and therefore prevents methylation, while H3K36me3 repressive modification stimulates DNMT3a activity [43].

Methylation of H3K27 and H3K9 and ubiquitination of histone H2A play an important role in gene suppression by compacting the chromatin. These reactions are mediated by Polycomb repressive complexes (PRCs) — PRC1 and PRC2 [51]. PRC2 di- or trimethylates Lys 27 of the H3 histone through complex's subunits EZH1 and EZH2 with enzymatic properties, PRC1 uses the ubiquitin ligases RING1A and RING1B to monoubiquitylate Lys 119 of the H2A histone [51, 52]. PRC2, in particular its product H3K27me3, is shown to have a dependence with DNA methylation. PRC2 and DNA methylation shares a role of maintaining gene repression. DNA methylation is believed to attenuate PRC2 binding, while the lack of DNA methylation attracts and guides the complex, creating a negatively correlated relationship between two epigenetic mechanisms [52]. Furthermore, disturbance in these epigenetic mechanisms has an impact for different types of cancer [53]. It has been shown that genes with CGI promoters already silenced by PRC have

higher chances to become methylated in cancer [46]. One specific PRC, known as PRC4, has been observed uniquely in cancer or stem cells (adult and embryonic) and proposed to be a part of oxidative damage response in cancer leading to suppression of gene expression [54].

The classical understanding is that DNA methylation of promoter regions always leads to gene expression downregulation, but recently more studies argue with this paradigm. Interestingly, in some research projects, mostly related to cell differentiation, it was shown that certain genes with completely unmethylated CGI promoter regions are not able to produce fully functional transcripts, because RNA polymerase II cannot be recruited [48] or some TFs (with and without MBD) actually prefer sequences with methylated CpG [47, 55]. Although there are chances of false-positive results and gene regulation patterns in stem cells are known to be different than in differentiated cells. In case of TFs which bind methylated sequences, there is just a small number of such proteins recognized up to now [56]. On the other hand, ability of DNA methylation to silence gene expression, simply, might depend on a density of DNA methylation marks in the promoter region. Low density can be insufficient enough to suppress the expression, so methylated cytosine marks act more like a "traffic light" rather than a firewall [57, 58]. Furthermore, promoters with low CpG content (and thus most likely without CpG islands) are often methylated, but it was noted that transcription might remain to be active [57, 58].

Relationship and the causality between promoter DNA hypermethylation and upregulation of gene expression is still very unclear. Research projects usually show some correlation between promoter hypermethylation and increased gene expression, but in most cases the focus is put on the classical regulation pattern, since it is the most common and significant in many studies. However, scientists are motivated to re-evaluate and expand the picture of a classical regulation pattern [59]. Methylation of TF binding sites that leads to the prevention of TF binding cannot be considered as a universal mechanism of transcription regulation [58].

DNA methylation can also occur in gene bodies, which are poor in CpGs [46]. It is believed that heavily methylated gene bodies have increased expression in dividing cells, while in slowly or non-dividing (for example, brain cells) cells such relationship between expression and methylation is opposite [43, 55]. Others believe that gene body methylation may have a role in gene splicing [46]. The exact impact of DNA methylation in gene body is still unclear.

Methylation patterns of distant regulatory elements, such as enhancers, are more tissue-specific, and also altered in cancer tissues, compared with normal [60]. Enhancers have

a strong connection with promoters and their methylation highly correlates with gene expression in cancer cells and can even predict the changes in the expression profiles of cancer genes [60]. However, it is challenging to associate enhancer methylation with an expression of a particular gene.

Even though many studies have been conducted on DNA methylation and gene expression, relationships between the two processes are highly complex and not so straightforward.

1.2.2. DNA methylation changes in cancer

Cancer is more than a genetic disease, and epigenetic modifications, such as DNA methylation, get more and more attention. Epigenetic mechanisms (DNA methylation together with histone modifications, non-coding RNAs and nucleosome positioning) is an additional layer of regulation of gene expression in carcinogenesis. Altered DNA methylation can increase the risk to develop cancer, drive it or contribute to its progress. In general, global DNA hypomethylation and loci-specific hypermethylation is an epigenetic hallmark for all human cancers [59, 61].

Promoters of tumor suppressor genes (TSGs) are often the target for DNA hypermethylation in cancer [50]. Usually promoter and the region around are unmethylated in normal cell, while gene body together with intergenic region are known to be methylated. On the other hand, in a cancer cell, promoter is hypermethylated to suppress the expression of TSG, while gene bodies are hypomethylated (Figure 8).

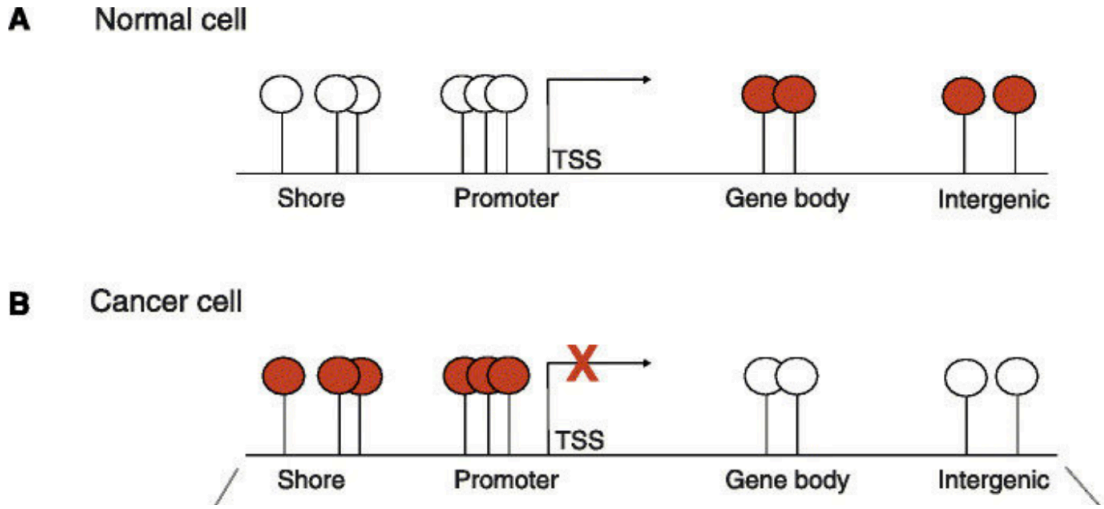


Figure 8 **DNA methylation in normal and cancer cells.** (A) In a normal cell, when gene transcription is active, methylation (red bubbles) occurs in gene body and intergenic region, while the shore of the CpG island and the promoter are unmethylated. (B) In a cancer cell, when gene transcription is suppressed, the shore and the promoter are methylated, while gene body and intergenic regions are left unmethylated. Adapted from [38].

This is one of the examples how DNA methylation can affect cancer development (Figure 9B). *p16*^{INK4a} is a TSG gene with one of the most commonly hypermethylated CGIs in a large number of human cancers [62]. *p16*^{INK4a} encodes a cyclin-dependent kinase inhibitor, involved in cell cycle processes, which is a common feature for TSGs [63]. *p16*^{INK4a} gene is overexpressed in different tumors in order to downregulate the expression of genes, causing cell proliferation by inhibiting the S phase [63]. Inhibition of cell proliferation is an opposite of what cancer cells aim to do, which leads to a methylation strategy.

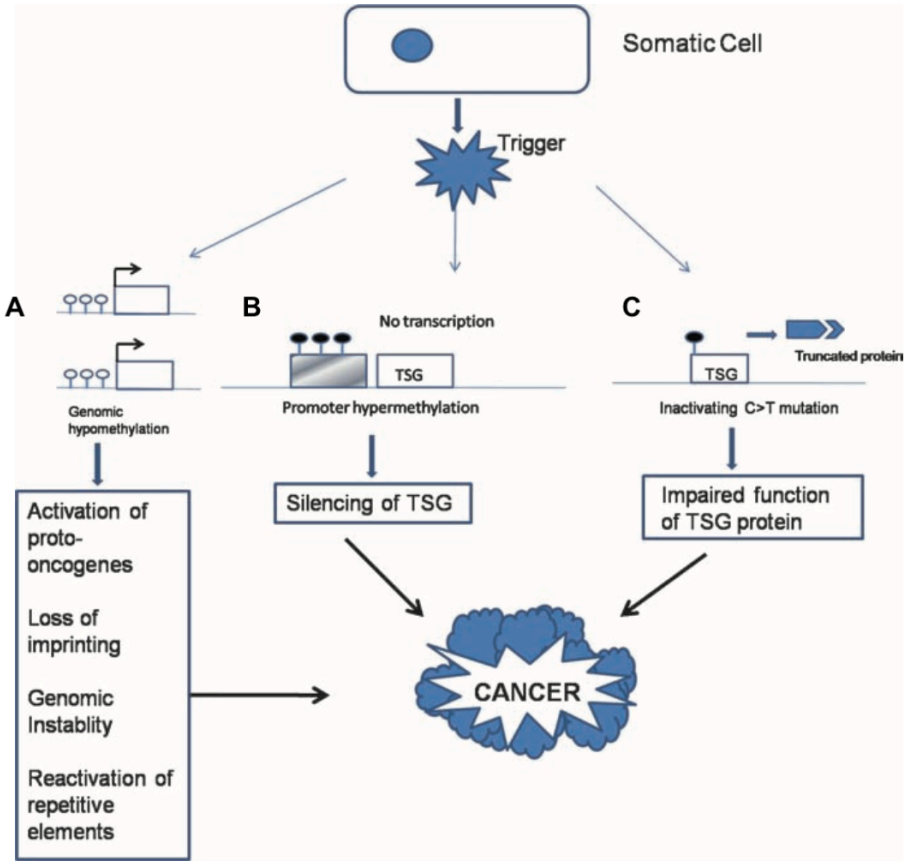


Figure 9 **Three examples of how changes in methylation can affect cancer development.** (A) Genomic hypomethylation of regulatory regions of oncogenes leads to their activation and increased gene expression. As well as loss of imprinting, genomic instability and reactivation of repetitive elements. (B) Promoter hypermethylation often leads to downregulation of TSG transcription. (C) Methylated cytosine residues are unstable and can be spontaneously converted to thymine. This can lead to the inactivation of TSGs. TSG – tumor suppressor gene. Adapted from [49].

Another important example of DNA methylation in cancer is hypomethylation of oncogenes (Figure 9A) [64]. Oncogenes encode proteins of oncogenic pathways, involved in cell proliferation, angiogenesis, immortality and metastasis. Other oncogenes encode inhibitors of TSGs and apoptosis [65]. Proliferation and survival processes are necessary for human organism, but just to ensure tissue homeostasis and development [65]. Regulatory sequences and promoters of oncogenes in normal cells are hypermethylated to prevent

unnecessary transcription. In cancer, expression of oncogenes is activated by removing DNA methylation marks [64]. This way, cancer cells avoid apoptosis and are able to divide indefinitely, grow new blood vessels and, eventually, invade adjacent tissues.

Hypomethylation in cancer is also used to activate transposons, such as SINE and LINE (Figure 9A) [66]. In healthy cells, repetitive and transposable elements are heavily methylated and located in heterochromatin, but in tumor cells genome integrity is disturbed and these elements are able to induce additional mutations, for example, by frequent insertions into exons, introns, causing compromised transcription or generating antisense transcripts [64, 66].

Inactivating mutations is the third example of DNA methylation affecting cancer development (Figure 9C). As mentioned in previous section, methylated cytosine can spontaneously mutate to thymine by deamination [37]. If a mutation is not repaired by repair machinery, it can cause the termination of gene transcription, resulting in a truncated, inactive protein. Therefore, cell functions can be disturbed, contributing to the formation of cancer.

Since aberrant DNA methylation is a hallmark for various cancer types, it is one of the most studied epigenetic change and, lately, genome-wide DNA methylation studies can be an opportunity to identify unexplored epigenetic changes [38, 59]. Using constantly improving high throughput methylation profiling technologies scientists are producing methylation maps (known as methylomes) for different cancers in order to determine how DNA methylation distributes throughout the whole genome [38, 48]. Methylomes can help to uniquify each type of cancer because methylation changes are usually cancer-specific, which could be helpful for early diagnosis or prognostics. Unfortunately, some cancer types, including prostate cancer, are exceptionally heterogenic, which complicates studies intending to generate a unique epigenetic landscape for PCa [7].

Example of a methylome for 10 different types of cancer can be seen in figure 10.

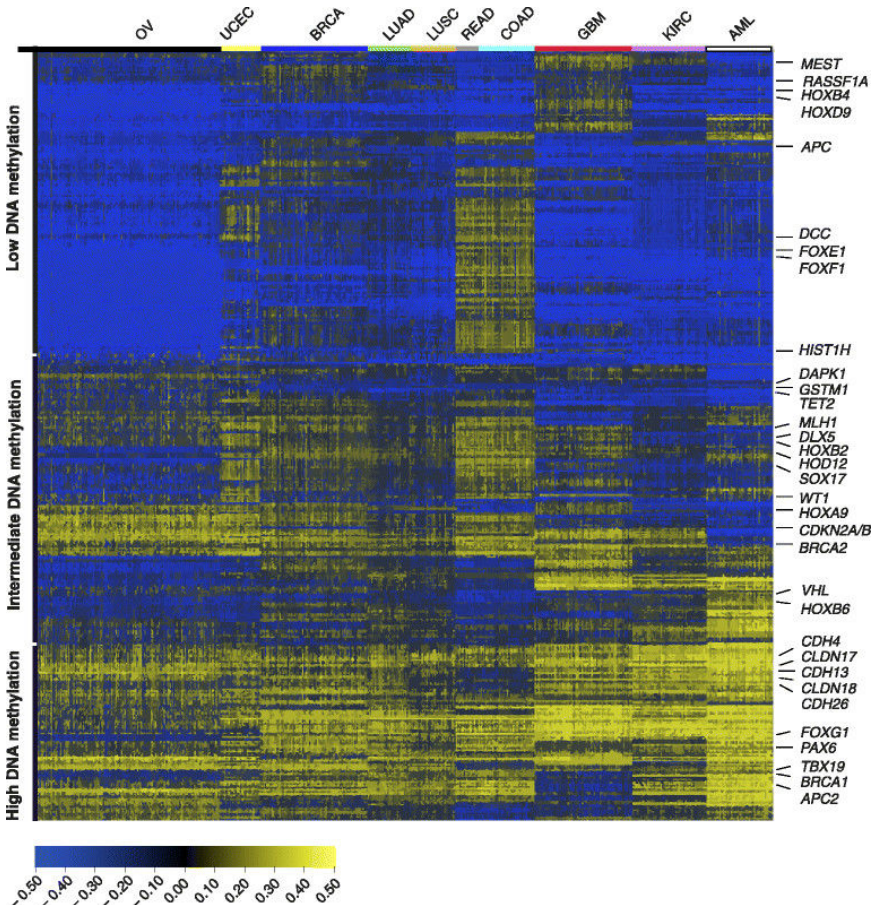


Figure 10 **Example of DNA methylation map (methylome) for ten cancer cohorts from the Cancer Genome Atlas (TCGA).** There is a distinguishable difference in methylation profiles in different cancer types. Homogeneity within the group is relatively high. Methylation patterns in different tissue of origin are distinct. Highest overall methylation can be observed in colorectal cancer, while lowest – in kidney cancer. 27k Illumina data represents 24,980 CpG sites tested in 2,224 tumor samples. CpG sites, associated with single-nucleotide polymorphisms (SNPs) and located on chromosome X and Y, were removed. Beta values vary from -0.5 (dark blue) to 0.5 (yellow). Tumor types: OV – ovary, UCEC – uterus, BRCA – breast, LUAD and LUSC – lung, READ and COAD – colorectal, GBM – brain, KIRC – kidney, AML. No prostate cancer data. Some of the gene names are indicated on the right. Adapted from [38].

DNA methylation in prostate cancer

DNA methylation changes can be cancer-specific, but prostate cancer has variable collection of genetic and epigenetic changes across different PCa cases. Scientists are working to find PCa-specific variations in DNA methylation patterns in order to produce a better picture of molecular changes in PCa.

CpG islands are more likely to be hypermethylated in prostate cancer and silencing of the genes because of the promoter hypermethylation is more common than by DNA mutation [67]. In a study by Kirby and colleagues, more than a half of all significant CpG dinucleotides were methylated in prostate cancer samples [34]. More than 100 genes, by far, have been shown to be inactivated by DNA methylation in promoter regions in PCa [68]. It is suggested that aberrant methylation occurs as a result of dysregulation of DNMT3A2 and

DNMT3B DNA methyltransferases [34, 69]. Methylation in PCa usually leads to promoter inactivation and silencing of the genes, such as DNA repair or TSGs [11, 67]. Many pathways, such as DNA damage repair, hormonal responses, cell invasion and cell cycle control, are being disrupted, because of aberrant DNA methylation [11]. Such changes in normal prostate epithelium may result in cancer formation, since DNA methylation occurs in early stages of prostate carcinogenesis (Figure 11) [11, 67].

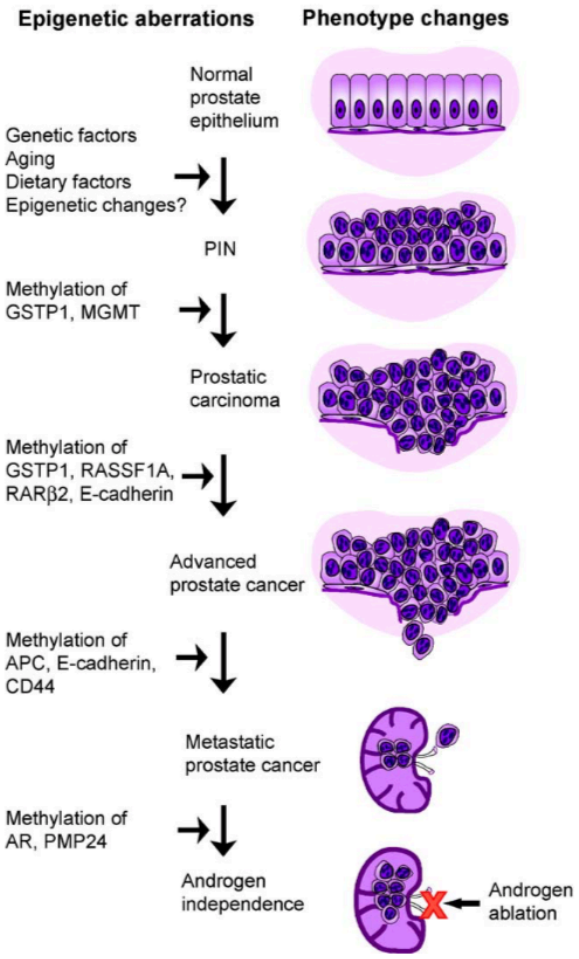


Figure 11 **Epigenetic changes in prostate cell that contributes to cancer formation** Various genetic factors, aging and probably diet all together with epigenetic changes can trigger cancer formation from normal prostate epithelium. Prostatic intraepithelial neoplasia (PIN) is pre-cancerous tissue, which later changes into prostatic adenocarcinoma after, for example, methylation of *GSTP1* and *MGMT*. If there are more genetic and epigenetic changes in cells, prostate tumor can progress into an advanced form of cancer (metastatic and androgen-independent). For such change methylation of *APC*, *E-cadherin* or *AR* might be responsible, but many changes leading to cancer formation are still unknown. Adapted from [11].

GSTP1 is a gene, well known to be hypermethylated in prostate cancer. Gene is responsible for protection of prostate cells from carcinogenesis and DNA adducts [67]. In about 90% of PCa cases the absence of *GSTP1* is caused by DNA hypermethylation [12, 68]. Its methylation has been found already in PIN (over 50% of precursor lesions), which

describes *GSTP1* as a biomarker for early diagnosis [11, 12, 68]. Normally, GSTP1 levels elevate in PIA because of the increased oxidative stress, but after methylation response to stress is suppressed in PIA and PIN [11, 70].

As described in previous paragraphs, specific DNA methylation patterns in pre-malignant prostate tissue can be associated with deletions, tandem duplications or interchromosomal translocations following in cancerous tissue [15]. Highly methylated CpG sites are more susceptible for somatic base mutations during PCa development [15].

It also has been shown that DNA methylation is important for the development of aggressive prostate cancer [15]. For example, alterations in DNA methylation are more frequent than mutations or copy number changes in CRPC [71]. With disease progression global DNA hypomethylation increases, and overall 5mC content is lower in metastases. Furthermore, AR silencing by DNA hypermethylation is shown to be linked with progression to CRPC [68].

A big part of the research projects that investigate DNA methylation in prostate cancer focus on individual genes, their functions or even individual methylation sites [72], but genome-wide DNA methylation techniques are able to investigate many genes simultaneously. Such studies could help to cluster PCa into several groups, even though PCa is very different in each case. For instance, scientists from TCGA research network analyzed 333 prostate carcinomas and managed to distinguish 7 subtypes of PCa [73], based on oncogenic drivers, such as fusions and mutations (Figure 12). Later they noted that DNA methylation changes in cancer samples are diverse, compared with normal samples. Most importantly, independent clustering by methylation changes, showed distinct and matching methylation patterns with several subtypes, such as *ERG*-positive tumors [73].

DNA methylation patterns are also being associated with various clinical parameters for diagnostic and prognostic reasons. For PCa Gleason score is an important example, since it is still the best predictor of aggressiveness [74]. Geybels et al. created an epigenetic signature of Gleason score, consisting of 52 differentially methylated CpG sites. The group used TCGA methylation data and proposed their epigenetic signature to improve the risk prediction for PCa (in particular Gleason score 7 tumors) outcomes [74].

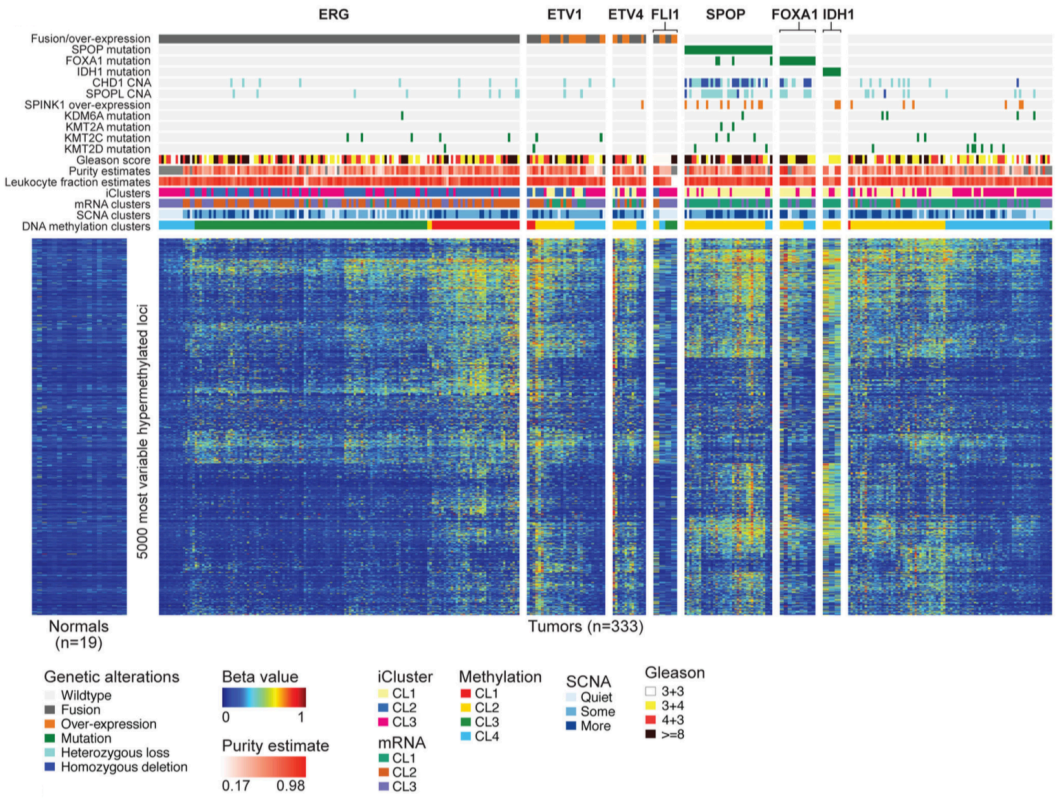


Figure 12 **An attempt to group prostate cancer into several subtypes, according to molecular changes, and connect such subtypes to alterations in DNA methylation patterns.** Even though PCa is very heterogenic, scientists were able to distinguish 7 subtypes, based on mutations or chromosome fusions (subtypes are indicated above the heatmap). They also showed that DNA methylation pattern in PCa is altered compared to normal prostate tissue (left). Independent clustering of methylation values showed some connections to distinguished subtypes, for example, *ERG*-positive tumors connects to a nearly unique methylation pattern (cluster 1). Adapted from [73].

De-methylating drugs

One of the possible therapies to encounter harmful DNA hypermethylation are drugs that remove DNA methylation marks. For example, DNMT inhibitor (DNMTi) 5-azacytidine (5-Aza-CR), which has already been approved as an epigenetic chemotherapeutic agent. 5-Aza-CR replaces cytosine in DNA and traps DNA methyltransferases, resulting in their degradation and loss of DNA methylation [68]. Unfortunately, de-repression of genes is not conclusive in different cases [50]. Blattler et al. in their research found that loss of DNA methylation after the treatment with de-methylating drugs does not result in an increase in active histone marks (such as H3K4me3 and H3K27ac) at promoters [50]. This can be one of the factors, why full recovery of gene expression is not achieved.

1.2.3. Current methods for detection and analysis of DNA methylation

Methods, that are used to detect DNA methylation, can be grouped into three categories, based on the main principle of the method: immunoprecipitation of methylated

cytosine with methyl-binding proteins or antibodies, cleavage of genomic DNA by methylation-sensitive restriction enzymes and distinction of methylated and unmethylated cytosine after conversion by bisulphite [39, 41, 75]. Methylation arrays and sequencing-based technologies, together called high-throughput technologies, provide genome-scale DNA methylation maps, which give a huge amount of information on aberrant methylation patterns in various diseases, including cancer [38]. One of the most popular and widely used DNA methylation detection techniques are DNA methylation arrays Illumina Infinium HumanMethylation450 BeadChip (HM450) and Infinium HumanMethylation27 BeadChip (27k). Methods, such as MeDIP-ChIP, MeDIP-Seq, bisulphite sequencing and many others are also widely used. Which method will be chosen depends on many aspects, such as cost, the amount and quality of the DNA sample, required sensitivity, robustness of the method and others [39].

Bisulphite sequencing is a method capable to detect differentially-methylated regions on a whole-genome scale. It is based on a sequencing of DNA sequence after bisulphite treatment. Unmethylated cytosine residues are converted into uracil by deamination, while methylated cytosine remains untouched and is recognized as cytosine during sequencing [39]. This method is good for detection of DNA methylation in repetitive sequences and assignment of methylation state to the specific alleles [76]. Unfortunately, bisulphite sequencing is expensive and sequencing data is challenging to process because of the reduction and complexity of genome sequence [39].

Therefore, DNA methylation microarrays can be a way to profile DNA methylation across human genome regions. This technique is based on hybridization of bisulphite-converted DNA to the microarrays. For example, Infinium chemistry does not require methylated DNA to be captured by immunoprecipitation, which provides access to most target sites [41]. Such assays focus on specific regions in genome where DNA methylation occurs most frequently and has impact on gene expression. Regions include gene promoter regions, 3’ untranslated regions (3’UTRs) and enhancer regulatory elements [39]. The fact that regions of less importance can be excluded and there is a possibility to examine large number of samples simultaneously, give an advantage of saving money and time, while still providing biologically significant information about gene regulation [39, 76].

Infinium methylation arrays contain beads that have target-specific probes attached to them. These probes are designed to examine individual CpG sites [41]. Quantitative genotyping of bisulphite-converted DNA is a measure for DNA methylation [41].

Infinium HumanMethylation27 BeadChip (27k) contains probes targeting 27,578 CpG sites [75]. 27k technology focuses mainly on proximal promoter regions across human genome. Targeted sites are located 1 kb upstream and 500 bp downstream of TSS of 14,475 genes with an average two probes for one gene [75].

Illumina Infinium HumanMethylation450 BeadChip (HM450), on the other hand, is a more recent method that targets not only promoter regions, but also many other sites, important for gene regulation. HM450 contains 485,577 probes targeting 21,231 (99%) of RefSeq genes, 26,658 (96%) of CpG island regions, 26,249 (92%) CpG island shores, 24,018 (86%) CpG island shelves, 62,600 Hidden Markov Model-defined (HMM) CpG islands, 11,754 FANTOM 4 promoters, 16,232 DMRs, 80,538 informatically-predicted enhancers and other biologically significant sites [41]. Targeted genes have multiple (average of 17.2) probes associated with them and one probe might target a region that belongs to several content categories [41].

Illumina Infinium HM450 array correlates with 27k showing an R^2 of >0.95, since over 94% of regions present on 27k are included in the HM450 [41].

2. DATA AND METHODS

2.1. Research project design and workflow

As a preparation for the project, available literature was studied in order to find available DNA methylation datasets and review whether there are similar research projects published (Figure 13). PubMed search engine, maintained by the United States National Library of Medicine (NLM) at the National Institutes of Health (NIH), was used [77]. Keywords *"DNA methylation in prostate cancer"*, *"computational analysis of prostate cancer"* were used in the searches.

The final step in the foundation for the main analysis was finding three DNA methylation datasets. Data analysis then was divided into three main parts (Figure 13). The outcome from each part was used in the following parts, creating a workflow that leads to the conclusions of this research project.

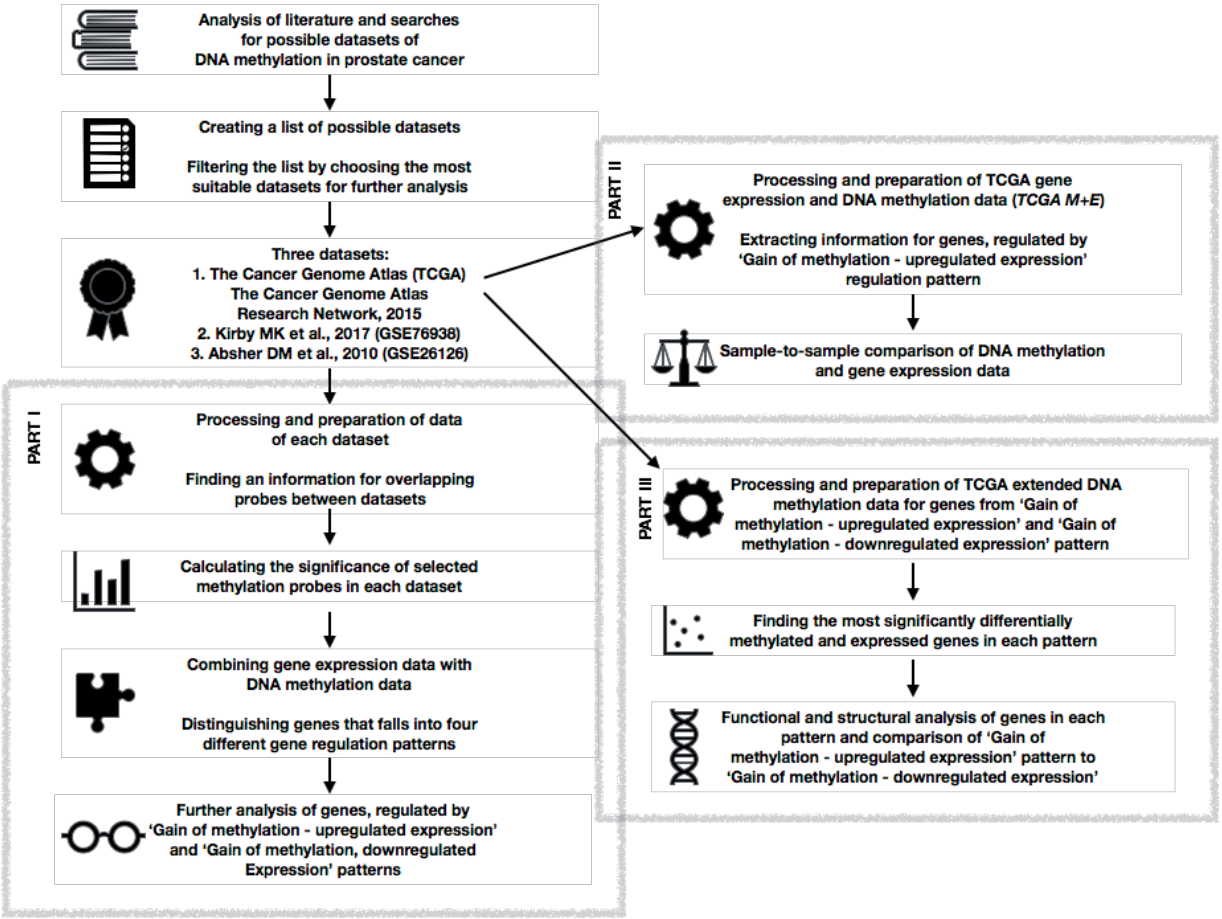


Figure 13 **Workflow of the research project.** Analysis is divided into three main parts. In preparation for the project, analysis of the literature and selection of DNA methylation datasets were done.

PART I of the project focused on the analysis of three DNA methylation datasets: *Absher*, *Kirby* and TCGA. First of all, data had to be processed and prepared for further analysis. Datasets were connected with each other by finding overlapping probes between all three. Information on these probes was filtered and statistical analysis, comparing DNA methylation in cancer and normal prostate tissue samples, was performed for each dataset. After that, DNA methylation data was combined with gene expression data in order to find the connections between them. In this study the connection between probes and gene expression was made based on the reference for the Illumina probes. Genes that have hypermethylated or hypomethylated probes associated with them, were divided into groups, according to DNA methylation and gene expression connection. Four such relationship patterns (gene regulation patterns) were distinguished: 'Gain of methylation — upregulated expression', 'Gain of methylation — downregulated expression', 'Loss of methylation — upregulated expression' and 'Loss of methylation — downregulated expression'. Numbers of genes, following the same patterns in all three DNA methylation datasets, were found. The latter focus was put on two patterns with gain of methylation, followed by upregulated or downregulated expression. Sets of genes that follow these patterns were investigated using gene set enrichment analysis (GSEA).

The goal of the PART II of the project was to compare DNA methylation and gene expression values in same samples (sample-to-sample comparison) and uncover whether the overcompensation of gene expression occurs for genes, following 'Gain of methylation — upregulated expression' pattern. TCGA DNA methylation and gene expression dataset (TCGA *M+E*) was used for this analysis. Probes with methylation gain and upregulated expression from PART I analysis were used to filter the data in TCGA dataset. Samples were sorted according to the methylation values and expression values then were matched for each sample.

Going further, the main focus of the PART III has been put on 'Gain of methylation — upregulated expression' regulation pattern and later comparing the findings with classical and well established 'Gain of methylation — upregulated expression' regulation pattern. Significantly differentially methylated and expressed genes were found following each pattern. Those lists of genes later were narrowed down by keeping the genes that are associated with probes that have only the gain of methylation. These genes were then analyzed and compared using GSEA. UCSC genome browser (http://genome.ucsc.edu) was used to visualize the positions of the probes, associated with top 10 genes with most significant methylation changes in each pattern [78]. Besides the visualization of probes,

distances between transcription start sites (TSSs) and methylated positions were calculated and the overlap with CpG islands was evaluated. UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables) was used to download TSSs for all the genes in both patterns [79]. In addition, visualization of TSSs was done using a DataBase of Transcriptional Start Sites (DBTSS) [80, 81].

2.2. Data selection

2.2.1. DNA methylation data

One of the main recourses of publicly available genome-wide methylation data along with gene expression data in various types of cancer is the Cancer Genome Atlas (TCGA). This recourse was considered directly, because it contains more than 500 PCa samples together with 50 benign tissue samples, which makes the dataset very promising [73]. TCGA cohort originated from a research project, published in 2016, that analyzed 333 primary prostate carcinomas [73]. Data on more PCa samples to a TCGA Network repository was added later.

To find and download other DNA methylation in PCa datasets the Gene Expression Omnibus (GEO) database was used [82, 83]. GEO is an international repository, curated by a division of the National Library of Medicine of the National Center for Biotechnology Information (NCBI), that archives and distributes high-throughput functional genomic data, such as microarray and next-generation sequencing (NGS) data [82]. Keyword phrase *"prostate cancer methylation"* was used in searches. Potential datasets were selected to narrow down the possibilities, and later datasets for the project were chosen based on the number of samples (normal tissue and cancer), availability and the method data was produced with.

2.2.2. Gene expression data

As mentioned before, gene expression data in PCa and normal tissue samples from TCGA was used, since the data was generated from the same samples as DNA methylation data. This TCGA gene expression data was used in PART II (Figure 13) of the project and distinct as *TCGA gene expression* dataset. In order to better distinguish different TCGA datasets, TCGA DNA methylation and gene expression on the same samples dataset has been referred as TCGA *M+E* dataset.

For other parts of the research gene expression in PCa and normal tissue samples combined data from *five-study-cohort* analysis (as referred in [84]) was used. Dataset consists

of five distinct sets of gene expression records: *Bertilsson* (116 PCa, 40 normal, 14,149 unique genes, Illumina Human HT-12 WG-DASL microarray), *Chen* (65 PCa, 71 normal, 12,497 unique genes, Affymetrix Human Genome U133A microarray), *Taylor* (131 PCa, 29 normal, 18,294 unique genes, Affymetrix Human Exon 1.0 ST microarray), *TCGA* (497 PCa, 52 normal, 20,504 unique genes, Illumina HiSeq2000/Genome Analyzer IIX) and *Prensner* (78 PCa, 38 normal, 23,712 unique genes, Illumina Genome Analyzer) [84].

2.3. Programming languages

2.3.1. Python

Python programming language was used throughout the research project. Particularly, an application for Python, called IPython (Interactive Python, http://ipython.org), was the main tool for data processing and analysis [85]. Python is a high-level language and is executed by an interpreter, which makes the language numerically not that efficient and slower than compiled, lower-level languages, such as C, C++ or Fortran [85, 86]. However, compiled languages are more complicated to work with and therefore Python is being used by many less-skilled bioinformaticians and scientists. Even if Python requires more time to run the code, it saves the time of the scientists, because the language is easier to learn and use. Python and other high-level languages are also known as scripting languages because they allow to combine modular tools in a script [86, 87].

Other advantages of Python programming language are free availability, active developer community and its many scientific libraries [86, 87]. Free availability is an obvious advantage, which creates a wide circle of people that choose Python as their programming language. This leads to an active community that is able to provide a fast, but informative answers to risen questions while working on a project. Scientific libraries allow to extend the capability of Python and solve more advanced tasks [86]. For example, write graphical applications or work with a specific scientific data [85].

IPython

IPython is an interactive shell that allows an easy editing, parallel computing, recording of a work session and other features. Code written for IPython must be a valid Python code, but IPython supports projects with their own syntax [85]. Such projects includes *Sage*, *the Pymerase* project and others [85]. One of the most significant advantages of IPython compared to basic Python is that it stores inputs (`In`) and outputs (`Out`) of a session and they are available to use throughout a work session [85]. Furthermore, working

interactively allows to directly manipulate objects and code [85]. IPython is also a very good environment for learning, since it traces back mistakes in a code and provides information about such errors (and even suggests how to fix them) better than many other coding environments [85].

NumPy

NumPy is a Python library, which is essential for matrix-related calculations and contains some of the statistical tools [86]. NumPy array (also known as *ndarray* or N-dimensional array) is a standard numerical data representation – "a multidimensional, uniform collection of elements" [88]. An array can be characterized by the type of elements and shape (*M*x*N*) [88].

Arrays is a useful way to manage big datasets, such as methylation or gene expression data. Elements in an array are grouped into vectors by a process, called *vectorization-Iets*, which allows to perform `for` loops more efficiently and faster than applying the loop for each element [88]. The output of operations, given in an array, is also more understandable. When two arrays are used, element-wise subtraction, for example, can be easily performed. In case of different array shape with a common shape dimension, NumPy can adjust one of the array to perform an operation — the operation is *broadcast* [88].

2.3.2. R

R programming language was used for statistical tests throughout the project (https://www.r-project.org) [89]. R is a freely available programming language and software environment created by Ross Ihaka and Robert Gentleman with first version released in 1995 [90]. R language was created based on another programming language for statistical analysis, called S [91]. R allows looping, branching and usage of functions, because the core of the language is an interpreted computer language [92]. Statistical procedures include parametric and nonparametric tests, clustering, time series analysis, nonlinear regression models, linear models and others [92].

The software has many libraries filled with different statistical packages that can solve different tasks as mentioned before [93]. The *basic packages* that provide main functions of R are *base*, *compiler*, *datasets*, *grDevices*, *graphics*, *grid*, *methods*, *parallel*, *splines*, *stats*, *stats4*, *tcltk*, *tools* and *utils* [92, 93]. Other 15 packages can be put in *recommended packages* group. These packages are: *boot*, *class*, *cluster*, *cadetools*, *foreign*, *KernSmooth*, *lattice*, *MASS*, *Matrix*, *mgcv*, *nlme*, *nnet*, *rpart*, *spatial* and *survival* [92, 93]. The last group of packages is *contributed packages*. As the name describes, contributed

packages are created by contributors, since R code is open to every user [93]. Packages from various contributors expand the capabilities and functions of R. For example, Bioconductor has almost 1500 packages for bioinformatics (DNA microarray, SNP, sequencing and other data) [94, 95].

Limma is a R Bioconductor package, created to analyze gene expression microarray data, for example, to assess differential expression [96]. The goal of *Limma* is to fit a linear model to the expression data for each gene (`lmFit()`) [96]. In this project cancer samples were compared with normal samples, so the contrast step was performed in order to compare the coefficients using `contrasts.fit()`. After applying the linear model, *Limma* summarizes the results, performs hypothesis tests and adjusts the p-value for multiple testing, using `decideTests()` and `topTable()` [96].

2.4. Gene set enrichment analysis

Gene Set Enrichment Analysis (GSEA) is an analytical method for gene expression data interpretation based on evaluation of microarray data at the gene set level [97]. GSEA aim is to identify genes that are overrepresented and can be associated with a particular phenotype. Gene sets are defined using published information about biochemical pathways, co-expression and other prior biological knowledge [97]. In GSEA method, enrichment score (ES) is first calculated to reflect overrepresentation degree of the gene set in the entire ranked list (step 1). In step 2, statistical significance of enrichment score is estimated using empirical phenotype-based permutation test. Finally, for step 3, adjustment for multiple hypothesis testing is done, when the whole library of gene sets is evaluated. The outcome of this step is a normalized enrichment score (NES) and each of such score has a corresponding false discovery rate (FDR) [97].

There are many tools to perform GSEA and one of the web-based tools is Enrichr, first introduced in 2013 by Ma'ayan laboratory. Enrichr uses three methodologies for computation of enrichment: Fisher exact test (resulting in p-value), correction to the Fisher exact test (resulting in a z-score) and a combination of p-value and z-score, resulting in a combined score *c* [98]. Enrichr is able to visualize the GSEA results by clustering into a clustergram or a grid as well as creating a bar graph or a network, which makes the tool easy and enjoyable to use [98, 99]. Grids, created for each category, have an index of significance associated with them [98].

A tool provides enrichment scores for different categories with many libraries. There are 35 gene-set libraries in total. Transcription category includes *ChEA* (ChIP enrichment

analysis), *Epigenomics Roadmap HM ChIP-seq*, *ENCODE TF ChIP-seq* and *Histone Modifications*, *Transcription factor PPIs* (protein-protein interactions) and others. Pathways category contains *KEGG*, *WikiPathways*, *Reactome*, *Kinase Perturbations from GEO* and others. Enrichr also analyses ontologies, which includes *GO Cellular Component*, *GO Biological Process*, *GO Molecular Function* as well as *Jensen tissues*, *compartments* and *diseases* and a few others. There are also libraries that contain differentially expressed genes after disease, pathogen or drug perturbation, such as *OMIM disease*, *Virus MINT*, *MSigDB Oncogenic Signatures* and others [98, 99]. This category also includes *Achilles fitness* library, which represents context-specific genetic dependencies, by cataloging genes that affect the survival of certain cell lines. Such genes are identified, using RNAi knockout system and later CRISPR-Cas9 [100]. Differential expression is also analyzed in different cell types and Enrich calculates enrichment scores using *Human Gene Atlas*, *Cancer Cell Line Encyclopedia*, *Tissue Protein Expression* and other libraries. One more category, called miscellaneous, consists of *Chromosome Location*, *HomoloGene*, *Pfam InterPro Domains* and a few other libraries. Some of these libraries are derived from other tools, but some are available only in Enrichr [98, 99].

In a present project GSEA using Enrichr web-tool was applied two times: first time in PART I for overlapping genes, following each of two regulation patterns (‘Gain of methylation — upregulated expression’ and ‘Gain of methylation — downregulated expression’), and second time in PART III for genes, following only those two patterns.

2.5. Genome browser

UCSC (the University of California, Santa Cruz) Genome browser is a database and genome annotation display tool, first released in 2001 (http://genome.ucsc.edu) [78]. This web-based tool allows to visualize various tracks with specific annotations, such as gene sequences (for example, GENCODE v24 and NCBI RefSeq), CpG islands, GC percent, integrated regulation from ENCODE, DNase I hypersensitivity clusters from ENCODE, RepeatMasker, gene expression information (such as Genotype-Tissue Expression (GTEx)), single nucleotide polymorphisms (SNPs) and others [78, 101]. The browser also has a function to upload custom tracks (for example, using BED format), which was used in this project to display differentially methylated and non-differentially methylated positions for top 10 most significantly hypermethylated genes in each of two (‘Gain of methylation — upregulated expression’ and ‘Gain of methylation — downregulated expression’) regulation patterns.

UCSC Table browser is a tool that allows to download Genome browser graphical-based data as a text [79]. The option is very beneficial when data is needed to perform additional calculations, which was used in this project to calculate the distances between TSSs and methylated positions.

2.6. DBTSS

DataBase of Transcriptional Start Sites (DBTSS) is a database, containing information about exact positions of TSSs in genomes of various human tissues and/or cell lines and other species (https://dbtss.hgc.jp) [80, 81]. TSS positions vary in different tissues, which plays a role in gene transcription regulation. DBTSS was used to visualize TSS for top 10 most significantly hypermethylated genes in each of two regulation patterns: ‘Gain of methylation — upregulated expression’ and ‘Gain of methylation — downregulated expression’.

The data in DBTSS is obtained using a technique developed by the authors of the database — TSS-seq. DBTSS contains 491 million TSS tag sequences for 20 human adult and embryonic tissues and 26 lung cancer cell lines in addition to 7 other different cell lines. Database also integrates RNA-seq, BS-seq, ChIP-seq data of histone modifications and the binding of RNA polymerase II, as well as binding of several TFs in cell lines and single nucleotide variation (SNV) data [81].

3. RESULTS

3.1. Datasets of DNA methylation in prostate cancer

After analyzing literature and databases, 17 possible DNA methylation in prostate cancer datasets were chosen for further selection (Table 1).

Three datasets (numbers 1, 2 and 11 in Table 1) were selected for this research project [34, 69, 73]. These datasets contain DNA methylation data on more than a hundred samples each, data is publicly available and collected using Illumina Infinium HumanMethylation450 BeadChip (HM450) (datasets 1 and 2) or Infinium HumanMethylation27 BeadChip (27k) platforms (dataset 11). HM450 and 27k were chosen, because they share the part of the probes thus information from three datasets can be connected.

Table 1 **List of possible datasets.** Information was collected by analyzing literature and databases. Datasets, used in the project, marked by *.

Nr.	Dataset	Samples		Platform	Availability
		Cancer	Benign		
1*	**The Cancer Genome Atlas (TCGA) The Cancer Genome Atlas Research Network, 2015**	**503**	**50**	**Illumina Infinium HumanMethylation450 BeadChip array (HM450)**	**TCGA Data Portal [73]**
2*	**Kirby MK et al., 2017**	**73**	**63**	**Methylation HM450**	**GSE76938 [34]**
3	Shiah Y et al., 2016	172	-	Methylation HM450	GSE80685 [102]
4	Babikova EA and Generozov, 2015	12	12	Methylation HM450	GSE74013 [103]
5	Internation Cancer Genome Consortium (ICGC) Bristow R et al., 2012	102	-	DNA methylation array	ICGC Data Portal [104, 105]
6	Ramalho-Carvalho J and Esteller M, 2017	25	5	DNA methylation array	GSE52955 [106]
7	Kim SJ et al., 2010	12	12	Infinium HumanMethylation27 BeadChip (27k)	GSE26319 [107]
8	Boutros PC, 2014	143	-	Methylation HM450	GSE55479 [108, 109]
9	Hovens CM et al., 2013	4	4	Methylation HM450	GSE47915 [110]
10	Kron K et al., 2009	20	-	Agilent ChIP-on-ChIP methylation	GSE15298 [111]
11*	**Absher DM et al., 2010**	**95**	**86**	**Infinium 27k**	**GSE26126 [69]**

12	Jarrard D and Yang B et al., 2012	4	5	Human Encode 384K HG18 methylation array	GSE38982 [112]
13	Börno ST et al., 2012	51	53	Methylation MeDIP-Seq	GSE35342 [113]
14	Aryee MJ and Yegnasubramanian S, 2012	8	4	Methylation HM450	GSE38240 [114]
15	Aryee MJ and Yegnasubramanian S, 2012	18	21	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	GSE38241 [114]
16	Lin PC et al., 2012	13	7	Illumina HiSeq 2000	GSE41701 [15]
17	Goh LK and Sim HG, 2012	87	24 (BPH)	Illumina GoldenGate Methylation Cancer Panel I	GSE39603 [115]

Dataset number 1, shortly TCGA dataset, contains information about 503 PCa and 50 benign tissue samples. What makes this dataset very informative is that DNA methylation analysis and gene expression data was collected on the same samples (497 PCa and 35 benign tissue) [73]. Frozen samples were evaluated by multiple pathologists to determine the ratio of tumor cells in a sample and detect if there is any significant RNA degradation [73].

Dataset number 2, shortly *Kirby* dataset, consists of 73 prostate cancer and 63 benign tissues. Samples were collected at Stanford University Medical Center between 1999 and 2007 from patients undergoing radical prostatectomy. Each sample was evaluated by pathologists to determine the percentage of prostate cancer epithelial cells. DNA methylation data published by Kirby et al. [34].

Dataset number 11, shortly *Absher* dataset, consists of 95 primary prostate tumors and 86 healthy prostate tissue samples [69]. Samples were collected at Stanford University Medical Center and flash-frozen and then evaluated by pathologists as usual to quantify cancerous cells in samples. Specimens with at least 90% of cancerous epithelium were used for DNA and RNA extraction. Samples with no tumor epithelium were labeled as normal [69].

3.2. PART I: DNA methylation and gene expression patterns and their significance

3.2.1. DNA methylation gain and loss in TCGA, *Kirby* and *Absher* DNA methylation datasets

In order to investigate methylation patterns in prostate cancer, information about matching methylation probes in three datasets TCGA, *Kirby* and *Absher* was extracted. 11,375 matching probes were found. 4,557 hypermethylated probes and 1,786 hypomethylated probes in PCa, compared with normal tissue samples, are overlapping between three datasets. These overlapping probes are associated with 3,326 and 1,502 unique genes for DNA methylation gain and loss, respectively. In three datasets, number of probes correlates with the number of genes, associated with them (Figure 14).

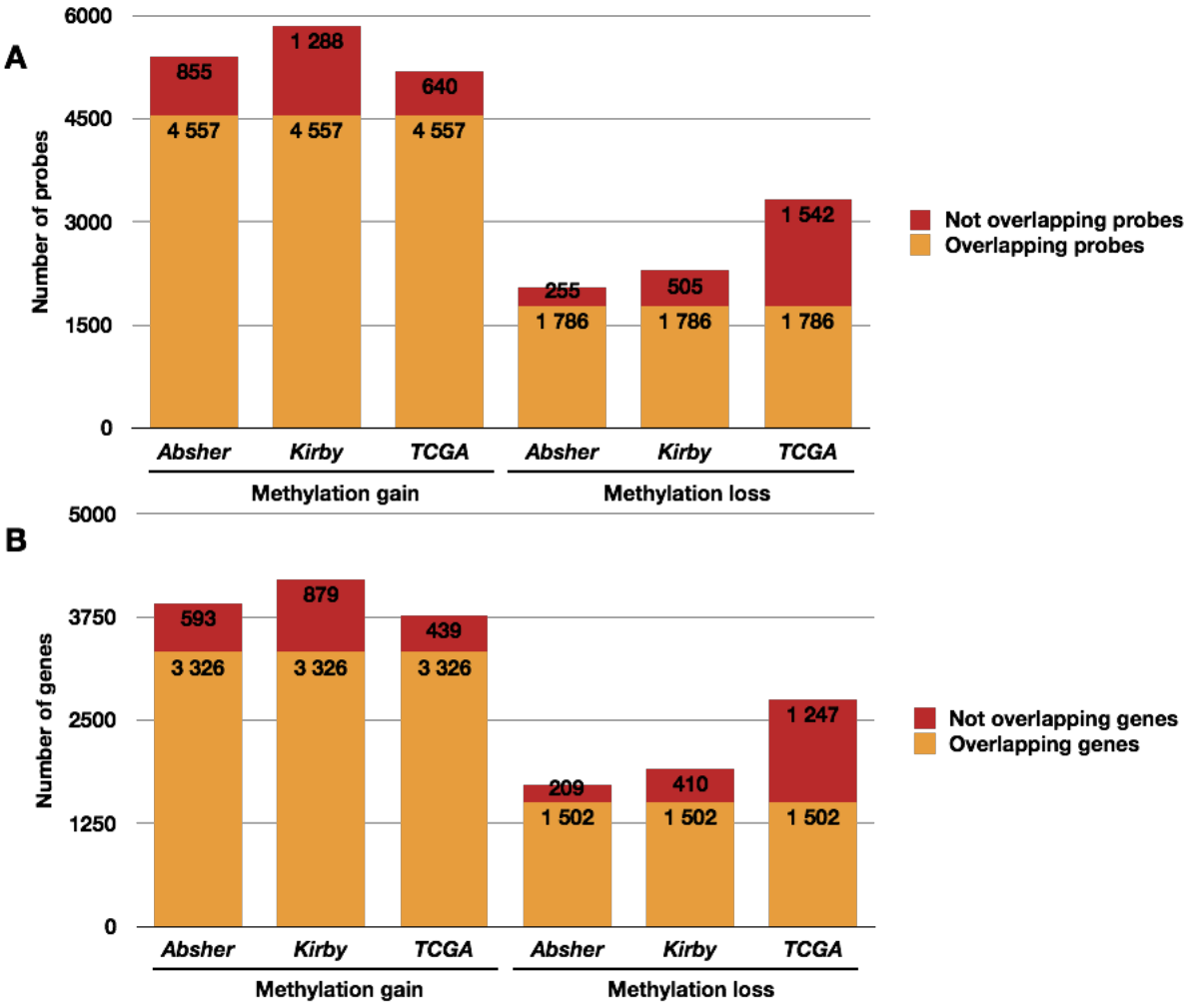


Figure 14 **Number of probes (A) and number of genes (B) with methylation gain or loss in three DNA methylation dataset: *Absher*, *Kirby* and TCGA.** Red indicates not overlapping probes (A) and genes (B), associated with probes, while yellow — overlapping.

In all three datasets there are more probes that gained methylation in PCa, compared with normal tissue, than probes with methylation loss (Figure 14). Although the number of probes and genes with methylation gain are similar, most methylation gain can be observed in Kirby dataset: 5,845 probes out of 11,375 (47,58%). These probes were associated with 4,205 unique genes. Such genes, associated with top 5 most significantly hypermethylated probes, are *RND2*, *SOSTDC1*, *TMLHE*, *MCF2* and *SLC2A2*. The TCGA dataset has the smallest number of hypermethylated probes (5,197). It is 45,69% of all probes and probes are associated with 3,765 genes. 5 most significantly hypermethylated probes in this dataset are associated with genes *CYBA*, *SOSTDC1*, *FLT4*, *GSTP1* and *TPM4*.

Even though *Absher* dataset lays in the middle according to the number of hypermethylated probes, but this dataset contains the smallest number of probes with methylation loss. It has 2,041 hypomethylated probes (17,94% of all probes), which are associated with 1,711 unique genes. Top 5 hypomethylated probes are associated with genes *SCGB2A2*, *GPR160*, *URB*, *DARC* and *ZNF436*. TCGA, on the other hand, has the largest number of probes with methylation loss: 3,328 probes (29,26% of all probes). They are associated with 2,749 unique genes, which is more than a thousand genes more, compared with *Absher* dataset. Top 5 genes with most significantly hypomethylated probes are *DARC*, *ANAPC2*, *ATP4A*, *CFLAR* and *NLRP10*.

SOSTDC1 and *FLT4* are among the genes, associated with top 5 most significantly hypermethylated positions in both *Absher* and *Kirby* datasets, while gene *CYBA* in *Absher* and TCGA. Considering genes, associated with top 5 most significantly hypomethylated probes, gene *DARC* reoccurs in all three datasets, while *SCGB2A2* and *GPR160* in *Absher* and *Kirby* and, finally, gene *CFLAR* in *Kirby* and TCGA datasets.

The smallest number of non-differentially methylated probes has TCGA dataset: 2,850 (25.05% of all probes). This dataset has the largest number of samples, which can affect the statistical significance of the results.

The overlap, consisting of 4,557 hypermethylated probes, is a significant part of all hypermethylated probes in each dataset, as well as genes, associated with these probes. In TCGA overlapping probes make 87.69% of all 5,197 hypermethylated probes, while genes, associated with those probes are 88.34% of 3,765 unique genes. In *Kirby* dataset overlap is 77.96% for probes and 79.10% for genes. In *Absher* dataset it is 84.20% for probes and 84.87% for genes. Additionally, with the hypomethylated probes, 1,786 probes overlap makes 53.67% of all hypomethylated probes, while overlap of 1,502 unique genes is 54.64%

in TCGA dataset. In *Kirby* 77.96% is for probes and 78.56% for genes, while 87.51% for probes and 87.78% are in *Absher* dataset.

3.2.2. DNA methylation and gene expression patterns in TCGA, *Kirby* and *Absher* datasets

After combining DNA methylation data with gene expression data, it was found that 1,058 genes are overlapping between three datasets for 'Gain of methylation — upregulated expression' DNA methylation and gene expression pattern, 1,858 genes overlap for 'Gain of methylation — downregulated expression' pattern. For 'Loss of methylation — upregulated expression' and 'Loss of methylation — downregulated expression' 721 and 544 genes are overlapping, respectively (Figure 15). In order to simplify the presentation of the results, abbreviations for each regulation pattern will be used: UPUP for 'Gain of methylation — upregulated expression', UPDOWN for 'Gain of methylation — downregulated expression', DOWNUP for 'Loss of methylation — upregulated expression' and, lastly, DOWNDOWN for 'Loss of methylation — downregulated expression'.

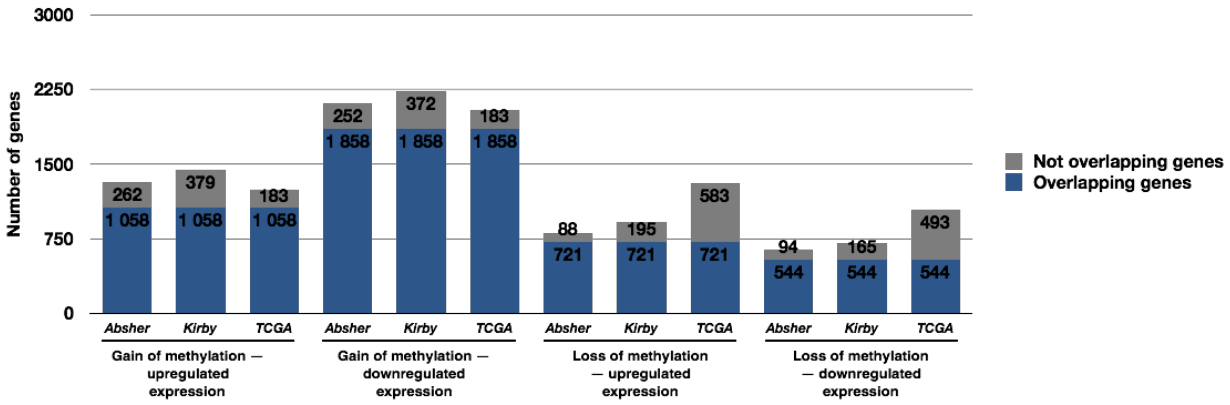


Figure 15 **Number of genes, following 'Gain of methylation — upregulated expression' (UPUP), 'Gain of methylation — downregulated expression' (UPDOWN), 'Loss of methylation — upregulated expression' (DOWNUP) and 'Loss of methylation — downregulated expression' (DOWNDOWN) regulation patterns in *Absher*, *Kirby* and TCGA DNA methylation datasets.** Grey indicates number of not overlapping genes, while blue represents overlapping genes, between three datasets.

The pattern with a biggest share of genes is UPDOWN for all three datasets (Figure 15). *Kirby* dataset has most of the genes — 2,230 (Table S1). A little bit less genes are in UPUP pattern, where *Kirby* dataset, again, has the largest number of genes of all datasets — 1,437 genes. DOWNUP pattern has less genes, compared with previous regulation patterns, except of TCGA dataset. For DOWNUP pattern TCGA has 1,304 genes, which is more than, for example, in UPUP pattern — 1,241 genes. DOWNDOWN pattern has least of the genes, but TCGA here also has an exceptionally large number of genes — 1,037.

Overlap is a big part for all patterns and datasets, with and exception of TCGA for DOWNUP and DOWNDOWN patterns, where overlap smallest — 55.29% and 52.46%, respectively. In other cases, the overlap varies from 73.63% for UPUP pattern in *Kirby* dataset to 91.03% for UPDOWN pattern in TCGA dataset (Table S1).

3.2.3. Consistency of genes, associated with multiple probes

Genes can have multiple probes, associated with them. Gene is named as inconsistent, when probes for the gene are both for gained and lost methylation. 1,243 genes in *Absher*, 1,514 genes in *Kirby* and 1,319 genes in TCGA dataset are associated with multiple probes that are hypermethylated in PCa, while 232, 381 and 575 genes in *Absher*, *Kirby* and TCGA respectively are associated with hypomethylated probes in PCa. 7 genes in *Absher* dataset are associated with both hypermethylated and hypomethylated probes. In *Kirby* dataset it is 16 genes and in TCGA — 21 genes (Table S2). This overlap represents inconsistent genes.

Analysis of genes, associated with multiple probes, in four DNA methylation and gene expression patterns showed that the largest numbers of genes with multiple probes are for UPDOWN pattern (Table 2). *Kirby* dataset has most of those genes — 994. Smallest numbers of genes, associated with multiple probes are for DOWNDOWN regulation pattern, where *Absher* dataset has only 97 of those genes. Numbers of genes with multiple probes correlates with number of all genes for each of four regulation patterns in all three datasets.

TCGA is a dataset with the biggest number of inconsistent genes — 12. However, 12 genes are only 1.34% and 5.04% of all genes with multiple probes for UPDOWN and DOWNDOWN regulation pattern, respectively. The minimum number of inconsistent genes is 3 genes and they can be observed in *Absher* dataset for patterns with upregulated expression. Genes *GNAS* and *OSBPL5* recur in all three datasets for UPUP and DOWNUP regulation patterns, which strongly supports their inconsistency. For patterns with downregulated expression, on the other hand, genes *PEG10* and *SNRPN* are persistent in all three datasets.

Inconsistency is not an issue in three selected DNA methylation datasets, since the number of inconsistent genes is very low. Inconsistent genes make up a small part of all genes, associated with multiple probes: less than 3% for UPUP and DOWNUP regulation patterns, less than 2% for UPDOWN and less than 7% for DOWNDOWN regulation pattern.

Table 2 **Number of genes, associated with multiple probes, in *Absher*, *Kirby* and *TCGA* datasets for each of four DNA methylation and gene expression patterns, and inconsistent genes that are associated with probes that both, gained and lost methylation.** Number in brackets indicates what part of all genes, associated with multiple probes, is inconsistent genes.

		Absher	***Kirby***	**TCGA**
Gain of methylation — upregulated expression (UPUP)	Genes with multiple probes	338	510	416
	Inconsistent genes	3 genes (0.89% of all genes with multiple probes): *GNAS*, *MGMT*, *OSBPL5*	6 genes (1.18%): *ATP10A*, *CD96*, *GNAS*, *MEST*, *NTM*, *OSBPL5*	9 genes (2.16%): *C22orf45*, *CHFR*, *GNAS*, *MED12L*, *MGMT*, *NTM*, *OSBPL5*, *PRH1*, *PRR4*
Gain of methylation — downregulated expression (UPDOWN)	Genes with multiple probes	748	994	893
	Inconsistent genes	4 genes (0.53%): *BCL2*, *CCND1*, *PEG10*, *SNRPN*	10 genes (1.01%): *CCND1*, *IGF2*, *INS-IGF2*, *PEG10*, *RAB32*, *RUNX3*, *SEMA3B*, *SGCE*, *SNRPN*, *ZIM2*	12 genes (1.34%): *BCL2*, *C21orf29*, *GNASAS*, *IGF2*, *INS-IGF2*, *KCNQ1*, *MEG3*, *NTRK1*, *PEG10*, *RUNX3*, *SGCE*, *SNRPN*
Loss of methylation — upregulation of expression (DOWNUP)	Genes with multiple probes	133	216	330
	Inconsistent genes	3 genes (2.26%): *GNAS*, *MGMT*, *OSBPL5*	6 genes (2.78%): *ATP10A*, *CD96*, *GNAS*, *MEST*, *NTM*, *OSBPL5*	9 genes (2.73%): *C22orf45*, *CHFR*, *GNAS*, *MED12L*, *MGMT*, *NTM*, *OSBPL5*, *PRH1*, *PRR4*
Loss of methylation — downregulation of expression (DOWNDOWN)	Genes with multiple probes	97	161	238
	Inconsistent genes	4 genes (4.12%): *BCL2*, *CCND1*, *PEG10*, *SNRPN*	10 genes (6.21%): *CCND1*, *IGF2*, *INS-IGF2*, *PEG10*, *RAB32*, *RUNX3*, *SEMA3B*, *SGCE*, *SNRPN*, *ZIM2*	12 genes (5.04%): *BCL2*, *C21orf29*, *GNASAS*, *IGF2*, *INS-IGF2*, *KCNQ1*, *MEG3*, *NTRK1*, *PEG10*, *RUNX3*, *SGCE*, *SNRPN*

3.2.4. Gene set enrichment analysis for overlapping genes

Genes from UPUP and UPDOWN regulation patterns were selected for further analysis with a focus on UPUP pattern. 1,058 overlapping genes for UPUP were investigated using gene set enrichment analysis (GSEA) tool Enrichr.

Top 5 most significant *GO Molecular Functions* for this gene set included 'transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding' (adjusted $p < 0.05$, combined score $c = 38.43$), 'RNA polymerase II core promoter proximal region sequence-specific DNA binding' ($p < 0.05$, $c = 32.70$), 'RNA polymerase II transcriptional activity, metal ion regulated core promoter proximal region sequence-specific binding' ($p < 0.05$, $c = 27.52$), 'RNA polymerase II distal enhancer sequence-specific DNA binding' ($p < 0.05$, $c = 24.91$) and 'G-protein coupled peptide receptor activity' ($p < 0.05$, $c = 23.03$). Less significant results have combined enrichment score lower than 23. Two clusters can be distinguished for *GO Molecular Function* category: one, related to transcriptional activity together with RNA polymerase activity, and second, related to G-protein coupled receptor activity (Figure 16).

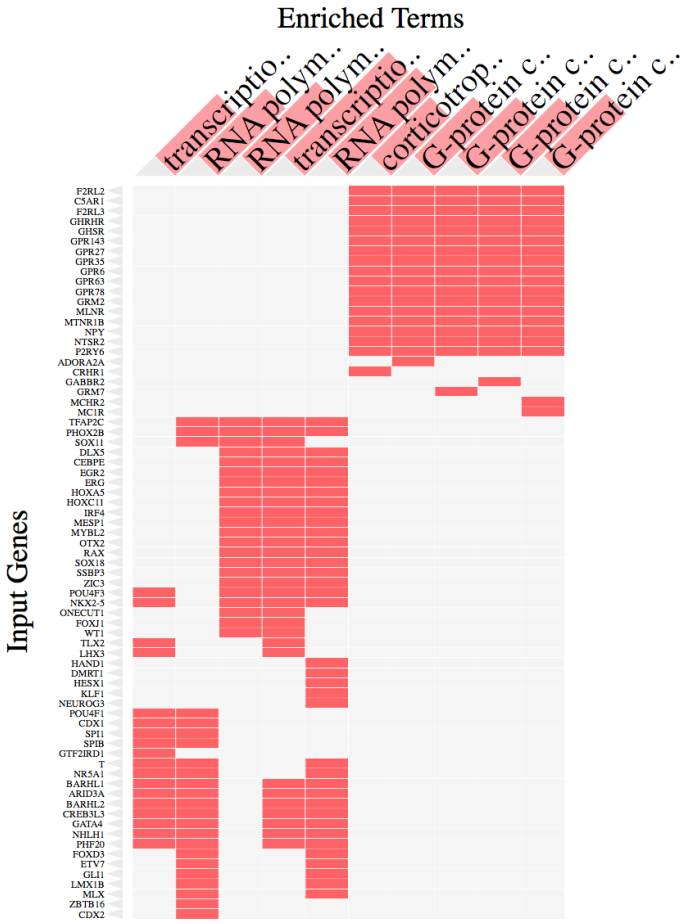


Figure 16 **Clustergram of enriched terms from *GO Molecular Function* category with genes, following 'Gain of methylation — upregulated expression' (UPUP) pattern, as an input.** Two clusters can be distinguished: transcription and RNA polymerase activity related, and G-protein coupled receptor activity related.

ENCODE TF ChIP-seq and *ChEA* categories revealed gene set relationships with proteins SUZ12 and EZH2 (adjusted $p < 0.001$). In category *ENCODE TF ChIP-seq* category

combined enrichment scores are c = 116.01 and c = 61.79 for SUZ12 and EZH2, respectively. In *ChEA* SUZ12 in results appears four times among top 10 results with combined scores c = 139.68, c = 121.59, c = 90.88 and c = 86.05. EZH2 appears two times with c = 104.84 and c = 82.00. Both proteins are components of Polycomb group (PcG) complex, which is responsible for methylation of histones (H3K9me and H3K27me). Results from *Epigenomic Roadmap HM ChIP-seq* and *ENCODE Histone Modification* strengthened previous observations by significantly associating gene set with H3K27me3 histone modification ($p < 0.05$).

Additionally, Enrichr analysis returned a result for *Achilles fitness decrease* specifically in prostate. However, combined score is only 17.39, but adjusted p-value is significant ($p < 0.05$). Furthermore, the most significant result for *KEGG 2016* category was 'neuroactive ligand-receptor' (adjusted $p < 0.001$, combined score 57.23).

1,858 genes, recurring in all three datasets for the UPDOWN pattern, were used for GSEA as an input. Generally, the results were more statistically significant with higher combined scores than results from genes, following UPUP regulation pattern.

Very differently than in UPUP analysis, most significant UPDOWN results from categories *GO Molecular Function* and *GO Biological Process* are associated with extracellular matrix or potassium and calcium ion channel activity. Top 5 most significant ($p < 0.001$) molecular functions are 'Integrin binding involved in cell-matrix adhesion' (c = 45.88), 'integrin binding' (c = 43.72), 'delayed rectifier potassium channel' (c = 38.57), 'voltage-gated potassium channel activity' (c = 37.78) and 'calcium ion sensor activity' (c = 36.50), while 'extracellular matrix disassembly' (c = 147.65), 'extracellular matrix assembly' (c = 147.01), 'collagen fibril organization' (c = 141.77), 'fibronectin fibril organization' (c = 141.21) and 'basement membrane organization' (c = 140.60) are top 5 most significant ($p < 0.001$) biological processes.

ENCODE TF ChIP-seq and *ChEA*, as in previous analysis with genes from UPUP regulation pattern, also revealed statistically significant ($p < 0.001$) gene set relationships with proteins EZH2 (five results of top 10 in *ENCODE TF ChIP-seq* category with combined scores c = 92.67, c = 75.93, c = 43.41, c = 40.56 and c = 38.59) and SUZ12 (two results out of top 10 in *ENCODE TF ChIP-seq* category with c = 75.41 and c = 37.62). Furthermore, categories *ENCODE Histone Modification* and *Epigenomics Roadmap HM ChIP-seq* also showed a strong and more statistically significant association with H3K27me3 histone modification ($p < 0.001$), compared with genes from UPUP regulation pattern.

However, differently than with the UPUP genes, *KEGG 2016* category as a most significant result returned 'Pathways in cancer' (adjusted $p < 0.001$, $c = 62.74$), which is important, since input genes are the genes, that are differentially methylated and expressed in prostate cancer. Second result 'neuroactive ligand-receptor interaction' matched with the first result in analysis of UPUP gene set ($p < 0.001$, $c = 62.61$). The third hit 'focal adhesion' can be connected with previous results from *GO Molecular Function* and *GO Biological Process* categories ($p < 0.001$, $c = 58.69$).

3.3. PART II: Sample-to-sample comparison of gene expression and DNA methylation of genes in 'Gain of methylation — upregulated expression' regulation pattern

In order to confirm or deny the hypothesis that there is an overcompensation in expression of genes that follow the pattern 'Gain of methylation — upregulated expression' (UPUP), DNA methylation and gene expression data, retrieved from the same 532 samples (497 PCa and 35 normal tissue samples) from TCGA dataset, were used for analysis. 1,772 probes with gained methylation and upregulated expression from PART I analysis are associated with 1,501 unique genes in TCGA *M+E* dataset. 1,013 gene names matched in both, methylation and expression, parts of TCGA *M+E* dataset and were used in further analysis.

After sorting methylation values from lowest to highest and then matching the expression values, according to the samples, arrays of values were divided into two groups for normal tissue samples and PCa samples. 497 cancer samples were divided into 248 samples with lower methylation and 249 samples with higher, while 35 normal samples were divided into 17 with lower methylation and 18 with higher. The average methylation and corresponding expression levels in groups of samples were then calculated and compared.

For cancer samples, expression levels of 442 genes (43.63% out of all 1,013 analyzed genes) were corresponding to methylation levels fittingly to the UPUP regulation pattern — average expression was higher in the sample group with higher average methylation and lower in the sample group with lower DNA methylation. On the other hand, data from other 571 genes (56.37%) correlated in an opposite manner — lower expression is observed in the higher methylation sample group and higher expression — in the group with lower methylation. For normal samples, expression of 511 genes (50.44%) was lower in the group of samples with higher average methylation, while expression of 495 genes (48.86%) is higher, when DNA methylation was correspondingly higher.

Additionally, after comparing two groups of PCa samples and two groups of normal tissue samples, it was found that expression of 275 genes (27.14% of all 1,013 genes) is higher in higher DNA methylation sample group both in PCa and normal tissue samples. Yet expression of 164 genes (16.19%) is higher when DNA methylation is higher only in cancer samples and the trend does not last in normal samples. However, it is important to consider that the number of normal tissue samples is significantly lower, compared with PCa samples.

Overall, the findings show no clear sample-to-sample correlation of DNA methylation and gene expression.

3.4. PART III: Analysis of 'Gain of methylation — upregulated expression' regulation pattern and comparison with 'Gain of methylation — downregulated expression' pattern

3.4.1. Genes, following only UPUP and UPDOWN regulation patterns

23,265 probes, associated with 1,058 unique genes, following 'Gain of methylation — upregulated expression' (UPUP), were investigated by distinguishing them according to their methylation status. 9,569 of 23,265 probes gained methylation, while 5,131 probes lost. These probes again can be associated with 1,014 and 864 unique genes, respectively. 856 genes are shared, which means that they have probes with both methylation gain and loss associated with them. More importantly, 158 genes can be distinguished that are associated with probes (1,560), which are only hypermethylated or non-differentially methylated. This set of genes is taken as a representation of UPUP-only regulation pattern. 127 genes out of 158 (80.38%). have more than a half of their probes significantly hypermethylated. Top 10 genes, with the most significant methylation changes in this set, are (from most significant to least): *TNFSF4*, *GDAP1L1*, *CD226*, *MGP*, *CPT1B*, *LTK*, *KLK13*, *GSC*, *TSPAN16* and *ZAR1*.

On the other hand, for 1,858 genes from 'Gain of methylation — downregulated expression' (UPDOWN) pattern there are 18,399 probes of total 41,221 with methylation gain and 9,544 probes with methylation loss. These probes can be again associated with 1,787 and 1,548 genes, respectively. 1,541 unique genes are associated with both upmethylated and downmethylated probes. For this pattern there are 246 genes that have probes only with gain of methylation (or non-differentially methylated) and this set of gene is representing UPDOWN-only regulation pattern. 205 genes out of 246 (83.33%) have more than half of the probes significantly upmethylated, similarly to the UPUP-only genes. Top 10 most significantly methylated genes in UPDOWN-only set are (from most significant to

least): *PLA2G3*, *WFDC2*, *SPARCL1*, *RGN*, *TMEM106A*, *COL3A1*, *KLF8*, *LTC4S*, *LTC4S*, *EFS* and *SCGB3A1*.

The average p-values of probes, associated with genes, representing each of two patterns, were compared. Figure 17 shows that top 10 genes from UPDOWN pattern have more significant methylation changes than genes, fitting only UPUP. The trend persists when visualizing the average p-values from all the genes, following each of the regulation pattern (Figure 18).

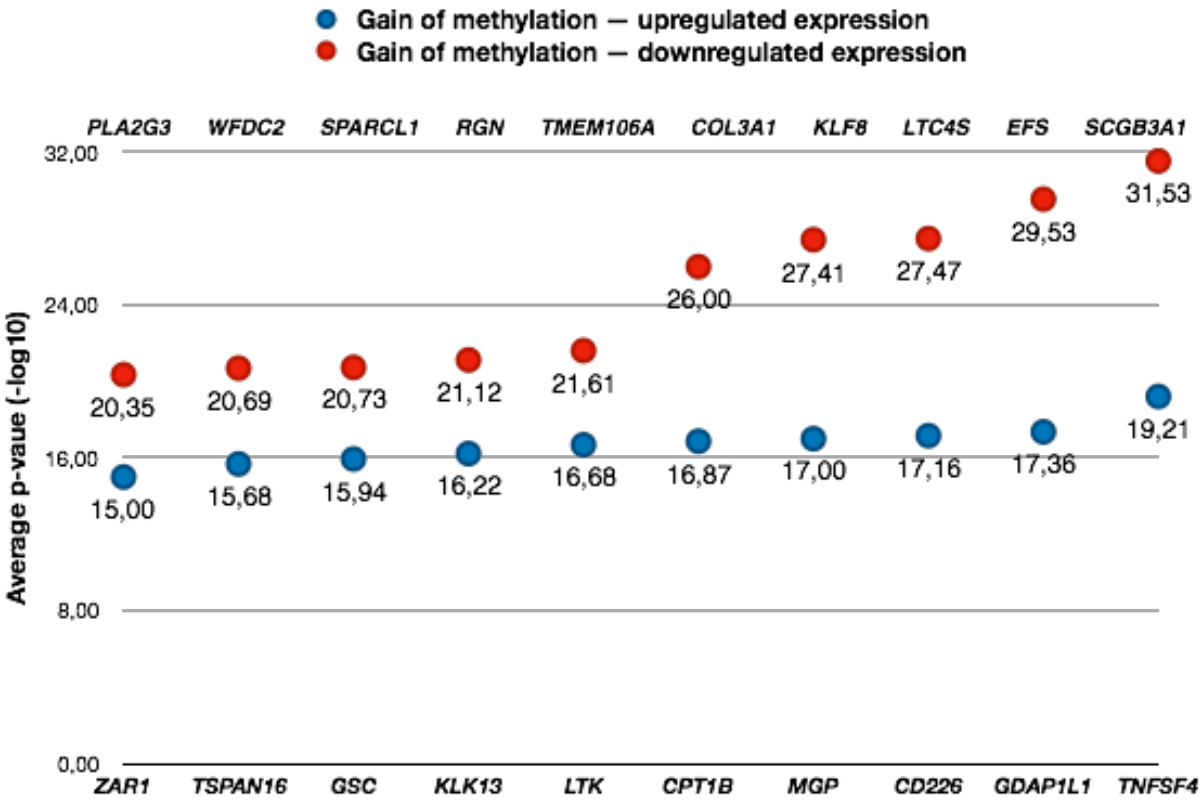


Figure 17 **The average p-values of probes, associated with top 10 genes, following only 'Gain of methylation — upregulated expression' (UPUP) (blue) and only 'Gain of methylation — downregulated expression' (UPDOWN) (red) regulation patterns.** The names of top 10 genes from UPDOWN pattern are listed (from the least to the most significant) above the graph, each matching the data point below, while genes, following UPUP pattern are listed below, each matching the data point above.

From 158 genes, following only UPUP regulation pattern, the highest average p-value (-log10) is 19.21 for a gene *TNFSF4*, while lowest is 2.06 for a gene *PLA2R1*. Comparing with 246 genes, following only UPDOWN pattern, the highest average p-value is 31.53 for a gene *SCGB3A1*, whereas the lowest is 1.80 for a gene *ANP32E*. It is worth to mention that only two genes from UPDOWN-only pattern have lower average p-values (1.80 and 1.92) than the lowest average p-value (2.06) in UPUP-only regulation pattern, and there are 13 genes with a higher average p-value compared with the highest (19.21) in UPUP-only. The

average of all average p-values for genes in UPUP-only regulation pattern is equal to 7.44, while for genes in UPDOWN it is higher and equal to 8.94.

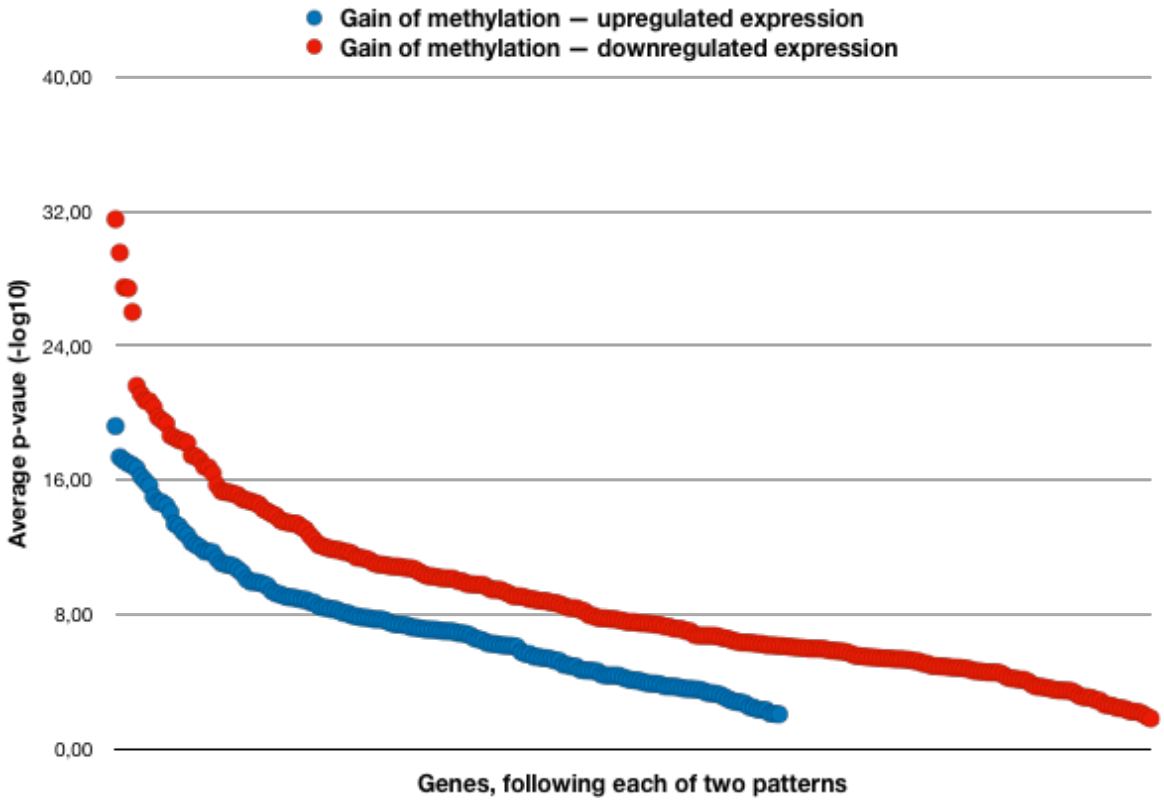


Figure 18 **The average p-values of probes, associated with 158 genes, following only 'Gain of methylation — upregulated expression' (UPUP) (blue) and 246 genes, following only 'Gain of methylation — downregulated expression' (UPDOWN) (red) regulation pattern.** Average p-values for genes, following each pattern are visualized from the most to the least significant.

3.4.2. Visualization of differentially and non-differentially methylated positions in Genome Browser

Differentially and non-differentially methylated probes were visualized using Genome Browser in order to see how they are distributed in the context of gene and other noteworthy features, such as CGIs. Top 10 genes, listed in a previous paragraph, with most significantly hypermethylated probes were visualized for UPUP-only and UPDOWN-only regulation patterns.

The distribution of probes in UPDOWN-only pattern can be distinguished in two different groups. The first group includes genes with all significantly hypermethylated probes that are usually clustered together around TSS (genes *WFDC2* and *PLA2G3*). Gene *WFDC2* has a methylated region that is overlapping with CGI and TSS (Figure 19). Least significant position Pos11 stands out from the cluster.

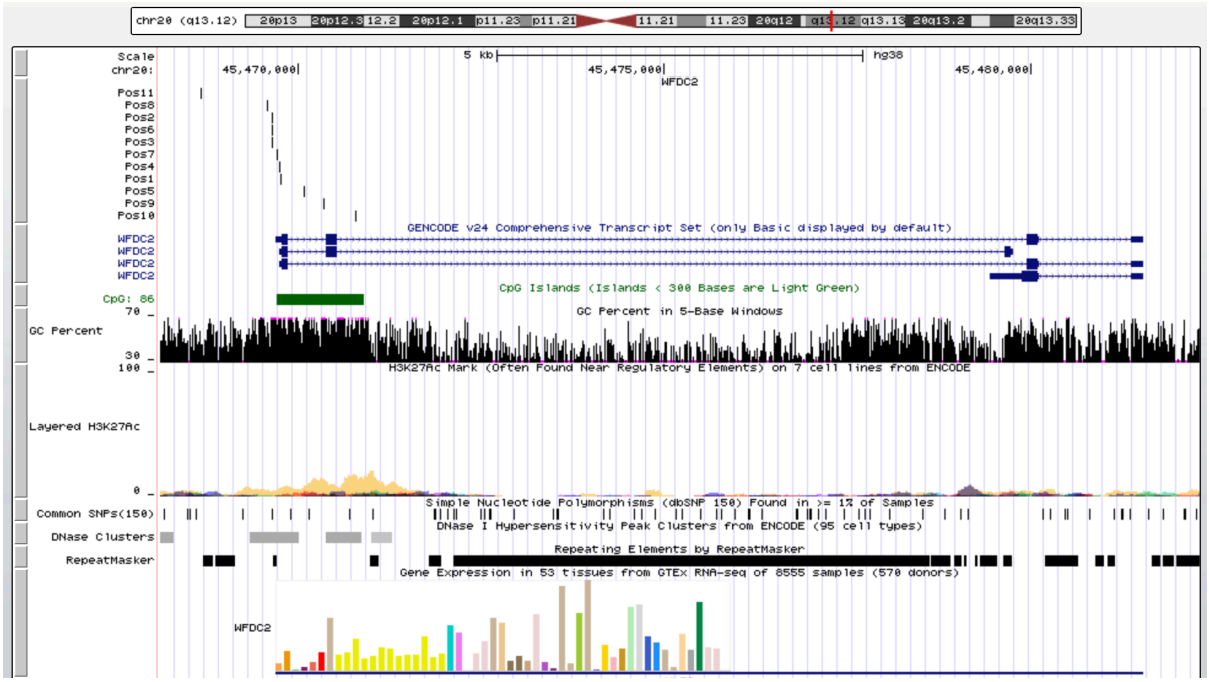


Figure 19 **Visualized methylation probes for a gene *WFDC2* from UPDOWN-only regulation pattern.** Pos1-11 stand for differentially hypermethylated positions from most (Pos1) to least (Pos11) significantly methylated.

The second group includes genes with mostly significantly hypermethylated probes clustered together around TSS also and just a few non-differentially methylated positions that are usually more scattered (genes *SPARCL1*, *RGN*, *TMEM106A*, *COL3A1*, *KLF8*, *LTC4S*, *LTC4S*, *EFS* and *SCGB3A1*). This group of genes usually have a cluster of hypermethylated positions overlapping with a CpG island. Example of such gene can be *RGN*, which is presented in Figure 20. Here all 6 differentially hypermethylated positions are clustered around TSS and overlapping with a CpG island. In addition, region overlaps with DNase I hypersensitivity cluster. GTEx RNA-seq data shows highest expression of *RGN* gene product in adrenal gland. Non-differentially methylated positions are scattered around methylated cluster. *RGN* can be a typical example of UPDOWN pattern. Distribution of probes for gene *KLF8* is very similar to RGN and is shown in Figure S1. This gene has 15 hypermethylated position overlapping with a CpG island in TSS.

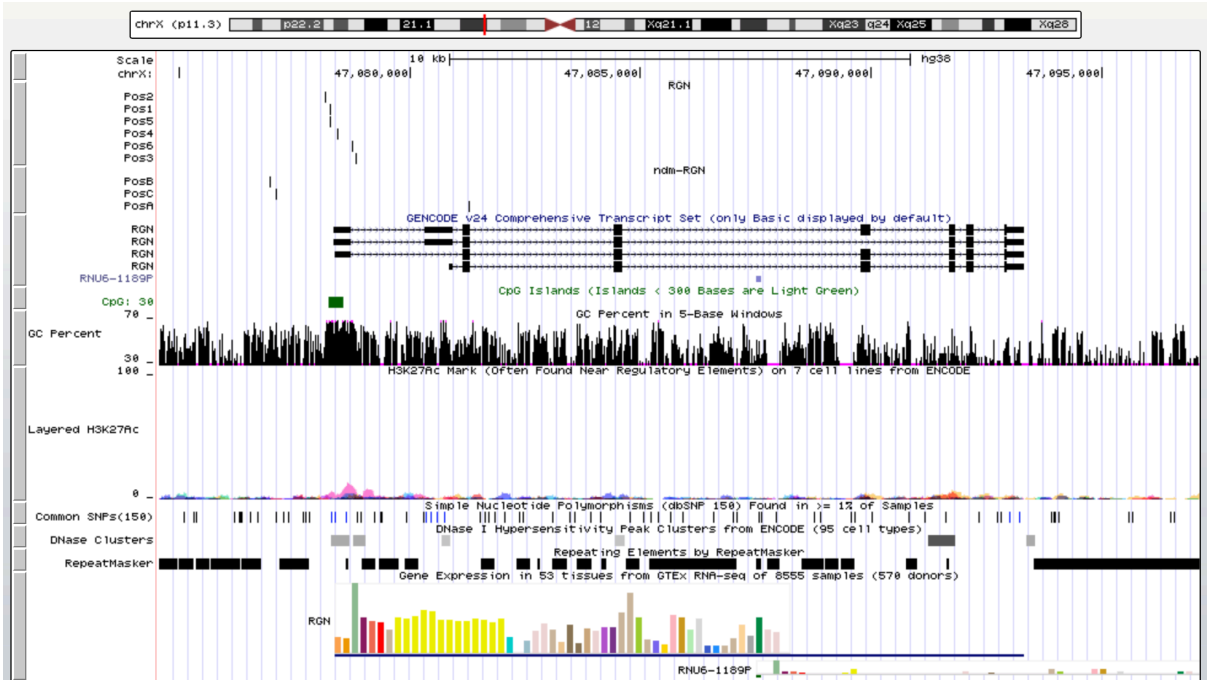


Figure 20 **Visualized methylation probes for a gene *RGN* from UPDOWN-only regulation pattern.** Pos1-6 stand for differentially hypermethylated positions from most (Pos1) to least (Pos6) significantly methylated. PosA, B and C are non-differentially methylated positions.

The distribution of the probes for genes, following UPUP-only pattern, is much more complex. None of the top 10 genes have all probes significantly hypermethylated — all genes have at least one non-differentially methylated probe, associated with it. Nonetheless, one group of genes is similar to the UPDOWN, because they have a cluster of hypermethylated positions around TSS and a few non-differentially methylated positions scattered alongside the gene. This group includes genes, such as *TNFSF4*, *GDAP1L1*, *CPT1B*, *LTK* and *ZAR1*. A view of gene *GDAP1L1* (as well as *TNFSF4*) is very similar to *KLF8* from UPDOWN pattern, but *GDAP1L1* TSS region lacks a CpG island, according to the Genome Browser (Figure S2). However, the region has a high GC percent. Gene *LTK* (*CPT1B* and *ZAR1* as well), on the other hand, has a CpG island, overlapping with TSS and all methylated positions cluster around this region (Figure 21). These features could suggest the regulatory function of the region, as mentioned before. *LTK* also has a DNase I hypersensitivity cluster in TSS.

Gene *TNFSF4* is an interesting example. The gene has a very noticeable alternative TSS and hypermethylation cluster is clearly overlapping with only one of the start sites (Figure 22).

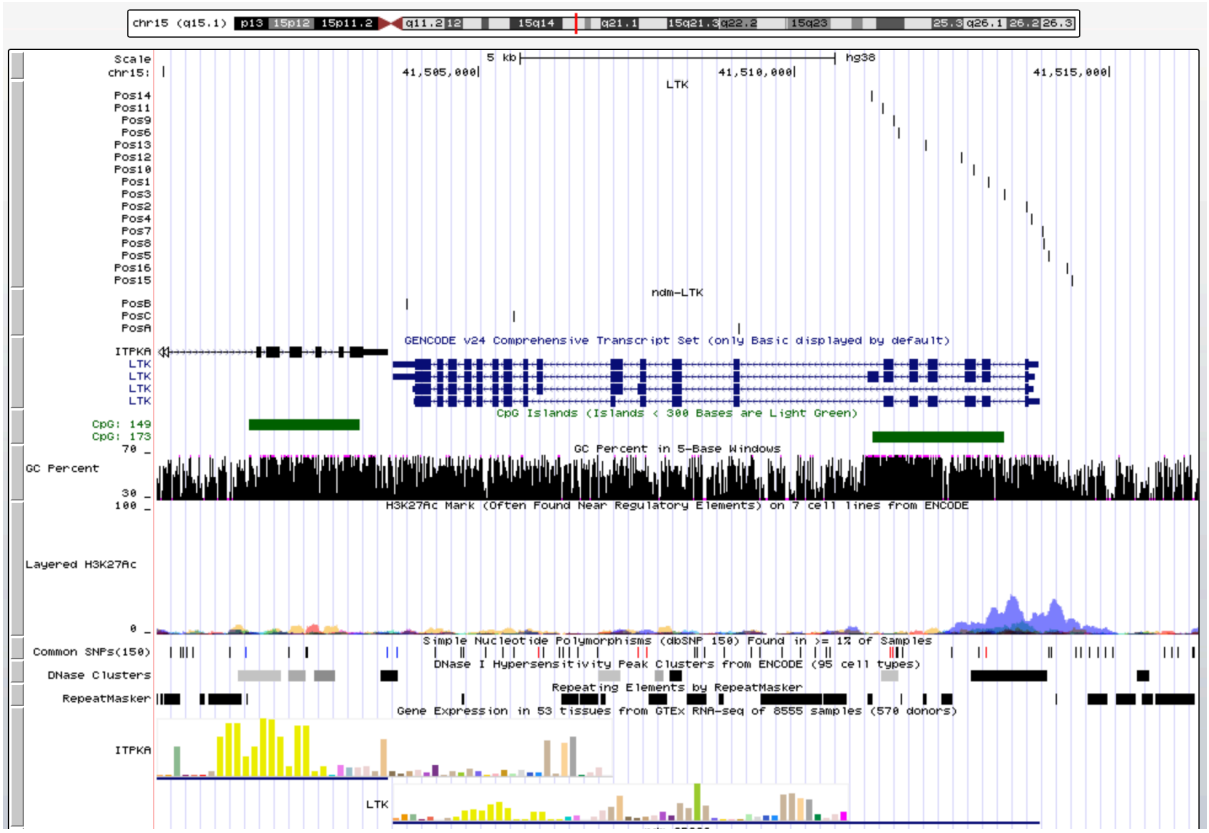


Figure 21 Visualized methylation probes for a gene *LTK* from UPUP-only regulation pattern. Pos1-16 stand for differentially hypermethylated positions from most (Pos1) to least (Pos16) significantly methylated. PosA, B and C are non-differentially methylated positions.

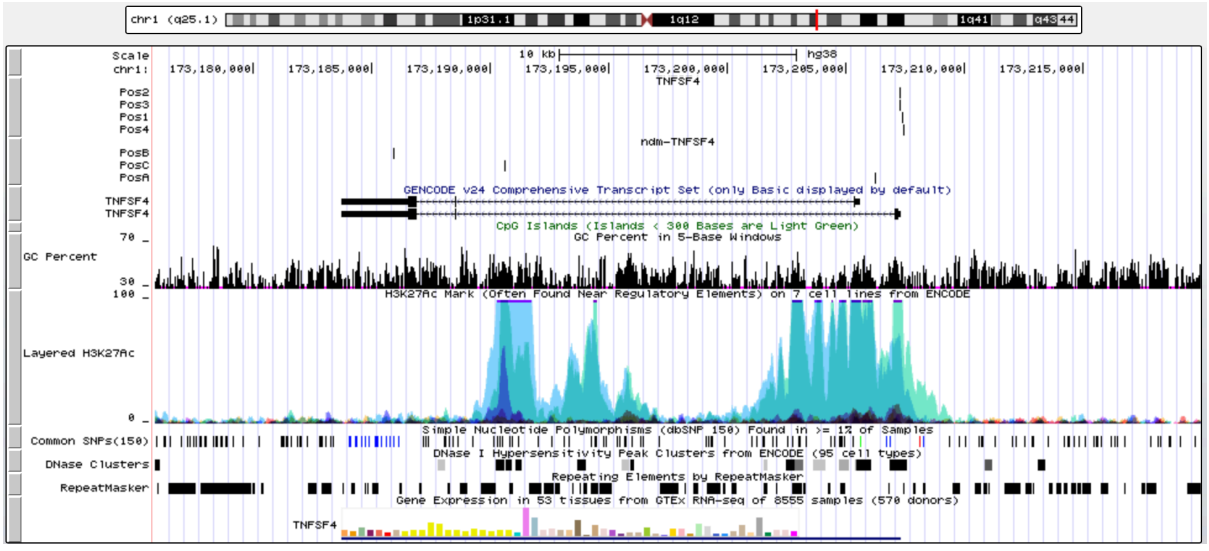


Figure 22 Visualized methylation probes for a gene *TNFSF4* from UPUP-only regulation pattern. Pos1-4 stand for differentially hypermethylated positions from most (Pos1) to least (Pos4) significantly methylated. PosA, B and C are non-differentially methylated positions.

The second group of genes from UPUP-only pattern are genes more non-differentially methylated positions than significantly hypermethylated. Genes *CD226* and *TSPAN16* have only one, but very significantly methylated CpG position and five non-differentially methylated.

Genes *MGP*, *KLK13* and *GSC* could make up a third group, which contains complicated distribution patterns for the methylation probes. For example, gene *GSC* have hypermethylated probes that could be grouped in two clusters with one cluster overlapping with one CGI and another cluster near the second CGI (Figure 23). These two hypermethylated clusters could be related to transcription regulation of the gene, but 7 non-differentially methylated positions are found between two hypermethylation clusters and near TSS.

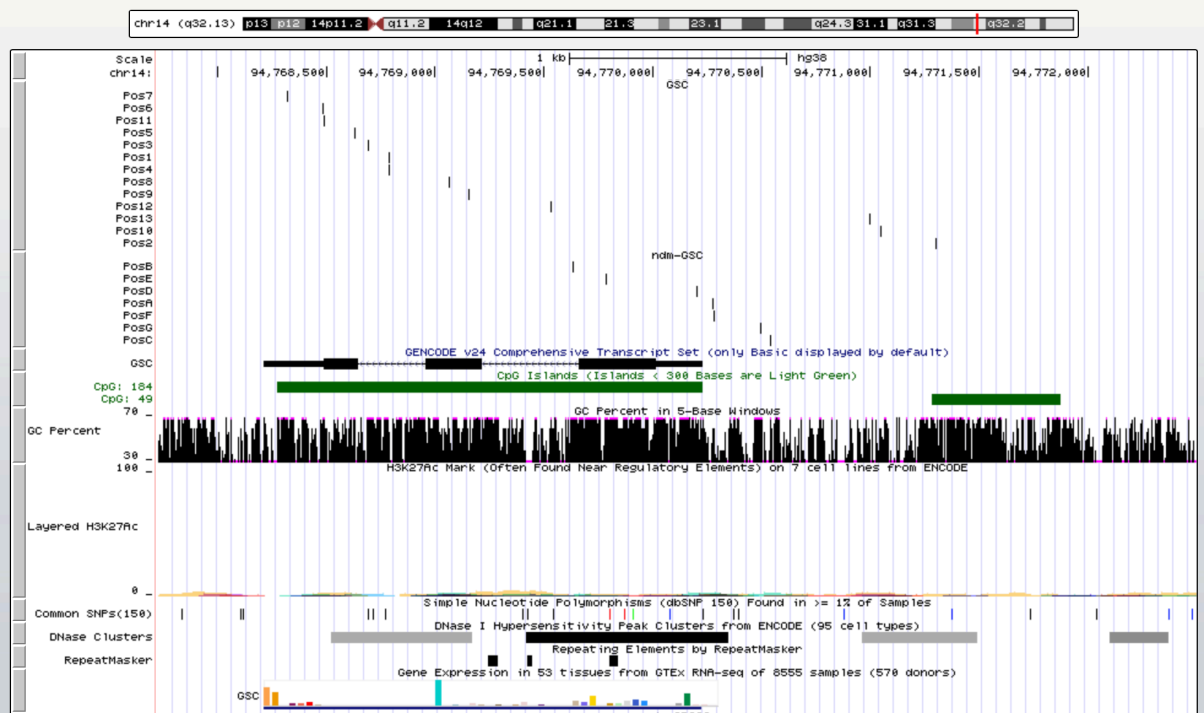


Figure 23 **Visualized methylation probes for a gene *GSC* from UPUP-only regulation pattern.** Pos1-13 stand for differentially hypermethylated positions from most (Pos1) to least (Pos13) significantly methylated. PosA-G are non-differentially methylated positions.

3.4.3. Connection between methylated positions and transcription start sites

The number of transcription start sites for each gene in two patterns of interest varies from 1 to the maximum of 34 (Table S3). Both, UPUP-only and UPDOWN-only, regulation patterns include genes with one, two, three or four TSSs. Most of the genes have only one TSS (75.16% of all UPUP-only genes and 68.70% UPDOWN). Two TSSs have 19.11% and 19.92% UPUP and UPDOWN-only genes, respectively.

In order to calculate the distances between TSS and methylated positions, four closest TSS were picked, if a gene has four or more TSS. Results shows that methylated positions,

associated with genes from UPUP-only pattern, are closer to their TSS, compared with UPDOWN-only pattern (Figure 24).

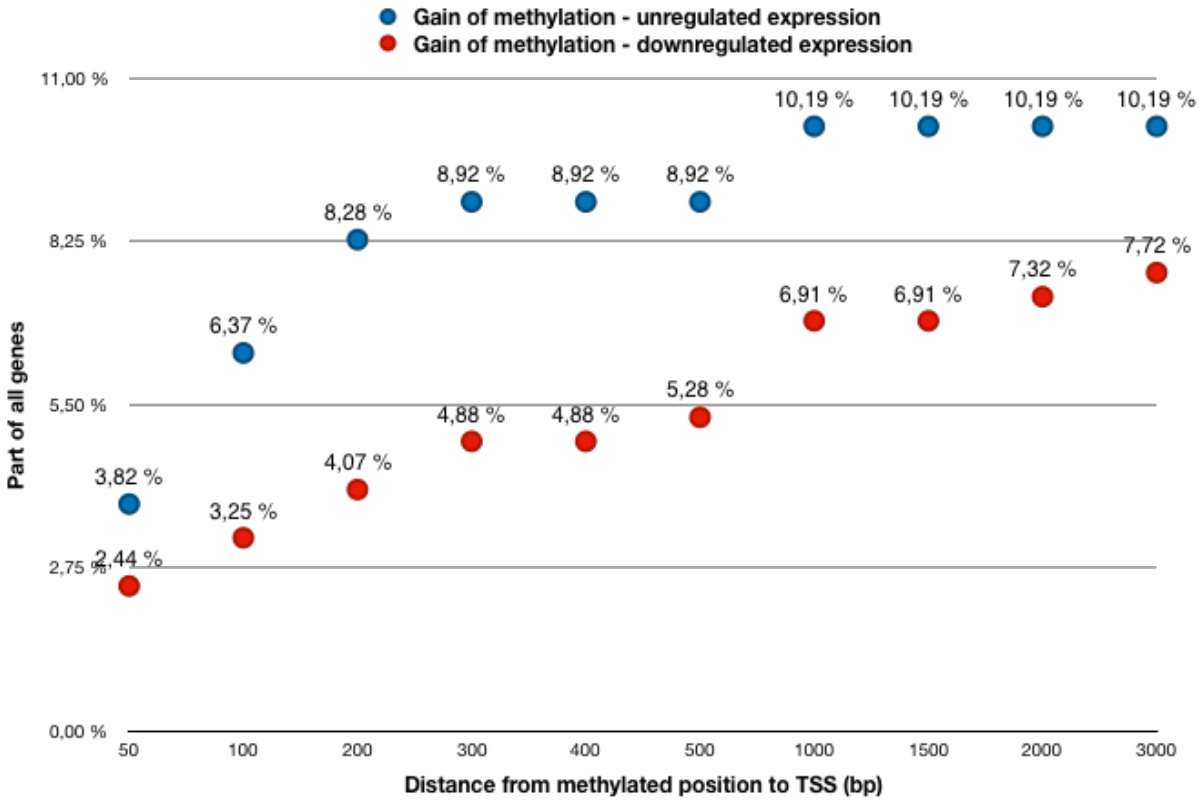


Figure 24 **Part of all genes in 'Gain of methylation — upregulated expression' (UPUP) only and 'Gain of methylation — downregulated expression' (UPDOWN) only patterns that have hypermethylated probes in a distance of 50 to 3000 bp upstream or downstream transcription start site (TSS).**

TSSs of top 10 most significantly hypermethylated genes from UPUP-only and UPDOWN-only were illustrated using DataBase of Transcriptional Start Sites (DBTSS).

For genes, following UPDOWN-only regulation pattern, TSSs in prostate are usually the same as in other tissues. For example, gene *WFDC2* has most frequently used TSSs at the beginning of the gene sequence and overlapping with a CGI (Figure 25). From Genome Browser analysis, it is known that the cluster of hypermethylated positions for this gene is in this particular region (Figure 19). Similar pattern can be observed with a gene *KLF8* (Figure S3) as well as *EFS*, *TMEM106A* and *SCGB3A1*. For genes *LTC4S*, *COL3A1* and *SPARCL1* only difference is the lack of a CpG island in TSS.

With UPUP-only pattern genes TSS distribution is more complex, in comparison with other tissues, and interesting differences in normal prostate tissue can be observed. The view of TSSs of gene *LTK*, for example, is similar to previously explained for genes from UPDOWN pattern, although *LTK* is not expressed in all tissues (Figure S4).

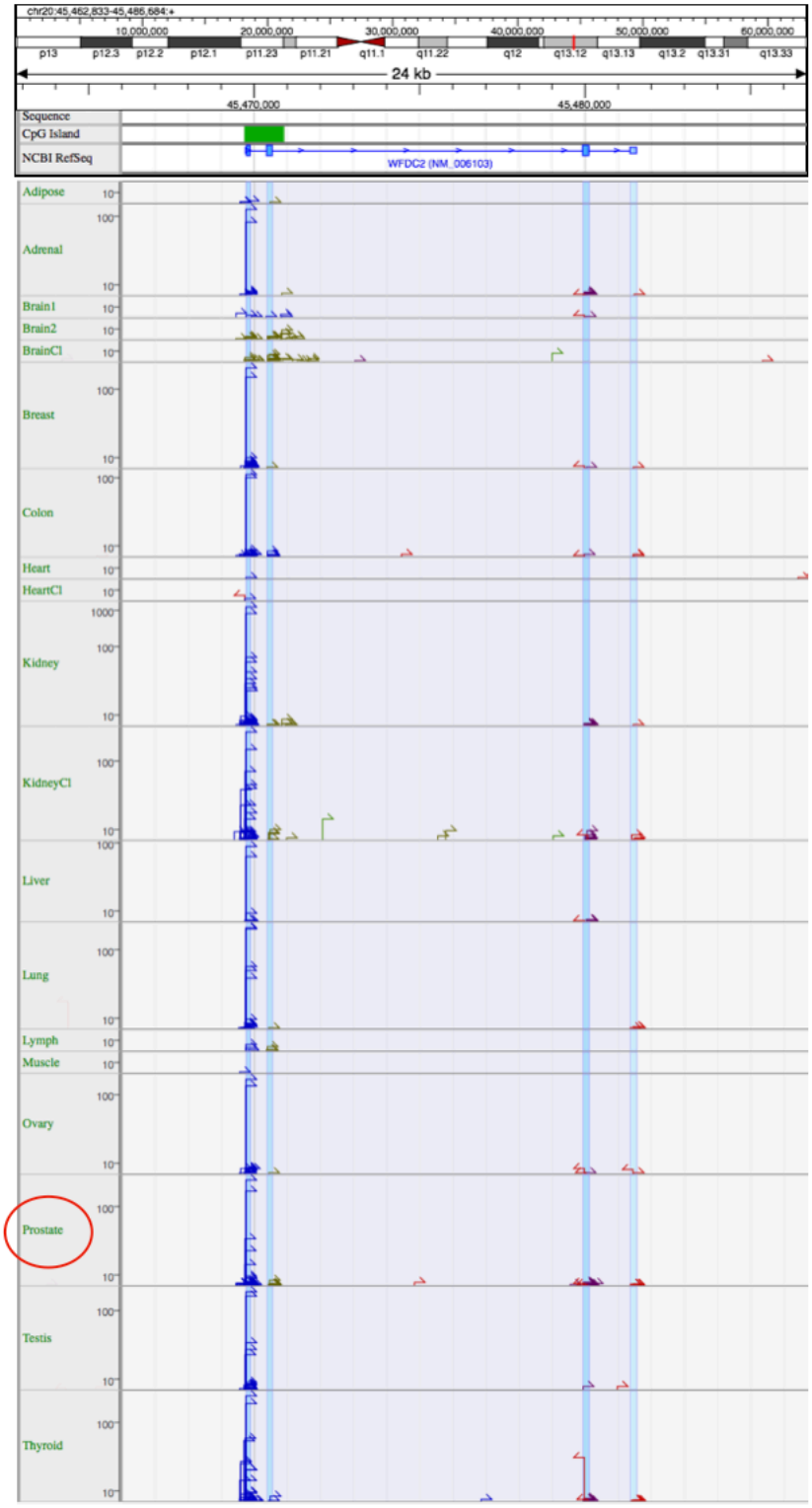


***Figure 25* Transcription start sites (TSSs) for a gene *WFDC2* in various adult human tissues, including prostate.** Upper part of the figure shows position of the gene in genome and CpG island, associated with the gene. Lower part illustrates different TSSs, where the height of the peak corresponds to the frequency of a particular TSS used in transcription.

On the other hand, genes, like *TNFSF4* and *CD226*, have alternative TSSs in prostate, which are not matching with the clusters of methylation probes and are located downstream the gene (Figure 26). In a case of *GDAP1L1* gene, TSS is located downstream (in fact, before

the second exon) in prostate tissue and brain tissue, which means that transcription machinery skips the first exon of the common transcript form (Figure S5).

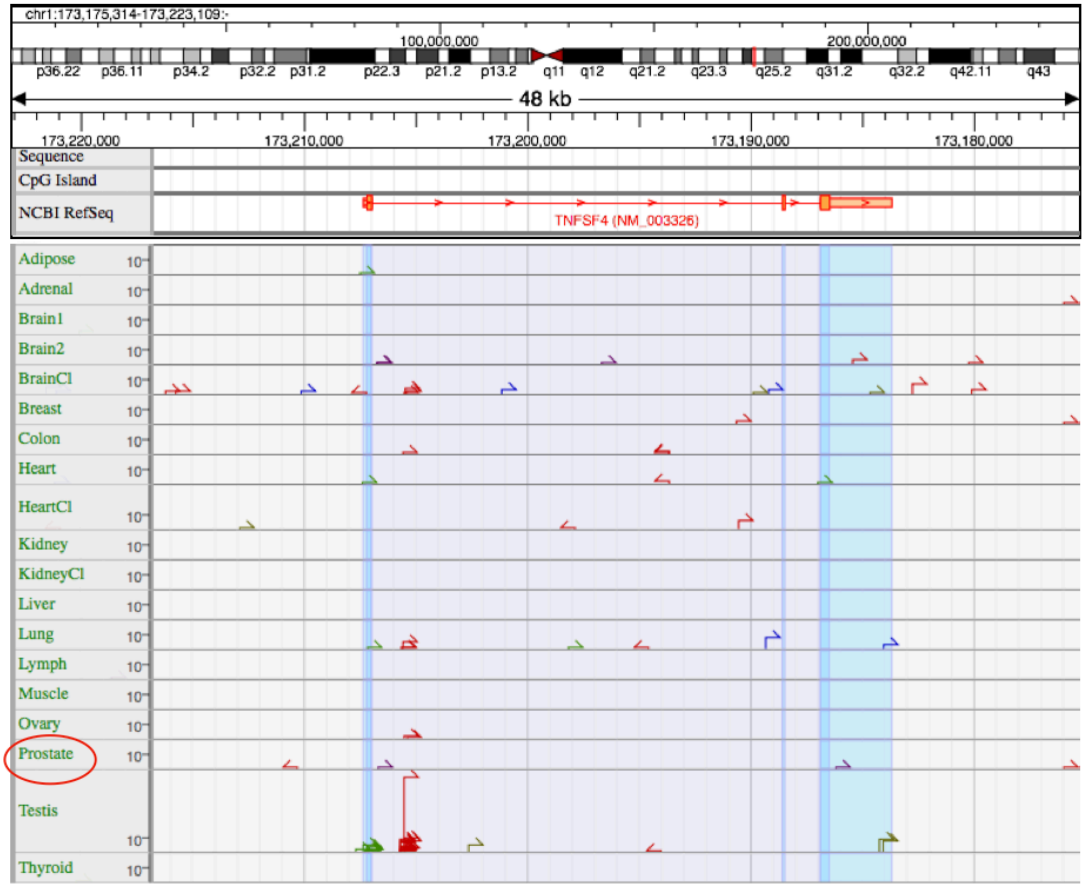


Figure 26 **Transcription start sites (TSSs) for a gene *TNFSF4* in various adult human tissues, including prostate.** Upper part of the figure shows position of the gene in genome. Lower part illustrates different TSSs, where the height of the peak corresponds to the frequency of a particular TSS used in transcription.

3.4.4. Connection between hypermethylated positions and CpG islands

Gene methylation status can also be considered according to where methylated position is located: CpG islands, shores or shelves. 110 out of 158 genes (69.62%), following UPUP-only pattern, all have significantly hypermethylated probes, associated with them and located in islands, shelve or shore. For UPDOWN-only pattern there are 132 out of 246 genes (53.66%) with probes in these regions.

In the context of hypermethylated positions, 1,368 out of all 1,560 probes (88%), from UPUP-only pattern, are located in islands, shores or shelves together. 841 (54%) are in CGIs, 490 (31%) are in shores and 37 probes (2%) are in shelves. On the other hand, for UPDOWN-only pattern there are 1,883 out of all 2,373 probes (79%) in CGIs, shores or shelves together, whereas 1,086 (46%) are in islands, 714 (30%) — in shores and 83 (3%) probes are located in shelves.

There are 9.87 and 9.63 hypermethylated probes per gene on average for UPUP-only and UPDOWN-only patterns, respectively.

3.4.5. Gene set enrichment analysis

A set of 158 UPUP-only genes was used as an input for gene set enrichment analysis in Enrichr.

Top 5 most significant results from *GO Molecular Function* for UPUP-only genes are 'transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding' (adjusted $p < 0.001$, combined score $c = 36.82$), 'RNA polymerase II core promoter proximal region sequence-specific DNA binding' ($p < 0.05$, $c = 33.13$), 'sequence-specific DNA binding' ($p < 0.05$, $c = 31.25$), 'transcriptional activator activity, RNA polymerase II distal enhancer sequence-specific DNA binding' ($p < 0.001$, $c = 28.24$) and 'transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding' ($p < 0.001$, $c = 27.47$). Compared with a previous analysis with all UPUP genes, results with UPUP-only genes are similar — related to the activation of transcription, but with higher enrichment scores. All top 10 results from *GO Biological Process* category were found significant (adjusted $p < 0.05$ with combined score higher than 54.72) and linked to the positive transcription regulation.

For UPUP-only genes, proteins SUZ12 and EZH2 appeared as top results as they did in analysis of all UPUP genes. In *ENCODE and ChEA consensus TFs from ChIP-X* category SUZ12 is the first and the second result with adjusted p-value lower than 0.001, $c = 55.32$ and $c = 22.00$, respectively. EZH2 is a third most significant result with $p < 0.001$ and $c = 21.82$. In a distinct *ChEA* category Polycomb proteins EED and RNF2 are top results ($p < 0.001$, $c = 108.13$ and $c = 105.59$, respectively), which supports relationship with transcription suppression by Polycomb complexes. *Epigenomic Roadmap HM ChIP-seq* and *ENCODE Histone Modification* results, again, showed associations with H3K27me, but significant only in *ENCODE* category ($p < 0.001$).

The first result from *Achilles fitness decrease* is prostate, but for UPUP-only gene set not significantly. However, interestingly, for *KEGG 2016* category the third most significant result is 'Transcriptional misregulation in cancer' ($p < 0.05$, $c = 12.69$).

A set of 246 genes, following only UPDOWN regulation pattern, was used as an input for Enrichr.

Results from *GO Molecular Function* category are more similar to the results from all UPUP and UPUP-only genes than to the results from analysis of all UPDOWN genes, mostly

related to the transcription activity. Top 5 most significant results include ‘RNA polymerase II core promoter proximal region sequence-specific DNA binding’ (adjusted $p < 0.05$, $c = 37.07$), ‘transcriptional activator activity RNA polymerase II core promoter proximal region sequence-specific binding’ ($p < 0.05$, $c = 29.51$), ‘transcriptional activator activity, RNA polymerase II distal enhancer sequence-specific binding’ ($p < 0.05$, $c = 25.48$), ‘chemokine activity’ ($p < 0.05$, $c = 23.70$) and ‘RNA polymerase II core promoter proximal region sequence-specific DNA binding, bending’ ($p < 0.05$, $c = 21.55$). Results, though, are less significant and with lower combined scores. Same tendency stands with GO Biological Process category, where all top 10 most significant ($p < 0.05$, combined scores higher than 60.16) results are related to positive regulation of transcription, repeating UPUP-only results.

ENCODE and ChEA Consensus returned protein SUZ12 as the most significant result (adjusted $p < 0.001$, $c = 65.11$). Other results are much less significant or not significant at all, for example, the second hit is protein NANOG with combined score 14.56. ChEA also returned the protein EED ($p < 0.001$, $c = 55.44$), which was previously seen in UPUP-only gene set analysis. Another Polycomb protein JARID2 also appeared as a significant result two times among top 10 results ($p < 0.001$, $c = 55.20$ and $c = 51.77$). Furthermore, H3K27me3 and H3K9me3 were again the results for *Epigenomics Roadmap HM ChIP-seq* category, but more significant, compared with UPUP-only genes ($p < 0.05$).

However, for UPDOWN-only genes *KEGG* showed no significant results and no associations with prostate in *Achilles fitness* category.

4. DISCUSSION

Majority of the research projects tend to use high throughput DNA methylation data to select gene candidates in order to test their ability to act as biomarkers in different types of cancer, including prostate cancer. Some research groups then compare methylation status of target genes with their gene expression to confirm or deny that DNA methylation changes are correlating with gene expression changes in cancer. For example, Kim et al. used genome-wide DNA methylation and gene expression data to select new target genes and then validated them experimentally [116], while Geybels et al. applied an opposite strategy, validating target genes in TCGA dataset [117]. This study aims to investigate different DNA methylation patterns epigenome-wide in PCa and simultaneously link the changes in methylation with gene expression data. Then, using various computational approaches, genes that fall into groups of different regulation patterns are investigated in order to find an explanation behind the pattern — how can DNA methylation impact gene expression. This is not a common setup for a study of DNA methylation in cancer, which increases the value of the present project.

Furthermore, DNA methylation is generally associated with suppression of gene expression by inhibition of promoter activity [48]. Combined methylome and expression profile studies show correlation between promoter of particular genes DNA methylation and expression suppression in PCa [74, 117, 118]. This was also proved in a present study — UPDOWN regulation pattern has highest number of genes. Although, surprisingly, there was a large number of genes with hypermethylated promoters, but upregulated expression, which disagrees with a classical statement. The study focuses on this non-classical UPUP regulation pattern, proving its robustness and significance in DNA methylation data analysis.

4.1. Data selection

Three DNA methylation datasets, TCGA, *Kirby* and *Absher*, were selected, because they have a large number of samples, both cancer and normal tissue [119]. Although TCGA dataset has a significantly smaller number of normal tissue samples, but dataset is particularly valuable, because not only it has data of DNA methylation, but also gene expression data, derived from the same samples. Furthermore, all three datasets are publicly available for downloading. More importantly, DNA methylation data was collected using Illumina DNA methylation arrays. 27k and HM450 correlates and share a big part of methylation probes (94% of the regions from 27k are included in HM450) and this allows to connect data more easily [41]. Another reason, why this particular *Absher* dataset was included, is that data from

27k helped to filter HM450 data. 27k probes target the promoter regions and changes in promoter DNA methylation status have been associated with changes in gene expression [46].

Lastly, only three out of total 17 datasets found were used in this study (Table 1). Other datasets, such as number 3, 5, 8, 13 and 17, could be used in the future to validate DNA methylation results from this project or conduct similar studies. For example, dataset number 13 is particularly interesting, because sequencing data could disclose the methylation status of those regions that are not covered by 27k or HM450 methylation arrays.

4.2. PART I

Generally, the fact that more than 65% of all matching probes between *Absher*, *Kirby* and TCGA datasets were statistically significantly differentially methylated in prostate cancer, compared with normal tissue samples, shows the importance and frequency of DNA methylation process in PCa. In this study, more methylation gain than loss was observed in PCa samples. These findings support previous studies, where stable hypermethylation of gene promoter regions in prostate cancer has been observed and particularly promoter hypermethylation more than hypomethylation has been associated with differences in gene expression [114]. It has been proposed that promoter CGI-specific DNA hypermethylation (one of the epigenetic hallmarks of human cancers [59, 61]) can occur because of the oxidative damage that forces to form specific silencing complexes, such as PRC4, found only in cancer and stem cells [54]. GSEA analysis could confirm the relationship between both UPUP and UPDOWN gene sets and PRCs, since connections with proteins EZH2 (part of the PRC4, PRC2 and is known to be highly expressed in metastatic PCa [54, 120]), SUZ12 (part of PRC2 complex and its overexpression is known to be involved in carcinogenesis [121]) and histone modification H3K27me3 were made.

The numbers of significant probes and genes, associated with them, are similar in all three datasets (Figure 14 and Figure 15). The overlap of hypermethylated probes and genes is no less than 77%, while for DNA methylation and gene expression patterns the overlap is higher than 73% for patterns with hypermethylation. However, there is an exception with TCGA dataset for patterns with methylation loss (DOWNUP and DOWNDOWN), where number of probes and genes is higher, and overlap is smaller, compared with other datasets. Nonetheless, these results still demonstrate robust changes in DNA methylation in PCa, compared with normal tissue.

Genome-wide studies have shown that downregulation of gene expression in PCa is more common than upregulation, which this project has demonstrated as well with results from UPDOWN and DOWNDOWN regulation patterns (Figure 15) [114]. Despite that the number of genes in non-classical UPUP pattern is large enough in all three datasets to suspect that it can be important in PCa and definitely encourages to explore the pattern in additional analyses. Furthermore, GO from GSEA analysis disclosed substantial differences between molecular functions of UPUP and UPDOWN gene sets, which supports the robustness of UPUP regulation pattern. Thus, further focus was put on UPUP pattern and findings were compared with classical UPDOWN pattern in order to understand possible DNA methylation role in gene expression upregulation.

The question of gene inconsistency in DNA methylation was raised, hypothesizing that a significant number of genes might have both hypermethylated and hypomethylated probes associated with them, resulting in an increased number of genes in UPUP regulation pattern. In this case, gene expression upregulation might have been caused by DNA hypomethylation instead of hypermethylation. Assumption proved to be wrong, when just a small number of genes was found carrying this feature. 12 genes as the maximum was found or no more than 6.21% of all genes with multiple probes in certain pattern were inconsistent (Table 2).

4.3. PART II

Since a surprisingly large number of genes with upregulated expression in UPUP pattern was observed, the question of overcompensation in gene expression data can be raised. The concept of the overcompensation in this study can be described as a possibility that some samples in gene expression dataset might have had an unusually high expression of hypermethylated genes, while in other samples expression of the same genes would be very low. This would lead to the overall upregulation of gene expression in a dataset, because a very high expression “compensates” the low expression. The gain of DNA methylation is then connected with unreasonably increased expression, resulting in UPUP pattern. If this proves to be true, the connections between gene expression and DNA methylation for UPUP pattern would not be robust enough to state that promoter hypermethylation affects the increase in gene expression. Unfortunately, such phenomenon is not discussed in a literature describing gene expression differences. To investigate the possibility of overcompensation DNA methylation and gene expression values from same samples, so called sample-to-sample analysis, was necessary to be able to prove the authenticity of the regulation pattern.

The distribution of UPUP pattern genes between two groups of PCa samples was 44% and 56%, which indicates that almost half of all genes have higher DNA methylation followed by higher gene expression. However, a similar ratio was observed in normal prostate tissue samples (50% and 49%), which could indicate that overcompensation of gene expression does not occur in PCa. Additionally, PCa have unique genes (almost 20% of all investigated genes) with higher expression in the sample group with higher methylation, demonstrating the robustness of UPUP regulation pattern in PCa. DNA methylation is altered together with gene expression in PCa, compared with normal prostate.

This analysis of DNA methylation and gene expression in same PCa and normal tissue samples could be more accurate if samples were divided into more groups, based on levels of methylation. Smaller groups would allow to see smaller differences in gene expression and if it actually correlates with DNA methylation. Furthermore, samples can also be grouped according to the expression levels which then could be compared, but it is difficult to do comparisons for genes with multiple methylation probes. Moreover, the higher number of normal tissue samples could give more robust information about DNA methylation and gene expression patterns in normal prostate.

4.4. PART III

PART III aims to investigate possible explanations behind UPUP pattern — how can DNA methylation be associated with expression upregulation in PCa. First of all, only TCGA methylation dataset was used to extend the list of probes, associated with UPUP and UPDOWN genes. This means that all the probes from HM450 DNA methylation array, excluded in PART I, were added and analyzed in PART III. The reason behind this decision was to better cover promoter regions in order to have a more robust image of methylation in promoter regions. In PART I the number of probes was limited, because 27k data was used and, since this array has only an average of two probes per gene (compared with an average of 17.2 in HM450) [41, 75], it could limit the information about methylation. For instance, a gene can have all probes hypermethylated in 27k data, leading to a conclusion that a gene is hypermethylated in PCa, while in HM450 data the same gene would be associated with hypermethylated, hypomethylated and non-differentially methylated probes, leading to an ambiguous methylation status. Furthermore, extended list of probes allows to understand whether there are other regions involved in regulation. Lastly, for the PART III analysis it was chosen to filter out and use genes only with probes that gained methylation (or have non-differentially methylated probes) to avoid genes, with ambiguous methylation status, since

the list of probes was extended. However, the dissimilarities between methylation of UPUP-only, UPDOWN-only and excluded UPUP, UPDOWN genes could be also explored in order to discover more differences between HM450 and 27k DNA methylation data.

Top 10 genes from each pattern were chosen for further analysis, assuming that they represent the patterns best, because all probes are hypermethylated more significantly compared with all the other genes in the patterns. For example, gene *EFS* and *SPARCL1* from UPDOWN-only pattern were shown to be suppressed in PCa [122, 123]. Moreover, in another epigenome-wide study promoter hypermethylation was demonstrated for genes *TMEM106A* and *SCGB3A1*, where hypermethylation in latter also correlated with lower expression [117]. *ZAR1* from UPUP-only, on the other hand, was demonstrated to actually follow UPUP pattern in neuroblastoma [124] and *MGP* to have high expression in PCa [125].

Visualization of top 10 genes allowed to see instantly, how hypermethylated and non-differentially methylated probes are distributed, if they form a hypermethylated cluster and, finally, how they can be associated with other features, such as CGIs. After seeing that the presence of CpG islands in the promoter region and their overlap with methylated positions are different for different top 10 genes in the images from Genome Browser, the overlap was investigated for all genes, and results were opposing. More UPUP-only genes have hypermethylated positions in CGIs, shores or shelves. However, this contradicts with the literature, where silencing protein complexes were shown to be recruited more to the CGI-promoters, compared with promoters without [54].

CGIs are important for gene expression regulation, but when they are found in the promoter region near TSS, this CpG-rich region is usually a target for DNA methylation leading to the downregulation of gene expression [48]. This study again showed the opposite results, supplementing observation from analysis of the CpG overlaps. UPUP-only pattern actually has more genes, associated with probes that are closer to TSSs (Figure 24). This indicates that there are more reasons (or at least as many as for the classical UPDOWN pattern) to state that gene expression upregulation of genes from UPUP regulation pattern can be influenced by DNA hypermethylation in the promoter region.

Unfortunately, the possible mechanisms behind such connections have not been shown as much as it is for UPDOWN pattern. One possible mechanism could be that the low density of DNA methylation marks in the promoter region is not be sufficient to silence gene expression [57, 58]. The results from this study prove to be opposite, as it is known that UPDOWN-only and UPUP-only both have approximately 10 hypermethylated probes per gene and, as mentioned previously, positions from UPUP pattern are close to TSS and

located in CpG-rich regions. Furthermore, the number of hypermethylated positions in hypermethylated cluster usually is not very different comparing UPUP and UPDOWN genes. On the other hand, gene *KLF8* from UPDOWN pattern, for instance, has 15 hypermethylated positions very close to each other, while *LTK* form UPUP has 16 positions, but with lower density (Figure 21, Figure S1).

In addition, DNase I hypersensitivity clusters are found in promoter regions for most of the top 10 genes for both patterns. It is known to contribute to the regulatory role of the promoter region and, consequently, have an impact for carcinogenesis [126, 127]. Unfortunately, since there are no obvious differences between the distribution of clusters in UPUP and UPDOWN genes, DNase I clusters are not able to explain the differences in gene expression.

One of the most interesting findings from DBTSS analysis was alternative TSSs for genes *TNFSF4*, *GDAP1L1* and *CD226* (Figure 22 and Figure S5). Alternative TSSs could indicate different gene expression setting in prostate cancer. Probes of *TNFSF4* overlapping only with one TSS suggested that upregulation of gene expression can occur as a consequence of switching TSS whilst methylation occurs in the insignificant region. Such phenomenon in PCa has been shown in several studies [128, 129] Later it was confirmed that alternative TSS downstream is indeed being used, but already in a normal prostate tissue (Figure 26), which also discloses the drawback of the HM450 methylation array— probes cover non-specific TSS in prostate and PCa. DNA methylation mechanism for some genes, apparently, might not be PCa-specific and expression of genes with alternative TSSs results in upregulation. Although, since none of the methylation probes are covering an alternative TSS, there is no way to tell if DNA hypermethylation occurs in this specific region in PCa, compared with a normal prostate tissue. The same TSSs of gene *TNFSF4* and gene *GDAP1L1* were also noticed in brain (and brain only), which promotes a hypothesis of an association between gene expression patterns in brain and prostate. For instance, androgen deprivation therapy was shown to have effects on brain function, meaning resemblances of molecular mechanisms [130] and the similarities between expression in brain and testis has been noticed before [131].

Differently than in previous enrichment analysis in PART I, associations between genes from both UPUP-only and UPDOWN-only regulation patterns and protein EED was found. This protein, together with EZH2, is known to be a part of cancer-specific PRC4 and plays a role in cell proliferation [54, 120]. This connection has been observed only here, because the as an input for GSEA pure sets of genes, only with hypermethylation, were used.

CONCLUSIONS

While UPUP pattern is obscure, there is no solid evidence against it and results clearly indicate a potential, which encourages to continue the studies in order to understand the pattern and its mechanisms in prostate cancer. Moreover, it is important to take into account gene expression data alongside DNA methylation and do not make early assumptions about DNA hypermethylation always followed by gene silencing.

Following conclusions for each goal of the project can be formulated:

1. After analysis of the literature 17 promising DNA methylation datasets were found and 3 of them, TCGA, *Absher* and *Kirby*, were selected to be investigated in this study.
2. Four DNA methylation and gene expression patterns were distinguished and genes, overlapping between three datasets, were found. 'Gain of methylation — upregulated expression' (overlapping 1,058 genes), 'Gain of methylation — downregulated expression' (1,858 genes), 'Loss of methylation — upregulated expression' (721 genes) and 'Loss of methylation — downregulated expression' (544 genes).
3. Gene expression overcompensation in a dataset as an explanation for 'Gain of methylation — upregulated expression' regulation pattern is unlikely.
4. Analysis of genes, following only 'Gain of methylation — upregulated expression' revealed that methylated positions from this pattern are more likely to be in CpG islands, shores or shelves and closer to TSSs, compared with positions from 'Gain of methylation — downregulated expression'. Alternative TSSs was proposed as a possible explanation of increased expression for genes *TNFSF4*, *GDAP1L1* and *CD226* or as a limitation of methylation detection systems.

REFERENCES

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F: **Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012**. *Int J Cancer* 2015, **136**(5):E359-386.
2. **Cancer in Norway 2016 — Cancer incidence, mortality, survival and prevalence in Norway** [https://www.kreftregisteret.no/en/]
3. Siegel RL, Miller KD, Jemal A: **Cancer statistics, 2016**. *CA Cancer J Clin* 2016, **66**(1):7-30.
4. Siegel RL, Miller KD, Jemal A: **Cancer Statistics, 2017**. *CA Cancer J Clin* 2017, **67**(1):7-30.
5. Siegel RL, Miller KD, Jemal A: **Cancer statistics, 2018**. *CA Cancer J Clin* 2018, **68**(1):7-30.
6. Vickers AJ, Sjoberg DD, Ulmert D, Vertosick E, Roobol MJ, Thompson I, Heijnsdijk EA, De Koning H, Atoria-Swartz C, Scardino PT *et al*: **Empirical estimates of prostate cancer overdiagnosis by age and prostate-specific antigen**. *BMC Med* 2014, **12**:26.
7. Massie CE, Mills IG, Lynch AG: **The importance of DNA methylation in prostate cancer development**. *J Steroid Biochem Mol Biol* 2017, **166**:1-15.
8. Feldman BJ, Feldman D: **The development of androgen-independent prostate cancer**. *Nat Rev Cancer* 2001, **1**(1):34-45.
9. Powers GL, Marker PC: **Recent advances in prostate development and links to prostatic diseases**. *Wiley Interdiscip Rev Syst Biol Med* 2013, **5**(2):243-256.
10. Chung LW, Huang WC, Sung SY, Wu D, Odero-Marah V, Nomura T, Shigemura K, Miyagi T, Seo S, Shi C *et al*: **Stromal-epithelial interaction in prostate cancer progression**. *Clin Genitourin Cancer* 2006, **5**(2):162-170.
11. Li LC, Okino ST, Dahiya R: **DNA methylation in prostate cancer**. *Biochim Biophys Acta* 2004, **1704**(2):87-102.
12. Nelson WG, De Marzo AM, Isaacs WB: **Prostate cancer**. *N Engl J Med* 2003, **349**(4):366-381.
13. Buyyounouski MK, Choyke PL, Kattan MW, McKenney JK, Srigley JR, Barocas DA, Brimo F, Brookland RK, Epstein JI, Fine SW *et al*: **Prostate**. In: *AJCC Cancer Staging Manual.* Edited by Amin MB, Edge S, Greene F, Byrd DR, Brookland RK, Washington MK, Gershenwald JE, Compton CC, Hess KR, Sullivan DC *et al*, 8 edn. New York: Springer; 2017: 723-734.
14. Orsted DD, Bojesen SE: **The link between benign prostatic hyperplasia and prostate cancer**. *Nat Rev Urol* 2013, **10**(1):49-54.
15. Lin PC, Giannopoulou EG, Park K, Mosquera JM, Sboner A, Tewari AK, Garraway LA, Beltran H, Rubin MA, Elemento O: **Epigenomic alterations in localized and advanced prostate cancer**. *Neoplasia* 2013, **15**(4):373-383.
16. Vandekerkhove G, Chi KN, Wyatt AW: **Clinical utility of emerging liquid biomarkers in advanced prostate cancer**. *Cancer Genet* 2017.
17. Esposito M, Guise T, Kang Y: **The Biology of Bone Metastasis**. *Cold Spring Harb Perspect Med* 2017.
18. Jimenez-Andrade JM, Mantyh WG, Bloom AP, Ferng AS, Geffre CP, Mantyh PW: **Bone cancer pain**. *Ann N Y Acad Sci* 2010, **1198**:173-181.
19. Lonergan PE, Tindall DJ: **Androgen receptor signaling in prostate cancer development and progression**. *J Carcinog* 2011, **10**:20.
20. Velcheti V, Karnik S, Bardot SF, Prakash O: **Pathogenesis of prostate cancer: lessions from basic research**. *Ochsner J* 2008, **8**(4):213-218.

21. Barbieri CE, Rubin MA: **Genomic rearrangements in prostate cancer**. *Curr Opin Urol* 2015, **25**(1):71-76.
22. Perner S, Demichelis F, Beroukhim R, Schmidt FH, Mosquera JM, Setlur S, Tchinda J, Tomlins SA, Hofer MD, Pienta KG *et al*: **TMPRSS2:ERG fusion-associated deletions provide insight into the heterogeneity of prostate cancer**. *Cancer Res* 2006, **66**(17):8337-8341.
23. Graham MK, Meeker A: **Telomeres and telomerase in prostate cancer development and therapy**. *Nat Rev Urol* 2017, **14**(10):607-619.
24. Kovtun IV, Murphy SJ, Johnson SH, Cheville JC, Vasmatzis G: **Chromosomal catastrophe is a frequent event in clinically insignificant prostate cancer**. *Oncotarget* 2015, **6**(30):29087-29096.
25. Chan JM, Gann PH, Giovannucci EL: **Role of diet in prostate cancer development and progression**. *J Clin Oncol* 2005, **23**(32):8152-8160.
26. Lilja H, Ulmert D, Vickers AJ: **Prostate-specific antigen and prostate cancer: prediction, detection and monitoring**. *Nat Rev Cancer* 2008, **8**(4):268-278.
27. Stamey TA, Yang N, Hay AR, McNeal JE, Freiha FS, Redwine E: **Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate**. *N Engl J Med* 1987, **317**(15):909-916.
28. Thompson IM, Ankerst DP, Chi C, Goodman PJ, Tangen CM, Lucia MS, Feng Z, Parnes HL, Coltman CA, Jr.: **Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial**. *J Natl Cancer Inst* 2006, **98**(8):529-534.
29. Stephenson AJ, Kattan MW, Eastham JA, Dotan ZA, Bianco FJ, Jr., Lilja H, Scardino PT: **Defining biochemical recurrence of prostate cancer after radical prostatectomy: a proposal for a standardized definition**. *J Clin Oncol* 2006, **24**(24):3973-3978.
30. Thompson IM, Pauler DK, Goodman PJ, Tangen CM, Lucia MS, Parnes HL, Minasian LM, Ford LG, Lippman SM, Crawford ED *et al*: **Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter**. *N Engl J Med* 2004, **350**(22):2239-2246.
31. Okotie OT, Roehl KA, Han M, Loeb S, Gashti SN, Catalona WJ: **Characteristics of prostate cancer detected by digital rectal examination only**. *Urology* 2007, **70**(6):1117-1120.
32. D'Elia C, Cerruto MA, Cioffi A, Novella G, Cavalleri S, Artibani W: **Upgrading and upstaging in prostate cancer: From prostate biopsy to radical prostatectomy**. *Mol Clin Oncol* 2014, **2**(6):1145-1149.
33. Tan N, Margolis DJ, McClure TD, Thomas A, Finley DS, Reiter RE, Huang J, Raman SS: **Radical prostatectomy: value of prostate MRI in surgical planning**. *Abdom Imaging* 2012, **37**(4):664-674.
34. Kirby MK, Ramaker RC, Roberts BS, Lasseigne BN, Gunther DS, Burwell TC, Davis NS, Gulzar ZG, Absher DM, Cooper SJ *et al*: **Genome-wide DNA methylation measurements in prostate tissues uncovers novel prostate cancer diagnostic biomarkers and transcription factor binding patterns**. *BMC Cancer* 2017, **17**(1):273.
35. Epstein JI, Allsbrook WC, Jr., Amin MB, Egevad LL, Committee IG: **The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma**. *Am J Surg Pathol* 2005, **29**(9):1228-1242.
36. Pierorazio PM, Walsh PC, Partin AW, Epstein JI: **Prognostic Gleason grade grouping: data based on the modified Gleason scoring system**. *BJU Int* 2013, **111**(5):753-760.

37. Long MD, Smiraglia DJ, Campbell MJ: **The Genomic Impact of DNA CpG Methylation on Gene Expression; Relationships in Prostate Cancer**. *Biomolecules* 2017, **7**(1).
38. Witte T, Plass C, Gerhauser C: **Pan-cancer patterns of DNA methylation**. *Genome Med* 2014, **6**(8):66.
39. Kurdyukov S, Bullock M: **DNA Methylation Analysis: Choosing the Right Method**. *Biology (Basel)* 2016, **5**(1).
40. Day JJ, Sweatt JD: **DNA methylation and memory formation**. *Nat Neurosci* 2010, **13**(11):1319-1323.
41. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL *et al*: **High density DNA methylation array with single CpG site resolution**. *Genomics* 2011, **98**(4):288-295.
42. Gardiner-Garden M, Frommer M: **CpG islands in vertebrate genomes**. *J Mol Biol* 1987, **196**(2):261-282.
43. Moore LD, Le T, Fan G: **DNA methylation and its basic function**. *Neuropsychopharmacology* 2013, **38**(1):23-38.
44. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M *et al*: **The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores**. *Nat Genet* 2009, **41**(2):178-186.
45. Schulz WA, Steinhoff C, Florl AR: **Methylation of endogenous human retroelements in health and disease**. *Curr Top Microbiol Immunol* 2006, **310**:211-250.
46. Jones PA: **Functions of DNA methylation: islands, start sites, gene bodies and beyond**. *Nat Rev Genet* 2012, **13**(7):484-492.
47. Zhu H, Wang G, Qian J: **Transcription factors as readers and effectors of DNA methylation**. *Nat Rev Genet* 2016, **17**(9):551-565.
48. Siegfried Z, Simon I: **DNA methylation and gene expression**. *Wiley Interdiscip Rev Syst Biol Med* 2010, **2**(3):362-371.
49. Lim DHK, Maher ER: **DNA methylation: a form of epigenetic control of gene expression**. *The Obstetrician & Gynaecologist* 2010, **12**(1):37-42.
50. Blattler A, Yao L, Witt H, Guo Y, Nicolet CM, Berman BP, Farnham PJ: **Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes**. *Genome Biol* 2014, **15**(9):469.
51. Margueron R, Reinberg D: **The Polycomb complex PRC2 and its mark in life**. *Nature* 2011, **469**(7330):343-349.
52. Reddington JP, Perricone SM, Nestor CE, Reichmann J, Youngson NA, Suzuki M, Reinhardt D, Dunican DS, Prendergast JG, Mjoseng H *et al*: **Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes**. *Genome Biol* 2013, **14**(3):R25.
53. Kondo Y: **Shall we crosstalk? - The relationship between DNA methylation and histone H3 lysine 27 trimethylation**. *Bioessays* 2014, **36**(2):128.
54. O'Hagan HM, Wang W, Sen S, Destefano Shields C, Lee SS, Zhang YW, Clements EG, Cai Y, Van Neste L, Easwaran H *et al*: **Oxidative damage targets complexes containing DNA methyltransferases, SIRT1, and polycomb members to promoter CpG Islands**. *Cancer Cell* 2011, **20**(5):606-619.
55. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, Das PK, Kivioja T, Dave K, Zhong F *et al*: **Impact of cytosine methylation on DNA binding specificities of human transcription factors**. *Science* 2017, **356**(6337).

56. Hu S, Wan J, Su Y, Song Q, Zeng Y, Nguyen HN, Shin J, Cox E, Rho HS, Woodard C *et al*: **DNA methylation presents distinct binding sites for human transcription factors**. *Elife* 2013, **2**:e00726.
57. Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D: **Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome**. *Nat Genet* 2007, **39**(4):457-466.
58. Medvedeva YA, Khamis AM, Kulakovskiy IV, Ba-Alawi W: **Effects of cytosine methylation on transcription factor binding sites**. *BMC Genomics* 2013, **15**(119).
59. Marzese DM, Hoon DS: **Emerging technologies for studying DNA methylation for the molecular diagnosis of cancer**. *Expert Rev Mol Diagn* 2015, **15**(5):647-664.
60. Aran D, Sabato S, Hellman A: **DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes**. *Genome Biol* 2013, **14**(3):R21.
61. Hoque MO: **DNA methylation changes in prostate cancer: current developments and future clinical implementation**. *Expert Rev Mol Diagn* 2009, **9**(3):243-257.
62. Esteller M: **CpG island hypermethylation and tumor suppressor genes: a booming pesent, a brighter future**. *Oncogene* 2002, **21**:5427-5440.
63. Romagosa C, Simonetti S, Lopez-Vicente L, Mazo A, Lleonart ME, Castellvi J, Ramon y Cajal S: **p16(Ink4a) overexpression in cancer: a tumor suppressor gene associated with senescence and high-grade tumors**. *Oncogene* 2011, **30**(18):2087-2097.
64. Ehrlich M: **DNA hypomethylation in cancer cells**. *Epigenomics* 2009, **1**(2):239-259.
65. Van Tongelen A, Loriot A, De Smet C: **Oncogenic roles of DNA hypomethylation through the activation of cancer-germline genes**. *Cancer Lett* 2017, **396**:130-137.
66. Robertson KD: **DNA methylation, methyltransferases, and cancer**. *Oncogene* 2001, **20**(24):3139-3155.
67. Yang M, Park JY: **DNA methylation in promoter region as biomarkers in prostate cancer**. *Methods Mol Biol* 2012, **863**:67-109.
68. Graca I, Pereira-Silva E, Henrique R, Packham G, Crabb SJ, Jeronimo C: **Epigenetic modulators as therapeutic targets in prostate cancer**. *Clin Epigenetics* 2016, **8**:98.
69. Kobayashi Y, Absher DM, Gulzar ZG, Young SR, McKenney JK, Peehl DM, Brooks JD, Myers RM, Sherlock G: **DNA methylation profiling reveals novel biomarkers and important roles for DNA methyltransferases in prostate cancer**. *Genome Res* 2011, **21**(7):1017-1027.
70. Nakayama M, Bennett CJ, Hicks JL, Epstein JI, Platz EA, Nelson WG, De Marzo AM: **Hypermethylation of the Human Glutathione S-Transferase-π Gene (GSTP1) CpG Island Is Present in a Subset of Proliferative Inflammatory Atrophy Lesions but Not in Normal or Hyperplastic Epithelium of the Prostate**. *The American Journal of Pathology* 2003, **163**(3):923-933.
71. Friedlander TW, Roy R, Tomlins SA, Ngo VT, Kobayashi Y, Azameera A, Rubin MA, Pienta KJ, Chinnaiyan A, Ittmann MM *et al*: **Common structural and epigenetic changes in the genome of castration-resistant prostate cancer**. *Cancer Res* 2012, **72**(3):616-625.
72. Tang Y, Jiang S, Gu Y, Li W, Mo Z, Huang Y, Li T, Hu Y: **Promoter DNA methylation analysis reveals a combined diagnosis of CpG-based biomarker for prostate cancer**. *Oncotarget* 2017, **8**(35):58199-58209.
73. Cancer Genome Atlas Research N: **The Molecular Taxonomy of Primary Prostate Cancer**. *Cell* 2015, **163**(4):1011-1025.
74. Geybels MS, Wright JL, Bibikova M, Klotzle B, Fan JB, Zhao S, Feng Z, Ostrander EA, Lin DW, Nelson PS *et al*: **Epigenetic signature of Gleason score and prostate cancer recurrence after radical prostatectomy**. *Clin Epigenetics* 2016, **8**:97.

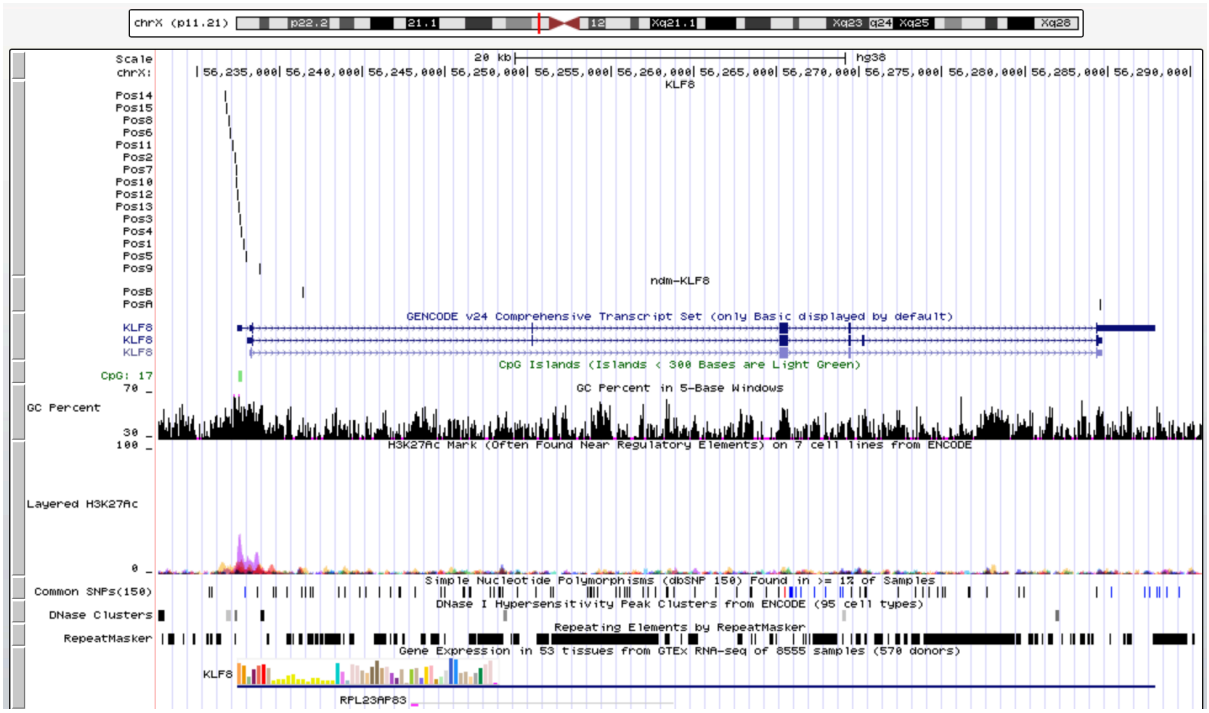
75. Bibikova M, Le JM, Barnes B, Zhou L, Shen R, Gunderson KL: **Genome-wide DNA methylation pro ling using In nium® assay**. *Epigenomics* 2009, **1**(1):177-200.
76. Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerstrom-Billai F, Jagodic M, Sundberg CJ, Ekstrom TJ, Teschendorff AE, Tegner J *et al*: **An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform**. *Epigenetics* 2013, **8**(3):333-346.
77. **PubMed** [https://www.ncbi.nlm.nih.gov/pubmed/]
78. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC**. *Genome Res* 2002, **12**(6):996-1006.
79. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool**. *Nucleic Acids Res* 2004, **32**(Database issue):D493-496.
80. Yamashita R, Sugano S, Suzuki Y, Nakai K: **DBTSS: DataBase of Transcriptional Start Sites progress report in 2012**. *Nucleic Acids Res* 2012, **40**(Database issue):D150-154.
81. Suzuki A, Wakaguri H, Yamashita R, Kawano S, Tsuchihara K, Sugano S, Suzuki Y, Nakai K: **DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data**. *Nucleic Acids Res* 2015, **43**(Database issue):D87-91.
82. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M *et al*: **NCBI GEO: archive for functional genomics data sets--update**. *Nucleic Acids Res* 2013, **41**(Database issue):D991-995.
83. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Res* 2002, **30**(1):207-210.
84. Rye MB, Bertilsson H, Andersen MK, Rise K, Bathen TF, Drablos F, Tessem M-B: **Cholesterol Synthesis Pathway Genes in Prostate Cancer are consistently downregulated when tissue confounding is minimized**. 2017.
85. Perez F, Granger BE: **IPython: A System for Interactive Scientific Computing**. *Computing in Science & Engineering* 2007, **9**(3):21-29.
86. Ekmekci B, McAnany CE, Mura C: **An Introduction to Programming for Bioscientists: A Python-Based Primer**. *PLoS Comput Biol* 2016, **12**(6):e1004867.
87. Fourment M, Gillings MR: **A comparison of common programming languages used in bioinformatics**. *BMC Bioinformatics* 2008, **9**:82.
88. Van Der Walt S, Colbert SC, Varoquaux G: **The NumPy array: a structure for efficient numerical computation**. *Computing in Science and Engineering* 2011, **13**(2):22-30.
89. **R: A language and environment for statistical computing** [https://www.r-project.org]
90. Ihaka R, Gentleman R: **R: A Language for Data Analysis and Graphics**. *Journal of Computational and Graphical Statistics* 1996, **5**(13):299-314.
91. Becker RA, Chambers JM, Wilks AR: **The New S Language**. London, UK: Chapman & Hall; 1988.
92. **The R FAQ** [https://cran.r-project.org/doc/manuals/R-FAQ.html#Citing-this-document]
93. Kohl M: **Introduction to statistical data analysis with R**. In., 1st edn: bookboon.com; 2015.
94. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al*: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**(10):R80.

95. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T *et al*: **Orchestrating high-throughput genomic analysis with Bioconductor**. *Nat Methods* 2015, **12**(2):115-121.
96. Smyth G: **Limma: linear models for microarray data**. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor.* Edited by Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. New York: Springer; 2005: 397-420.
97. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545-15550.
98. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A: **Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool**. *BMC Bioinformatics* 2013, **14**(128):128.
99. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A *et al*: **Enrichr: a comprehensive gene set enrichment analysis web server 2016 update**. *Nucleic Acids Res* 2016, **44**(W1):W90-97.
100. Cowley GS, Weir BA, Vazquez F, Tamayo P, Scott JA, Rusin S, East-Seletsky A, Ali LD, Gerath WF, Pantel SE *et al*: **Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies**. *Sci Data* 2014, **1**:140035.
101. Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D *et al*: **The UCSC Genome Browser database: 2018 update**. *Nucleic Acids Res* 2018, **46**(D1):D762-D769.
102. Taylor RA, Fraser M, Livingstone J, Espiritu SM, Thorne H, Huang V, Lo W, Shiah YJ, Yamaguchi TN, Sliwinski A *et al*: **Germline BRCA2 mutations drive prostate cancers with distinct evolutionary trajectories**. *Nat Commun* 2017, **8**:13671.
103. Babikova EA, Generozov EV: **Epigenetic analysis of normal prostate tissue and prostate adenocarcinoma**. In. Gene Expression Omnibus; 2015.
104. Barbieri CE, Demichelis F, Rubin MA: **Molecular genetics of prostate cancer: emerging appreciation of genetic complexity**. *Histopathology* 2012, **60**(1):187-198.
105. Esgueva R, Park K, Kim R, Kitabayashi N, Barbieri CE, Dorsey PJ, Jr., Abraham C, Banerjee S, Leung RA, Tewari AK *et al*: **Next-generation prostate cancer biobanking: toward a processing protocol amenable for the International Cancer Genome Consortium**. *Diagn Mol Pathol* 2012, **21**(2):61-68.
106. Ramalho-Carvalho J, Graca I, Gomez A, Oliveira J, Henrique R, Esteller M, Jeronimo C: **Downregulation of miR-130b~301b cluster is mediated by aberrant promoter methylation and impairs cellular senescence in prostate cancer**. *J Hematol Oncol* 2017, **10**(1):43.
107. Kim SJ, Kelly WK, Fu A, Haines K, Hoffman A, Zheng T, Zhu Y: **Genome-wide methylation analysis identifies involvement of TNF-alpha mediated cancer pathways in prostate cancer**. *Cancer Lett* 2011, **302**(1):47-53.
108. Fraser M, Sabelnykova VY, Yamaguchi TN, Heisler LE, Livingstone J, Huang V, Shiah YJ, Yousif F, Lin X, Masella AP *et al*: **Genomic hallmarks of localized, non-indolent prostate cancer**. *Nature* 2017, **541**(7637):359-364.
109. Shiah YJ, Fraser M, Bristow RG, Boutros PC: **Comparison of pre-processing methods for Infinium HumanMethylation450 BeadChip array**. *Bioinformatics* 2017, **33**(20):3151-3157.
110. Naeem H, Wong NC, Chatterton Z, Hong MK, Pedersen JS, Corcoran NM, Hovens CM, Macintyre G: **Reducing the risk of false discovery enabling identification of**

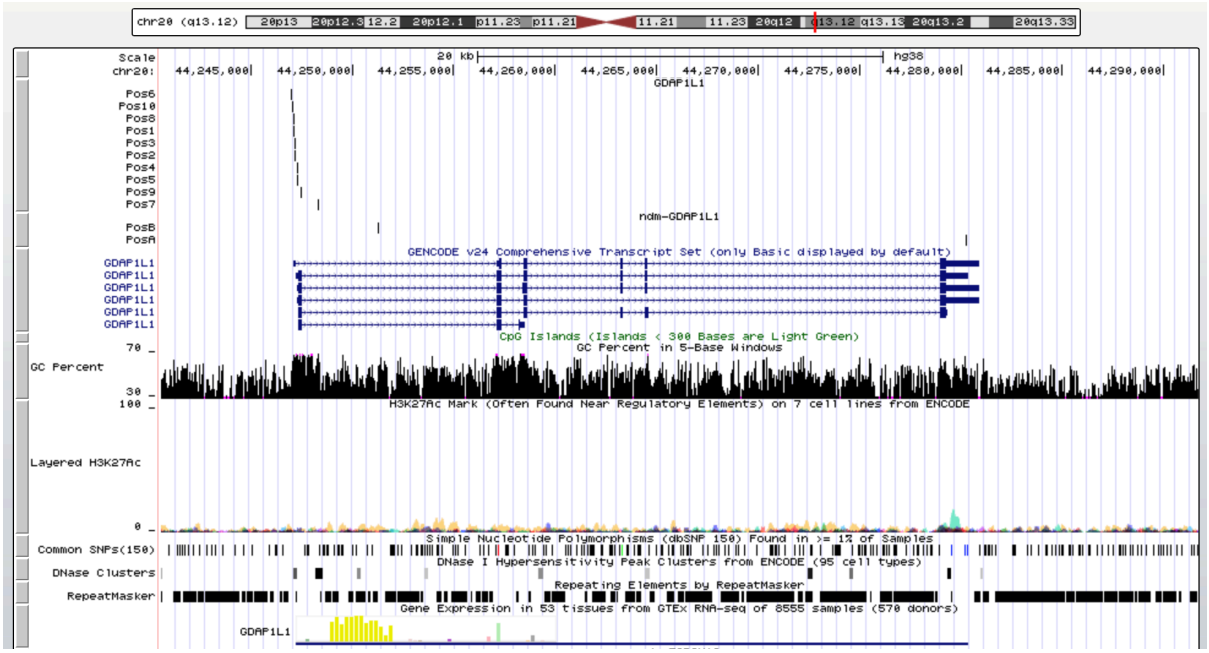
biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics* 2014, **15**:51.
111. Kron K, Pethe V, Briollais L, Sadikovic B, Ozcelik H, Sunderji A, Venkateswaran V, Pinthus J, Fleshner N, van der Kwast T *et al*: **Discovery of novel hypermethylated genes in prostate cancer using genomic CpG island microarrays**. *PLoS One* 2009, **4**(3):e4830.
112. Jarrard D, Yang B: **Methylation profiling defines a widespread field defect in histologically normal prostate tissue associated with prostate cancer**. In. Gene Expression Omnibus; 2012.
113. Borno ST, Fischer A, Kerick M, Falth M, Laible M, Brase JC, Kuner R, Dahl A, Grimm C, Sayanjali B *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):1024-1035.
114. Aryee MJ, Liu W, Engelmann JC, Nuhn P, Gurel M, Haffner MC, Esopi D, Irizarry RA, Getzenberg RH, Nelson WG *et al*: **DNA methylation alterations exhibit intraindividual stability and interindividual heterogeneity in prostate cancer metastases**. *Sci Transl Med* 2013, **5**(169):169ra110.
115. Goh LK, Liem N, Vijayaraghavan A, Chen G, Lim PL, Tay KJ, Chang M, Low JS, Joshi A, Huang HH *et al*: **Diagnostic and prognostic utility of a DNA hypermethylated gene signature in prostate cancer**. *PLoS One* 2014, **9**(3):e91666.
116. Kim JW, Kim ST, Turner AR, Young T, Smith S, Liu W, Lindberg J, Egevad L, Gronberg H, Isaacs WB *et al*: **Identification of new differentially methylated genes that have potential functional consequences in prostate cancer**. *PLoS One* 2012, **7**(10):e48455.
117. Geybels MS, Zhao S, Wong CJ, Bibikova M, Klotzle B, Wu M, Ostrander EA, Fan JB, Feng Z, Stanford JL: **Epigenomic profiling of DNA methylation in paired prostate cancer versus adjacent benign tissue**. *Prostate* 2015, **75**(16):1941-1950.
118. Singh AN, Sharma N: **Identification of key pathways and genes with aberrant methylation in prostate cancer using bioinformatics analysis**. *Onco Targets Ther* 2017, **10**:4925-4933.
119. Biau DJ, Kerneis S, Porcher R: **Statistics in brief: the importance of sample size in the planning and interpretation of medical research**. *Clin Orthop Relat Res* 2008, **466**(9):2282-2288.
120. Bracken AP, Pasini D, Capra M, Prosperini E, Colli E, Helin K: **EZH2 is downstream of the pRB-E2F pathway, essential for proliferation and amplified in cancer**. *EMBO J* 2003, **22**(20):5323-5335.
121. Cao R, Zhang Y: **SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex**. *Mol Cell* 2004, **15**(1):57-67.
122. Sertkaya S, Hamid SM, Dilsiz N, Varisli L: **Decreased expression of EFS is correlated with the advanced prostate cancer**. *Tumour Biol* 2015, **36**(2):799-805.
123. Xiang Y, Qiu Q, Jiang M, Jin R, Lehmann BD, Strand DW, Jovanovic B, DeGraff DJ, Zheng Y, Yousif DA *et al*: **SPARCL1 suppresses metastasis in prostate cancer**. *Mol Oncol* 2013, **7**(6):1019-1030.
124. Sugito K, Kawashima H, Yoshizawa S, Uekusa S, Hoshi R, Furuya T, Kaneda H, Hosoda T, Konuma N, Masuko T *et al*: **Non-promoter DNA hypermethylation of Zygote Arrest 1 (ZAR1) in neuroblastomas**. *J Pediatr Surg* 2013, **48**(4):782-788.
125. Levedakou EN, Strohmeyer TG, Effert PJ, Liu ET: **Expression of the matrix Gla protein in urogenital malignancies**. *Int J Cancer* 1992, **52**(4):534-537.

126. Cockerill PN: **Structure and function of active chromatin and DNase I hypersensitive sites**. *FEBS J* 2011, **278**(13):2182-2210.
127. M DA, Weghorn D, A DA-C, Coulet F, Olson KM, DeBoever C, Drees F, Arias A, Alakus H, Richardson AL *et al*: **Identifying DNase I hypersensitive sites as driver distal regulatory elements in breast cancer**. *Nat Commun* 2017, **8**(1):436.
128. Strand SH, Switnicki M, Moller M, Haldrup C, Storebjerg TM, Hedegaard J, Nordentoft I, Hoyer S, Borre M, Pedersen JS *et al*: **RHCG and TCAF1 promoter hypermethylation predicts biochemical recurrence in prostate cancer patients treated by radical prostatectomy**. *Oncotarget* 2017, **8**(4):5774-5788.
129. Kim JH, Dhanasekaran SM, Prensner JR, Cao X, Robinson D, Kalyana-Sundaram S, Huang C, Shankar S, Jing X, Iyer M *et al*: **Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer**. *Genome Res* 2011, **21**(7):1028-1041.
130. Chao HH, Uchio E, Zhang S, Hu S, Bednarski SR, Luo X, Rose M, Concato J, Li CS: **Effects of androgen deprivation on brain function in prostate cancer patients - a prospective observational cohort analysis**. *BMC Cancer* 2012, **12**(371):371.
131. Guo JH, Huang Q, Studholme DJ, Wu CQ, Zhao Z: **Transcriptomic analyses support the similarity of gene expression between brain and testis in human as well as mouse**. *Cytogenet Genome Res* 2005, **111**(2):107-109.

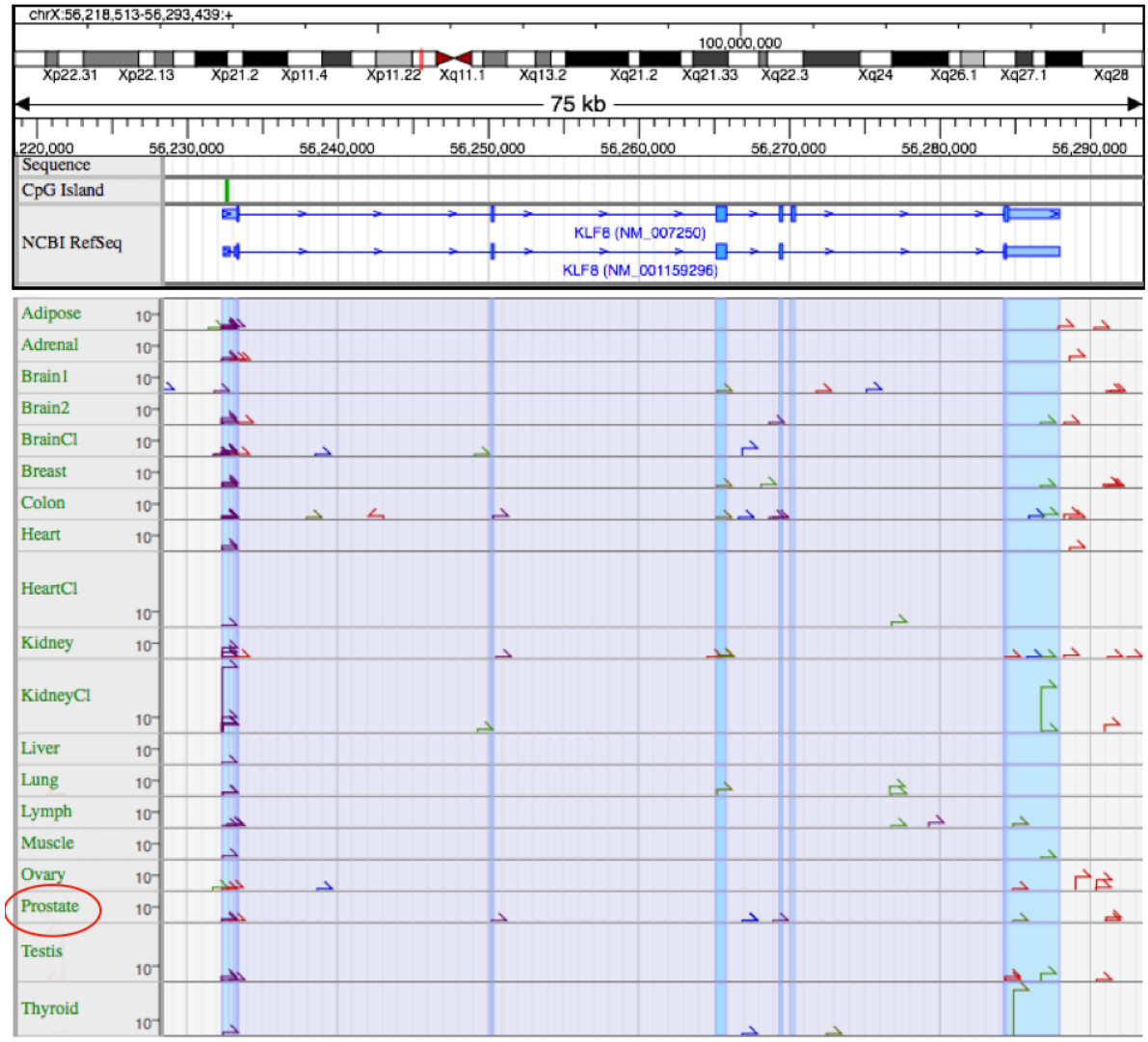
SUPPLEMENTARY MATERIAL



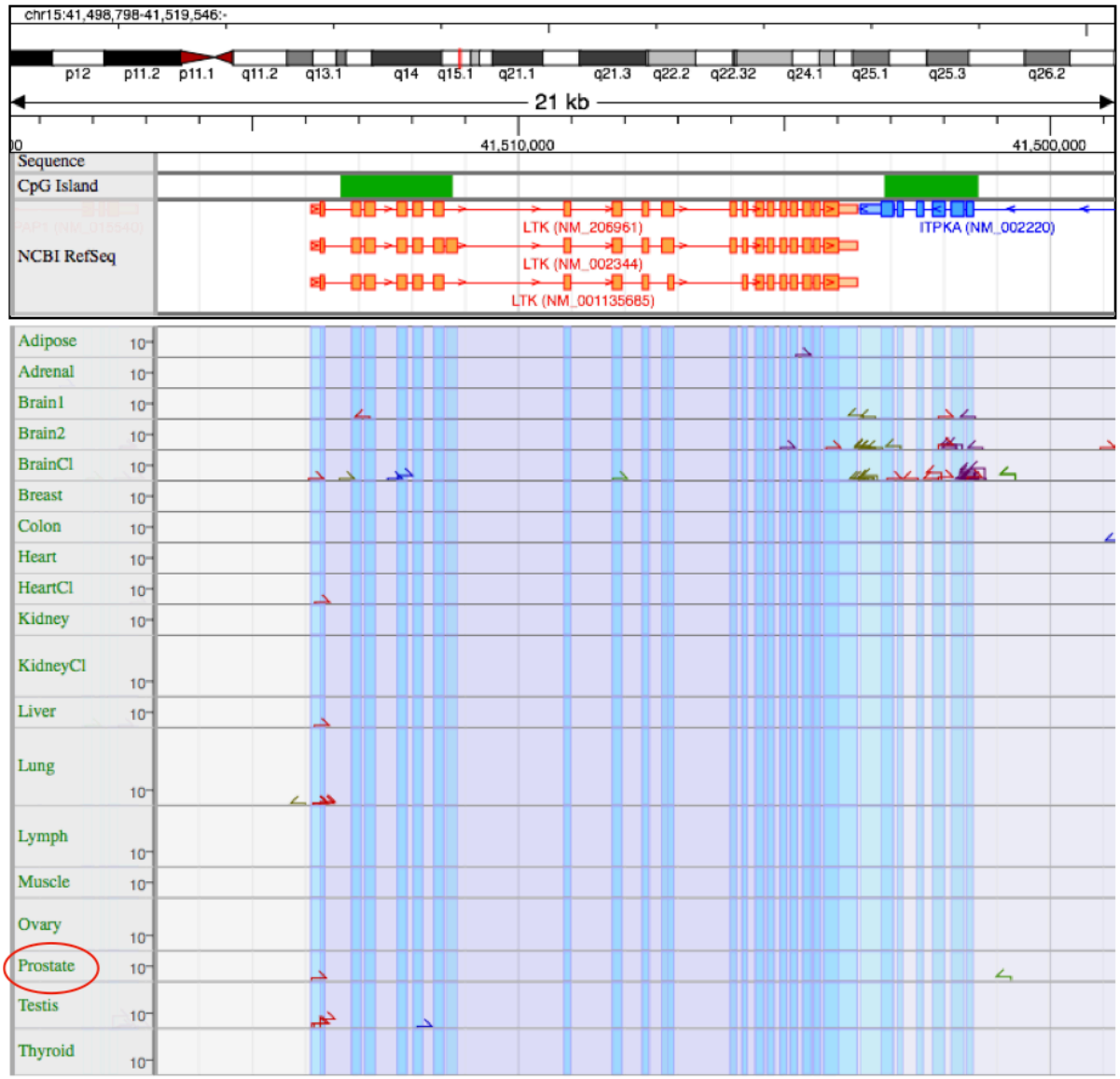
Supplementary Figure S1 **Visualized methylation probes for a gene *KLF8* from UPDOWN-only regulation pattern.** Pos1-15 stand for differentially hypermethylated positions from most (Pos1) to least (Pos15) significantly methylated. PosA and B are non-differentially methylated positions.



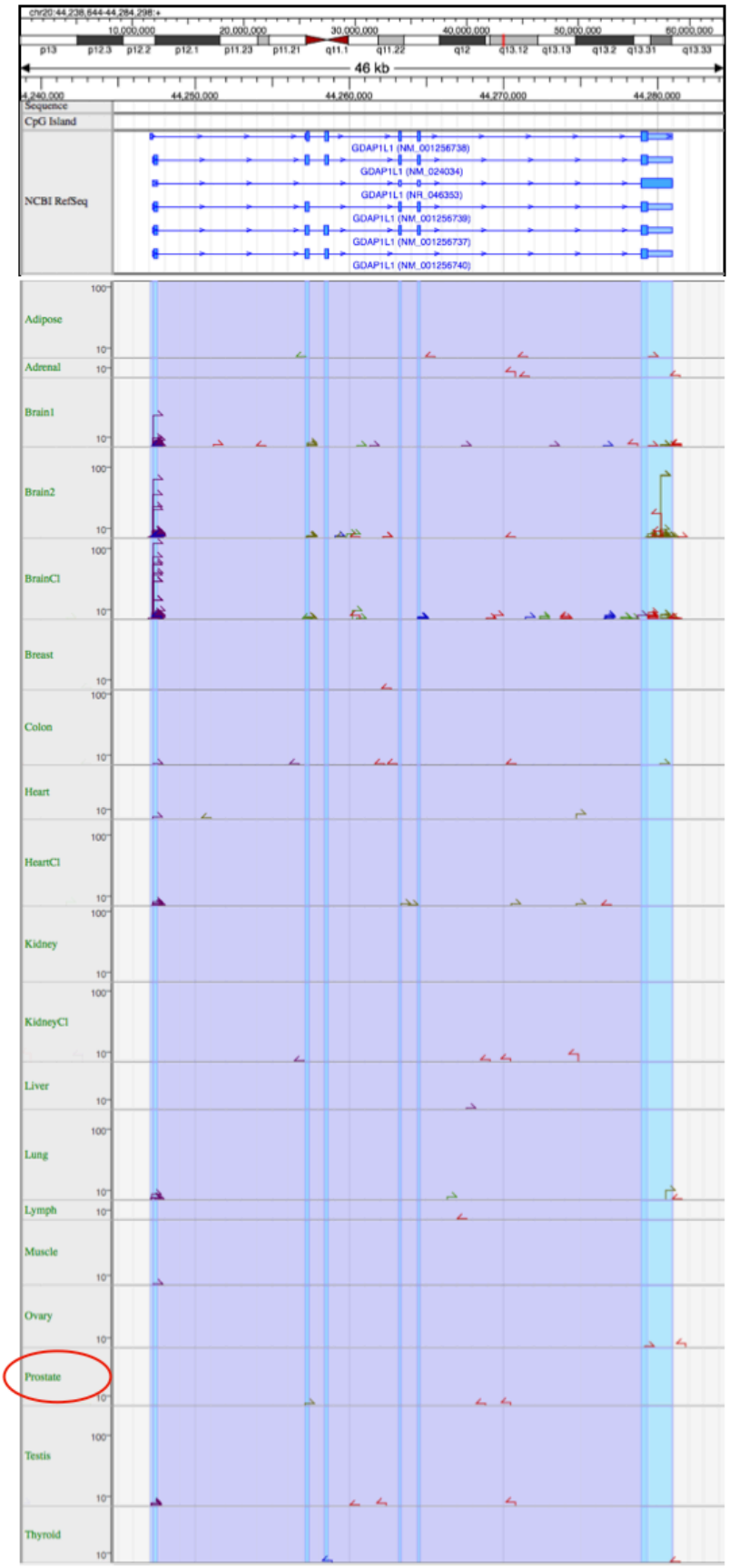
Supplementary Figure S2 **Visualized methylation probes for a gene *GDAP1L1* from UPUP-only regulation pattern.** Pos1-10 stand for differentially hypermethylated positions from most (Pos1) to least (Pos10) significantly methylated. PosA and B are non-differentially methylated positions.



Supplementary Figure S3 **Transcription start sites (TSSs) for a gene *KLF8* in various adult human tissues, including prostate.** Upper part of the figure shows position of the gene in genome and CpG island, associated with the gene. Lower part illustrates different TSSs, where the height of the peak corresponds to the frequency of a particular TSS used in transcription.



Supplementary Figure S4 **Transcription start sites (TSSs) for a gene *LTK* in various adult human tissues, including prostate.** Upper part of the figure shows position of the gene in genome and CpG island, associated with the gene. Lower part illustrates different TSSs, where the height of the peak corresponds to the frequency of a particular TSS used in transcription.



Supplementary Figure S5 **Transcription start sites (TSSs) for a gene *GDAP1L1* in various adult human tissues, including prostate.** Upper part of the figure shows position of the gene in genome. Lower part illustrates different TSSs, where the height of the peak corresponds to the frequency of a particular TSS used in transcription.

Supplementary Table S1 **Number of genes in each dataset, following four different DNA methylation and gene expression patterns, and number of genes, following each pattern, that are overlapping between datasets.** Number in brackets shows what part of all genes for each pattern and in that datasets is an overlap.

	Absher	***Kirby***	**TCGA**	**Overlap**
Gain of methylation — upregulated expression (UPUP)	1320 (overlap is 80.15%)	1437 (73.63%)	1241 (85.25%)	1058
Gain of methylation — downregulated expression (UPDOWN)	2110 (88.06%)	2230 (83.31%)	2041 (91.03%)	1858
Loss of methylation — upregulated expression (DOWNUP)	809 (89.12%)	916 (78.71%)	1304 (55.29%)	721
Loss of methylation — downregulated expression (DOWNDOWN)	638 (85.27%)	709 (76.72%)	1037 (52.46%)	544

Supplementary Table S2 **Number of genes in *Absher*, *Kirby* and TCGA datasets that are associated with multiple probes with methylation gain or loss and overlapping genes that are associated with both, hypermethylated and hypomethylated probes.**

	Absher	***Kirby***	**TCGA**
Gain of methylation	1243	1514	1319
Loss of methylation	232	381	575
Overlap (gain and loss of methylation)	7 genes: *BCL2*, *CCND1*, *GNAS*, *MGMT*, *OSBPL5*, *PEG10*, *SNRPN*	16 genes: *ATP10A*, *CCND1*, *CD96*, *GNAS*, *IGF2*, *INS-IGF2*, *MEST*, *NTM*, *OSBPL5*, *PEG10*, *RAB32*, *RUNX3*, *SEMA3B*, *SGCE*, *SNRPN*, *ZIM2*	21 genes: *BCL2*, *C21orf29*, *C22orf45*, *CHFR*, *GNAS*, *GNASAS*, *IGF2*, *INS-IGF2*, *KCNQ1*, *MED12L*, *MEG3*, *MGMT*, *NTM*, *NTRK1*, *OSBPL5*, *PEG10*, *PRH1*, *PRR4*, *RUNX3*, *SGCE*, *SNRPN*

Supplementary Table S3 **Number of genes in each regulation pattern with a certain number of transcription start sites (TSS).** Number in brackets shows the part of all genes for each number of TSS.

Number of TSS	UPUP-only genes	UPDOWN-only genes
1	118 (75.16% of all genes)	169 (68.70% of all genes)
2	30 (19.11%)	49 (19.92%)
3	6 (3.82%)	17 (6.91%)
4	2 (1.27%)	6 (2.44%)
5	-	3 (1.22%)
7	-	1 (0.41%)
18	1 (0.64%)	-
34	-	1 (0.41%)
Number of genes in total	157	246