Tampere University

Paula Mosallaei

IN SEARCH FOR EVIDENCE OF TRANSCRIPTIONAL REGULATION OF UNANNOTATED TRANSCRIPTS IN PROSTATE AND PROSTATE CANCER

Faculty of Medicine and Health
Technology
Master’s Thesis
- 2022

ABSTRACT

Paula Mosallaei: In Search for Evidence of Transcription Regulation of Unannotated Transcripts in Prostate and Prostate Cancer
Master's Thesis
Tampere University
Bioinformatics
Supervisors: Sinja Taavitsainen, Matti Nykter
Examiners: Sinja Taavitsainen, Matti Nykter, Juha Kesseli
December 2022

Prostate cancer (PC) is the third cause of cancer deaths in men in European Union. As such, PC exerts a substantial burden on the European health care systems and society. Efforts to find more effective ways of diagnosing and treating prostate cancer are of great value. Non-coding RNAs play an essential role in tumorigenesis, including the development of prostate cancer. Discoveries of novel transcripts driving oncogenesis or transition towards castration resistance offer new potential diagnostic tools or therapeutic targets and advance the understanding of PC evolution. The objective of this thesis was to study the patterns of interplay between the genomic and epigenomic data in previously found unannotated transcripts to determine whether the transcripts are subject of multi-layered transcriptional regulation. The relationships between expression, chromatin accessibility, and methylation were studied. The results were integrated with results from another project, involving promoter prediction and incidence of transcriptional activity-associated histone modifications. The presence of binding sites of AR, the main player in PC, within the promoters was also investigated. Each individual step of analysis and multilayer data integration provided evidence for regulation of a small subset of unannotated transcripts. However, identification of individual putative novel transcripts was not successful. The study was not exhaustive, and more analyses could be done in attempt to find biologically significant novel transcripts.

Keywords: prostate cancer, novel transcripts, non-coding transcripts, promoter

PREFACE

This document template conforms to the Guide to Writing a Thesis in Technical Fields at Tampere University (2019).

Before you lies the master thesis “In Search for Evidence of Transcription Regulation of Unannotated Transcripts in Prostate and Prostate Cancer”. It has been written to fulfil the graduation requirements of the Bioinformatics program at Tampere University. I performed the research work and wrote the thesis between January to November 2022.

I would like to thank my supervisors, Professor Matti Nykter, and PhD candidate Sinja Taavitsainen for the guidance and support during last year. I admire your knowledge and experience, but also the kindness and understanding. I appreciate the opportunity I was given by joining your research team. I would like to thank Ebrahim Afyounian for giving directions when I was lost the meanders of statistical analysis and Python coding, and for his patience. Finally, I thank my husband Milad Mosallaei, who has provided me with a word of encouragement every time I started doubting myself.

Tampere, 19 December 2022

Paula Mosallaei

CONTENTS

1. INTRODUCTION 1
2. BACKGROUND 2
   2.1 Prostate cancer 2
   2.2 Transcription 5
   2.3 Non-coding RNA 8
   2.4 Promoter 10
   2.5 Chromatin accessibility 14
   2.6 Epigenetics 18
      2.6.1 DNA methylation 19
      2.6.2 Histone modifications 23
      2.6.3 Transcription factors 26
   2.7 Previous work and the starting point 31
3. MOTIVATION AND OBJECTIVES 32
4. MATERIALS AND METHODS 34
   4.1 Samples and datasets 34
   4.2 ATAC-seq data processing 34
   4.3 Finding distances from TSS to nearest ATAC-seq peak 35
   4.4 Quantification of ATAC-seq peaks in individual samples 35
   4.5 Correlation between accessibility and expression 36
   4.6 MeDIP-seq data processing and correlation between methylation and expression 36
   4.7 Promoter prediction based on sequence composition 37
   4.8 Studying the presence of histone modifications associated with transcriptional activity within estimated promoters 38
   4.9 Transcription factor AR binding sites 38
5. RESULTS 39
   5.1 Co-occurrence of transcripts and ATAC-seq peaks on the genome-wide level 39
   5.2 The relationship between chromatin accessibility and expression in individual samples 41
      5.2.1 Sample-specific ATAC-seq peaks 41
      5.2.2 Co-occurrence of ATAC-seq peaks and expression 43
      5.2.3 Correlation between accessibility and expression 45
      5.2.4 Thresholding and filtering 48
   5.3 Integration of DNA methylation data 50
      5.3.1 Correlation between methylation level and expression 56
      5.3.2 Correlation in ON and OFF transcripts 58

5.3.3 Thresholding of ON and OFF non-coding unannotated and annotated transcripts 61
5.3.4 Thresholding of protein-coding ON transcripts 62
5.4 Promoter prediction * 62
5.4.1 Correlations in ON transcripts with predicted promoters.... 64
5.4.2 Shortlisted unannotated transcripts with promoter predictions.... 68
5.4.3 GC content of the shortlisted transcripts with promoter predictions 69
5.5 Histone mark detection 70
5.5.1 Histone marks associated with transcriptional activity*.... 70
5.5.2 Activity marks, chromatin accessibility and expression 71
5.5.3 Activity marks in shortlisted transcripts.... 75
5.5.4 Binding sites of transcription factor AR and shortlisted transcripts 79
6. DISCUSSION.... 84
7. CONCLUSIONS.... 88
REFERENCES.... 89
APPENDIX A: GTRD EXPERIMENTS USED IN HISTONE MARK ANALYSIS.... 95
APPENDIX B: THE FRACTIONS OF "ON" AND "OFF" TRANSCRIPTS IN EACH TRANSCRIPT GROUP.... 97
APPENDIX C: THE RELATIONSHIP BETWEEN THE NUMBER OF EXPRESSED TRANSCRIPTS AND THE STRENGTH OF CORRELATION BETWEEN CHROMATIN ACCESSIBILITY AND EXPRESSION.... 98
APPENDIX D: THE RELATIONSHIP BETWEEN THE NUMBER OF ATAC-SEQ PEAKS IN PROMOTERS AND THE STRENGTH OF CORRELATION BETWEEN EXPRESSION AND CHROMATIN ACCESSIBILITY 102
APPENDIX E: THE MOST EXPRESSED UNANNOTATED TRANSCRIPTS WITH ACCESSIBLE PROMOTERS AND LOW METHYLATION.... 106
APPENDIX F: THE MOST EXPRESSED UNANNOTATED TRANSCRIPTS WITHOUT ATAC-SEQ PEAKS AND WITH HIGH METHYLATION IN PROMOTERS 107
APPENDIX G: THE MOST EXPRESSED UNANNOTATED TRANSCRIPTS WITHOUT ATAC-SEQ PEAKS AND LOW METHYLATION IN PROMOTERS 109

LIST OF FIGURES

Figure 1. *Anatomy of the prostate (Adapted from [2])* 3
Figure 2. *Reference-based transcriptome assembly. (Adapted from [32])* 7
Figure 3. *De novo transcriptome assembly. (Adapted from [32])* 8
Figure 4. *A schematic illustration of the basal transcription machinery assembly. (Adapted from [26])* 11
Figure 5. *Schematic illustration of the most common core promoter elements. (Adapted from [44])* 13
Figure 6. *Layers of chromatin organization. (Adapted from [26])* 19
Figure 7. *A representation of a nucleosome. Green molecules correspond to H2A, orange molecules to H2B, red molecules to H3 and blue molecules to H4. The gray strand represents genomic DNA, whereas the brown molecule represents the linker histone H1. (Adapted from [26])* 24
Figure 8. *Genomic localization of histone H3 modifications associated with active transcription. Colored peaks visualize ChIP-seq signal shapes which would correspond to individual histone marks in a eukaryotic gene. The arrow represents TSS. (Adapted from [82])* 25
Figure 9. *Standard sequence logo plot (Adapted from [94])* 29
Figure 10. *Distribution of the distances from TSS to the nearest ATAC-seq peak* 40
Figure 11. *The percentages of transcripts with the nearest ATAC-seq peak within different distances from TSS. The percentages are rounded to the nearest integer.* 41
Figure 12. *The number of ATAC-seq peaks in individual samples.* 42
Figure 13. *Violin plots representing the distributions of the number of peaks in each sample group. The blue shapes represent the density of the peak numbers in each sample group. The red dots represent the median numbers of peaks, and the black dots represent numbers of peaks in each individual sample.* 43
Figure 14. *P-values from Spearman's correlation tests between ATAC-seq peak intensity and expression for all ON transcripts. The dashed red line marks the significance level 0.01.* 46
Figure 15. *P-values from Spearman's correlation tests for ON transcripts with TPM higher than sample-specific upper quartiles. The dashed red line marks the significance level 0.01.* 50
Figure 16. *A bar plot presenting the percentages of transcripts whose promoter region was methylated in some degree with distinction of the conditions. The values are rounded to one decimal place.* 51
Figure 17. *The percentages of transcripts with a peak in their promoter and high methylation level or low methylation level.* 52
Figure 18. *The percentages of highly methylated transcript promoters with and without an ATAC-seq peak.* 54
Figure 19. *The percentages of highly methylated transcripts low and high expression level.* 54
Figure 20. *Percentages of promoters with low methylation level and an ATAC-seq peak or no ATAC-seq peak.* 55
Figure 21. *Percentages of promoters with low methylation level and high expression or low expression.* 55
Figure 22. *P-values from Spearman's correlation tests between methylation and expression. The red dashed line marks the significance level.* 57

Figure 23. *P-values from Spearman's correlation tests between methylation and expression in ON transcripts. The red dashed line marks the significance level 0.01.* 59

Figure 24. *P-values from Spearman's correlation test between methylation and expression in OFF transcripts. The red dashed line marks the significance level 0.01.* 60

Figure 25. *The percentages of predicted promoters in each transcripts group. The numbers are rounded to one decimal place.* 63

Figure 26. *Mean percentages of ON and OFF transcripts, whose previously defined promoters intersected promoter predictions.* 64

Figure 27. *P-values from Spearman's correlation test between accessibility and expression for ON transcripts whose previously defined promoters overlapped promoter predictions. The red dashed line marks the significance level.* 66

Figure 28. *P-values from Spearman's correlation tests between methylation and expression for ON transcripts, whose previously defined promoters intersected promoter predictions.* 67

Figure 29. *Percentages of promoters hitting predictions within shortlisted groups.* 68

Figure 30. *Mean GC content of promoters intersecting promoter predictions. Only promoters of previously shortlisted transcripts were studied. GC content presented as a percentage.* 69

Figure 31. *Percentages of transcripts whose promoters or bodies overlapped at least one activity-associated histone mark locus.* 71

Figure 32. *Percentages of active transcripts whose promoters or bodies covered loci of all activity-associated histone modifications.* 71

Figure 33. *The fractions of ATAC-seq peaks accompanied by an activity associated histone mark within the same promoter, and the fractions of activity histone marks accompanied by an ATAC-seq peak within the same promoter.* 72

Figure 34. *Mean percentages of transcripts with activity-associated histone marks, which were expressed.* 73

Figure 35. *Fractions of the transcripts with an ATAC-seq peak and activity histone mark within their promoters, whose expression was high (TPM > 0.9 quantile) and low (TPM < 0.75 quantile).* 74

Figure 36. *Shortlisted transcripts with histone marks within their promoters in BPH samples.* 75

Figure 37. *The percentages of shortlisted transcripts with histone marks and promoter predictions in BPH samples.* 76

Figure 38. *Shortlisted transcripts with histone marks within their promoters in PC samples.* 77

Figure 39. *The percentages of shortlisted transcripts with histone marks and promoter predictions in PC samples.* 77

Figure 40. *Shortlisted transcripts with histone marks within their promoters in CRPC samples.* 78

Figure 41. *The percentages of shortlisted transcripts with histone marks and predicted promoters in CRPC samples.* 79

Figure 42. *Shortlisted transcriptionally active transcripts with AR binding sites within their promoters in BPH samples.* 80

Figure 43. *Percentages of transcripts with AR binding sites and promoter predictions in BPH samples.* 80

Figure 44. *Shortlisted transcriptionally active transcripts with AR binding sites with their promoters in PC samples.* 81

Figure 45. *Percentages of transcripts with AR binding sites, which had promoter prediction in PC samples.* 82

Figure 46. *Shortlisted active transcripts with AR binding sites in their promoters in CRPC samples.* *82*

Figure 47. *Percentages of transcripts with AR binding sites, which had promoter prediction in CRPC samples.* *83*

LIST OF TABLES

Table 1. *Mean Spearman's and Pearson's correlation estimates for each transcript group in each condition... 45*

Table 2. *Mean Spearman's and Pearson's correlation estimates for transcripts with TPM higher than sample-specific upper quartiles. 49*

Table 3. *The mean values of sample mean methylation levels in each transcript group in each condition, rounded to four decimal places. 52*

Table 4. *Mean Spearman's and Pearson's correlations between methylation and expression.. 56*

Table 5. *Mean Spearman's and Pearson's correlations between methylation and expression for ON and OFF transcripts. 58*

Table 6. *Mean correlation estimates between accessibility and expression for ON transcripts whose previously defined promoters overlapped promoter predictions. 65*

Table 7. *Mean correlation estimates between methylation and expression for ON transcripts, whose previously defined promoters intersected promoter predictions. 66*

Table 8. *GTRD experiments used in analysis of histone marks falling within promoters of the studied transcripts. Column "Used for" shows which sample group was studied using given experiment. Column "Associated with" explains the functional association of the enrichment of a given histone mark... 96*

Table 9. *The percentages of ON and OFF genes of all transcripts in each transcript group, rounded to two decimal places. Plus and minus signs in the parentheses represent the sense assumption for unannotated transcripts... 97*

Table 10. *Highly expressed ON unannotated transcripts with low methylation level, whose TPM was higher than 30.00.. 106*

Table 11. *Unannotated highly methylated OFF transcripts with particularly high expression level (TPM > 30.00)... 108*

Table 12. *Unannotated OFF transcripts with high expression and low methylation, which were overexpressed (TPM > 30.00) in more than five samples or whose expression level was exceptionally high (TPM > 100.00). Column "Mean TPM" contains the mean TPM computed from values in all samples, in which given transcript passed the filtering. Column "Max TPM (Sample)" shows the largest expression of a given transcript among the samples, in which the transcript passed the filtering, and the identifier of the sample in which the expression reached this maximum value............ 111*

LIST OF SYMBOLS AND ABBREVIATIONS

AD	Activation domains
AR	Androgen receptor
ATAC-seq	Assay for transposase-accessible chromatin couples with high-throughput sequencing
bp	Base pair
BPH	Benign prostatic hyperplasia
cCRE	Candidate Cis-Regulatory Element
ChIP-seq	Chromatin immunoprecipitation using high-throughput sequencing
CRPC	Castration-resistant prostate cancer
DBD	DNA-binding domain
DHT	Dihydrotestosterone
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
GTRD	Gene Transcription Regulation Database
HTS	High-throughput sequencing
lncRNA	Long non-coding RNA
Mb	Megabase
MeDIP-seq	Methylated DNA immunoprecipitation coupled with sequencing
NDR	Nucleosome-depleted regions
nt	Nucleotide
PC	(Primary) prostate cancer
PSA	Prostate-specific antigen
PWM	Position Weight Matrix
QSEA	Quantitative Sequencing Enrichment Analysis
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
SWI/SNF	Switching/ non-fermenting chromatin remodelling complexes
TAD	Topologically associated domain
TF	Transcription factor
TFBS	Transcription factor binding site
TPM	Transcripts per million
TSS	Transcriptional start site
TTS	Transcription termination site
UMI	Unique molecular identifiers

1. INTRODUCTION

The prostate gland is an exocrine organ accessory to the male reproductive system. Prostate cancer (PC) is a non-cutaneous malignancy of the prostate gland. It is the third cause of cancer deaths in men in European Union, and as such exerts a substantial burden on the European health care systems and society. Therefore, it is important to find effective ways to not only diagnose and treat the disease, but also to assess the risks of PC development and progression into castration-resistant form. The main challenge in PC research is the genetic heterogeneity, which increases as the disease progresses into more advanced stages.

Recent studies have found that tumorigenesis, also in PC, is driven by non-coding long RNAs. Since the landscape of biologically significant non-coding genes has not been fully explored, studying so far unannotated transcripts present in prostate cancer might provide new insights into disease evolution and progression to castration-resistant stage. Ultimately, non-coding transcripts can become diagnostic markers, potentially more precise than the current ones, or targets of therapeutic strategies. Finally, exploring the landscape of non-coding genes involved in PC might improve our general understanding of the disease.

In a previous study of a patient cohort including cases of benign prostate hyperplasia, untreated PC, and castration-resistant prostate cancer, two large subsets of unannotated transcripts had been identified. The objective of this thesis was to analyse transcriptomic data and several levels of epigenomic data to investigate whether they provide evidence that the newly identified transcripts encompassed a group of genes subject to transcription regulation. Transcription regulation is a highly complex process, which requires coordinated modulations on several levels of DNA structure and is a sign of biological significance of its subject. Unannotated transcripts being targets of transcription regulatory mechanisms would constitute a set of putative novel transcripts, whose role in tumorigenesis could be further studied. Thus, the relationships between transcript expression, chromatin accessibility, DNA methylation, histone modifications and AR binding sites were studied in the unannotated transcripts on a genome-wide scale and in a sample-specific manner. To the knowledge of the author, a similar study has not been performed to date.

2. BACKGROUND

2.1 Prostate cancer

The prostate gland is an exocrine organ accessory to the male reproductive system, 4-cm in diameter, located under bladder and around urethra. Anatomically, it consists of four zones, which can be seen in Figure 1. Histologically, the prostate is a branched duct gland. Superior to the base of the prostate, are seminal vesicles, which join with vas deferens to form ejaculatory ducts. The function of prostate is to secrete alkaline prostatic fluid, which increases volume to the semen, promotes sperm motility, and enhances the chance of conception [1]–[3] Coagulation of the semen is prevented by a glycoprotein, prostate-specific antigen (PSA), secreted into the glandular ducts. Normally, the level of PSA is much larger in the prostate than in the serum. However, development of cancer in the prostate degenerates the structure of the ducts, leading to active secretion of PSA out of the prostate, and consequently to dramatically elevated serum levels. Due to this fact, PSA is used as a biomarker in PC diagnosis and to follow the effectiveness of treatment. [3], [4]

Prostate cancer is a non-cutaneous malignancy of the prostate gland [5]. Approximately 70-80% of prostate cancer cases arise from the peripheral zone of the prostate, but reportedly, the most aggressive cancers and the ones with tendencies of spreading arise from the central zone (roughly only 2.5 % of all PC cases) [3]. Typically, PC does not present specific symptoms related to prostate anatomical region, although it may occur alongside obstructive symptoms of age-related benign prostatic hyperplasia (BPH) [2], which is a non-malignant disease characterized by proliferation of the epithelial and stromal components of the prostate [6]. Genomic studies have revealed not only interpatient molecular heterogeneity between prostate cancer cases, but also intrapatient heterogeneity [7]–[9]. Each PC case reveals its own unique molecular composition, and within the tumor there are several disease foci, which either constitute separate tumors of independent origin or lineages resulting from evolutionary branches from a common ancestor [8].

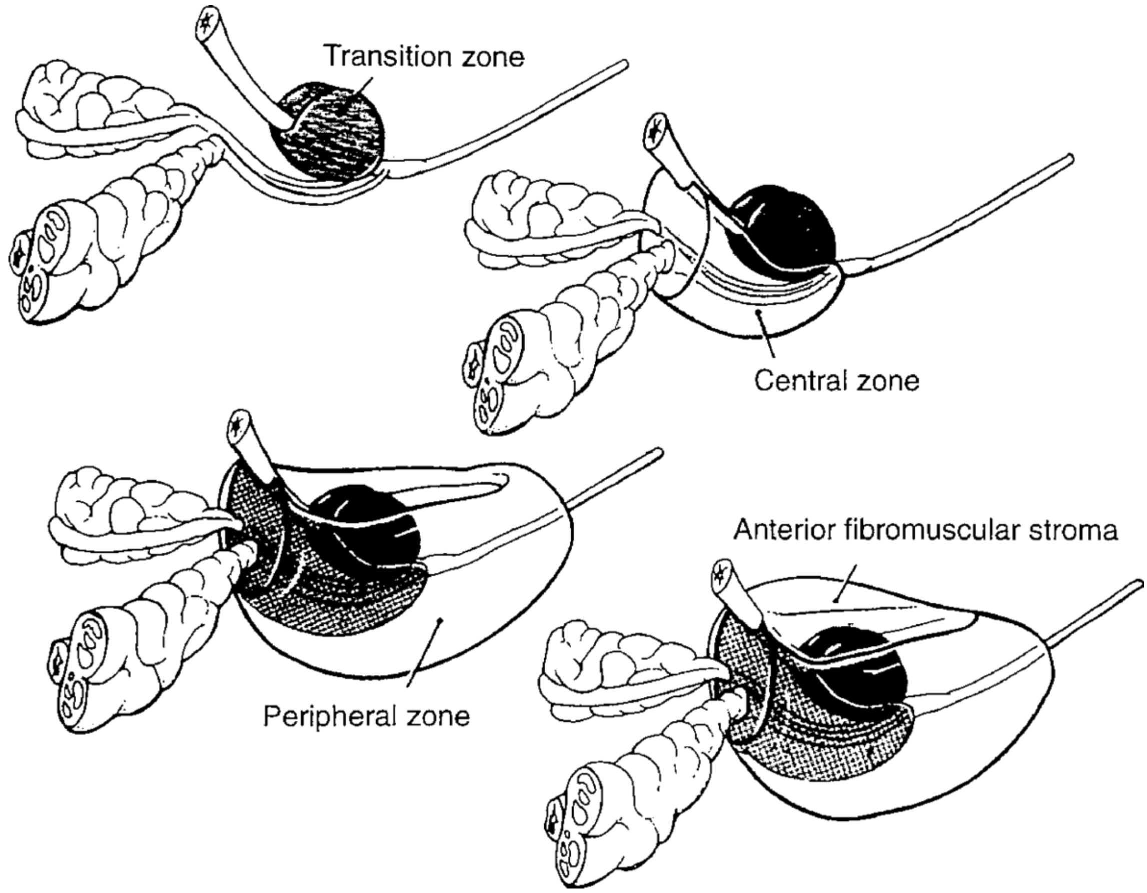


Figure 1. *Anatomy of the prostate (Adapted from* [2]*)*

From the early cell divisions in the zygote until death, human cells accumulate somatic point mutations in response to mutagenic stresses. The majority of such point mutations do not result in changes of the cell functioning and can arise due to aging. Certain mutations, however, are proliferatively advantageous and drive pathological changes in the cell, showing some level of order. The early driver events, often shared between multiple cancers, include a small number of mutations and lead to development of pre-cancer states. The transition to cancer involves increased number of aberrant genes, and further evolution to more aggressive forms is shaped by rare driver mutations, characteristic for a given malignancy. The first mutations contributing to cancer evolution may happen decades before the diagnosis. [10]

The evolution of prostate cancer in most cases is related to androgens and androgen receptor (AR) signalling pathway. Androgens, particularly testosterone and 5α-dihydrotestosterone (DHT), play a central role in normal development and functioning of the prostate gland. They bind to AR, inducing its transcriptional activity. AR, in turn, is responsible for inducing transcription of genes involved in proliferation and apoptosis. PSA expression is also regulated by AR activity. This androgen/AR interaction has been proven to remain crucial in prostate cancer. Because of PC's dependence on the AR signalling pathway, the standard therapy for patients with advanced disease or metastatic PC is androgen deprivation therapy. It can be performed either as surgical (orchiectomy) or medical castration, or as combined therapy, resulting in full androgen blockage. Medical castration involves either luteinizing hormone-releasing hormone agonists/antagonists, which block testosterone production in testicles, or anti-androgens, which prevent androgen interaction with AR by competitive binding. Although initially androgen deprivation therapy proves effective, which is indicated by no detectable presence of PSA in the serum, the disease eventually progresses to lethal castration resistant prostate cancer (CRPC). The transition to the advanced stage is a result of molecular changes in the cancer tissue, which lead to AR signalling reactivation and maintenance without the testicular androgens. It is achieved through AR overexpression, signal transduction cascades, AR somatic mutations, modulation of the AR coregulators, ligand-independent AR activation, and steroidogenesis. CRPC tissues are able to synthesize testicular androgens from adrenal androgens and cholesterol. Thus, despite the lack of testicular androgen production, the levels of androgens inside a CRPC tumor are similar to the levels in a prostate of a healthy man, or even higher. [3]–[5]

Next-generation AR inhibitors abiraterone and enzalutamide are increasingly used in the treatment of CRPC, however, some cases relapse with aggressive AR-negative forms of castration resistance. The exact mechanisms behind the emergence of such PC variants are being studied and a histological classification has not yet been established, although several distinct forms have been mentioned in literature: neuroendocrine prostate cancer [9], [11], [12], AR indifferent forms [12], [13], as well as forms dependent on different transcription factors (TFs), e.g. fibroblast growth factor, the glucocorticoid receptor or the pluripotent stem cell TF SOX2 [12].

In effort to better understand the mechanisms driving the evolution of cancer from normal prostate to PC to castration resistance and to define molecular signatures of the malignancies, genomic alteration studies have been conducted widely [14]–[21]. However, these studies focus mainly on known protein-coding genes, while transcriptome studies from last two decades have revealed existence of tremendous number of non-coding ribonucleic acid (RNA) molecules in all investigated organisms, including human. [22], [23] In fact, only 1.5% of the genome displays the ability to code for proteins.[24]

2.2 Transcription

Transcription is the process of converting DNA (deoxyribonucleic acid)-encoded genetic information into RNA-encoded information. It is a mechanism which is highly regulated through interactions of different factors, which will be further described in this section. In eukaryotes, transcription can be performed by one of the three major types of RNA polymerases, depending on a species of RNA being synthesized [25]–[27]:

- RNA polymerase I transcribes ribosomal RNA (rRNA) genes.
- RNA polymerase II is responsible for transcription of protein-coding genes, but also lncRNAs, snRNAs, and sniRNAs.
- RNA polymerase III is recruited for transcription of 5S rRNA, small non-coding RNAs and transfer RNAs.

The genes transcribed by RNA polymerases I and III are often referred to as "housekeeping genes" due to the fact that they encode RNA molecules responsible for the basic functions of the cell. [26] Because this thesis focuses partly on protein-coding genes and lncRNAs, the following sections will focus on the transcription process and machinery specific for RNA polymerase II. [27]

The set of all transcripts (RNAs) synthesized in a given moment in a given cell population (or tissue) is called a transcriptome. The composition of the transcriptome depends on the cell type, developmental stage, and the environment of the cells, and therefore it is subject to dynamic changes. [27] Investigating transcriptome provides information about the functioning of a cell type or tissue. The current method of choice to study the transcriptome is RNA sequencing (RNA-seq), with the widely used platform being Illumina. [28]–[30] However, most experiment designs of RNA-seq do not sequence RNA molecules directly, but rather the complementary DNA (cDNA), flanked by adapters appropriate for a given application. The design depends on the type of RNA being of interest. [28] Typical workflow of RNA-seq experiment comprises the following steps: (1) extraction and purification of RNA from cell or tissue, (2) preparation of sequencing library, including fragmentation and linear or PCR amplification, (3) RNA sequencing, (4) RNA-seq data processing, and (5) data analysis. The workflow introduces several sources of bias, the most prominent being amplification, and especially PCR amplification, which need to be addressed during data processing. [30]

Next to whole-transcriptome sequencing, different protocols are available for specific research questions, e.g. targeted RNA-seq for predetermined group of genes of interest, or single-cell RNA-seq for chosen cell types. Special methods can also be used to study alternative splicing and gene fusions, small RNAs, and circular RNAs. One useful technique, especially in case of single-cell sequencing, is molecular labelling using unique molecular identifiers (UMIs) to enable distinction of different products of PCR cycles. UMIs can be either intentionally designed sequences or random nucleotides, and they are usually inserted within adapter sequences. Moreover, when multiple cell types are being sequenced simultaneously, a barcode made of oligonucleotide beads for each of them can be introduced [28]

While studying the abundance of transcripts can already increase understanding of cell state, RNA-seq reads usually are very short, and do not represent complete transcripts. A deeper insight into gene activity as well as identifying novel genes and gene isoforms is only possible with transcriptome assembly. It means reconstructing full-length genes from short RNA-seq reads. Mainly, two techniques have been defined to conduct assembly: *de novo* and reference-guided assembly, and for each there are available bioinformatics tools. As the name suggests, reference-guided assemblers use a reference genome to align RNA-seq reads with help of algorithms allowing spliced alignments (e.g. HISAT or STAR) and based on that reconstruct individual transcripts. [31] The steps of such approach can be seen in Figure 2. Step a) shows RNA-seq reads (grey blocks) aligned to a reference genome, using spliced alignment. Part b) represents a connectivity or splice graph, which includes all possible isoforms at a locus. Steps c) and d) show how alternative paths through the graph (blue, red, yellow and green) lead to merging compatible reads into isoforms. [32] The widely used tools performing reference-based assembly are Cufflinks, Bayesembler, StringTie, TransComb, and Scallop. [31] The existence of so many reference-guided assemblers is dictated by the ambiguity of the read mapping. This ambiguity arises from the alternative gene splicing and potential mapping of a single read to multiple locations in the genome. Thus, every assembler is based on a strategy of selecting the genomic sites the ambiguous reads should be mapped to. For example, abovementioned Cufflinks's strategy is to generate the lowest number of transcripts which covers the highest percentage of mapped reads. Bayesembler, in turn, uses Bayesian likelihood estimation to find the most probable combination of transcripts. Finally, StringTie constructs transcripts based on a flow graph. As a result, each assembler produces different sets of transcripts from the same set of reads. [33] An assembler's performance also depends on the complexity of the studied genome, and thus appropriate algorithm should be chosen based on the application. [34] Also, the so-called ensemble assemblers are available, which generate consensus assemblies based on the results produced by different algorithms. Examples of ensemble assemblers are EvidentialGene and Concatenation. [33]

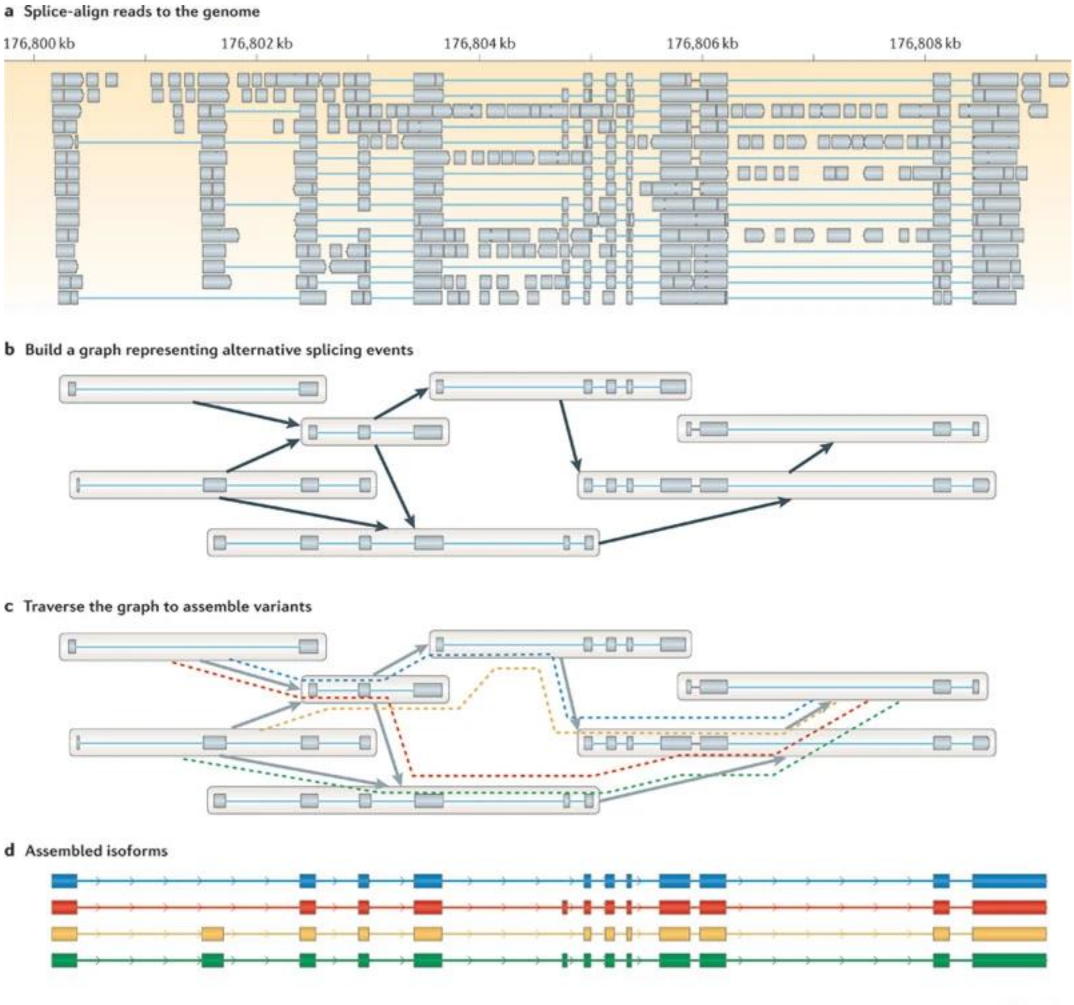

a Splice-align reads to the genome
176,800 kb
176,802 kb
176,804 kb
176,806 kb
176,808 kb
b Build a graph representing alternative splicing events
c Traverse the graph to assemble variants
d Assembled isoforms

Figure 2. *Reference-based transcriptome assembly. (Adapted from [32])*

On the other hand, de novo assemblers reconstruct transcripts solely based on the transcript sequences and their overlaps. The available tools for *de novo* assembly are Trinity and Oasis. [31] The steps of de novo assembly are visualized in Figure 3. In the first step, shown in part a), each read is split into substrings of length k (k-mers). The Figure uses 5-mers as an example. Then, in b) step, De Bruijn graph is formed from all the k-mers to map overlaps between them. In step c), the adjacent nodes of the graph are collapsed into a single node when the first node has an out degree of one and the second node has an in degree of one. Step d) includes traversing four alternative paths (blue, red, yellow and green) through the graph to identify isoforms, similarly to the reference-based approach. In the final step e), isoforms are assembled. [32]

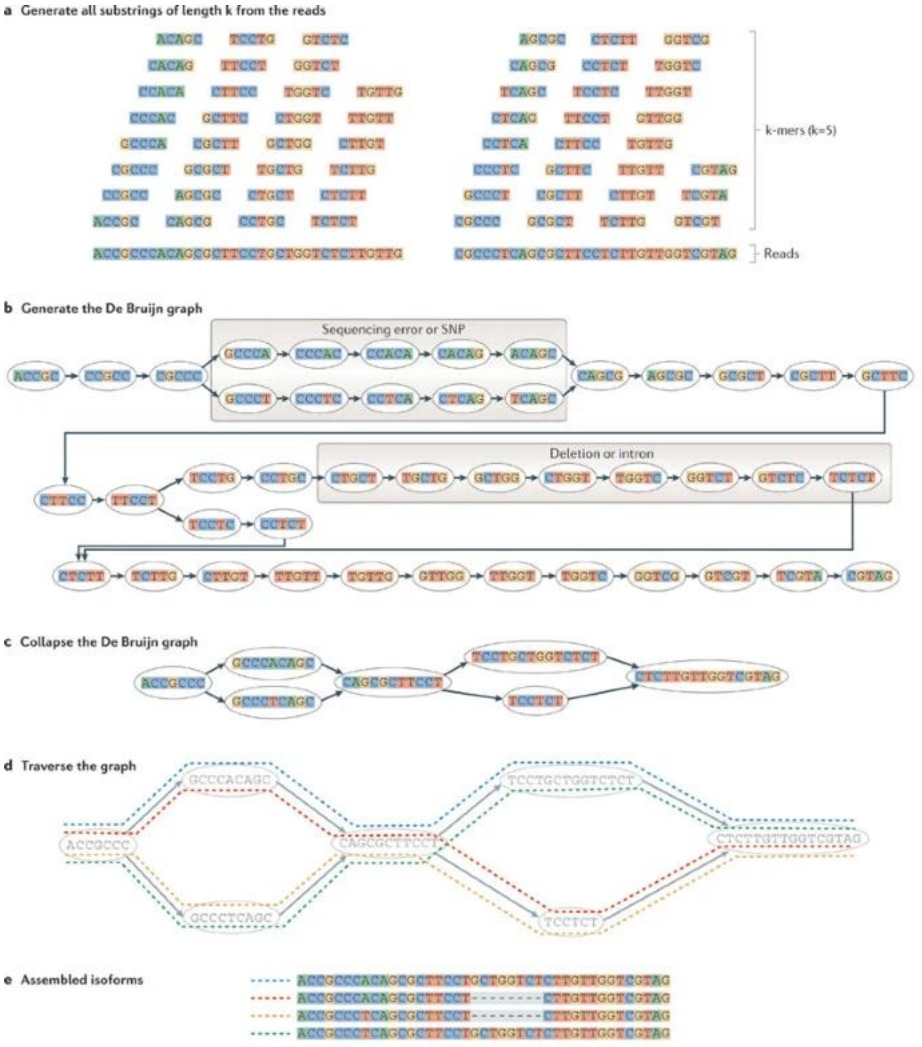


Figure 3. *De novo transcriptome assembly. (Adapted from [32])*

2.3 Non-coding RNA

Non-coding transcripts are genomic elements which do not encode proteins but are transcribed in a regulated manner. Numerous studies have demonstrated that non-coding RNAs have their own distinct functions and operating mechanisms, and they are involved in development, differentiation, and metabolism. [22], [24], [35], [36] Moreover, it has been reported that non-coding elements may have a greater cell specificity than coding genes. Finally, most of the single nucleotide polymorphisms (SNPs) in human diseases fall within the non-coding regions of the genome, which suggests the importance of those regions in emergence of pathological conditions. [36]

Different classes of non-coding transcripts have been defined, and many of them are highly conserved: siRNAs, miRNAs, and piRNAs, however, long non-coding RNAs (lncRNAs) are poorly conserved.[24] Those transcripts are defined as polyadenylated non-coding transcripts longer than 200 nt (nucleotides) that are transcribed by RNA polymerase II and are associated with epigenetic signatures common to protein-coding genes. These signatures will be discussed later in this chapter. However, this group of transcripts is characterized by the heterogeneity in terms of genomic origin and based on that several subtypes of lncRNA have been distinguished. Among the processes in which lncRNAs are involved is epigenetic transcriptional regulation, modulating tumor suppressor activity, regulation of mRNA processing and translation, and RNA-RNA interactions. [22]

The discovery of the biological significance of non-coding RNAs has led to studies on their involvement in tumorigenesis. The developments in the field have revealed that non-coding RNAs, especially lncRNAs, drive evolution of cancers. [22]–[24] Investigation of lncRNAs in prostate cancer resulted in identification of several lncRNAs specific to PC. *PCGEM1* and *PRNCR1* were found to be regulated by AR and overexpressed in prostate cancer. [37], [38] As a highly PC-specific lncRNA, *PCA3* has been demonstrated to have a potential as a PC diagnostic marker. [39] *PCAT-1* was identified as a transcriptional repressor in prostate cancer. [40] *SChLAP1* was shown to negate the tumor-suppressive functions of the SWI/SNF complex, which means that this lncRNA plays an important part in tumor progression to lethal stage.[41] *PCAT5* has been found to be an ERG-regulated oncogene that impacts cell proliferation pathways.[23] *EPCART* was found to enhance mobility and proliferation of prostate cancer cells. [42]

Although the evidence of the biological functionality of lncRNAs has been accumulating over last two decades, researchers have emphasized that the non-coding transcriptome has not been fully explored and there is space for discoveries of novel functional non-coding RNAs. [23], [43] At the same time, the large amount of non-coding genomic material has raised questions whether transcription is always a result of an orchestrated, meaningful biological process, or is part of it a product of a “leaky transcriptional system”, non-specific RNA-polymerase activity? And how can we identify a new gene? [22]

Prensner and Chinnaiyan, 2011 listed features of a distinct lncRNA. According to them, what makes a transcript distinguishable from the transcriptional background is: (1) expression in a tissue-specific manner; (2) the presence histone marks associated with transcriptional activity (especially H3K4me3 at the gene promoter, and H3K36me3 throughout the gene body); (3) transcription through RNA polymerase II; (4) regulation by well-established TFs; (5) polyadenylation, and (6) frequent splicing of multiple exons via canonical genomic splice site motifs. [22] In other words, there must be evidence on different levels that the expression of a given transcript undergoes similar regulatory modifications to those accompanying protein-coding gene expression.

2.4 Promoter

Returning to the transcription and its regulation, for the RNA polymerase II to initiate the process, the presence of a specific sequence in the vicinity of a gene, called promoter, is required. [44] Gene expression is orchestrated by specific sequences of DNA, which are called regulatory elements. Promoters are such regulatory elements that contain specific motifs, or short sequence features, recognized and bound by TFs, which enable the assembly of pre-initiation complex, or basal transcription machinery. [45], [46] Often, two segments of promoter sequence are discussed, core promoter region, being in the close vicinity to the transcription start site (TSS), and extended promoter region. [45] The core promoter is described in latest literature as stretching from -40 to +40 bp from the TSS. [44], [47] TSS is usually referred to as position +1 and it is the first nucleotide transcribed to RNA, defining the 5'-end of a gene. [26], [44] The extended promoter has been defined differently in the literature, but in this study, I followed the definitions used by Uusi-Mäkelä and colleagues [48]: -1000 to +100 bp from TSS, and a wider -2000 to +100 bp.

Even though TSS is referred to as the first transcribed nucleotide, in fact there is no single specific nucleotide which serves this function. Instead, there are many TSSs situated near each other within some 70 bp stretches of genomic DNA. Furthermore, genes can have more than one TSS region, and this is the case for most of the human protein-coding genes, which results in complex proteome. [26]

The basal transcription machinery forms at the TSS from various regulatory proteins with RNA polymerase II at its core. This requires previous binding of activator proteins to enhancer regions. In promoters containing TATA-box motif, they recruit chromatin remodelling factors, which subsequently allows associating of TATA-binding protein (TBP) with TATA-box. TBP is the essential component of the machinery, as it bends DNA, and binds TAFs (TBP-associated factors), forming TFIID. Genomic DNA wraps around TFIID in a similar fashion it is wrapped around a nucleosome, and this is a signal for other TFs, TFIIA, TFIIB, TFIIE, TFIIF, TFIIH, and RNA polymerase II that they can associate to form the rest of the transcription machinery. The formation of the machinery is illustrated schematically in Figure 4.

A subunit of TFIIH hydrolyses ATP needed to open double helix and separate DNA strands and passes the template strand to RNA polymerase II. TFIIE stabilizes the melted DNA, so that polymerase can proceed along the strand in the 3' – 5' direction. Carboxy terminal domain (CTD) of the RNA polymerase II is phosphorylated by TFIIH, and different phosphorylation patterns ignite different steps of transcription by associating different factors. And so, transition to elongation, polyadenylation, and transcription termination are triggered by those patterns. The primary transcript undergoes 5' capping and addition of Poly(A) tail so that it is protected from digestion by exonucleases. [25]–[27] Transcription continues until transcription termination site (TTS) located at the 3'-end of the gene, where RNA polymerase II disassociates from the DNA template. [26]

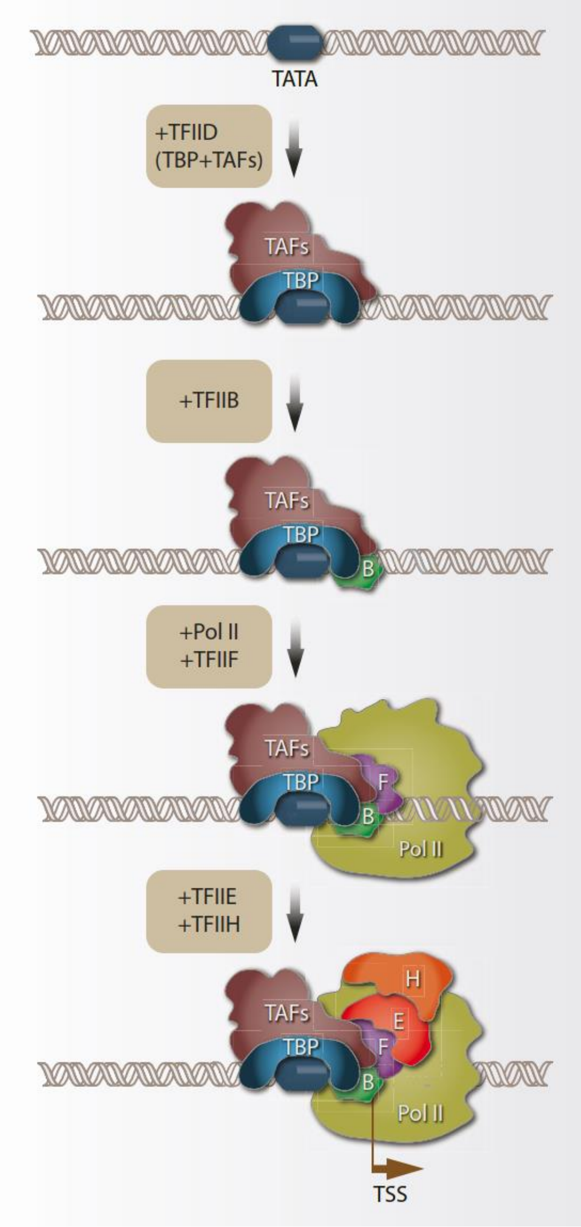


Figure 4. *A schematic illustration of the basal transcription machinery assembly. (Adapted from* [26]*)*

The structure of core promoters has been studied extensively due to its importance in the process of transcription and the predictability of its location. Consequently, they are the best-characterized regulatory sequences. [49] It has been found that some motifs are frequently conserved in the promoter sequence, however, some promoter may not include any of them and have more unique structure.[44], [47]. These motifs are responsible for binding the components of the pre-initiation complex. The most common ones are [44]:

At the TSS:

1. Inr – initiator, the most prevalent element of core promoters, encompasses TSS and is bound by subunits of TFIID, essential for the transcription initiation. Its presence is crucial for the downstream functional elements: DPE, MTE, Bridge I, and Bridge II.
2. TCT – the polypyrimidine initiator motif, functional in rRNA transcription and involved in translation regulation.
3. XCPE1 and XCPE2 – the X gene core promoter element 1 and the X gene core promoter element 2. They drive RNA polymerase II transcription, the preceding one with assistance of co-activators, and the latter on its own.

Upstream from TSS:

4. TATA-box – the first core promoter motif that has been identified, located upstream from TSS. Its name is a short version of its consensus sequence. Initially thought to be always present in promoters, currently it has been recognized that only minority of promoters depends on the functionality of this motif. It binds TBP, a unit of TFIID.
5. BRE – the TFIIB recognition elements, present directly upstream or downstream from TATA-box, but lack of TATA-box does not rule out its presence. They bind TFBII and interact with TATA-box to regulate transcription.

Downstream from TSS:

6. DPE – downstream core promoter element, prevalent in developmental gene networks. It is a recognition site for TFIID.
7. MTE – the motif ten element, most often observed with DPE, however, they are also observed independently. Similarly to DPE, it is recognized by TFIID.
8. Bridge I and Bridge II – consist of sub-regions of DPE and MTE. Similarly to DPE and MTE, they are usually enriched in promoters without TATA-box.

9. DCE – the downstream core promoter, present in the promoters of the human adult β-globin. Distinct from the previous downstream motifs, bound by different sub-units of TFIID, and usually accompanied by TATA-box.

Figure 5 presents a schematic picture of the core promoter motifs listed above.

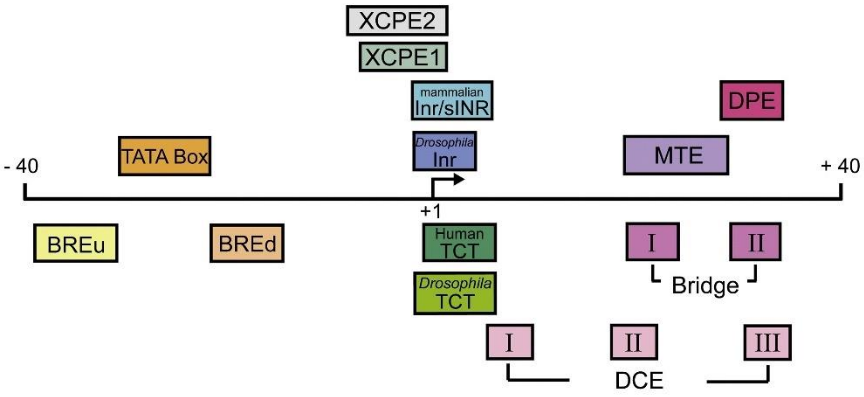


Figure 5. *Schematic illustration of the most common core promoter elements. (Adapted from [44])*

Proximal promoter regions might also include a CG-rich sequences upstream from the TSS. [50], [51]

Promoters are a subset of more generally defined enhancers. Enhancers contain binding sites of distal binding TFs, which recruit co-activator proteins to promote the transcriptional activity of their target gene. Unlike promoters, enhancers are located further away from the TSS both upstream and downstream. Still, they must be contained within the same topologically associated domain (TAD) as the gene they regulate. And since TADs are on average 1Mb long, some enhancers might be distant from their target gene. They are brought to the TSS of that gene through looping events of the DNA, which are mediated by complexes of protein cohesion and CCCTC binding factor, and which bring the enhancer-bound TFs into TSS's vicinity. The functioning of enhancers is not only dependent on cooperative binding of multiple regulatory proteins, but also on different epigenomic states. One enhancer may have opposite effects in different tissues. [26] Moreover, the number of enhancers, and the repertoire of the TFs recruited to them is dependent on the cell type, which plays the major role in cell specificity. Together, promoters and enhancers comprise *cis*-acting elements, whereas *trans*-acting proteins are those regulatory molecules which affect the transcription but come from another gene. [25]

In vivo identification of promoters or TSS sites is possible. Transient transfection promoter activity assay measures the effect of a putative promoter sequence on its target gene; however, the method has been criticized for omitting the genomic and epigenomic context as well as providing incomplete information about the promoter sequence itself. [45] The available techniques for *in vivo* TSS determination are OligoCap, CAGE, deepCAGE and PEAT, however, it has been pointed out that they are too costly to be widely applied, while the results are not always to researchers' satisfaction. [46] Moreover, the annotation of promoters and TSS is still incomplete, and limited to small number of species. [46] The advancing knowledge of the sub-sequences of the promoters, increasing availability of high-throughput sequencing (HTS), and the incorporation of machine learning approaches into bioinformatics led to numerous attempts of *in silico* prediction of promoter sequences, which would allow identification of promoters also for novel transcripts or for organisms which lack annotations. The algorithms have mostly focused on core promoter identification, and have been based on models using parameters such as (starting from the oldest approaches):

- enrichment of known promoter motifs, e.g., PromFind [52]
- thresholds computed from conversion tables based on the sequence features, e.g., EP3 [53] and PromPredict [54]
- TSS signal computed with help of linear chain conditional random fields [46].

The accuracy of the early algorithms was low.[46] After employment of machine learning and deep learning methods, the recall rates have improved, and algorithms using those methods are currently the best ones for promoter prediction. Nevertheless, discussion about possible ways of further enhancement continue. For example, it has been pointed out that reducing feature vector dimensionality by more strict feature selection or adding resampling step might improve machine learning algorithms' performance. [55]

2.5 Chromatin accessibility

For transcription to happen, the transcriptional machinery needs access to the sequence of promoter and enhancer of a gene to be transcribed. The fundamental structural organization of DNA is chromatin consisting of repeating nucleosomes, which are composed of approximately 147 bp of DNA wrapped around an octamer of histone proteins. In this form, DNA is packaged within the nucleus, is protected from damage, and can be equally distributed during cell division. Nonetheless, such organization limits sequence accessibility and constitutes an obstacle for cellular DNA-based processes, including transcription. [56] For the transcription to occur in an efficient way, the chromatin needs to be transformed into more accessible form. [57]

Chromatin accessibility can be defined as the degree to which nuclear macromolecules are able to physically contact chromatinized DNA. Approximately 2-3% of the genome is accessible at a given time point. Accessibility is defined by the presence and positioning of nucleosomes, and by the occupancy of other DNA-associated factors such as TFs and architectural proteins. [58] But it is the nucleosomes that play the central role in regulating the transcriptional competence of various chromatin regions. During transcription, they constitute a barrier for the DNA Polymerase II. If a nucleosome is located at the promoter region of a gene, it prevents the DNA Polymerase II from loading DNA template and the assembly of the pre-initiation complex is blocked at that site. In addition, a single nucleosome suffices to disable transcriptional elongation. [56]

Nucleosome density varies in different regions of DNA sequence. In the core promoter regions just upstream the TSS, there are short stretches of DNA where the nucleosome density is very low. These stretches are called nucleosome-depleted regions (NDRs). [59], [60] NDRs occur also in the regulatory regions of DNA such as enhancers. Among factors responsible for the maintenance of nucleosome depletion within NDRs are BRG1/BRM-associated factor (BAF) and promoter-proximally paused RNA polymerase. [58]

Sequences downstream from TSS typically contain regularly spaced nucleosomes (well-positioned nucleosomes) with the regularity gradually fading as the distance from TSS increases, however, such trend is not universally observed. In fact, there are two main configurations associated with promoters. Constitutive promoters, which usually do not contain a TATA-box, frequently display the nucleosome pattern described above. [59] Those promoters include bending-resistive sequences, which makes them thermodynamically unfavourable for nucleosome formation. [56], [59] NDRs of constitutive promoters provide non-competitive conditions for the transcriptional machinery to assemble, and they frequently contain TF binding sites. Certain constitutive promoters might also bind TFs to enhance the expression level of the genes they regulate. [59]

A different nucleosome configuration has been observed in inducible promoters. These promoters frequently comprise TATA-box, but that motif and the TSS are typically occluded by nucleosomes when the target genes are repressed. Activation of such genes is triggered by environmental or developmental stimuli, and in response an activator, whose binding site typically is exposed, is recruited to the promoter, after which nucleosomes upstream from TSS are disassembled and TATA-box is rendered accessible to

TBP and the transcriptional machinery. Returning to the silenced state means chromatin reinstatement, although gene repression is not unequivocal to nucleosome occlusion and can be induced by other mechanisms, which are discussed in the coming chapters. The dynamic expression range of genes regulated by inducible promoters is larger than that of constitutive promoters. [59]

To achieve the desired nucleosome rearrangement, activator of inducible promoters is assisted by chromatin modulating mechanisms: histone modifications, which alter nucleosome composition, chromatin remodelling, which changes nucleosome positioning, and linker histones. Histone modifications entail both post-translational histone alterations as well as histone variant replacements. They are the best-studied modulators of nucleosome positioning, and they play an essential role in transcriptional activity as they mark active and repressed genes by changing the way histones interact with DNA. Non-canonical histone variants are installed more frequently at promoters and enhancers and might enhance TF binding and initiation of chromatin remodeling. [58] Histone modifications will be discussed in more details in the later chapters. [56]

Chromatin shape is also modulated by histone chaperones and nucleosome remodelers. Histone chaperones are responsible for the delivery of histone dimers to the sites where nucleosomes should be assembled after replication, or where nucleosomes have been ejected during transcription and need to be reinstalled. [61] However, it is the nucleosome remodelers that play the key role in reshaping chromatin accessibility, as they can move, slide, evict or even assemble nucleosomes using energy from ATP hydrolysis. [56], [62] Typically they form multi-subunit macromolecular complexes, and based on their ATPase domain, nucleosome remodelers can be classified into four main groups: SWI/SNF (SWItch/Sucrose NonFermentable), ISWI (imitation switch), INO80 (inositol-requiring mutant 80), and CHD (Chromodomain Helicase DNA binding). [56], [63], [64]

The members of SWI/SNF remodeler family are primarily responsible for nucleosome ejection. [56], [64] Those remodelers do not act in sequence-specific manner, but rather are recruited by TFs to the site of activation. [64] ISWI remodelers facilitate the maturation of nucleosomes from the prenucleosomes (initially associated histones and DNA). In addition, the phasing of mature nucleosomes is maintained by them. [65] INO80 complexes are also responsible for nucleosome spacing and they cooperate with ISWI complexes in arranging the nucleosomes downstream from TSS. Another function of INO80 remodelers is histone variants replacement. [66] CHD family is the least understood from the families of chromatin remodelers, but some members of this family might facilitate transcription-related nucleosome turnover, while other members seem to perform inhibitory functions in the process of transcription. [56]

Linker histones and other architectural proteins also regulate chromatin accessibility. They are responsible for nucleosome arrangement and for heterochromatin formation. According to the current knowledge, linker histones modulate DNA nucleosome exit angle, which enables chromatin packaging into less accessible forms. Histone H1 belongs to the family of linker proteins, and most probably maintenance of heterochromatin depends on it. The higher order of chromatin fiber geometry is believed to also contribute to the nucleosome positioning. [58]

Furthermore, torsion arising during transcription also affects nucleosome barrier. The elongating activity of transcriptional machinery generates bidirectional torsional forces: positive torsion ahead of and negative torsion behind the elongating Polymerase II. This torsion reorganizes chromatin and destabilizes nucleosomes standing on the way of the Polymerase II to promote elongation, while simultaneously enhancing nucleosome reassembly in the regions already transcribed to maintain chromatin integrity. [56] It has been demonstrated that the bodies of actively transcribed genes are not accessible right from the initiation of the process, but instead they are gradually exposed as the elongation proceeds. In addition, a proper positioning of nucleosomes might promote transcription elongation. These discoveries suggest that nucleosomes do not play solely inhibitory role but are in the center of accessibility regulatory mechanisms. [58]

Although accessibility of promoters and enhancers is necessary for transcription to happen, their open state is not equivocal to transcriptional activity of the gene they regulate. Therefore, chromatin accessibility is required for transcription, but is not sufficient for the process to be initiated and does not determine the state of activity. However, studying the accessible regions provides an insight into the landscape of potential regulatory regions in the genome. [58]

There are several methods available for measuring chromatin accessibility, one of the most widely adopted being assay for transposase-accessible chromatin using sequencing (ATAC-seq). [58] This protocol was first introduced in 2013 by Buenrostro and his lab, and it uses genetically engineered Tn5 transposase to ligate sequencing adapters to the regions of open chromatin. PCR-amplified fragments are then paired-end sequenced. [67], [68] This method can be used with even a small amount of genomic material [68], and it has been reported to be simple, robust, and fast [58]. Unfortunately, only one tool for ATAC-seq peak calling has been developed, and the standard way to perform the analysis is to use tools originally designed for ChIP-seq and DNase-seq, especially MACS2 and HOMER, despite no systematic evidence that they produce accurate results. [67] Chromatin accessibility measurement is often paired with RNA abundance measurement to study the influence of accessibility on gene expression [69], but recently there is an increasing interest in integrating ATAC-seq data with other epigenetic data to better understand the mechanisms behind different states of given cell type [58].

Chromatin organization is not a static binary condition, but it is a dynamic continuum that varies along the genomic sequence. The different stages of chromatin accessibility are the molecular representation of epigenetics. [26]

2.6 Epigenetics

The term epigenetics refers to an extra layer of instruction upon DNA on how the genes are read and expressed. [70] Epigenetic changes are heritable, and they alter gene expression or phenotype through chromatin modifications, while keeping the genotype unchanged. [62], [70] The epigenome entails all chromatin modifications in a given cell type, and this includes DNA methylation, post-translational histone modifications, and binding of the transcription factors to the DNA, which interact and determine transcriptome and consequently proteome of a given cell type. [12], [26], [70]

All the levels of epigenomic chromatin organization can be interpreted as superimposed layers, visualized in Figure 6. They constitute a complex coordinated mechanism of gene activation and deactivation, which is essential in cellular differentiation and reprogramming. However, a major part of the epigenomic landscape is highly dynamic, as it responds to developmental and environmental stimuli. [26], [62] Many epigenomic changes can be reversed through inhibitory action of chromatin-modifying enzymes and modification reader proteins. [62]

Studying epigenetics in healthy and pathologically altered tissues has become possible with the emergence of new advances technologies, e.g., chromatin immunoprecipitation using high-throughput sequencing (ChIP–seq). Characterization of epigenomic profiles has been instrumental in establishing DNA regulatory elements such as promoters or enhancers but has also resulted in deeper understanding of disease progression paths. [62] Consequently, multiple studies have revealed that epigenetic abnormalities contribute to development of cancers, including prostate cancer. [12], [48], [60], [71] Studies on cancer epigenomic profiles carry a significant potential of clinical translation, as new biomarkers or therapeutic targets have been and still might be discovered in the future. [60]

The following three chapters describe different levels of epigenome, their role in transcriptional regulation, and how they are altered in cancer initiation and progression. Although epigenome comprises also other chromatin modifications, they are not discussed here, as it is beyond the scope of this work.

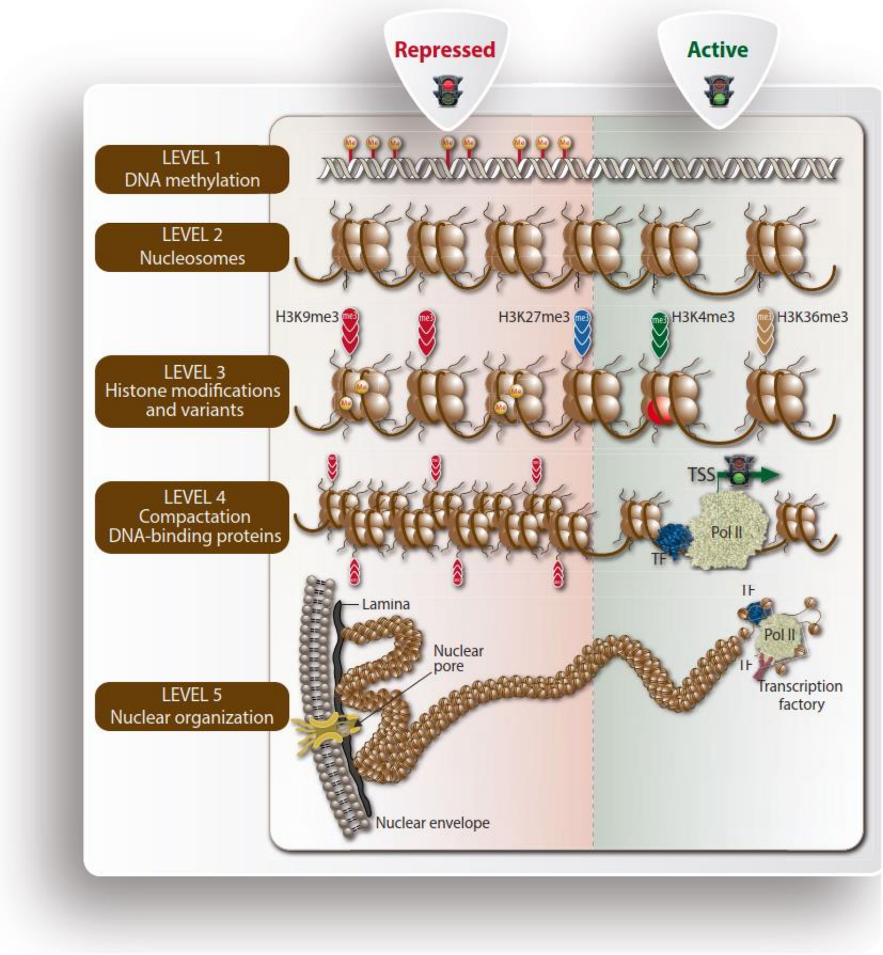


Figure 6. *Layers of chromatin organization. (Adapted from* [26]*)*

2.6.1 DNA methylation

The best-studied layer of chromatin organization is DNA methylation. It plays an essential role in the correct establishment of gene expression patterns. [51] DNA methylation means an epigenetic modification in which a methyl group is enzymatically added to DNA methyltransferase (DNMT) on the 5'-carbon of the pyrimidine ring in cytosine. [50] It is perpetuated through both mitotic divisions and meiotic divisions by maintenance DNA methyltransferases. [72] DNA methylation, next to transcription factors, is one of the elements of the epigenetic memory, which manifests itself in cells' ability to maintain the information about tissue characteristics. [73]

In mammals, DNA methylation is mainly observed in CpG dinucleotides within stretches of repetitive DNA, which are hypermethylated already during early embryogenesis, and in regions known as CpG islands. [26], [51], [72] CpG islands constitute 200 bp long regions containing more than 55% of GC and with an expected GC content to observed

GC content ratio greater than 0.65. [50] They are mainly associated with gene promoters. [50], [51] High DNA methylation level at the promoter rich in CpG islands is associated with transcriptional repression, and thus the methylation level is inversely correlated with the expression of the gene being regulated by that promoter. [26], [50], [51] Methylation of gene promoter regions is also known to play an essential role in genomic imprinting, e.g., silencing a parental allele, or in X-chromosome inactivation in females. [50], [51] CpG containing promoters maintain the phased positioning of nucleosomes. [59] The expression of genes with no CpG islands in their promoter is regulated by other mechanisms, e.g., binding of TFs to enhancers. [26]

Approximately 40% of CpG islands is intragenic or intergenic. [50] Intragenic CpG islands are found in highly expressed, active genes, and for those methylation level is positively correlated with the gene expression. [26], [50], [72] The intergenic CpG islands are involved in regulation of non-coding RNAs transcripton. [50]

While DNA methylation at a given locus is mostly a stable silencing mark, approximately 15-20% of all CpG islands in the human genome undergoes dynamic changes in methylation in a healthy organism. [26], [74] Those dynamically methylated regions are typically located away from TSS and overlap distal regulatory elements. They are referred to as differentially methylated regions (DMRs), and they gain or lose the methyl group in a lineage-specific manner and establish the identity of a tissue. [74]

Although DNA methylation has been studied extensively, the precise mechanism behind its influence on the chromatin remodelling processes and gene expression is not yet clear. However, three groups of epigenetic modifiers are known to modulate and interpret DNA methylation: writers, readers, and editors. Proteins from DNMT family, such as DNMT1, DNMT3A and DNMT3B, constitute the group of writers, and they establish and maintain methylation patterns during developmental stages and in the process of cellular differentiation. Readers comprise a wide group of proteins from various families: methyl-CpG-binding domain (MBD) proteins, the Kaiso family proteins and the SET- and Ring finger-associated (SRA) domain family proteins. Those proteins specifically bind CpG dinucleotides and are instrumental in the close regulatory interplay between DNA methylation and histone modifications. As a result, they influence chromatin packaging into heterochromatin, nucleosome remodelling, histone modifications, and higher order chromatin organization, which are discussed in other parts of this thesis. As discussed earlier, MBD proteins belong to SWI/SNF family of nucleosome remodelers. The last group of epigenetic modifiers, editors, is formed by proteins from the ten–eleven translocation (TET) protein family, whose activity has been suggested to lead to promotion of demethylation. [75]

In addition to the interactions with modifying proteins, DNA methylation is known to be involved in bidirectional relation with TFs. The presence of cytosine methyl groups is recognized by TFs, and it modulates TFs binding capability. Depending on a TF and the genomic context, binding of TFs to their recognition motif can be inhibited or enhanced by methylation [73], [74]. On the other hand, binding of certain TFs prevents cytosine methylation, e.g., at active promoters. [72], [74] It has not been established whether methylation is a consequence or the cause of TF binding. [74]

As mentioned above, DNA methylation is tightly connected to histone modifications. These two levels of epigenome regulate gene expression through close interactions. Generally, a pattern of dependency has been observed between histone acetylation and deacetylation, and DNA methylation. Acetylation of histones by histone acetyltransferases (HATs) co-occurs with lack of cytosine methyl groups, marking the regions of transcriptionally active chromatin. Deacetylation of histones by histone deacetylases (HDACs) and histone methylation by lysine methyltransferases (KMTs) typically is followed by DNA methylation, marking the regions of transcriptionally repressed chromatin. [26] Histone modifications are discussed in more detail in the next chapter of this section.

Abnormalities in methylome are result of aging [50], [76], but they have been also identified in cancer [51], [60], [76], [77]. Such changes include both gain and loss of DNA methylation, referred to as hyper- and hypomethylation, respectively, which result in aberrant transcriptional outcome. [76]

Among the abnormalities that attracted the most attention of researchers is hypermethylation in normally unmethylated gene promoter CpG islands, which results in transcriptional repression and loss of gene function, and often happens in tumor suppressor genes. Moreover, other hypermethylated genes might aggregate in the same signaling pathways, contributing to cancer initiation and complementing single driver genomic mutations. The gain of methylation in promoters has also been linked to repression of multiple microRNAs and other non-coding RNAs. This kind of changes might activate the modes of cellular signaling that promote invasiveness and metastasis. Downregulation of certain microRNA families might lead to overexpression of DNA methyltransferases, potentially allowing the gain of methylation in gene promoters. [60]

However, on the genome-wide level, the most widespread methylome aberration in cancers is DNA hypomethylation. [74] Hypomethylated regions span megabases of genomic DNA on multiple chromosomes. These large domains of losses and gains are associated with late-replicating, lamin-associated nuclear regions that contain the majority of the genes with bivalent, chromatin promoter domains, which are highly vulnerable to abnormal CpG island DNA hypermethylation in cancer. [60]

Aberrant methylation of certain genes has been associated with BPH and with prostate cancer progression. Promoter hypermethylation of critical tumor-suppressors such as *APC* and *RASSF1* have been reported in prostate cancer. Genome-wide hypomethylation has been observed in advanced and metastatic PC. [78] A study in 2021 [79], using whole genome bisulfite sequencing, found abnormalities in methylation profiles associated with Polycomb repressed regions and in promoters associated with bivalency. In addition, it was discovered that the changes enriched for binding motifs of AR and MYC, and that dynamic DNA methylation patterns observed in the normal luminal cell differentiation program were significant targets of aberrant methylation in PC. [79] Recently, Tonmoy and colleagues identified abnormal expression of lncRNAs related to poor prognosis of PC patients, resulting from aberrant methylation patterns. [80] These findings are in line with the general abnormal patterns reported in cancer methylomes.

The experimental methods of genome-wide characterization of DNA methylation can be divided into three main groups based on the mechanism used to detect methylated cytosines [81]:

1.) methods based on restriction enzymes,

2.) methods based on affinity enrichment,

3.) methods based on bisulfite conversion.

The methods most relevant in the context of this thesis are affinity-based methods, in particular methylated DNA immunoprecipitation (MeDIP). This technique takes advantage of an antibody specific for methylated cytosines to immunocapture methylated genomic fragments. [77] MeDIP is usually coupled with HTS to evaluate the relative enrichment of methylated DNA cross the genome. However, it is noteworthy that the analysis of data produced by MeDIP-seq is not straightforward – it requires normalization steps to compensate for the bias introduced by CpG-rich fragments, which results in underrepresentation of regions with low CpG content. Currently, methods based on bisulfite conversion coupled with sequencing are regarded as the gold standard for DNA methylation detection. [81]

The experimental methods are still sometimes coupled with methylation arrays. Despite the fact that the HTS offers a great advantage of genome-wide signal detection, it has not rendered the arrays completely irrelevant. Even though arrays produce information about only single-site methylations, their sensitivity, specificity, and reproducibility are on a level inaccessible yet by the HTS, and thus they are still widely used, for example in cancer methylome studies. [81]

2.6.2 Histone modifications

Methylated DNA attracts different transcriptional activator and repressor complexes, including histone modifying and chromatin remodeling enzymes that regulate chromatin structure. As discussed earlier, nucleosomes constitute a fundamental subunit of chromatin structure. A nucleosome contains an octamer formed by two copies of each of the core histones H2A, H2B, H3, and H4. Additional histone H1, belonging to the family of linker histones, binds to the DNA where the turns around a nucleosome start and end. [50] Such formation of a nucleosome and H1 is called chromatosome[26], and it can be seen in Figure 7. Unlike transcription factors, histones do not bind to specific sequences of genomic DNA, but rather to its phosphate-sugar backbone. [26]

The core histones are basic proteins composed of a globular domain and highly flexible unstructured N-terminal tails that protrude from the DNA wrapped nucleosome. The N-terminal tails are subject to post-translational modifications. Those modifications entail phosphorylation, acetylation, and methylation, but here the focus will be put on two latter ones, as to date, they are best characterized. [50], [82] Histone acetylation is a catalytic addition of acetyl-coA to the ε-amino group on lysine side chains of histone tails. Histone methylation means an enzymatic addition of a methyl group to the residue lysine and arginine. Lysines can be mono-, di-, and trimethylated while arginines can be mono- and symmetrically or asymmetrically dimethylated. [50] Individual instances of histone marks have been associated with different regions of transcriptional activity or repression, although the actual functionality of some of the marks has not been confirmed. [50], [82]

On a general level, as has been mentioned in the chapter regarding DNA methylation, high level of histone acetylation is characteristic for euchromatin, whereas histone deacetylation is typical for heterochromatin. [50] Furthermore, promoters of genes which are actively transcribed are associated with enriched trimethylation on histone H3 lysine 4 (H3K4me3) and lysine acetylation on histone H3 and H4. The bodies of actively transcribed genes, in turn, are marked with enriched H3K36me3 and H3K79me3. Active proximal enhancers have their own histone modifications: H3K27ac and high levels of H3K4me1 relative to H3K4me3. Histone marks usually associated with gene repression are H3K9 methylation, H3K27me3, and H4K20me3. [82] In addition, marks associated with both silencing and activation, H3K4me3 and H3K27me3, can be found within the same gene promoter, albeit on different N-terminal tails, determining poised genes – those being in an intermediate state. Such regions have been referred to as bivalent chromatin. [62] Histone modifications can be used to determine transcriptionally active promoters. [82]

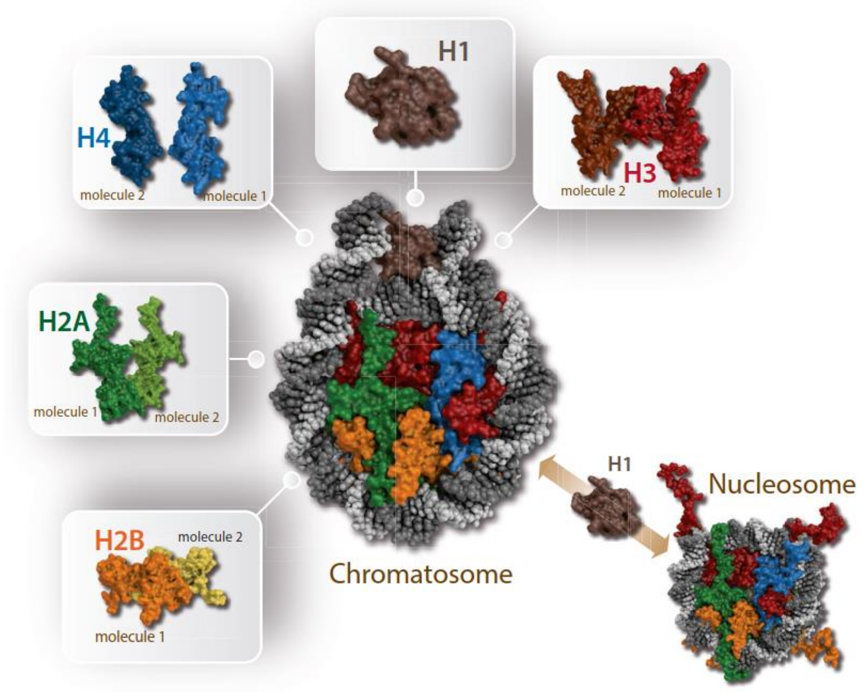


Figure 7. *A representation of a nucleosome. Green molecules correspond to H2A, orange molecules to H2B, red molecules to H3 and blue molecules to H4. The gray strand represents genomic DNA, whereas the brown molecule represents the linker histone H1. (Adapted from* [26]*)*

The approximated genomic locations of the histone marks associated with transcriptional activity are illustrated in a simplified way in Figure 8. In the figure, each colored peak is a visualization of what ChIP-seq signal shapes would look like for a given histone mark and where would it be situated in relation to NDRs, TSS (the arrow) and TTS.

Similarly to DNA methylation, histone modifications are modulated and interpreted by enzymes. And so, writers deposit histone marks with help of cofactors, and erasers, analogically, reverse the actions of writers. Readers, which frequently act as part of and in cooperation with large protein complexes, bind to modified histones and initiate appropriate reactions to the marks. Furthermore, readers contain multiple domains, which recognize different histone marks, which makes them capable of reading several marks at the same time. Due to this feature, it is believed that histone marks readers could be attracted by entire combinations of marks, and not by single modifications. [82] In the group of writers are histone acetyltransferases (HATs) that catalyze acetylation, and histone methyltransferase (HMT) enzymes, which perform histone methylation. Erasers include histone deacetylases (HDACs), which remove acetyl marks on lysine to restore the positive charge, and lysine-specific demethylase 1 (LSD1), an amine oxidase that catalyzes lysine demethylation and releases the product hydrogen peroxide. Arginine demethylation is not as straightforward and involves protein arginine deiminase 4 (PADI4). PADI4 does not perform a complete demethylation, so that histone replacement or further demethylation by aminotransferases is required. [50]

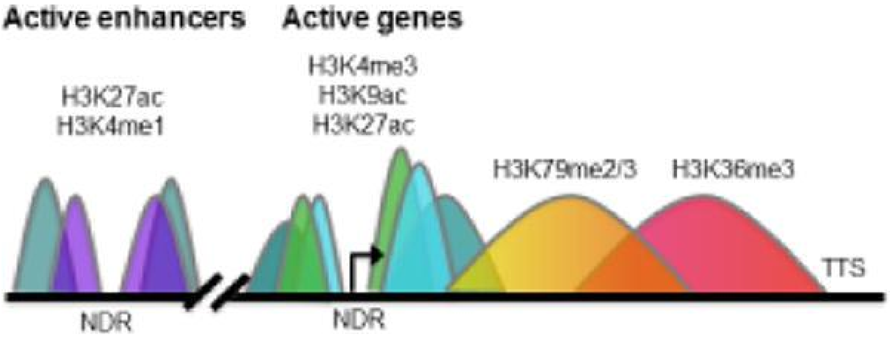


Figure 8. *Genomic localization of histone H3 modifications associated with active transcription. Colored peaks visualize ChIP-seq signal shapes which would correspond to individual histone marks in a eukaryotic gene. The arrow represents TSS. (Adapted from [82])*

There are three main modes of influence that histone marks exert on transcription. First, they prevent binding of DNA-associated proteins such as TFs, disabling their enhancing functions. Next, they may associate with proteins, which promote or repress gene activation. Finally, they directly affect chromatin structure. [50] Acetylation neutralizes the charge of histones, which breaks electrostatic interactions between histones and DNA, decreasing the level of chromatin packaging and increasing chromatin accessibility. Histone methylation, in turn, does not affect the charge of histones. [83] Moreover, histone marks can also impact one another. [82]

Abnormal accessibility to target DNA sequences causes loss or mutations of enzymes responsible for histone mark modulation, which, in turn, might lead to changes in histone modification patterns. Such changes in regulatory segments of DNA results in transcription dysregulation and alterations in gene expression programs. [50], [83] In addition, aberrant histone modification profiles are associated with genome instability and defective chromosome segregation. All these anomalies might contribute to cancer development. [83]

Chromatin immunoprecipitation followed by sequencing (ChIP–seq) has become the standard method for genome-wide profiling of DNA-binding proteins such as TFs, histone modifications or nucleosomes. [84], [85] The main steps of a ChIP-seq experiment include 1) crosslinking the DNA-binding protein of interest to DNA in vivo by treating cells with formaldehyde; 2) chromatin sonication into small (200-600 bp) fragments; 3) immunoprecipitation of DNA-protein complex with protein-specific antibody 4) DNA purification; 5) fragment amplification; 6) sequencing. ChIP-seq offers higher resolution, fewer artefacts, greater coverage and a larger dynamic range than its predecessors, such as ChIP–chip. [84] Nonetheless, this technique introduces also several difficulties. The most important one is the dependency on the quality of antibody. Moreover, prior to ChIP-seq experiment, one must know that a histone modification or DNA-binding protein is present to choose appropriate antibody. Finally, the protocol requires a significant amount of genetic material. [85]

Gene Transcription Regulation Database (GTRD) is an initiative started in 2011 that aims to provide uniform annotation and integrative analyses of all HTS data from publicly available repositories GEO and SRA that are related to transcription regulation and are widely utilized by the researchers in the field. It is a source of annotated and processed data from ChIP-seq, ChIP-exo, DNase-seq, MNase-seq, ATAC-seq and RNA-seq experiments available for non-commercial use. [86] Data from GTRD can be used for an integrative data analysis when a specific layer of epigenomic information would be useful to validate observed trends, but it is missing for the studied sample cohort.

Another chromatin remodeling mechanism involving histones is histone variant replacement. Histone variants are proteins that correspond to core histones, called canonical, but whose amino acid composition is encoded by different genes and thus differs from the canonical paralogues. [83], [87] As described above, the canonical histones are responsible for transcription regulation and chromatin organization. However, histone variants have many different functions, but in the context of transcription, they are recruited to the transcription initiation sites or termination sites to facilitate the respective stages of the process. [87]

2.6.3 Transcription factors

Gene transcription programs are driven by transcription factors, whose sequence-specific binding to DNA recruits activating or repressing coregulators. [82] Approximately 8%

of human protein-coding genes encode for transcription factors. [26] A transcription factor is a protein that recognizes and binds one or more specific motifs in the DNA sequence, called transcription factor biding sites (TFBSs), located in the regulatory regions of genes, and through this promotes or inhibits the transcription of the target gene. TFBSs are short, ranging from 5 to 20 bp and they are not necessarily exclusively bound by a single TF. [88] TF recruitment is either initiated by intracellular processes such as development and differentiation, or as a response to an extracellular stimulus. TF activity changes when its abundance in a cell is altered through transcription, translation or post-translational regulation, or when its binding sites accessibility changes. Furthermore, TF activity is affected by the availability of coactivators. [89] Despite the fact that certain TFs can bind to their cognate motifs within nucleosomal DNA, albeit often with lower affinity [59], over 90% of the regions containing known TFBSs are located within open chromatin. Thus, open chromatin typically reflects the presence of aggregate TF binding, and - as has been already stated before - marks putative regulatory regions associated to genes. [58] Binding of a single TF can facilitate chromatin opening for other TFs or prevent other TFs from binding. [89] Transcription factors enable cell differentiation by adding a new layer of expression instructions for cells containing identical genomic code. [58], [88]

In TFs structure, two important elements can be distinguished: one or multiple DNA-binding domains (DBDs), which anchor one or more activation domains (AD), distinct from DBDs. Structured DBDs are responsible for the recognition and binding of the cognate motifs, and they have been studied extensively. Their structural features have become the criterion of TF classification, and thus we have zinc-coordinating, basic helix-loop-helix, basic-leucine zipper, or helix-turn-helix DNA-binding transcription factors. [90] DBDs recognize the motifs via two protein-DNA mechanisms: base readout and shape readout. The first mechanism entails physical interactions between the amino acid side chains and the accessible edges of the base pairs, e.g., direct hydrogen bonds, water-mediated hydrogen bonds and hydrophobic contacts. The shape readout, in turn, includes recognition of the static and dynamic structural features of the DNA binding sites, such as sequence-dependent DNA bending, unwinding, and the electrostatic potential. [91]

The function of less studied ADs, in turn, is to cooperate with coactivators to modulate transcription. They are low-complexity, intrinsically disordered regions categorized based on their amino-acid composition into acidic, proline, serine/threonine, or glutamine rich ADs. One of the coactivators interacting with TF activation domains is Mediator complex. [90] Mediator complex, or Mediator of RNA polymerase II transcription, is a multi-subunit protein complex recruited by TFs to the enhances of transcriptionally active genes. It promotes the assembly of pre-initiation complex at the core promoter and constitutes a physical bridge between enhancer and promoter bound TFs. Furthermore, Mediator stimulates phosphorylation of the DNA Polymerase II, allowing the transition of the latter one from transcription initiation to elongation. In the context of TFs, the key role of Mediator complex is forwarding – mediating - signals from TFs present at gene regulatory regions to the transcription machinery. [92] However, not all TFs couple their activity with Mediator complex.

The members of a family of TFs called nuclear receptors do not prefer cooperation with Mediator complex, but rather act together with coactivators. Nuclear receptors belonging to ligand-induced TFs get activated through binding to specific ligands. Certain nuclear receptors reside in the cytoplasm, where they are associated with chaperones until the required ligand arrives and binds to the TF, releasing the chaperons. Activated TF-ligand complexes can be transported to the nucleus. One of the ligand-induced nuclear receptors is AR, which was discussed in chapter “Prostate Cancer”. However, most of the nuclear receptors are located and activated in the nucleus and do not require such complicated process to get triggered. Signal transduction of the nuclear receptors sometimes disrupt other signaling pathways, causing post-translational modifications of nuclear receptors or their coactivators. [26]

The significance of TFs for the efficiency of transcription is reflected in the fact that sole chromatin accessibility at the enhancers and promoters does not result in substantial RNA production. The abundance of transcripts is not even significantly increased when the transcriptional machinery has assembled on the promoter. It is the binding of an activating TF that triggers upscaled transcription. [26]

ChIP-seq, which was already mentioned in the context of experimental methods available for histone mark investigation, is currently also the “golden standard” to study TF binding sites on a genome-wide scale. [93] In addition, sequence-based computational methods have been developed to model TFBSs. A plethora of algorithms is currently available, and most of them are based on Position Weight Matrices (PWMs). [91], [93] PWM describes the preference for all four nucleotides at each position of TFBS motif in form of a 4xn matrix, where n is the length of the motif. [93] PWMs can be easily and intuitively visualized as TFBS motif logos. [91] A logo plot represents the relative frequency of each nucleotide by stacking characters corresponding to them on top of each other, with the height of each character proportional to its relative frequency. The characters are ordered by their relative frequency, and the total height of the stack is determined by the information content of the position. [94] An example of a typical sequence logo can be seen in Figure 9.

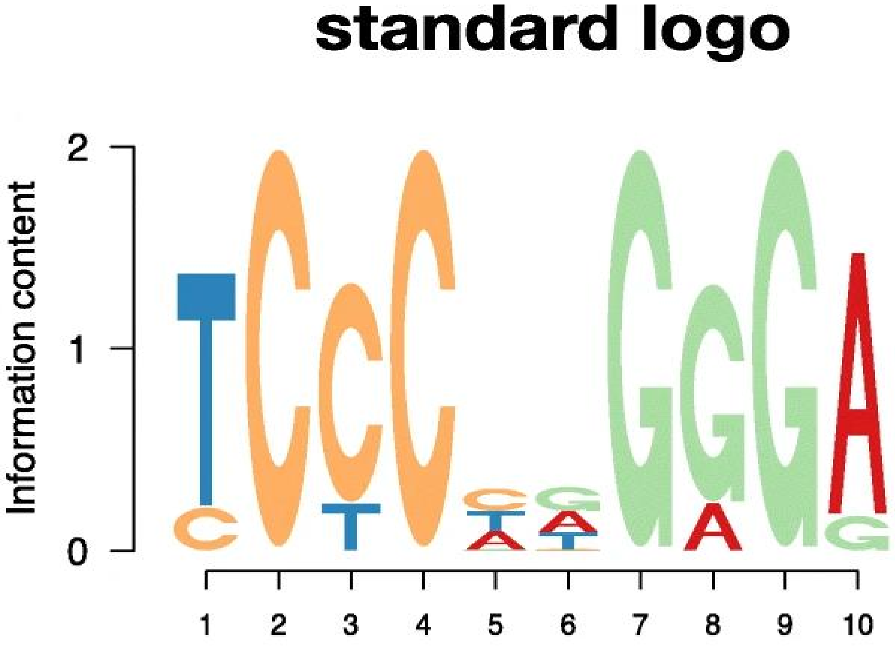


Figure 9. *Standard sequence logo plot (Adapted from* [94]*)*

Studying the genomic maps of TFBSs provides an insight into gene regulatory networks. [88], [89] Since genes and proteins do not work in isolation, but rather in coordinated systems, definition of the relations between genes and gene products has become a focal point in the field. Gene regulatory networks are the emerging results of such efforts, and they can be defined as the summarized activity of a TF set connected to its targeted genes. Maintaining cell-type specific transcriptional states and stimuli responses is possible due to the coordinated activity of gene regulatory networks. Nonetheless, the precise functional principles of gene regulatory networks are not yet understood. [89] TF recruitment at the core promoter regions rendered its sequence one of the focus points in the studies of TFBSs. (de Medeiros Oliveira et al., 2021)

Multiple databases collect motifs of known transcription factors, which have been inferred from in vitro and in vivo experiments, including ChIP-seq. [89] One such database is JASPAR, an open-access source of curated, non-redundant TF-binding profiles stored as PFMs for TFs across multiple species in six taxonomic groups. With the start of 2022, the nineth release of the database was published. [95] Data available in databases like JASPAR can be used to predict putative binding sites of TFs of interest in a studied genome. However, the presence of a binding motif in the sequence alone does not reflect the actual binding of a TF, and so such predictions are limited by high false discovery rates. [89]

Transcription factors have been associated with the evolution of human diseases, and to date more than 150 transcription factors have been identified to be directly responsible for nearly 300 diseases, and more discoveries are yet to come. Many TFs are encoded by oncogenes, e.g., MYC (MYC proto-oncogene, BHLH transcription factor), or by tumor suppressor genes, e.g., TP53, whose abnormal regulation and activity leads to diseases. [26] Certain TFs and their aberrant transcriptional activity have been identified to drive prostate cancer emergence and progression. [96] The most important TF in the context of PC is the nuclear receptor transcription factor, AR, responsible for the growth of PC in the initial stage of the disease [96], but also essential in the progression to castration resistance [97]. The cancer driving mechanism of AR was described in chapter “Prostate Cancer”.

Recent developments in prostate cancer studies revealed that three pioneer transcription factors facilitate AR-driven transcriptional programs: Forkhead Box A1 (FOXA1), Homeobox B13 (HOXB13), and GATA-binding factor 2 (GATA2). Normal prostate development and AR functioning are dependent on those TFs, however, they have been found to also promote AR oncogenic activity. Interestingly, there is evidence that AR-independent functionality of FOXA1, HOXB13, and GATA2 might inhibit some stages of PC development. One way or another, these three TFs are important in PC disease. [98]

ETS-related gene (ERG) is a member of the E-26 transformation-specific (ETS) family of transcription factors, which in normal conditions, is not expressed in prostate epithelial tissue. However, due to a gene fusion with the androgen-driven promoter of the TMPRSS2 gene, it is persistently overexpressed in many PCs, especially in the advanced tumors with high Gleason score. Furthermore, ERG overexpression is associated with metastasis and poor prognosis. [99]

Abovementioned transcription factor c-MYC (MYC) also plays a role in PC development, and its expression levels are high at the early stages of the disease, as well as in the advanced PC. MYC interacts with other transcription factors prominent in PC development, AR and FOXA1, but also with the DNA Polymerase II to disrupt normal AR transcriptional activity. Consequently, MYC activity leads to the emergence of cancer, and then to the progression into castrate-resistant and metastatic state. [100]

2.7 Previous work and the starting point

In 2015, Ylipää and colleagues [23] assembled and studied the transcriptome of BPH, PC and CRPC samples from Tampere Prostate Cancer cohort to characterize the differences between the pathological states. They were particularly interested in discovering novel long non-coding transcripts, specific for PC and CRPC. They annotated transcripts based on the exonic and intronic sequences in human reference transcriptomes, and classified transcripts into three groups: protein-coding, previously annotated lncRNA, and a large group of “novel loci of expression”, which comprised of 99 120 transcripts. Novel transcripts were further broken down into two categories: intragenic, which were found to be fully contained in an intron, and intergenic, which did not overlap with any exonic or intronic sequences. They applied strict filtering on the data of novel transcripts and focused on differentially expressed transcripts. While they described differences in the transcriptomic profiles of PC and CRPC, identified a small group of putative and defined novel lncRNA named PCAT5, they did not study the unannotated transcripts as a group. Nevertheless, the data produced by that study inspired questions: why is there so much transcription from the unannotated regions? Is there a group of unannotated transcripts, which are biologically meaningful, and thus, whose transcription is regulated?

The studies on the same cohort were continued, and in 2020, Uusi-Mäkelä and colleagues [48] performed an integrative analysis of chromatin accessibility with transcriptome, methylome, and proteome profiles of the same samples. They identified chromatin alterations related to the disease progression towards CRPC. Moreover, they investigated the TF binding patterns and used correlation studies to find putative regulatory elements for cancer-associated genes and described their influence on the cancer phenotype. However, they focused on the protein-coding genes. Nevertheless, the ATAC-seq and MeDIP-seq data obtained for the needs of their study provided the epigenomic information that opened new possibilities to investigate the unannotated transcripts identified previously by Ylipää et al. Both works provided the material for starting point for this thesis.

3. MOTIVATION AND OBJECTIVES

Prostate cancer is the third cause of cancer deaths in men in European Union, with the predicted mortality rate for 2022 being 9.49 per 100 000 citizens [101]. As such, PC constitutes a significant part of the cancer incidence as well as cancer-related mortality and exerts a substantial burden on the European health care systems and society. Thus, efforts to find more effective ways of diagnosing and treating prostate cancer are of great value. Recently, it has been discovered that non-coding RNAs play an essential role in tumorigenesis, including the development of prostate cancer. To enable further advances in the understanding of prostate cancer, it is crucial to investigate the non-coding genomic elements of the normal prostate and prostate cancer to find their involvement in the disease initiation and progression. Potential discoveries of novel transcripts driving oncogenesis or castration resistance not only provide new insights into PC evolution and enrich our knowledge of tumor heterogeneity, but also might result in new, highly specific diagnostic and prognostic markers, or therapeutic targets.

The objective of this thesis was to integrate transcriptomic data and several levels of epigenomic data to investigate whether they provide evidence that non-coding transcripts identified in the studied sample cohort are targeted by regulatory mechanisms similar to those orchestrating protein-coding gene transcription. Thus, the goal was to study if the epigenomic signatures characteristic for RNA Polymerase II transcription could be observed for the set of unannotated transcripts. The detailed aims of this investigation included:

1. Studying the relationship between chromatin accessibility and transcript expression
2. Examination of the relationship between expression, chromatin accessibility, and methylation
3. Verification of promoters of novel transcripts through promoter prediction algorithms *
4. Analysis of the presence of histone modifications related to transcriptional activity in the promoters of unannotated transcripts *
5. Investigation of the presence of AR TF binding sites in the promoters of the transcripts

Steps 3 and 4 marked with an asterisk were in part performed within a course project work. Some of the results have been reported previously, so they do not constitute a part of this thesis, and whenever this is the case, it is emphasized in the text. However, all the results are relevant in the context of this study, and thus are also presented in the thesis to provide a full picture of the analysis performed on this dataset.

To the knowledge of the author of this thesis, there is currently no publication focused on studying unannotated transcripts in the prostate and prostate cancer in a similar way, which means that this thesis sheds new light and highlights the possibilities in research conducted on the role of non-coding genomic elements in prostate cancer.

4. MATERIALS AND METHODS

This section explains the source and condition of the biological samples from which sequencing data was generated. Next, each chapter corresponds to a step of analysis and describes the type and source of data utilized at that stage, and the methods applied to analyse the data.

4.1 Samples and datasets

The sample group consisted of 10 benign prostatic hyperplasia (BPH) samples representing non-cancerous prostate tissue, 16 untreated prostate cancer (PC) samples, and 11 castration-resistant prostate cancer (CRPC) samples representing advanced prostate cancer. Samples were acquired from Tampere University Hospital as fresh frozen tissue specimen, obtained either through transurethral resection or radical prostatectomy [23]. The datasets from previous studies that were used in this thesis were RNA-seq, ATAC-seq, and MeDIP-seq.

4.2 ATAC-seq data processing

The transcriptome assembly from Ylipää et al., 2015, as well as their classification of transcripts were used to define the main object of interest in this thesis. Their whole transcriptome paired-end sequencing was performed on the Illumina HiSeq 2000 and the assembly was done with Cufflinks using NCBI 37.2/hg19 genome build. The transcripts categorized as intergenic and intragenic became the focus of this study, and the results of analysis of data related to them was compared to the results of coding and lncRNA transcripts for reference. For simplicity, coding and lncRNA transcripts will be further collectively referred to as “annotated transcripts”, and intergenic and intragenic transcripts will be referred to as “unannotated” transcripts.

In the study by Uusi-Mäkelä et al., 2020, the same transcriptome was aligned to GRCh38 using LiftOVer. The new alignment was then used in this thesis. Transcripts from mitochondrial genomic regions were filtered out.

Uusi-Mäkelä et al. also performed assay for transposase-accessible chromatin sequencing (ATAC-seq) from the same samples. ATAC-seq reads were aligned using Bowtie2 version 2.3.4.1 against GRCh38 reference genome, and peak calling was done with

MACS2 v2.1.0. Final ATAC-seq peak quantification included background correction, normalization and bias correction compensating for sample collection procedures, which all are described in their article [48]. ATAC-seq peaks quantified this way were used in the initial analysis steps.

4.3 Finding distances from TSS to nearest ATAC-seq peak

Majority of the analytical work in this and further steps was performed using R programming language version 4.1.2 in RStudio Workbench 2021.09.2 Build 382.pro1 with custom scripts. To find the closest peak from each transcripts' transcriptional start site (TSS) along with its distance, *bedtools closest -D* was used. The distance distributions with respect to TSS were further analysed in R.

4.4 Quantification of ATAC-seq peaks in individual samples

The transcript counts were manually normalized in R to obtain transcripts per million (TPM) counts. TPM can be obtained by applying the following formula [102]:

$$TPM = 10^6 * \frac{reads\ mapped\ to\ transcript/transcript\ length}{Sum(reads\ mapped\ to\ transcript/transcript\ length)}$$

TPM is dependent on the transcriptome composition in a sample, and it describes the relative abundance of a transcript among the sample population of sequenced transcripts. TPM is not suitable for differential analysis between samples, but it is a good measure for within-sample gene expression comparisons. [102] This thesis compares the transcriptome and epigenome profiles of different transcript groups to study their similarities within individual samples, and therefore, TPM was chosen as the measure of transcript abundance.

In addition, this step required sample specific ATAC-seq peak quantification. Quantification was performed by following the method described by Uusi-Mäkelä et al., 2020, but only until background correction step. Because the quantification was done for individual samples, median-of-ratios normalization or correction to account for acquisition bias were not needed. This processing was done using custom scripts in Python 3.9.7. Distribution of the distances to the nearest peak from TSS was again checked for individual samples. Then, the overlaps between the promoters and the sample specific peaks were called using *bedtools intersect -wo*. From this step onwards, whenever ATAC-seq peaks are mentioned, the individually quantified sample-specific peaks are meant.

After that, the peaks falling within the promoters of the transcripts were compared to an ATAC-seq blacklist from Buenrostro lab [105] containing artifact regions from mitochondrial homologs. The tool used for this purpose was *bedtools intersect*. The peaks which intersected the artifact regions were filtered out, and transcripts whose promoters contained those peaks were excluded from further analysis.

4.5 Correlation between accessibility and expression

Then, the correlations between ATAC-seq peak intensity and the expression levels were computed for each transcript type in each sample in R with *cor.test* command with *method* parameter set both to "pearson" and "spearman", corresponding to Pearson’s correlation and Spearman’s correlation, respectively. Pearson’s correlation, or Pearson’s product-moment correlation, is a measure of linear dependence between two variables, and is the most commonly used for studying the relationship between gene expressions. [103] However, it works well only for linearly related variables [103] with no or small number of outliers [104]. Spearman’s correlation, or Spearman’s rank correlation, is a nonparametric measure of monotonic dependence between two variables [103], suitable for heavy-tailed datasets [104]. Since there were significant outliers in the data (transcripts with very high expression or very high peak intensity), Spearman’s correlation was chosen as the main measure of relationship between chromatin accessibility and transcript abundance. The promoters were defined as 1000 bp upstream - 100 bp downstream from the TSS (which was adapted from Uusi-Mäkelä et al., 2020).

4.6 MeDIP-seq data processing and correlation between methylation and expression

MeDIP-seq reads were aligned to GRCh38 using Bowtie2, and subsequently quantified and normalized, however, that quantification was not used in this thesis. Instead, the result of QSEA (Quantitative Sequencing Enrichment Analysis) tool quantification was utilized in this thesis. Such quantified signal was then overlapped with the transcripts’ promoter regions. To obtain a methylation level of each promoter, a weighted average was computed based on the length of overlaps. Then, the correlations between methylation and expression were computed in R using *cor.test*.

4.7 Promoter prediction based on sequence composition

Three publicly available promoter prediction tools were chosen for promoter prediction: EP3, PromPredict, and TSSFinder. EP3 (Easy Promoter Prediction Program) is an algorithm created to identify the core regions of gene promoters in eukaryotic organisms, including human, written in Java and publicly available as an online tool, but also for download with a graphical user interface, or as a command line tool. It takes a FASTA file with a single sequence as its input. The algorithm calculates numerical profiles of the nucleotide sequence using experimentally validated conversion tables for sequence features, obtained from literature. The features that comprise the profile are stacking energy, propeller twist, nucleosome position preference, bendability, A-philicity, protein induced deformability, duplex stability disrupt energy, duplex stability free energy, DNA denaturation, DNA bending stiffness, B-DNA twist, protein-DNA twist, and stabilizing energy of Z-DNA. Then, averages of these profiles are computed for both the entire genome and for 400 bp windows. Thresholds are computed based on the whole genome profiles (for human, it is the whole genome average plus three standard deviations), and if a window's profile exceeds the thresholds, it is called as a putative core promoter. The formula for thresholds and the size of the window were established empirically by the creators of the tool. [53] The outcome is a list of 400 bp long core promoter predictions along with their genomic locations. The predictions are not linked to any genes. EP3 version 1.10 (the most recent version) with user interface was used to predict promoters. The tool is available at the site: http://bioinformatics.psb.ugent.be/webtools/ep3.

PromPredict is a promoter prediction algorithm implemented in PERL, utilizing the structural properties of DNA sequence. It was initially developed for bacterial organisms, and was gradually tweaked to be applicable to eucaryotes, including human. PromPredict evaluates relative free energy of neighboring regions within fragments 500 bp upstream and 500 bp downstream from TSS locations taken from public databases. Average free energy is computed for overlapping 100 nucleotide fragments (frameshift of one nucleotide). This average, and a relative free energy difference between neighboring fragments is compared to predefined thresholds determined from the entire 1001 nucleotide window based on its GC content, to call putative promoter sequence. The program's input is a FASTA file with a single sequence, and the output is a text file with the start and end of the predicted promoter regions along with the least stable position (lsp) in the predicted regions. The output promoter regions are not linked to genes. [54], [106], [107] Windows executable (genome sequence > 10MB) Version 1 was downloaded from the tool's website, and the program was used for promoter prediction. The online version of the tool and executables for download can be found at the site: http://nucleix.mbu.iisc.ernet.in/prompredict/prompredict.html.

TSSFinder is the newest of the three tools. It is not strictly speaking a promoter prediction program, but rather TSS prediction tool employing machine learning techniques. It applies linear chain conditional random fields to model the sequence structure of a proximal and core promoter region (2000 nucleotides) and based on that model it localizes TSS signal closest to the start codon of an annotated gene. It offers a pretrained model for human genome, as well as for other organisms.[46] As input files, it needs a FASTA file with the sequence of interest, as well as a BED file with a list of the genes of interest. It produces a new BED file with a list of predicted TSS positions and their genes, and another BED file with a list of TATA boxes localized within the analyzed promoters. It is available as an online tool or to be downloaded at the site: http://sucest-fun.org/wsapp/tssfinder/. Its precompiled package for Linux was downloaded and used for TSS prediction.

4.8 Studying the presence of histone modifications associated with transcriptional activity within estimated promoters

First, a list of histone marks associated with transcriptional activity was compiled from a literature survey. Next, the genomic loci of such modifications were obtained from Gene Transcription Regulation Database (GTRD) as bigBed tracks with MACS2 peaks for hg38. BigBed files were converted to BED files using USCS program *bigBedToBed* and were subsequently compared to the estimated ranged of promoters using *bedtools intersect -wo*. Appendix A presents the ChIP-seq experiments from GTRD database used, and the functional association of the histone marks selected for the analysis.

4.9 Transcription factor AR binding sites

The loci of binding sites were obtained from JASPAR2020 database [108], and the latest dataset was chosen. The presence of the motifs in hg38 genome was studied with *TFBSTools* R package version 1.32.0. The locations of the motifs were subsequently compared to the locations of promoters of each transcript group using *GenomicRanges* package's function *findOverlaps*.

5. RESULTS

5.1 Co-occurrence of transcripts and ATAC-seq peaks on the genome-wide level

As was mentioned in the Background section chapter 2.7, this study was possible thanks to two previously published studies conducted by members of the author's research group [23], [48]. To investigate whether the novel transcripts could be regulated by chromatin accessibility, RNA-seq and ATAC-seq data were used, and both datasets were obtained from the abovementioned studies.

Transcriptome assembly generated by Ylipää et al. contained transcript counts of 60 662 protein-coding transcripts, 49 517 LNC annotated transcripts, 66 280 newly identified unannotated intergenic transcripts, and 32 704 newly identified unannotated intragenic transcripts. After filtering out the transcripts mapping to mitochondrial genome, the number of protein-coding transcripts decreased to 60 625. As the vast majority of the novel transcripts did not have strand information, every step of analysis including novel transcripts, was performed twice, first assuming the transcripts were on plus strand, and then assuming they were on minus strand.

The relation between chromatin accessibility and expression was first studied on the genome-wide level. To see how many transcripts were located in the proximity of an ATAC-seq peak, the distribution of distances from TSS to the nearest peak was plotted. Figure 10 presents the distributions. The ATAC-seq peaks used in this preliminary analysis were from the set of unified peaks called and quantified by Uusi-Mäkelä in 2020. The graph in the left top corner presents the distribution of all transcripts when the unannotated transcripts with no strand had been assigned minus-sense strand, then next to it is the graph of all transcripts when the unannotated transcripts with no strand had been assigned plus-sense strand. The other graphs represent the distributions of each transcript type separately, with the distinction of strand assumptions for the unannotated transcripts. The distributions have similar shapes in all graphs, although the number of unannotated transcripts with a peak nearby their TSS is not as high as for annotated transcripts. Nevertheless, the density of distances reaches its peak around the TSS and decreases exponentially and symmetrically both upstream and downstream from the TSS in each transcript group.

Figure 11 presents the percentages of transcripts with peaks at different distances from TSS in form of pie charts. The orange parts represent the promoters, as defined in this study. The first significant observation can be made based on these two illustrations – the percentage of unannotated transcripts with an ATAC-seq peak in the promoter was lower than in the annotated transcripts, and lower than when all transcripts were considered together. This means that a lower fraction of promoters of unannotated transcripts was accessible in comparison to the annotated transcripts. Thus, already at this point we could hypothesize that a much lower number of unannotated transcripts would be expressed and would display other epigenomic marks of transcriptional activity than in the groups of protein-coding transcripts and LNC annotated transcripts. The pie charts also confirm the symmetrical shapes of the distribution density plots, as the percentages of peaks at the same distances downstream and upstream from TSSs are similar. In addition, the strand assumption does not change the distribution.

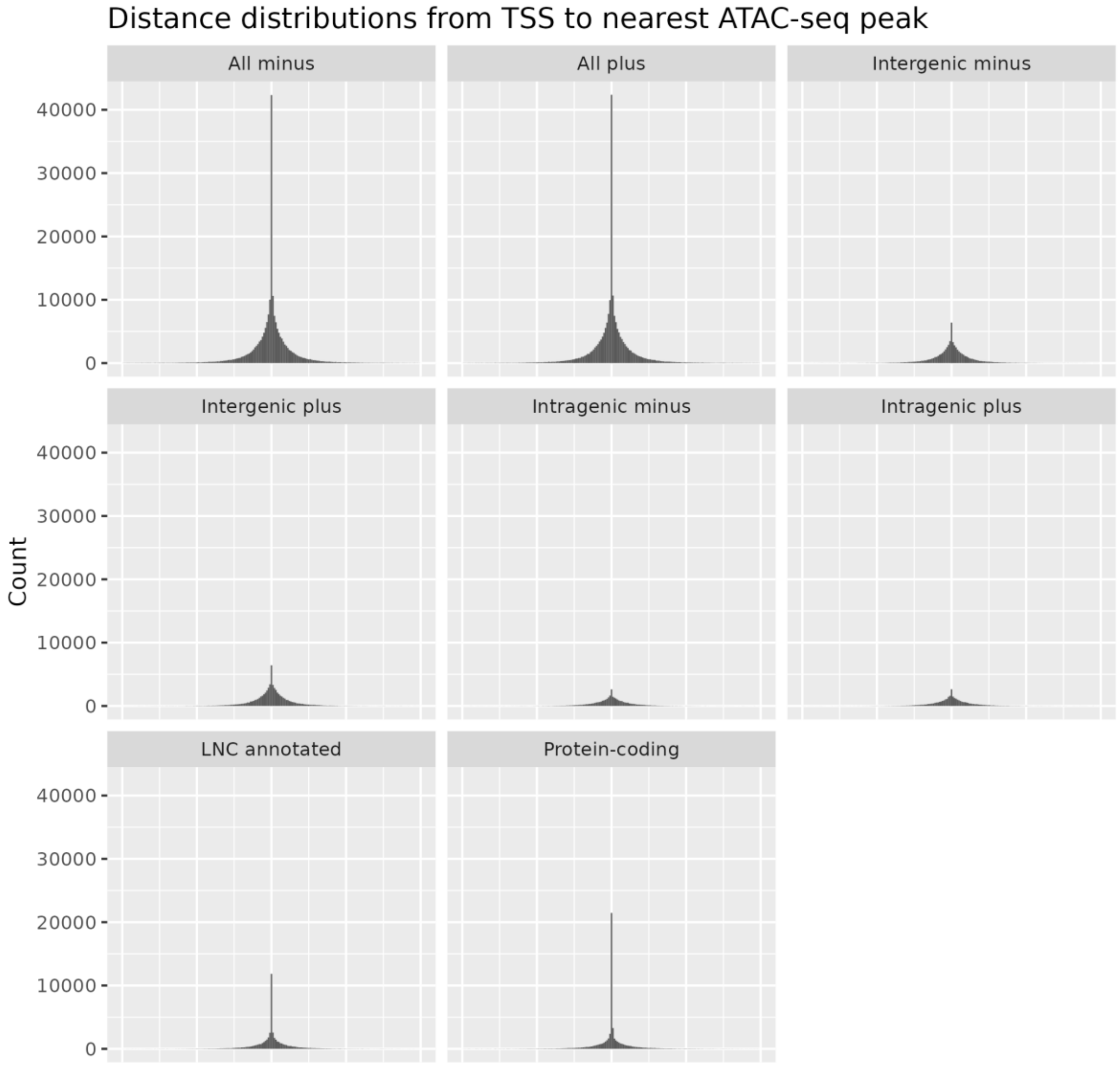


Figure 10. *Distribution of the distances from TSS to the nearest ATAC-seq peak*

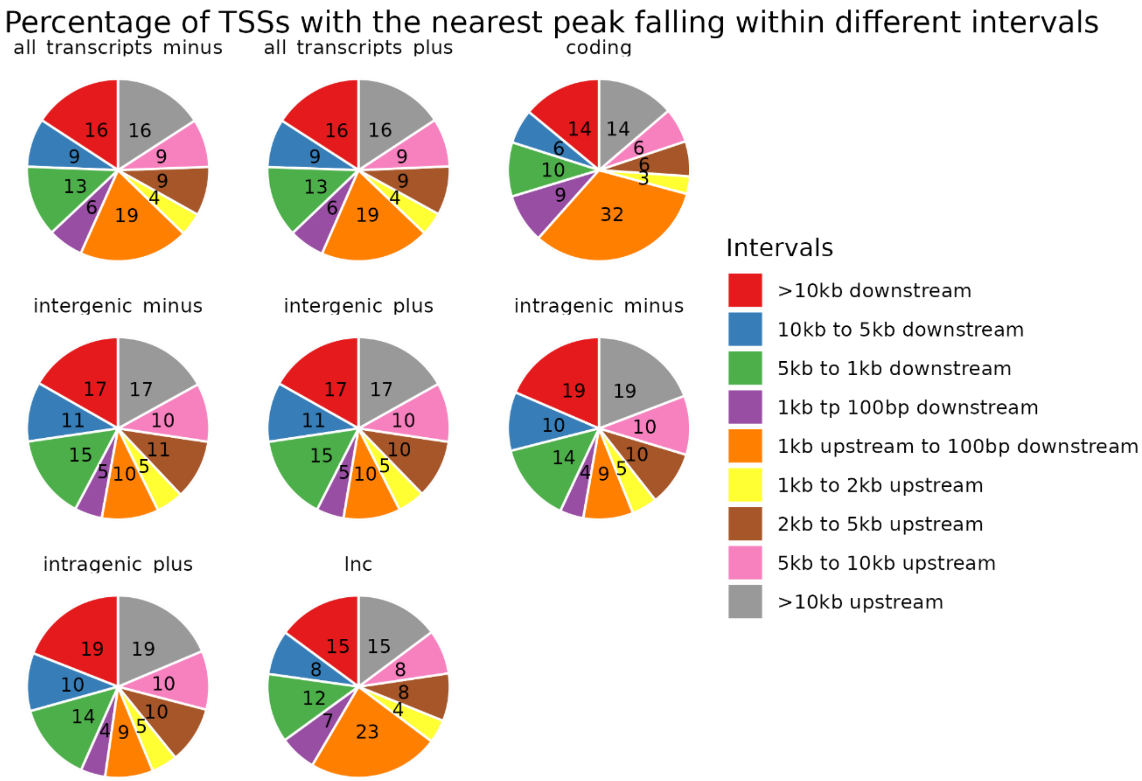


Figure 11. *The percentages of transcripts with the nearest ATAC-seq peak within different distances from TSS. The percentages are rounded to the nearest integer.*

5.2 The relationship between chromatin accessibility and expression in individual samples

5.2.1 Sample-specific ATAC-seq peaks

In order to investigate sample-specific occurrence of chromatin accessibility and expression, ATAC-seq peaks needed to be quantified in individual samples. From this point onward, the sample-specific ATAC-seq peaks were used in each step of the analysis. Figure 12 shows the number of obtained ATAC-seq peaks within each studied sample. As can be seen, the number varied a lot between individual samples, revealing samples with both plenty of accessible regions and scarcity of them within the same sample groups. Therefore, the numbers of peaks in each group were more closely investigated.

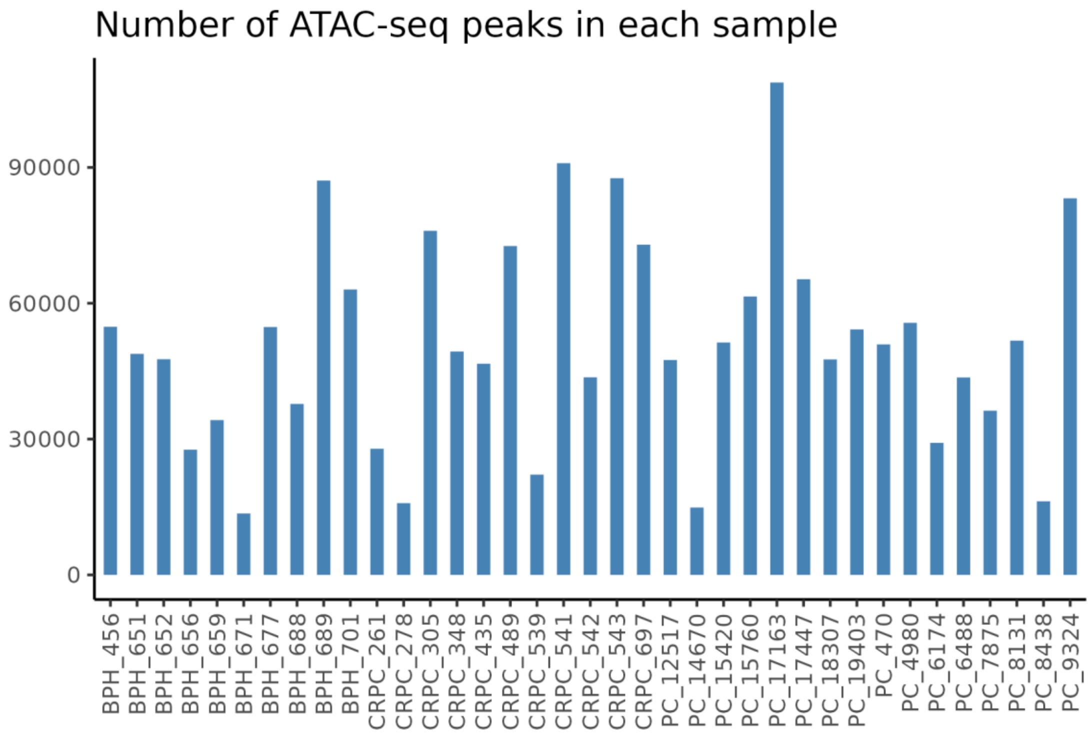


Figure 12. *The number of ATAC-seq peaks in individual samples.*

The similarity of the means between the sample groups was tested statistically. Shapiro normality test was performed for each sample group peak number distribution with the threshold set to 0.05. According to the test results, normality could be assumed for all three sample groups (p-values BPH: 0.9473; PC: 0.2253; CRPC: 0.3904). Based on the assumption of normality, one-way ANOVA test was used. The null hypothesis was that there was no significant difference between the numbers of peaks of the three sample groups, and the alternative hypothesis was that the mean of at least one group was different. Chosen threshold was 0.05. Obtained p-value was 0.733, much higher than the threshold, and thus based on this result there was no reason to suspect significant differences between the numbers of ATAC-seq peaks in the sample groups.

The distribution of the numbers of the peaks within each sample group was inspected also visually. Figure 13 presents graphical representations of the distribution in BPH, CRPC, and PC. It was observed that CRPC group displayed the highest level of variability, and PC group the lowest, even though this group contained the most significant outlier, PC 17163, with the highest number of peaks not only among PC samples, but in the entire cohort. The highest and the lowest number of peaks in BPH group were both lower than the respective values in the other two sample groups. Based on this plot, some initial conclusions were made. First, chromatin accessibility is a highly heterogenous feature among CRPC samples, and possibly contributes to or results from high genetic heterogeneity of advanced PC. While chromatin accessibility seems to be most homogenous in PC samples, it is BPH in which the extreme values are the lowest, which could mean higher level of regulation of chromatin accessibility and thus better control over proper gene expression.

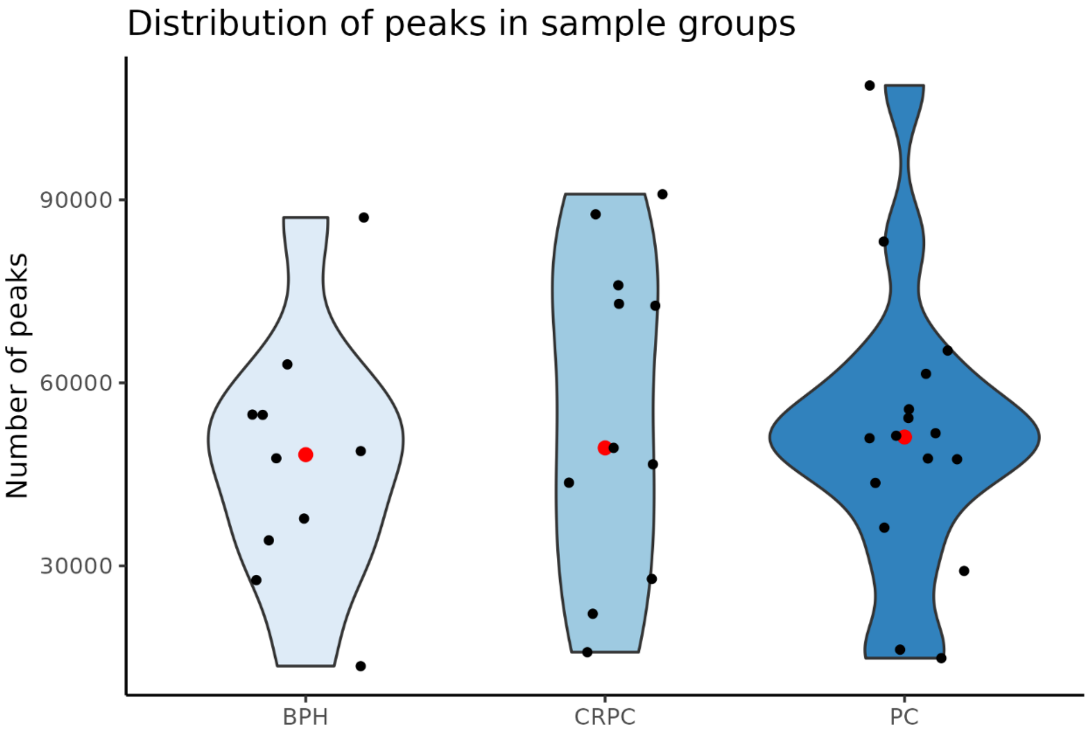


Figure 13. *Violin plots representing the distributions of the number of peaks in each sample group. The blue shapes represent the density of the peak numbers in each sample group. The red dots represent the median numbers of peaks, and the black dots represent numbers of peaks in each individual sample.*

5.2.2 Co-occurrence of ATAC-seq peaks and expression

To study whether chromatin accessibility and expression occurred together, the sample-specific peaks had to be associated with individual transcripts. Thus, the genomic locations of sample-specific ATAC-seq peaks were intersected with the estimated promoter regions. The transcripts whose promoter overlapped a peak, were marked as “ON” genes and the transcripts with no peak overlapping their promoter were marked as “OFF” genes. Those transcripts, whose promoters overlapped peaks, which fell within the blacklisted regions producing mitochondrial artifacts, were removed from the group of

ON genes. This meant removal of 361 protein-coding genes, 156 LNC annotated transcripts, 17 intergenic transcripts with plus-strand assumption and 37 with minus-strand assumption, and 9 intragenic transcripts with plus-strand assumption and 10 with minus-strand assumption.

The highest percentage of ON transcripts was found in protein-coding transcript set, and it was between 21 and 22%. The percentage of LNC annotated transcripts with a peak in their promoter was approximately 13% in all three conditions. Finally, the mean percentage of ON transcripts in unannotated transcripts oscillated between 2 and 3%, thus, it was much lower than in other transcript categories. These percentages were lower than the percentages from Figure 11. There, the unified set of peaks was used to study the distances to the nearest peak in every sample. The reason for this is the fact, that in reality less peaks were found in many of the samples, so less peaks could fall within promoters. The mean percentages of "ON" and "OFF" genes in each transcript group in each condition when sample-specific peaks were used, can be found in Appendix B.

As was explained in the Background section, chromatin accessibility allows protein-DNA interactions, and thus also more efficient transcription. Therefore, it was investigated whether on a general level the mean abundance of ON genes was higher than the mean abundance of OFF genes. A series of Wilcoxon rank sum tests was performed to test the hypothesis in each transcript type for each condition separately. In every case, testing confirmed that ON transcripts were significantly more expressed than OFF transcripts with significance level 0.01. In addition, the mean expression levels in unannotated transcripts did not differ between the two strand assumptions both for ON genes and for OFF genes.

Also, it was tested whether the number of peaks in gene promoters and the number of expressed genes (TPM > 0) in each sample were independent from one another. Chi-squared independence test showed with confidence level 0.01 that these two variables were not independent in any of the samples, but only for protein-coding transcripts and LNC annotated transcripts. For intergenic transcripts, these two variables were independent (p-value > 0.01) in BPH 656 and BPH 671 with plus-sense, and in BPH 656 with minus-sense assumption. The expression of intragenic transcripts turned out to be independent from the peaks in many more samples: with plus-sense assumption in samples BPH 656, BPH 671, BPH 677, BPH 689, CRPC 261, CRPC 539, PC 6174, and PC 9324; and with minus-sense assumption in the same BPH and CRPC samples, but in addition also in BPH 651, BPH 688, and CRPC 697, and in no PC sample.

Finally, the correlation between the total number of peaks and the total number of expressed transcripts in each sample group was tested with significance level of 0.01, however, the results of the tests were not significant.

These initial steps showed that chromatin was accessible in the vicinity of TSSs of a subset of unannotated transcripts, and that the expression of those transcripts was generally higher than of those transcripts, which were further away from the nearest peak. However, that subset contained much lower number of transcripts then the corresponding subset of protein-coding, or even LNC annotated transcripts. Moreover, in contrast to annotated transcripts, the number of expressed unannotated transcripts was not always dependent on the number of peaks in promoters.

5.2.3 Correlation between accessibility and expression

To study whether higher chromatin accessibility resulted in higher expression, the correlation between these two features was investigated. Spearman's correlation was primarily taken into consideration due to the fact that it is more robust against significant outliers, present in the dataset. First, the correlation between the ATAC-seq peak intensity and expression was tested for all ON transcripts. Table 1 presents the mean correlation estimates for each transcript group in each condition, and Figure 14 presents the distribution of the p-values.

Table 1. *Mean Spearman's and Pearson's correlation estimates for each transcript group in each condition.*

Transcript group	BPH	CRPC	PC
Spearman correlations			
Intergenic (+)	0.09	0.12	0.11
Intergenic (-)	0.09	0.11	0.12
Intragenic (+)	0.12	0.12	0.15
Intragenic (-)	0.09	0.11	0.13
Coding	0.18	0.21	0.19
LNC annotated	0.11	0.14	0.11
Pearson correlations			
Intergenic (+)	0.01	0.04	0.02
Intergenic (-)	0.02	0.04	0.03
Intragenic (+)	0.02	0.03	0.05
Intragenic (-)	0.01	0.01	0.06
Coding	0.00	0.01	0.00
LNC annotated	0.00	0.02	0.01

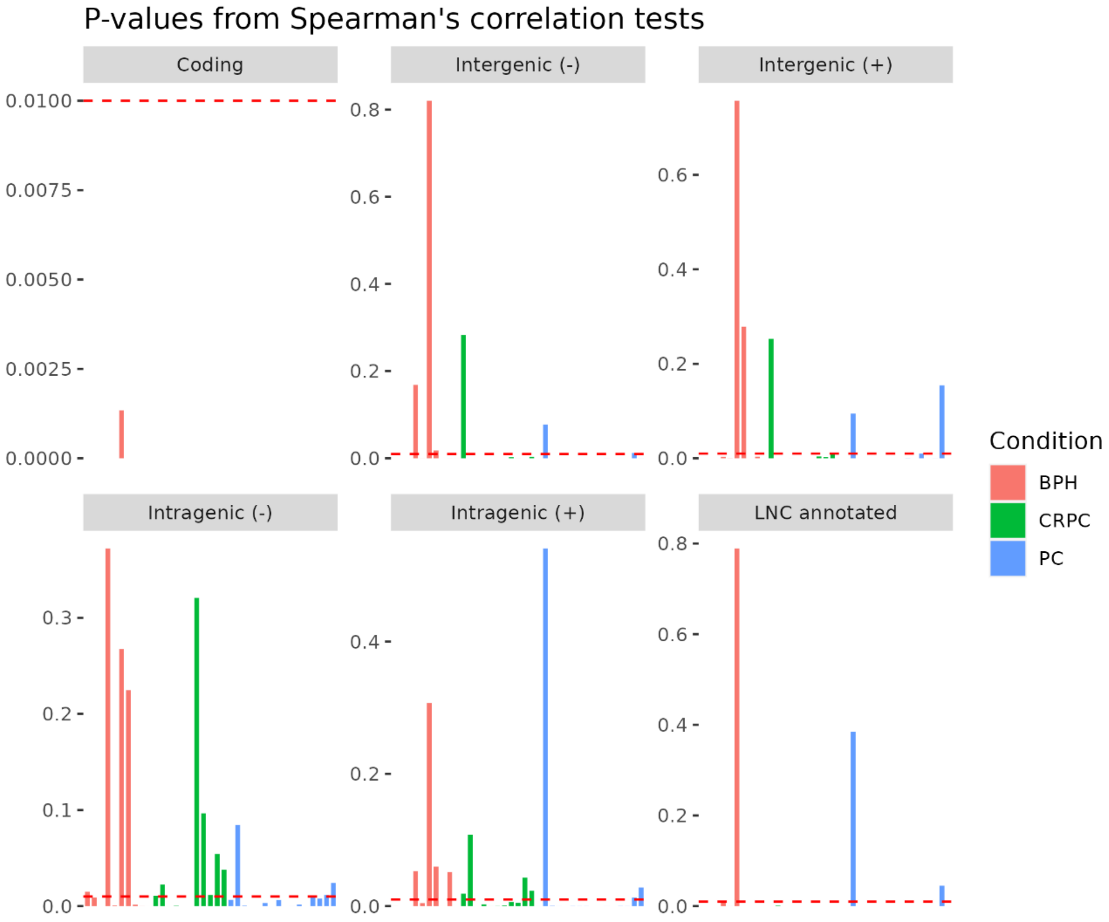


Figure 14. *P-values from Spearman's correlation tests between ATAC-seq peak intensity and expression for all ON transcripts. The dashed red line marks the significance level 0.01.*

Pearson's estimates were really low and rarely significant; thus, the further discussion focuses on the Spearman's correlations. All Spearman's correlation estimates computed for protein-coding transcripts were significant. They were mostly moderate, but always positive. In general, expression and chromatin accessibility positively correlated in protein-coding transcripts. Most of the results in the set of LNC annotated transcripts were significant, but there were exceptions: BPH 671, PC 14670, and PC 8438. Still, all estimates were positive, although rather weaker than in coding transcripts.

In all three sample groups most of the Spearman's correlations for intergenic transcripts were significant and were weak or moderate. With plus-sense assumption, there was only one negative correlation, which, however, was negligible (BPH 671: -0.0147, p-value 0.7568). With minus-sense assumption there were two negative correlations, both very weak, and one for the same sample as with the other assumption (BPH 671: -0.0107, p-value 0.8202, and CRPC 261: -0.0344, p-value 0.2831). On average, correlations were minimally higher with minus-sense assumption in PC samples, whereas in

CRPC samples correlations were minimally higher with plus-sense assumption, however, the differences were rather modest. The sense assumption did not change the mean correlation estimate in BPH samples. Expression in cancerous samples seemed to better correlate which chromatin accessibility than in BPH samples.

The patterns of correlations in intragenic transcripts differed slightly from those in intergenic transcripts, and the sense assumption altered the results more evidently. Overall, the correlations were weak to moderate. There was only one negative estimate with both sense assumptions, and it was for BPH 671, in which correlations were also negative for intergenic transcripts. The percentage of significant correlations with minus-sense assumption was higher in cancerous samples than in healthy samples. Also, in total, more p-values were insignificant, especially with minus-sense assumption, in comparison with intergenic transcripts. For all conditions, the estimates were higher in most of the samples with plus-sense assumption, and those differences were sometimes very big (e.g., for BPH 651, the difference between plus-sense correlation and minus-sense correlation was 0.1209, higher than many of the sample correlation estimates). Interestingly, in BPH, most of the minus-sense correlations were weaker than minus-sense correlations for intergenic transcripts, however, the opposite was observed for plus-sense correlations. Correlation estimates in PC and CRPC were not affected as much by the strand assumption. The average CRPC correlation for intragenic transcripts was similar to correlation for intergenic transcripts, while the average PC correlation was higher than the one for intergenic transcripts.

Indisputably, protein-coding transcripts displayed the strongest correlations. There were clear differences between coding and LNC annotated transcripts, but the estimates in intergenic and intragenic transcripts did not differ strongly form each other. Also, the sense assumption did not influence the correlations dramatically. The average correlations in unannotated transcripts were quite similar to the average correlations in LNC annotated transcripts and were even stronger in PC samples. There were observable inter-state differences. The expression of annotated transcripts in CRPC correlated with chromatin accessibility more strongly than in other two states. That was not the case for unannotated transcripts. The strongest correlations of intergenic and intragenic transcripts were mostly in PC samples. For all transcript groups, BPH displayed the weakest correlations.

In addition to the correlation analysis, the number of expressed transcripts was plotted against the reported Spearman's correlation estimates to see whether a higher number of expressed transcripts meant a stronger correlation between the accessibility and expression level. The discussed plots can be found in Appendix C. In annotated transcripts, there was almost no relationship between the investigated features. However, the situation was more interesting in unannotated transcripts. In BPH samples with plus- and minus-sense assumptions the slope of the trendline was steep and negative for both intergenic and intragenic transcripts, which meant that more expressed transcripts in a sample meant weaker correlation between the expression level and chromatin accessibility. In both cancer states with both sense assumptions, intergenic transcripts displayed weak negative dependency of the features. While for intragenic transcripts with plus-sense assumption the trend in all cancer samples was weakly positive, with minus-sense assumption it reversed to strongly negative, especially in CRPC.

Furthermore, the number of peaks in promoters was plotted against the Spearman’s correlation estimates to investigate whether more peaks near TSS meant stronger correlation between the accessibility and the expression level. The plots can be seen in Appendix D. In plots for annotated transcripts the trendlines had a steep positive slope in all states. The plots for unannotated transcripts looked differently. In intergenic transcripts in BPH and PC the correlation seemed to be stronger in samples with more peaks in promoters, however, the trend was exactly reversed in CRPC samples, and it was true for both sense assumptions. In intragenic transcripts in BPH samples the relationship was similar to intergenic transcripts in BPH samples, but trendline with a steep negative slope was observed not only in CRPC, but also in PC. Thus, in annotated transcript sets the higher number of peaks in promoters did mean a higher correlation between accessibility and expression, but in unannotated transcript sets such trend was true only in BPH, and in PC for intergenic transcripts.

5.2.4 Thresholding and filtering

Previous testing showed that on the general level chromatin accessibility in the promoter regions and expression were somewhat correlated. In attempt to better understand the dependency between the accessibility and expression, the correlation was also investigated for different subsets of each transcript type obtained by applying data-derived thresholds to the expression level and accessibility. The main goal was to find out if the correlation would improve with increasing expression levels and accessibility. Such trend was not observed in any of the transcript types. Instead, the correlations changed from weak to strong and jumped between negative and positive values without any clear pattern, and most often their p-values were insignificant. Perhaps, the dependency between the accessibility and expression cannot be simply explained through either linear or monotonic relationship.

As an example, Table 2 presents the mean correlations for the subset of transcripts whose TPM was higher than the sample-specific upper quartile in a given transcript group. This criterion was chosen in attempt to restrict the analysis only to those transcripts, for which chromatin accessibility co-occurred with observable expression.

Table 2. *Mean Spearman's and Pearson's correlation estimates for transcripts with TPM higher than sample-specific upper quartiles.*

Transcript group	BPH	CRPC	PC
Spearman's correlations			
Intergenic (+)	-0.04	0.03	0.01
Intergenic (-)	-0.06	0.01	0.02
Intragenic (+)	0.01	0.05	0.03
Intragenic (-)	0.04	-0.09	0.05
Coding	-0.01	0.02	0.02
LNC annotated	-0.06	-0.04	-0.04
Pearson's correlations			
Intergenic (+)	-0.05	0.00	-0.02
Intergenic (-)	-0.03	-0.01	-0.01
Intragenic (+)	0.01	0.00	0.02
Intragenic (-)	0.00	-0.07	0.08
Coding	-0.02	-0.01	-0.01
LNC annotated	-0.01	0.03	0.00

Figure 15 shows the p-values for Spearman's correlations from the same tests. As can be seen, some of the p-values in the protein-coding transcript set were significant, with the highest number in PC samples. There were less significant values in the set of LNC annotated transcripts. Nearly none of the estimates was significant for unannotated transcripts. Since the p-values are not significant, the results are not conclusive.

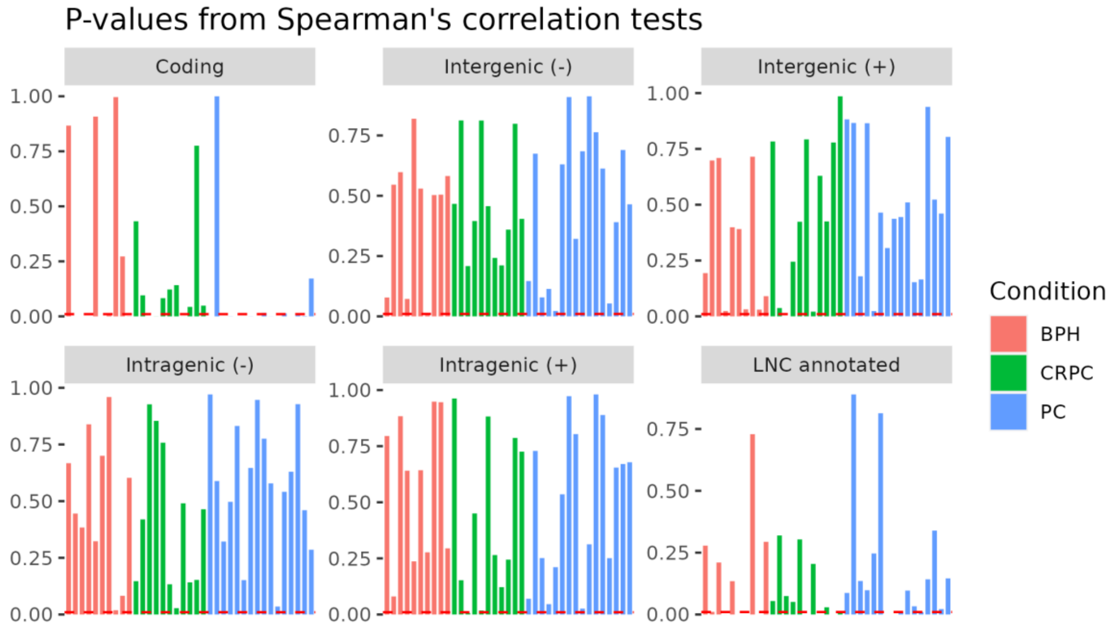


Figure 15. P-values from Spearman's correlation tests for ON transcripts with TPM higher than sample-specific upper quartiles. The dashed red line marks the significance level 0.01.

5.3 Integration of DNA methylation data

The next analysis was done to explore the effect of DNA methylation within promoters on the expression level of the transcripts. First, it was studied whether estimated promoters were methylated. The promoter ranges of transcripts were overlapped with methylation data, and a methylation level was computed for each of them as a weighted average of methylation levels of overlapping intervals. Since MeDIP-seq data was quantified and normalized using QSEA tool, methylation levels were values between 0 and 1. Two of the samples studied in earlier steps, CRPC 305 and CRPC 539, did not have MeDIP-seq data, so they had to be excluded from downstream analysis.

Some promoters displayed no level of methylation. PC seemed to have the highest proportion of transcripts with some level of methylation within promoters of each transcript group. The highest fraction of methylated promoters was identified within protein-coding transcripts set, and the lowest in intragenic transcript set. The fraction of methylated promoters in intergenic transcripts and LNC annotated transcripts was similar. Figure 16 presents the percentages for each transcript group in each condition.

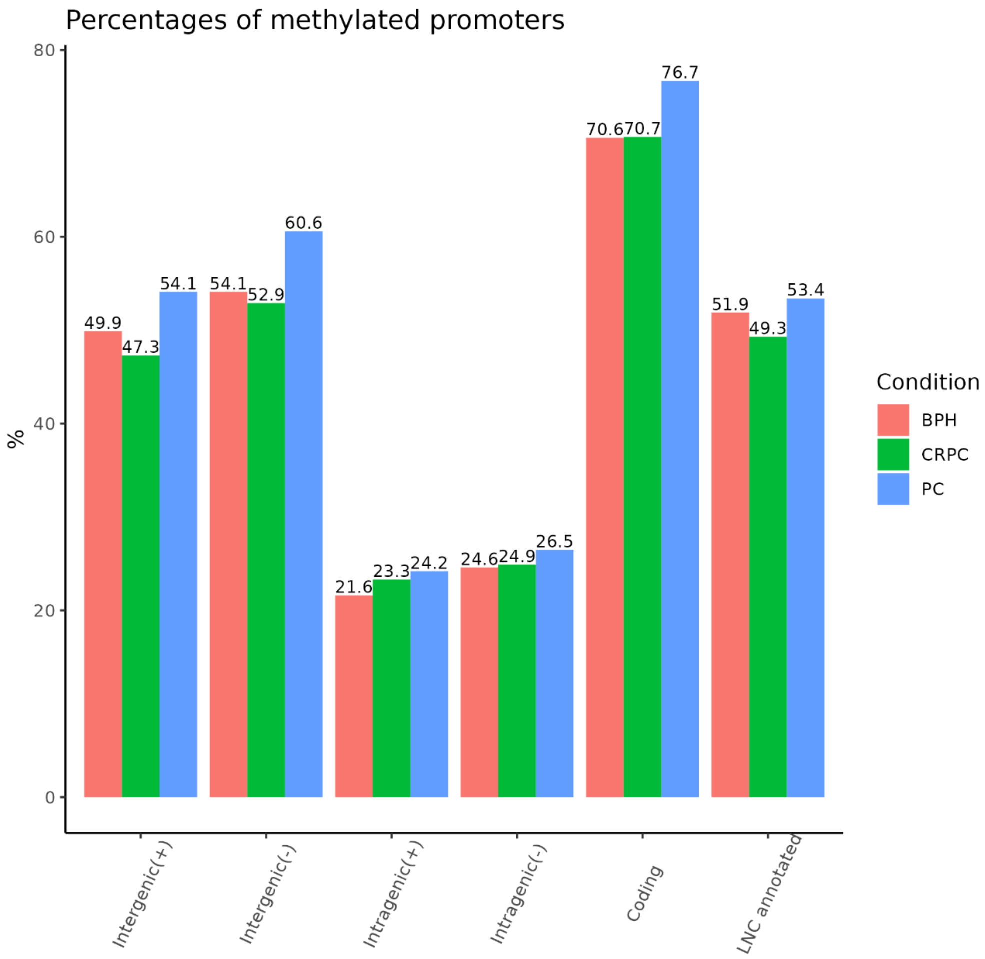


Figure 16. *A bar plot presenting the percentages of transcripts whose promoter region was methylated in some degree with distinction of the conditions. The values are rounded to one decimal place.*

Mean methylation levels were computed in each sample for transcript group for those transcripts, whose promoters were methylated. Then, mean of means in each condition was computed to compare inter-state methylation levels. The least methylated promoters were the promoters of protein coding transcripts. The methylation levels of promoters of LNC transcripts were slightly higher than those of coding transcripts. The most methylated promoters were found in intergenic and intragenic transcripts, and the methylation levels in those two groups were very similar. The differences between states in all transcript groups were negligible. Table 3 presents the methylation levels for each transcript group in each condition.

Table 3. *The mean values of sample mean methylation levels in each transcript group in each condition, rounded to four decimal places.*

State	Intergenic (+/-)	Intragenic (+/-)	Coding	LNC annotated
BPH	0.6338/ 0.6344	0.6319/ 0.6318	0.4342	0.4914
PC	0.6292/ 0.6304	0.6257/ 0.6255	0.4436	0.5009
CRPC	0.6220/ 0.6228	0.6174/ 0.6179	0.4389	0.4946

Next step was investigating whether the presence of an ATAC-seq peak in the promoter meant low or no methylation. In order to do that, the proportions of promoters with a peak and low or no methylation (below or equal to 0.25 quantile of methylation levels in a transcript group in a sample) and those with a peak and high methylation level (equal to or above 0.75 quantile of methylation levels in a transcript group in a sample) were compared in each transcript type. The average percentages in each condition can be seen in Figure 17.

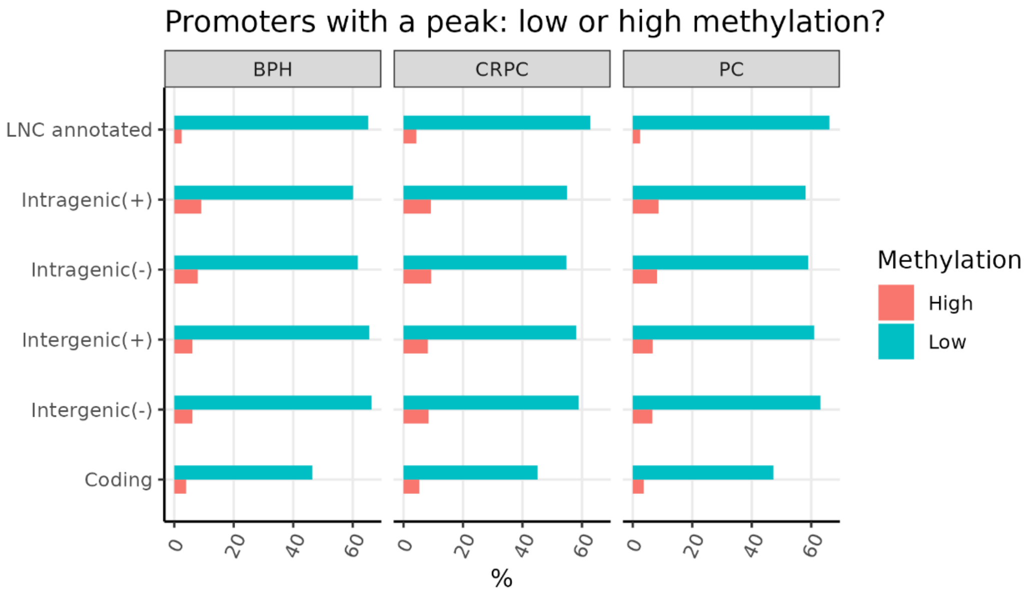


Figure 17. *The percentages of transcripts with a peak in their promoter and high methylation level or low methylation level.*

In every transcript group, there were many more transcripts with an ATAC-seq peak in their promoters with low methylation level than transcripts with ATAC-seq peak and high methylation level. In annotated transcripts, this relationship was the strongest in PC samples, while in unannotated transcripts in BPH samples. While in other transcript groups the proportion of transcripts with accessible promoter and low methylation level constituted the majority of those transcripts, the corresponding percentage of protein-coding transcripts was nearly the same as those which were moderately methylated. Thus, in this regard, unannotated transcripts were more similar to LNC annotated transcripts. Nevertheless, the percentages of unannotated transcripts with accessible promoter and high methylation level were much greater than the corresponding percentages of annotated transcripts. This proportion was always the largest in CRPC, although the difference between CRPC and two other conditions in intragenic transcripts was not as prominent as in other transcript types.

Also, the relation between the presence of an ATAC-seq peak, expression level, and methylation level was investigated. First, the methylated promoters were filtered into highly methylated ones and into those with low methylation level (the same definitions were used as described above. Then, it was studied what are the fractions of those subsets with and without a peak, and then also the fractions with high expression (equal or above 0.9 quantile of a transcript group in a sample) and with low expression (equal or lower than 0.75 quantile of a transcript group in a sample). The results for transcripts with highly methylated promoters can be seen in Figures 18 and 19, whereas the results for transcripts with low methylation level within their promoters in Figures 20 and 21.

In each transcript type, the vast majority of transcripts with high methylation did not have a peak in their promoter. Interestingly, this percentage was slightly higher in unannotated transcripts than in annotated transcripts, and the lowest fractions across all conditions were observed for the protein-coding transcripts. CRPC always had the highest fraction of highly methylated promoters with a peak, although the difference between CRPC and other conditions in unannotated transcripts was very small. The majority of highly methylated transcripts was characterized also by low expression level, especially in PC. The part of such transcripts was the smallest in LNC annotated transcripts. CRPC was the condition in which the fraction of highly methylated transcripts with low expression was the lowest, whereas the fraction of highly methylated transcripts with high expression was the largest, especially in unannotated transcripts. In PC and BPH, the percentage of highly methylated transcripts with high expression was the largest in LNC annotated transcripts. Such transcripts were the least abundant among the protein-coding transcripts.

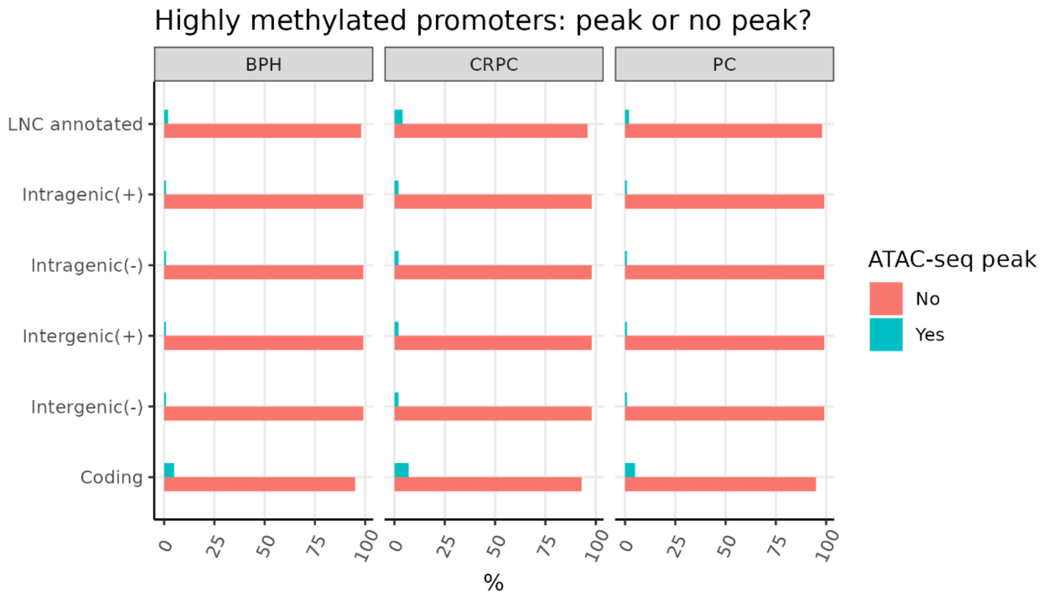


Figure 18. *The percentages of highly methylated transcript promoters with and without an ATAC-seq peak.*

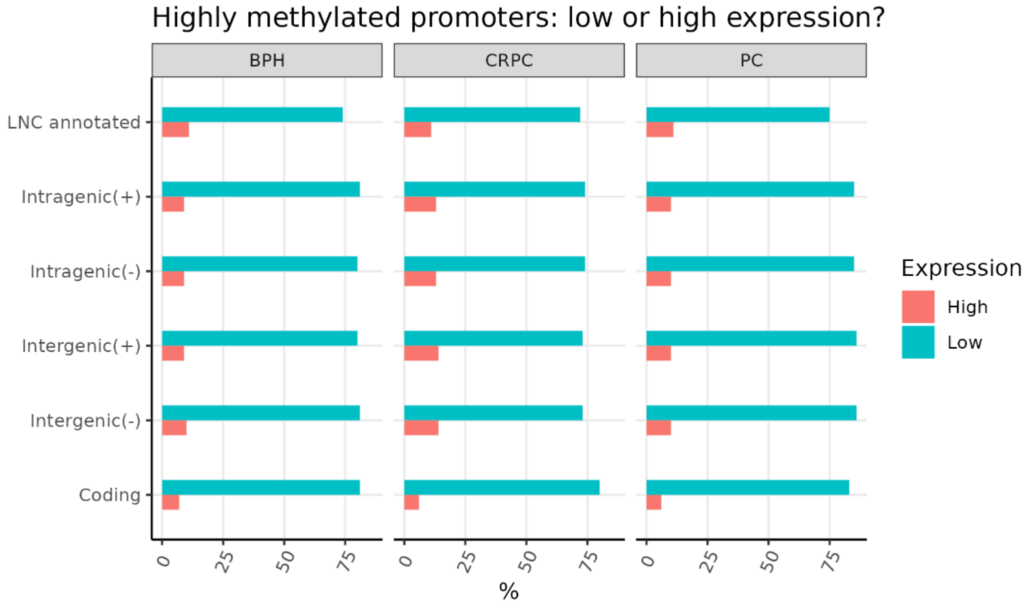


Figure 19. *The percentages of highly methylated transcripts low and high expression level.*

Surprisingly, low methylation level within a promoter did not necessarily mean that an ATAC-seq peak was present. Furthermore, the relationship between these two features was different in unannotated transcript sets than in annotated transcript sets. Approximately 55% of annotated transcripts with low methylation also had a peak, but the fraction of transcripts without a peak was lower by only a bit more than a dozen percentage points. In contrast, most of the unannotated transcripts with low methylation did not have a peak. Similarly, low methylation did not mean high expression in any of the transcript groups - most of the transcripts with low methylation were not highly expressed.

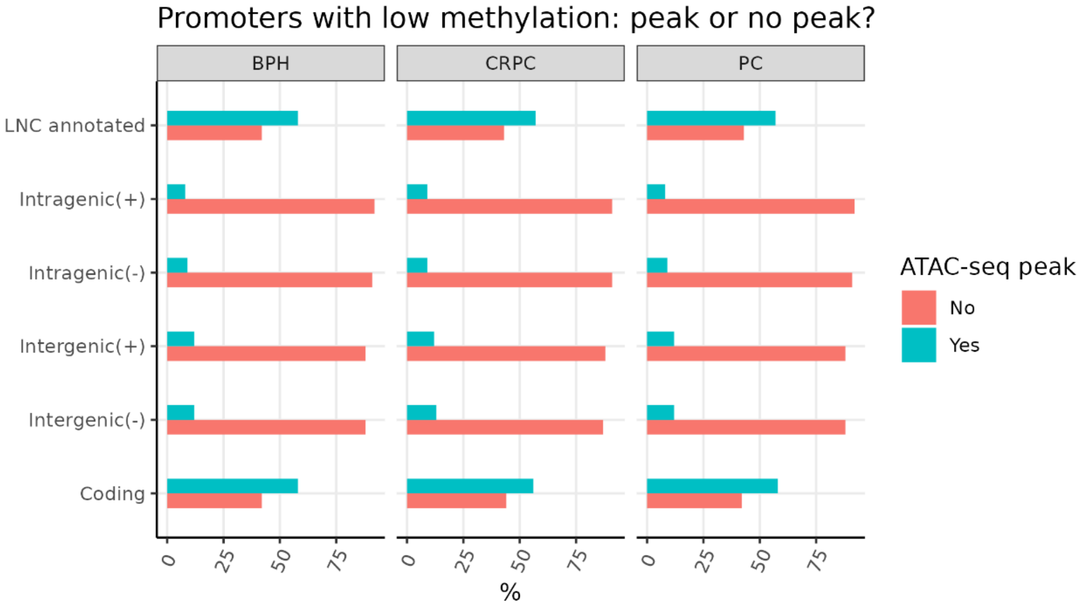


Figure 20. *Percentages of promoters with low methylation level and an ATAC-seq peak or no ATAC-seq peak.*

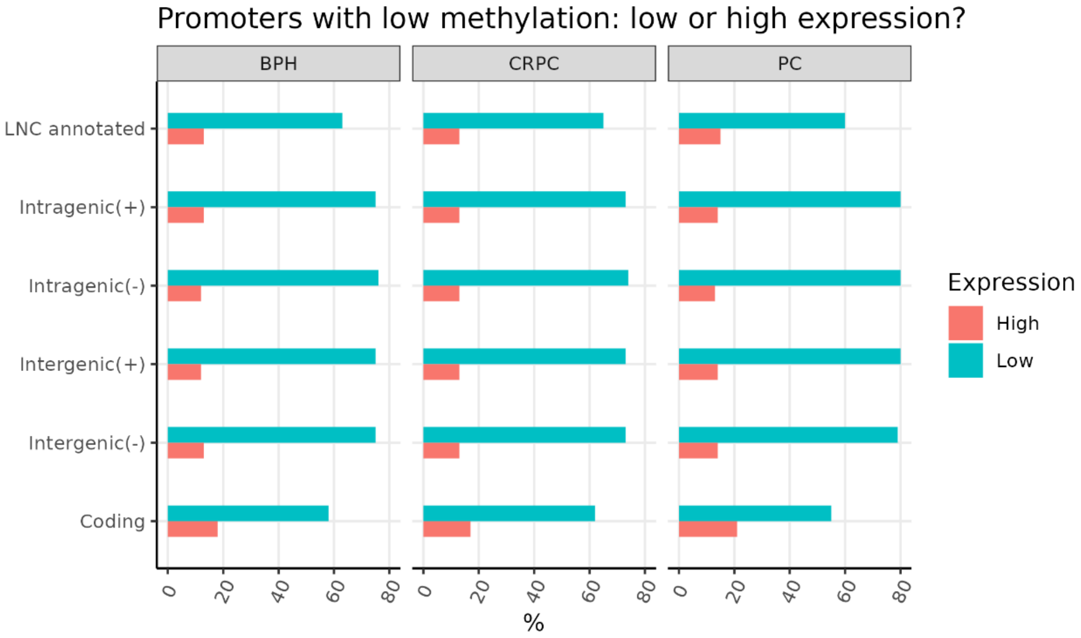


Figure 21. *Percentages of promoters with low methylation level and high expression or low expression.*

5.3.1 Correlation between methylation level and expression

To investigate whether higher methylation level of promoters meant lower expression, the correlation between these two features was computed. For the transcripts with methylated promoters, the correlation between methylation and expression was tested with significance level 0.01. The mean correlation estimates for each transcript set in each condition can be seen in Table 4, and Figure 22 presents the p-values from the Spearman's correlation tests.

Table 4. *Mean Spearman's and Pearson's correlations between methylation and expression.*

Transcript group	BPH	CRPC	PC
Spearman's correlation			
Intergenic (+)	-0.06	-0.01	-0.07
Intergenic (-)	-0.06	-0.02	-0.07
Intragenic (+)	-0.06	-0.01	-0.7
Intragenic (-)	-0.05	-0.02	-0.06
Coding	-0.24	-0.21	-0.27
LNC annotated	-0.16	-0.11	-0.18
Pearson's correlation			
Intergenic (+)	0.00	0.00	-0.01
Intergenic (-)	-0.01	0.00	-0.01
Intragenic (+)	-0.04	-0.03	-0.03
Intragenic (-)	-0.01	-0.01	-0.01
Coding	-0.03	-0.03	-0.03
LNC annotated	-0.01	-0.01	-0.01

Methylation level and expression level were quite strongly anti-correlated in protein-coding genes. All estimates were significant and much higher than in unannotated transcripts. Still, on average, PC samples had the strongest correlations and CRPC the weakest. Methylation and expression anti-correlated also in LNC transcripts. Once again, the strongest correlations were in PC samples and the weakest in CRPC samples. The average condition-specific estimates were lower than those of protein-coding transcripts by approximately ten percentage points, and higher than those of unannotated transcripts by nearly the same number of percentage points.

All Spearman's estimates for intergenic transcripts in BPH and PC samples were negative, and nearly all were significant, with both sense assumptions. Three estimates in CRPC samples were positive and insignificant, also with both sense assumptions. All individual correlations were rather weak, and on average the weakest ones were in

CRPC, and the strongest in PC. Minus-sense assumption produced marginally stronger correlations.

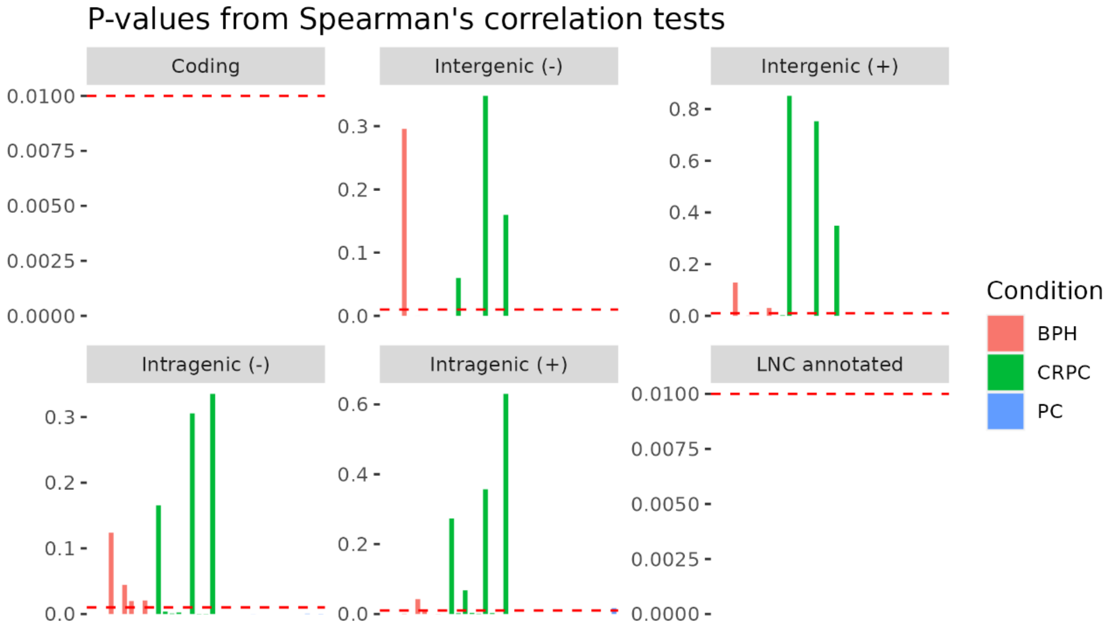


Figure 22. *P-values from Spearman’s correlation tests between methylation and expression. The red dashed line marks the significance level.*

The results for intragenic transcripts revealed similar pattern as in intergenic transcripts in almost all regards. Again, all BPH and PC samples produced negative Spearman’s estimates. In some BPH samples the correlation was insignificant, especially with plus strand assumption. As in the former transcript set, CRPC produced the weakest correlations on average, and in some samples methylation and expression were positively correlated, while in several other samples the correlation was insignificant. The studied features were the most strongly correlated in PC samples. The average condition-specific correlations were slightly lower for intragenic transcripts than for intergenic transcripts. Nevertheless, in both unannotated transcript groups the overall correlation between methylation and expression was weak and negative.

Based on these observations, it can be concluded that on the general level, the expression anti-correlated with methylation. But this feature seemed to be most characteristic for protein-coding genes and is not well pronounced in unannotated transcript groups or even in annotated non-coding transcripts. Moreover, in some samples, methylation and expression positively correlated. Since high methylation of inter- and intragenic CpG islands is involved in regulation of non-coding transcripts, the studied sets of transcripts might include subsets regulated in this way, which decrease the overall anti-correlation between methylation and expression.

5.3.2 Correlation in ON and OFF transcripts

In attempt to find subsets of transcripts positively and negatively regulated by methylation, data-derived thresholds were applied to the methylation levels and expression levels. First, the correlation between methylation and expression was studied in transcripts with ATAC-seq peaks in their promoters (“ON”) and those without peaks (“OFF”). The mean correlation estimates for each transcript group in each condition are presented in Table 5, and the p-values from Spearman’s correlation tests are presented in Figure 23 for ON transcripts, and in Figure 24 for OFF transcripts.

Table 5. *Mean Spearman’s and Pearson’s correlations between methylation and expression for ON and OFF transcripts.*

Transcript group	BPH	CRPC	PC
Spearman’s correlation – ON transcripts			
Intergenic (+)	-0.18	-0.08	-0.14
Intergenic (-)	-0.13	-0.07	-0.17
Intragenic (+)	-0.16	-0.13	-0.17
Intragenic (-)	-0.11	-0.09	-0.13
Coding	0.04	-0.03	0.03
LNC annotated	-0.03	-0.06	-0.04
Pearson’s correlation – ON transcripts			
Intergenic (+)	-0.07	-0.01	-0.03
Intergenic (-)	-0.05	-0.02	-0.04
Intragenic (+)	-0.07	-0.03	-0.08
Intragenic (-)	-0.02	-0.03	-0.04
Coding	0.01	0.00	0.01
LNC annotated	0.01	0.00	0.01
Spearman’s correlation – OFF transcripts			
Intergenic (+)	-0.04	0.00	-0.05
Intergenic (-)	-0.05	0.00	-0.05
Intragenic (+)	-0.05	0.00	-0.05
Intragenic (-)	-0.04	0.00	-0.05
Coding	-0.10	-0.04	-0.11
LNC annotated	-0.07	-0.01	-0.07
Pearson’s correlation – OFF transcripts			
Intergenic (+)	0.00	0.00	0.00
Intergenic (-)	-0.01	0.00	-0.01
Intragenic (+)	-0.04	-0.03	-0.03
Intragenic (-)	-0.01	-0.01	-0.01
Coding	-0.02	-0.01	-0.02
LNC annotated	-0.02	-0.01	-0.01

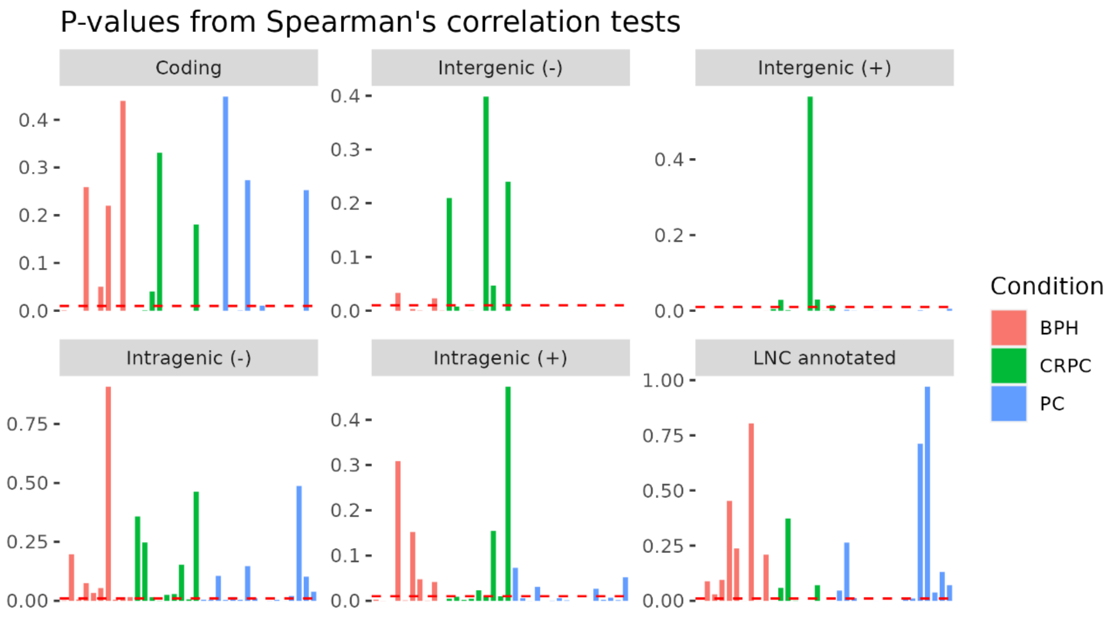


Figure 23. *P-values from Spearman’s correlation tests between methylation and expression in ON transcripts. The red dashed line marks the significance level 0.01.*

Correlation analysis in the category of ON transcripts produced quite surprising results. In annotated transcript sets, the mean correlations weakened in comparison to the mean correlations for the whole sets. Furthermore, for protein-coding transcripts, while weak, most of the correlations in BPH and PC samples were positive, and several were insignificant. The average correlation was negative only in CRPC samples. Average correlations for LNC transcripts were negative in all three conditions, but they were not much stronger than those for protein coding transcripts. Many estimates in this transcript group turned out to be insignificant, especially in BPH.

In intergenic transcripts, in nearly all samples expression and methylation anti-correlated much more strongly than when the entire intergenic set was studied. Only in CRPC 542 the correlation was positive, although weak (plus-sense 0.07/ minus-sense 0.06) and insignificant (p-value 0.0293/ 0.0468). In three more CRPC samples the correlation estimates were insignificant (p-values higher than 0.01). On average, correlation was the weakest in CRPC samples with both sense assumptions, and while with plus-sense assumption the condition with the strongest correlation was BPH, with minus-sense it was PC. Also, plus-sense assumption produced slightly stronger dependencies. But the overall trend for intergenic ON transcripts was moderate anti-correlation between methylation and expression. In intragenic ON transcripts, the trend was similar. In all samples methylation and expression anti-correlated, with strength comparable to intergenic transcripts. However, more of the estimates were insignificant across all conditions, with both strand assumptions.

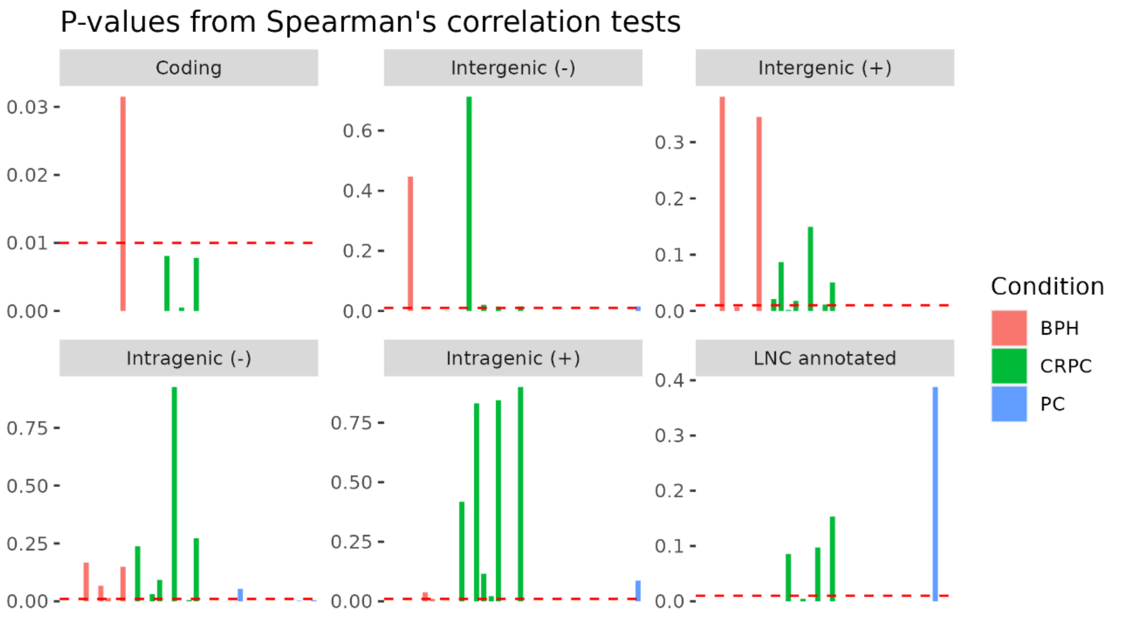


Figure 24. *P-values from Spearman's correlation test between methylation and expression in OFF transcripts. The red dashed line marks the significance level 0.01.*

In OFF transcripts, the trends were quite different. Correlation of coding OFF transcripts was much weaker than for all coding transcripts, but surprisingly, it was stronger than for ON transcripts, even though often with opposite sign. Only in individual CRPC samples and one PC sample OFF correlations were weaker than ON correlations. There was just one insignificant estimate for sample BPH 689, which was also positive (0.015, p-value 0.0315). Other positive correlations appeared only in individual CRPC samples. On a general level, the correlation between expression and methylation in OFF protein-coding transcripts was a bit higher than in other transcript sets, but it was still weak or moderate, and in several samples very weak.

In LNC annotated transcripts the trend was similar as in protein-coding transcripts. OFF correlations were in general stronger than ON correlations, except in most of CRPC samples and individual BPH and PC samples, but weaker than in the whole LNC annotated transcript set. There were more insignificant estimates than among coding transcripts, one in PC 8131, and several within CRPC samples. There were four positive estimates, all within CRPC. In general, correlation in LNC annotated OFF transcripts was weaker than in protein-coding transcripts, but slightly stronger than in unannotated transcripts. Nevertheless, it was rather weak.

In intergenic OFF transcripts, the correlation in most of the samples was even weaker than for the entire transcript group. Minus-sense assumption produced minimally stronger estimates than plus-sense assumption. All estimates were negative in BPH and

PC samples, but a few were positive in CRPC samples. Insignificant correlations were computed in some BPH and CRPC samples with plus-sense assumption, and in individual samples in all conditions with minus-sense assumption. The OFF correlations for intergenic transcripts were much weaker than ON correlations for this set. In general, the expression in OFF intergenic transcripts correlated weakly or very weakly with methylation level.

Similarly, in intragenic OFF transcripts, the correlations were also weaker than for the whole group of intragenic transcripts. Plus-sense assumption produced slightly stronger correlations. Most of the estimates in CRPC were insignificant. In summary, expression in OFF intragenic transcripts correlated weakly or very weakly with methylation level and was often insignificant.

Overall, OFF correlations were stronger than ON correlations in annotated transcripts, whereas this pattern was reversed in unannotated transcripts. Still, the correlation between expression and methylation was mostly negative, except for ON protein-coding transcripts. For those, the correlation was mostly positive, however, weak and often insignificant.

5.3.3 Thresholding of ON and OFF non-coding unannotated and annotated transcripts

In the next step, the methylation and expression within ON and OFF transcripts were subjected to data-derived thresholding in attempt to find subgroups in which transcription might be regulated by methylation. Because the initial correlation analysis showed that correlations in unannotated transcripts were closer to correlations in LNC annotated transcripts, these sets were studied together, whereas protein-coding transcripts were studied separately and less extensively.

First, LNC annotated and unannotated ON transcripts (transcripts with a peak in their promoter) were studied. Four subsets were extracted based on the expression and methylation levels: (1) transcripts with high methylation and high expression, (2) transcripts with low methylation and high expression, (3) transcripts with low methylation and low expression, and (4) transcripts with low expression and high methylation. Then, a similar analysis was performed on OFF transcripts (without a peak in promoter), but only three subsets were studied: (1) transcripts with high methylation and high expression, (2) transcripts with low methylation and high expression, and (3) transcripts with low expression and high methylation.

Unfortunately, correlation tests were mostly insignificant, or oftentimes they could not be computed due to low number of transcripts passing the thresholds, or due to expression being zero in too many transcripts. Thus, it was not possible to compare unannotated transcripts to LNC annotated transcripts based on the correlation between methylation and expression. Also, the results of each filtering constituted different percentages of the entire groups in unannotated transcripts and different in LNC annotated transcripts. However, the analysis produced some useful results, as it helped to define the initial lists of interesting unannotated transcripts, typically those with high expression (TPM > 30.00), often in multiple samples. Those lists can be found in Appendices E, F, and G.

5.3.4 Thresholding of protein-coding ON transcripts

Unlike for non-coding and unannotated transcripts, expression and methylation weakly correlated for protein-coding ON transcripts in BPH and PC samples. In attempt to explore the relationship between methylation and expression, two subgroups were extracted from protein-coding transcript set to study the trends in them: (1) highly expressed transcripts with low methylation, and (2) highly methylated transcripts with low expression. However, correlation tests of the first subsets did not produce a single significant result, and only 12 estimates were significant in the second subset.

5.4 Promoter prediction *

The results of promoter prediction and association of predictions with transcripts and their promoters was previously reported as a part of a course project, and thus they do not constitute a part of this thesis's work. However, integration of the results with other datasets and correlation analysis were done within the scope of this thesis. All mentioned results are relevant for the conclusion of this study, and therefore they are presented also here.

To validate the promoters of studied transcripts, it was decided to perform promoter prediction through genomic sequence feature analysis. Three programs were used to predict promoters: EP3, PromPredict, and TSSFinder. The total number of predictions produced by each tool genome-wide differed greatly (the lowest by EP3: 41 952, the largest by PromPredict: 5 262 288), and so did the average length of predicted promoters. While EP3 used predefined length of 400 bp to find putative promoters, and TSSFinder simply generated a list of TSS loci (1 bp), predictions produced by PromPredict differed in length, depending on the features of the predicted sequence, and the average length was 66 bp. Moreover, TSSFinder provided five pre-trained models to predict TSS loci. As the algorithm creators did not supply information on how to select the best fitting model, all five models were used to generate predictions. Then, all results were combined into a single list of predictions containing 125 949 unique TSS positions. Then, the results from all three tools were intersected and a unified list of unique predictions was obtained by filtering out the predictions which were not predicted by at least two tools, and by merging overlapping ranges. The list consisted of 31 496 unique intervals.

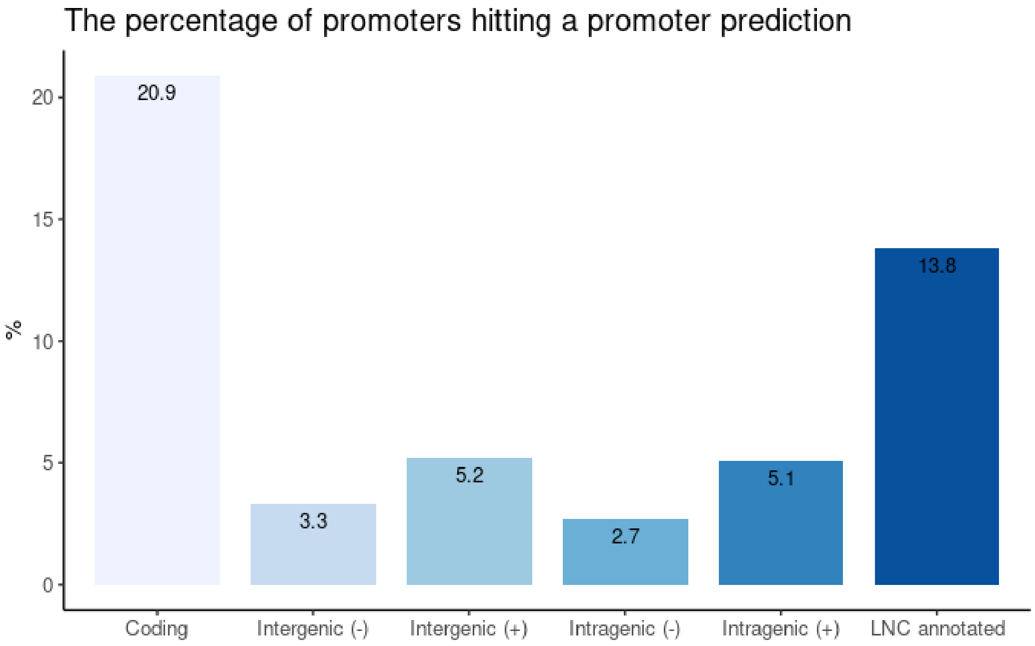


Figure 25. *The percentages of predicted promoters in each transcripts group. The numbers are rounded to one decimal place.*

Figure 25 shows the percentages of all transcripts in a given group constituted by the transcripts, whose previously defined promoters intersected promoter predictions. For simplicity, such promoters will be further referred to as predicted promoters, even though the predictions did not really cover the entire originally estimated ranges. Transcripts with predicted promoters were most prevalent among protein-coding transcripts, and the least prevalent among the unannotated ones, especially intragenic transcripts with minus-sense assumption. Generally, minus-sense assumption led to less predictions for unannotated transcripts.

Figure 26 presents the mean percentages of ON and OFF transcripts with prediction from each group in each condition. In all transcript groups, the proportion of ON transcripts with predictions was much larger than the corresponding proportion of OFF transcripts. On average, over 50% of promoters of annotated ON transcripts hit a promoter prediction across all conditions. In contrast, only approximately 12% of coding OFF transcript promoters and 8% of LNC annotated transcript promoters were found to overlap a prediction.

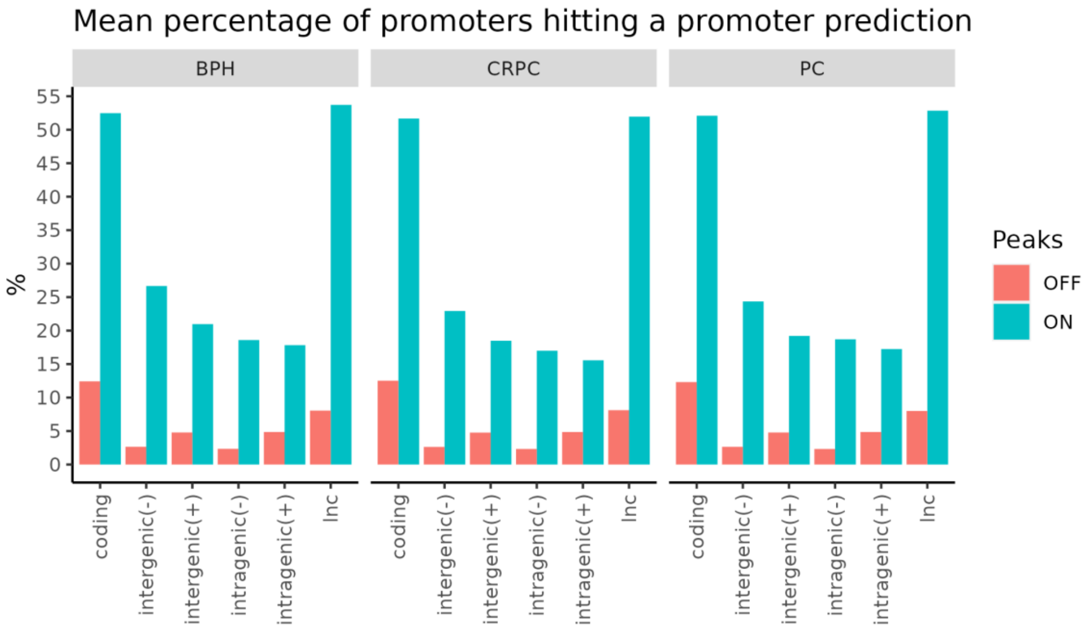


Figure 26. *Mean percentages of ON and OFF transcripts, whose previously defined promoters intersected promoter predictions.*

These proportions were substantially lower for unannotated transcripts, and there were noticeable variations between sense assumptions. Approximately 20-25% of promoters of intergenic ON transcripts, and less than 20% of promoters of intragenic ON transcripts intersected a prediction. Promoters of unannotated OFF transcripts constituted only several percent of the total numbers of OFF transcripts. Mean percentages were slightly higher in intergenic transcripts than in intragenic transcripts, and plus-sense assumption typically resulted in higher percentage. Based on the promoter predictions, approximately 20% of unannotated transcripts with accessible promoters seem to be preceded by sequences, whose structure resembles typical human core promoter. However, the direct influence of the putative promoters on the expression of those transcripts would need to be further confirmed by stronger evidence than just *in silico* promoter prediction.

5.4.1 Correlations in ON transcripts with predicted promoters

Next, the relationship between accessibility and expression was explored in ON transcripts with predicted promoters through correlation tests with significance level 0.01.

Table 6 presents mean correlations, and Figure 27 presents the p-values from Spearman's correlation tests.

Table 6. *Mean correlation estimates between accessibility and expression for ON transcripts whose previously defined promoters overlapped promoter predictions.*

Transcript group	BPH	CRPC	PC
Spearman's correlations			
Intergenic (+)	0.04	0.15	0.06
Intergenic (-)	0.00	0.08	0.04
Intragenic (+)	0.11	0.04	0.15
Intragenic (-)	0.06	0.14	0.10
Coding	0.10	0.15	0.12
LNC annotated	-0.01	0.06	0.01
Pearson's correlations			
Intergenic (+)	0.00	0.07	0.00
Intergenic (-)	-0.01	0.04	0.03
Intragenic (+)	-0.02	-0.03	0.02
Intragenic (-)	0.01	0.06	0.06
Coding	-0.01	0.00	0.00
LNC annotated	-0.01	0.04	0.00

In protein-coding transcripts, nearly all estimates were significant, positive, and moderate. Estimates in other groups, including LNC annotated transcripts, were mostly insignificant.

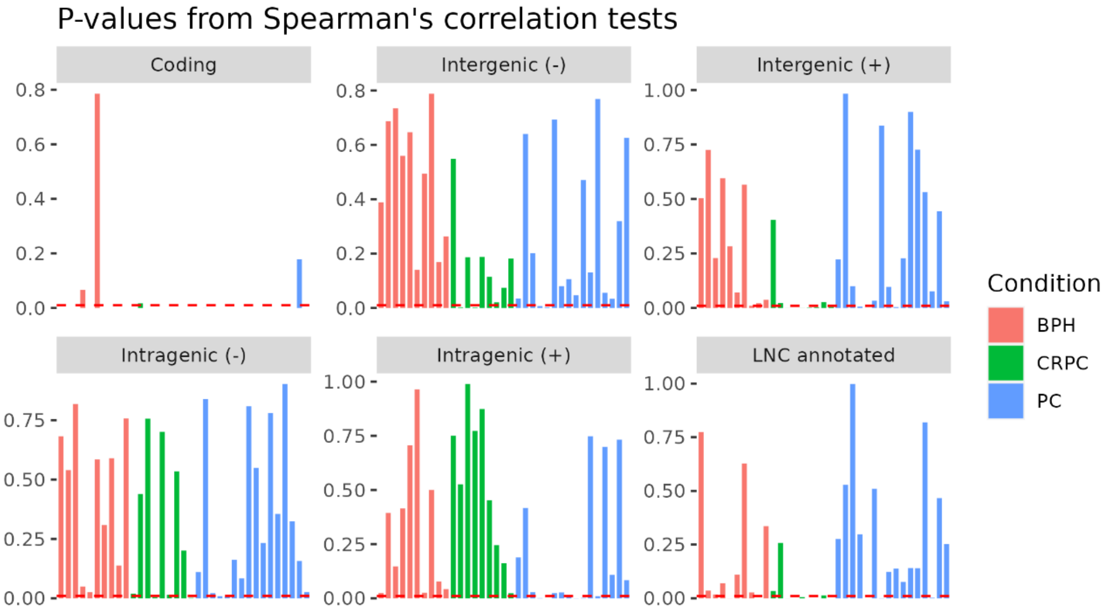


Figure 27. *P-values from Spearman's correlation test between accessibility and expression for ON transcripts whose previously defined promoters overlapped promoter predictions. The red dashed line marks the significance level.*

The relationship between methylation and expression in the subsets of ON transcripts with predicted promoters was studied as well. Table 7 presents the mean estimates for each condition, and Figure 28 shows the p-values from Spearman's correlation tests (significance level 0.01).

Table 7. *Mean correlation estimates between methylation and expression for ON transcripts, whose previously defined promoters intersected promoter predictions.*

Transcript group	BPH	CRPC	PC
Spearman's correlation			
Intergenic (+)	-0.18	-0.19	-0.15
Intergenic (-)	-0.10	-0.12	-0.20
Intragenic (+)	-0.14	-0.12	-0.18
Intragenic (-)	-0.09	-0.09	-0.14
Coding	0.10	0.03	0.09
LNC annotated	0.03	-0.02	0.00
Pearson's correlation			
Intergenic (+)	-0.07	-0.08	-0.04
Intergenic (-)	-0.03	-0.04	-0.11
Intragenic (+)	-0.03	-0.03	-0.07
Intragenic (-)	-0.01	0.01	-0.04
Coding	0.02	0.01	0.01
LNC annotated	0.01	0.00	0.00

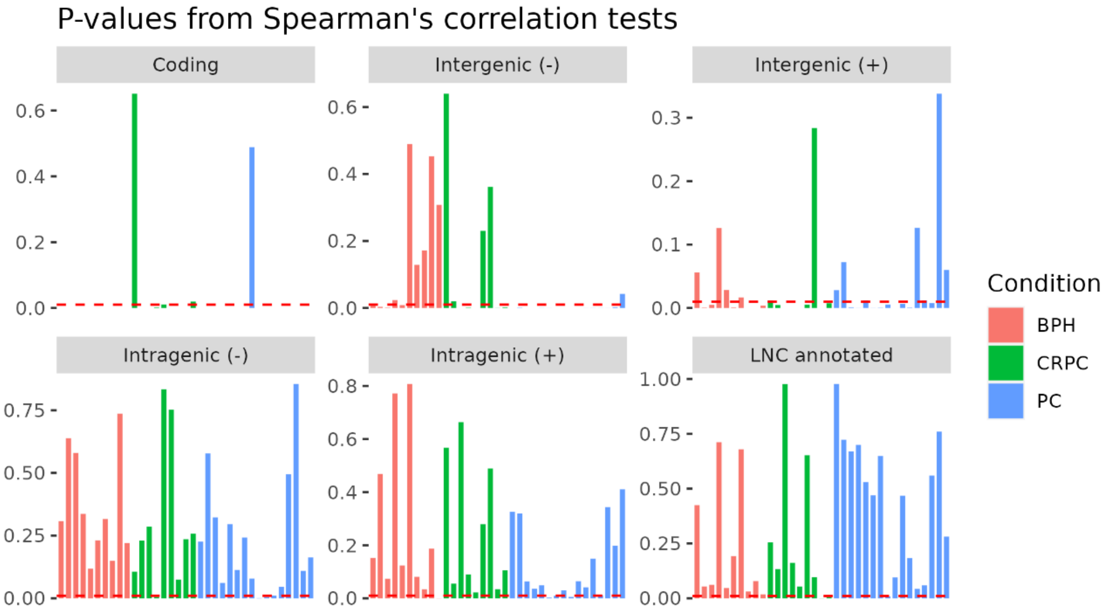


Figure 28. *P-values from Spearman's correlation tests between methylation and expression for ON transcripts, whose previously defined promoters intersected promoter predictions.*

Nearly all of the correlation estimates for protein-coding transcript set were significant, but surprisingly they were positive in all conditions. While the mean correlation in BPH and PC was moderate, the mean correlation in CRPC was weak. For LNC annotated transcripts, there was hardly any monotonic relationship between methylation and expression, and nearly all correlations were insignificant.

The results for intergenic transcripts depended on the sense assumption, both in terms of the number of significant p-values and the strength of correlations. Correlations in BPH and CRPC were stronger and more significant with plus-sense assumption, but in PC samples minus-sense assumption resulted in nearly all estimates being significant, and quite strong. The dependency in all individual samples was negative. The results for intragenic transcripts were consistently stronger with plus-sense assumption, and the largest mean was in PC samples. Several individual estimates with both sense assumptions were positive. Still, almost no estimates were significant.

The analysis resulted in significant results only for protein-coding and unannotated intergenic transcripts. It revealed that while methylation and expression positively correlated in protein-coding transcripts, they anti-correlated in intergenic transcripts.

5.4.2 Shortlisted unannotated transcripts with promoter predictions

Section “Integration of DNA methylation data” concluded with short-lists of potentially interesting unannotated transcripts. Figure 29 presents the percentages of six subsets of shortlisted transcripts, whose previously defined promoters covered a promoter prediction. For comparison, corresponding percentages of annotated transcripts passing similar thresholds were also computed and are shown in the Figure as well.

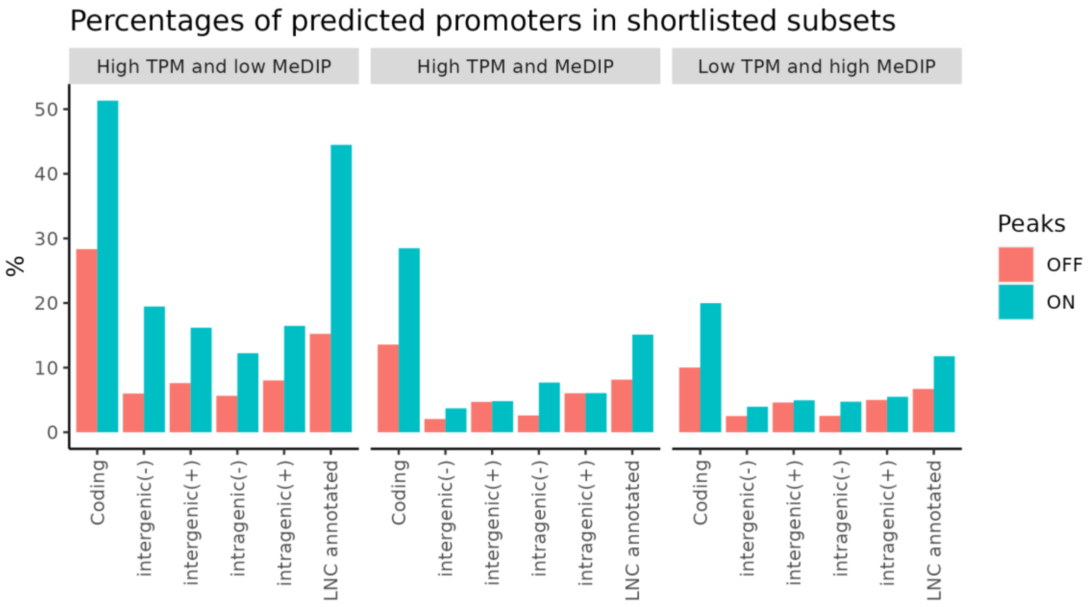


Figure 29. Percentages of promoters hitting predictions within shortlisted groups.

In all subsets, the part of predicted promoters was much higher in annotated transcripts than in unannotated transcripts. Protein-coding transcripts always had the highest rate of predicted promoters and reaching over 50% for ON transcripts with high expression and low methylation. Overall, the percentages of predicted transcripts were the highest in subsets of ON transcripts with high expression and low methylation across all transcript groups. Since supposedly these would be transcripts with accessible and unmethylated CpG-rich promoters, allowing efficient transcription, such result is reassuring. The lowest rates of prediction were found in OFF transcripts with low expression and high methylation. There were quite big differences between the percentages of predicted promoters in unannotated transcripts, depending on the sense assumption. Typically, plus-sense assumption resulted in higher percentage of predicted promoters, but not always, e.g., higher percentage of intergenic ON transcripts with high expression and low methylation had their promoter predicted with minus-sense assumption. In most cases, the prediction rates were higher for intragenic than for intergenic transcripts.

5.4.3 GC content of the shortlisted transcripts with promoter predictions

In addition to the promoter predictions, the output file from PromPredict contained the percentage of G and C bases of all analyzed 1000 nt windows, within which the predictions fell. Since methylation occurs within promoters with high GC content, the fraction of GC within predicted promoters was studied and integrated with methylation, accessibility and expression. Values produced by PromPredict were used to compute average GC content within all promoters of shortlisted transcripts, which intersected the predictions. The overlaps between the promoter ranges and the PromPredict windows were found, and weighted average was computed from the GC % of each overlapping window, weights being the lengths of the overlaps. Figure 30 presents the mean GC content of predicted promoters for ON and OFF shortlisted transcripts in every transcript group.

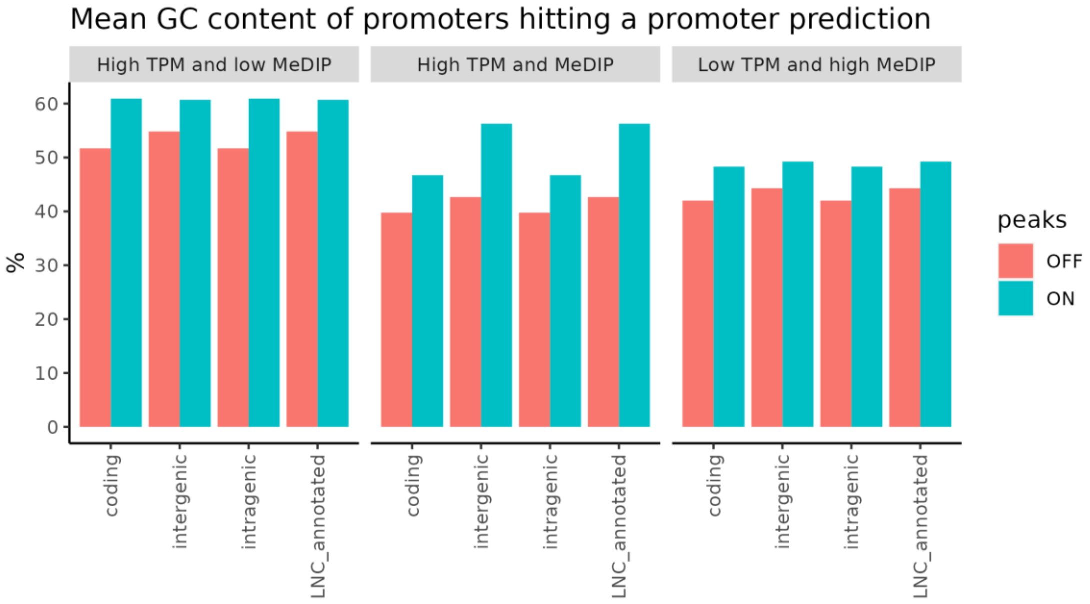


Figure 30. *Mean GC content of promoters intersecting promoter predictions. Only promoters of previously shortlisted transcripts were studied. GC content presented as a percentage.*

The largest GC contents were found within promoters of transcripts with high expression and low methylation. GC content of ON transcripts in all these subsets exceeded 60%, and the content of OFF transcripts crossed 50%. Interestingly, the mean GC contents were quite uniform across all transcript groups. This result further confirms that the promoters of these transcripts are rich in CpG islands, and their low methylation level allowed high expression. Surprisingly, the promoters of transcripts with low expression and high methylation, which were assumed to be transcripts whose promoter methylation suppressed the expression, were not particularly rich in GC bases, thus the hypothesis was not confirmed by this result. Promoters poorest in GC were the promoters of OFF transcripts with high expression and high methylation.

5.5 Histone mark detection

5.5.1 Histone marks associated with transcriptional activity*

To obtain another level of evidence for transcription regulation, the presence of histone marks associated with promoters/enhancers of transcriptionally active genes was studied within the estimated promoter ranges. The results of analysis detecting the presence of histone marks associated with transcriptional activity within transcripts promoters was previously reported as a part of a course project, and thus they do not constitute a part of this thesis's work. However, integration of the results with other datasets was done within the scope of this thesis. All mentioned results are relevant for the conclusion of this study, and therefore they are presented also here.

Each histone mark associated with transcriptionally active genes was present to some degree in the promoters or gene bodies of the studied transcripts. Transcripts with at least one histone mark were pooled together, and the percentages of the entire transcript groups they constituted in each condition can be seen in Figure 31. These transcripts will be further referred to as (transcriptionally) active transcripts, despite the fact that a histone mark locus falling within their promoters or bodies does not unambiguously determine their transcriptional activity.

In general, there were less transcripts with histone modifications in BPH samples than in cancerous samples, but that might be caused by the fact that more experiments for cancerous samples were included in the analysis. The percentages of unannotated transcripts with activity marks in cancerous tissue were more than two times greater than in benign tissue. This difference was smaller in annotated transcripts. However, both in cancer and BPH, protein-coding transcripts included the largest percentage of transcripts with activity marks, and intragenic transcripts included the lowest percentage of such transcripts. Overall, activity-associated histone marks were much less prevalent among unannotated than among annotated transcripts.

Figure 32 presents the percentage of active transcripts which covered the loci of all studied activity histone marks within given condition. Interestingly, there were much less transcripts with all activity marks in cancerous tissue, than in benign tissue. Consistently, the most abundant group of such transcripts was in protein-coding transcripts. There were barely any unannotated transcripts in cancerous tissue with all activity marks. Overall, it was quite rare for all activity marks to occur within the promoter and body of one transcript.

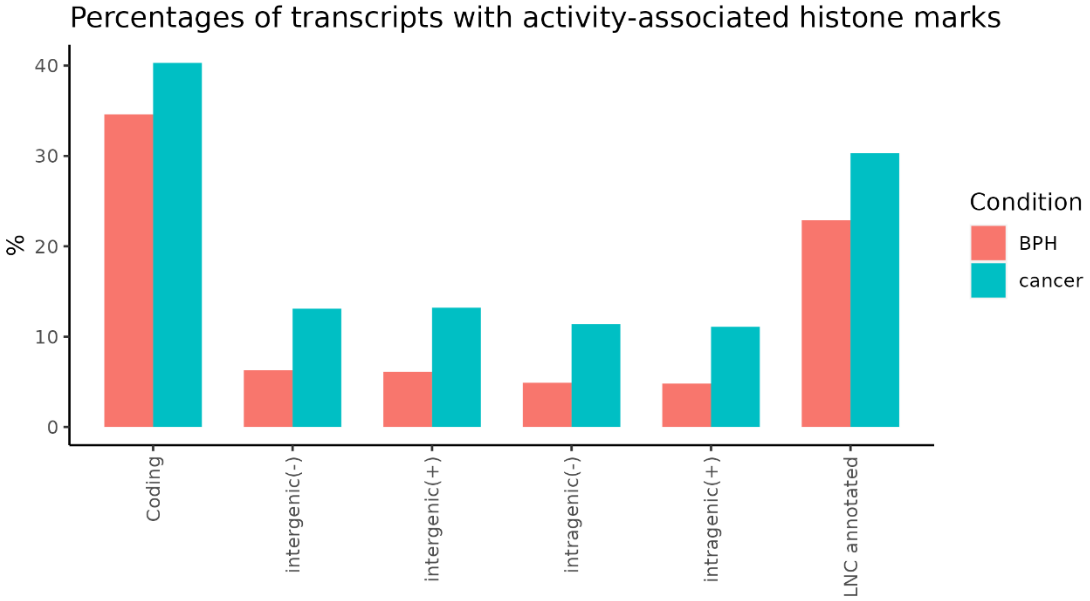


Figure 31. *Percentages of transcripts whose promoters or bodies overlapped at least one activity-associated histone mark locus.*

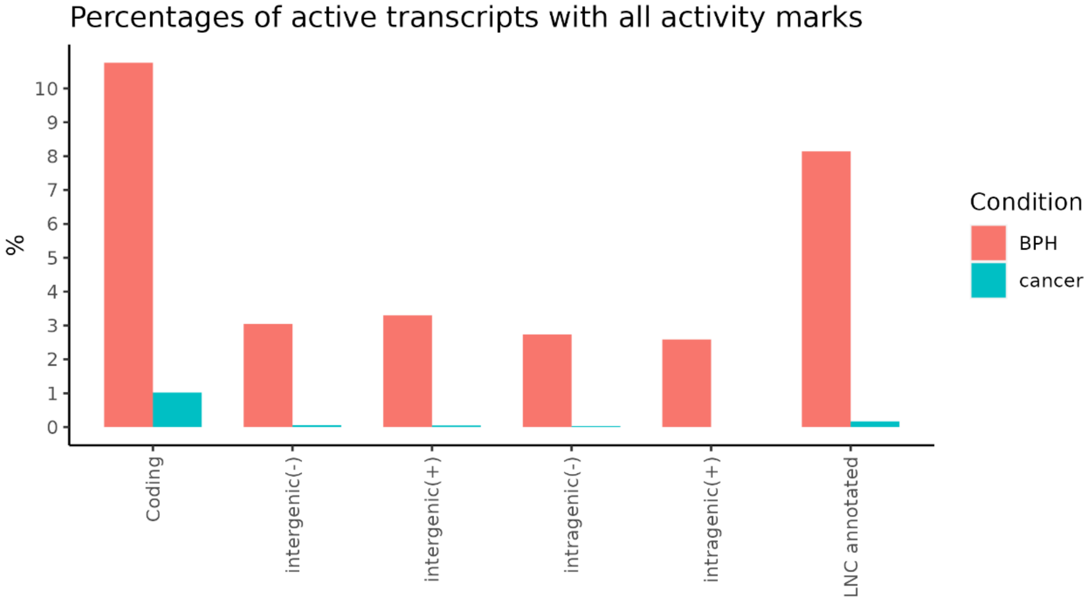


Figure 32. *Percentages of active transcripts whose promoters or bodies covered loci of all activity-associated histone modifications.*

5.5.2 Activity marks, chromatin accessibility and expression

In the next step, co-occurrence of transcript expression, and chromatin accessibility and presence of activity-associated histone marks within promoter regions was studied in attempt to find transcripts with multiple sources of evidence for transcription regulation. Since correlation analysis did not provide conclusive results in the previous analyses, another approach was taken. First, the fractions of ATAC-seq peaks in promoter regions accompanied by a histone mark within the same promoter, and the fractions of activity histone marks in promoters accompanied by an ATAC-seq peak within the same promoter were investigated. Figure 33 presents the average percentages in each condition for each transcripts group. The results differed significantly between annotated and unannotated transcripts, as well as between cancer and normal prostate. In BPH samples in annotated transcripts, accessible promoter was almost always accompanied by an activity histone modification, whereas only approximately half of the histone marks occurred within accessible promoters. The pattern was different for unannotated transcripts. Less than 60% of peaks in intergenic transcripts, and less than 50% of peaks in intragenic transcripts were accompanied by an activity mark, while only approximately 25% of histone marks occurred within accessible promoters for both groups.

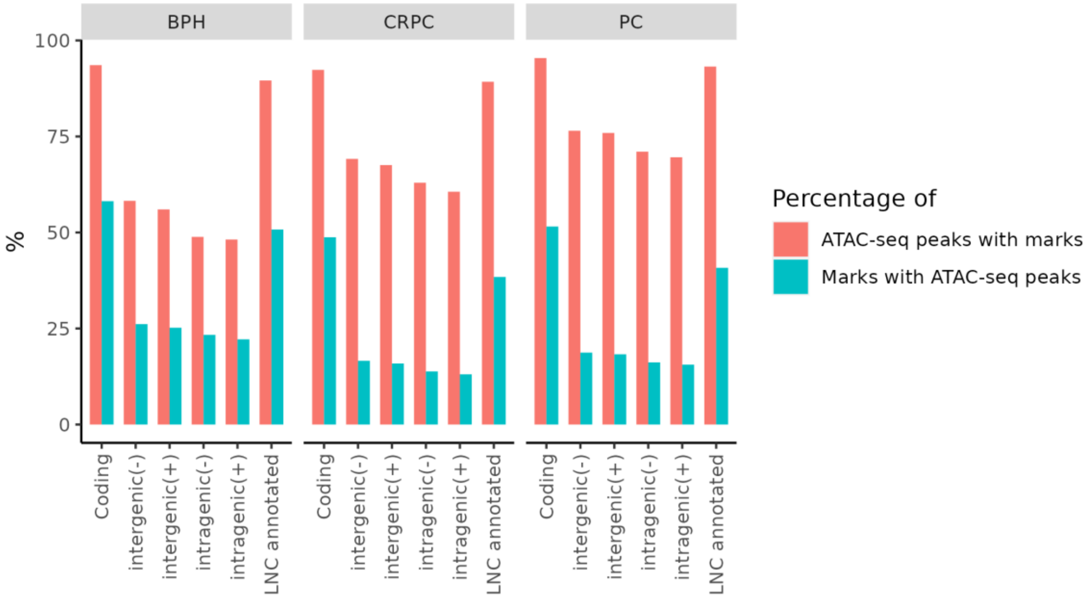


Figure 33. *The fractions of ATAC-seq peaks accompanied by an activity associated histone mark within the same promoter, and the fractions of activity histone marks accompanied by an ATAC-seq peak within the same promoter.*

Interestingly, the fractions of ATAC-seq peaks occurring together with histone marks were consistently even higher in cancer samples, especially in PC, across all transcript groups, also unannotated, while the fractions of histone modifications within open promoters were all lower in cancer than in BPH, especially in CRPC.

Then, since not all histone marks were accompanied by open chromatin, it was studied whether transcripts with a histone mark within their promoter were expressed (TPM > 0). Figure 34 presents the mean fractions of expressed transcripts in each condition within each transcript group.

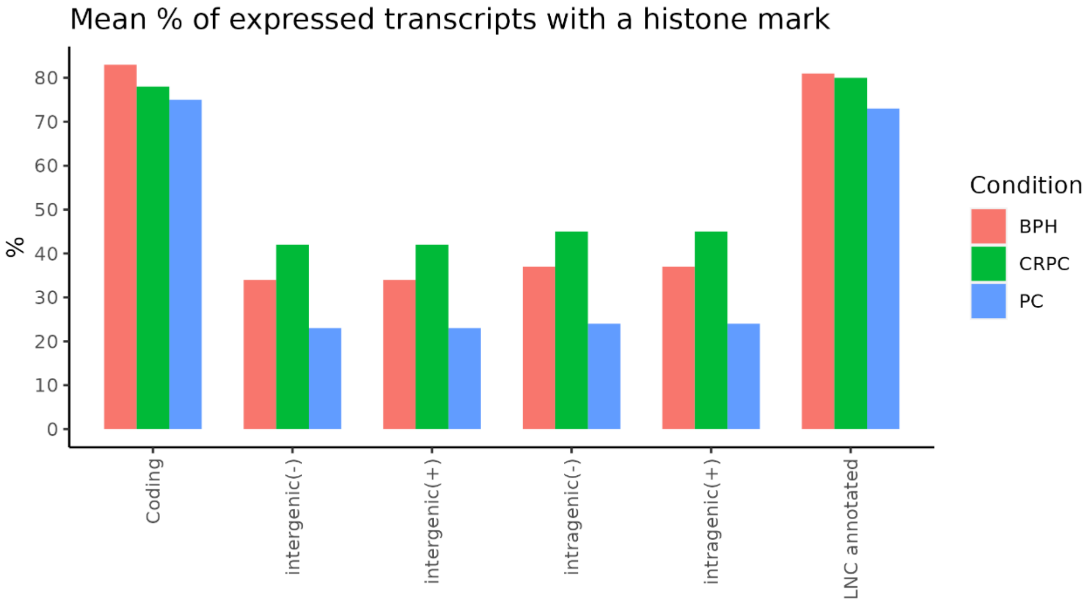


Figure 34. *Mean percentages of transcripts with activity-associated histone marks, which were expressed.*

Again, the observed patterns are different in annotated and unannotated transcripts. The majority (between 70 and 80%) of annotated transcripts marked with histone modifications were expressed. The largest fraction of such transcripts was observed in BPH samples, and the smallest in PC, however, the differences between conditions were rather modest. In contrast, inter-state differences were greater for unannotated transcripts, especially between PC and CRPC, which had the lowest and the largest fractions of expressed transcripts, respectively. Still, the percentage of expressed transcripts in CRPC was only 40% - much lower than in annotated transcripts. In BPH it was approximately 35-40%, whereas in PC only 25%.

To further explore the relationship between activity histone marks, chromatin accessibility, and expression, it was studied how many of the transcripts with both accessible promoter and a histone mark had high and how many had low expression (TPM > 0.9 quantile of a transcript group in a sample, and TPM < 0.75 quantile of a transcript group in a sample, respectively). Figure 35 presents obtained results.

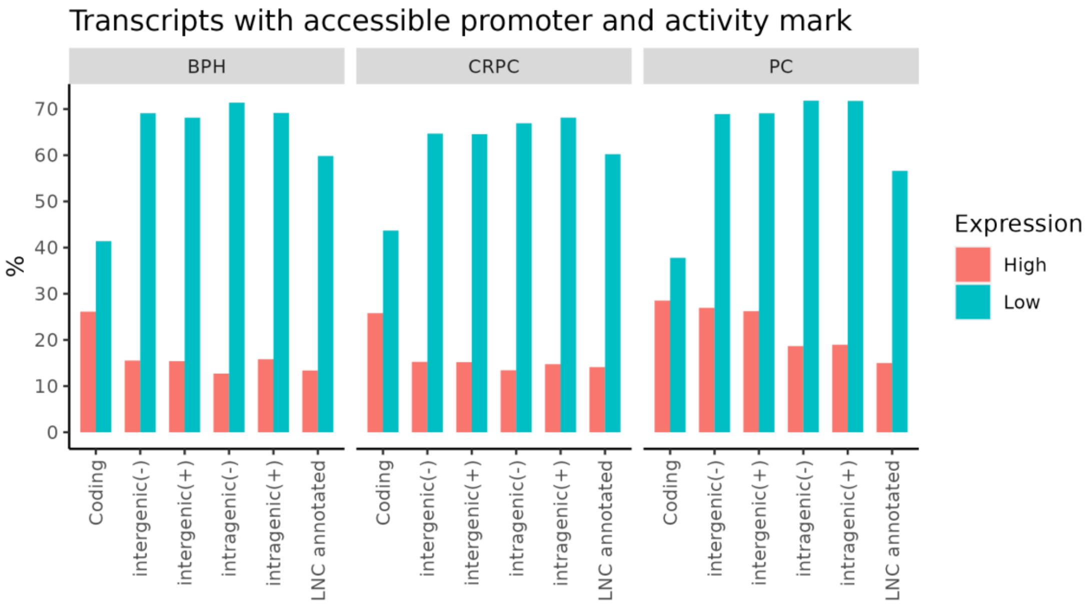


Figure 35. *Fractions of the transcripts with an ATAC-seq peak and activity histone mark within their promoters, whose expression was high (TPM > 0.9 quantile) and low (TPM < 0.75 quantile).*

Surprisingly, transcripts with high expression did not constitute a big fraction of any of the transcript groups. Modified histones within accessible promoters did not mean high expression. Quite the opposite, the majority of transcripts with such features had low or no expression across all conditions. The largest fraction of highly expressed transcripts was always within protein-coding transcripts. Despite PC having the lowest fraction of expressed transcripts overall, highly expressed transcripts were most numerous in PC samples. The fraction of highly expressed intergenic transcripts nearly reached the level of protein-coding transcripts in PC. Interestingly, the fractions of highly expressed LNC transcripts were lower than the corresponding fractions of unannotated transcripts in all conditions. Expectedly, unannotated transcripts with low or no expression were much more prevalent than annotated transcripts.

In summary, accessible promoter typically meant the presence of a transcriptional activity-related histone modification, especially in annotated transcripts and in cancer in all transcript groups. However, the presence of a histone mark was not equivalent to an accessible promoter. Activity histone marks were mostly observed in the promoters of expressed transcripts with annotations and were not indicative of expression of unannotated transcripts. Finally, histone modification and accessible promoter within the same estimated region did not guarantee high expression.

5.5.3 Activity marks in shortlisted transcripts

As was done with the promoter predictions, the presence of activity-associated histone marks was checked in the promoters of shortlisted transcripts, but this time only accessible promoters were studied. Protein-coding and LNC annotated transcripts were studied for reference. Figures 36, 38, and 40 present the percentages of transcripts passing various thresholds with at least one histone mark in their promoter in BPH, PC, and CRPC, respectively.

In BPH, histone marks were mainly found in transcripts with high expression and low methylation. Nearly all annotated transcripts passing these thresholds had a histone mark. High rates of histone marks were also observed for annotated transcripts with both high expression and high methylation. Nearly all protein-coding transcripts in this subset had a histone mark. Between 45-55% of unannotated transcripts with high expression and low methylation had also an activity-associated histone mark. In general, high methylation dramatically decreased the presence of histone marks in unannotated transcripts. As expected, histones marks in promoters of transcripts with low expression were rather rare in all transcript groups.

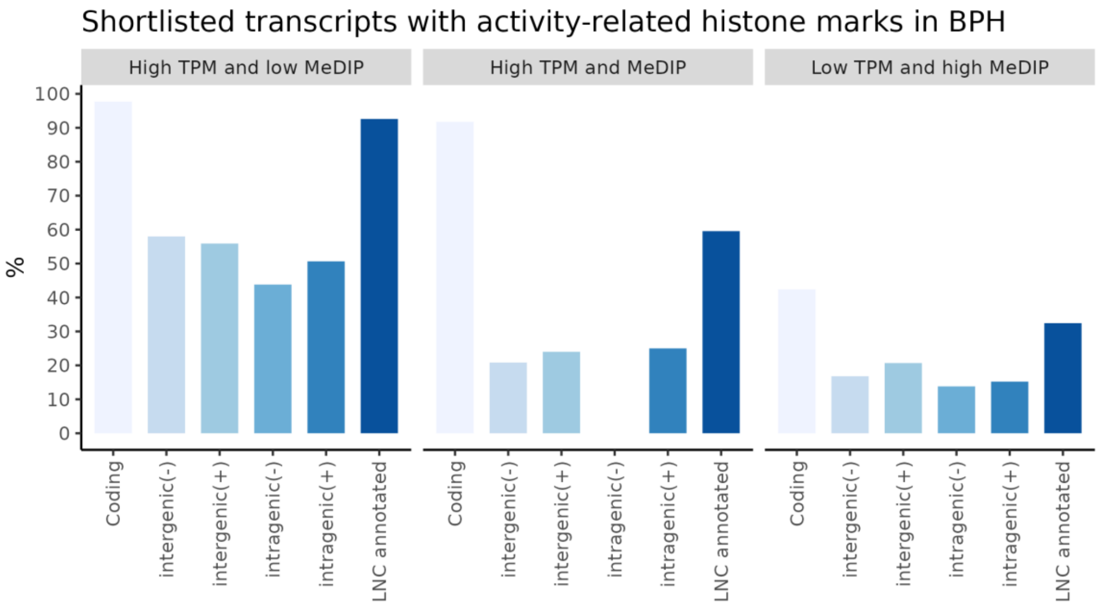


Figure 36. *Shortlisted transcripts with histone marks within their promoters in BPH samples.*

Then, it was checked how many of the transcript promoters within each subset were predicted. Figure 37 presents the results for BPH samples.

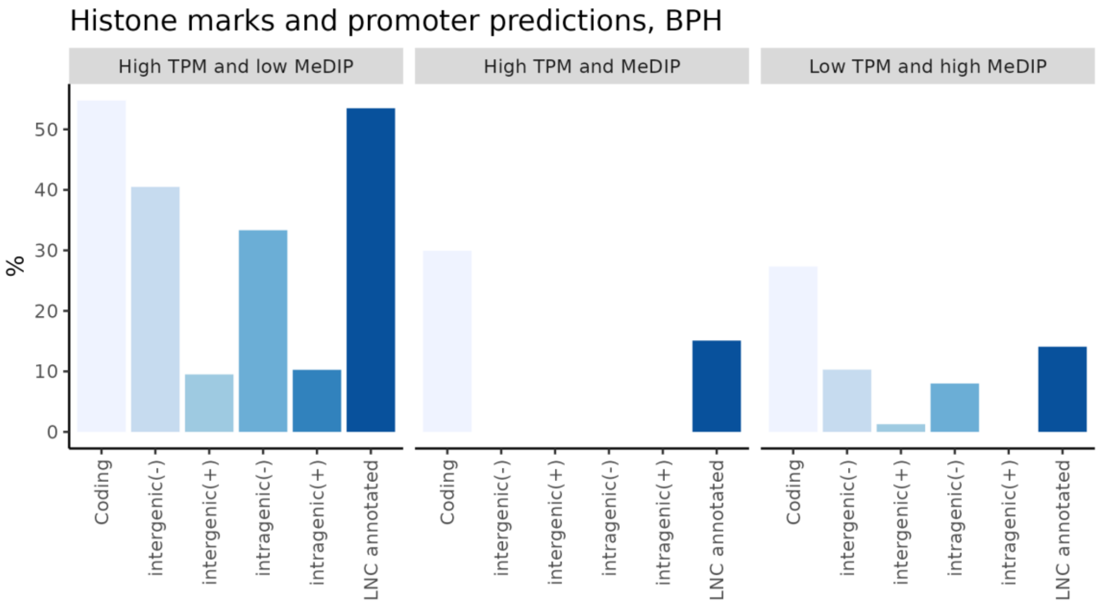


Figure 37. *The percentages of shortlisted transcripts with histone marks and promoter predictions in BPH samples.*

Most of the predicted promoters belonged to the transcripts with high expression and low methylation in all transcript groups. Interestingly, more promoters were predicted for unannotated transcripts with minus-strand assumption. Although some 30% of protein coding and 15% of LNC annotated transcripts with high expression and high methylation were predicted, no promoters of such unannotated transcripts were predicted.

Surprisingly, in PC samples histone marks were numerous not only within subsets of highly expressed transcripts with and low methylation. Also, high methylation and high expression occurred with histone marks frequently, in all transcript groups, but especially in protein-coding and intragenic transcripts with minus strand assumption. In those subsets, the percentages of unannotated transcripts were nearly as high as the percentages of annotated transcripts. In general, the percentages in PC samples were the largest across all three conditions. Again, the subset with low expression had the least histone marks, but still more than the corresponding subsets in BPH samples.

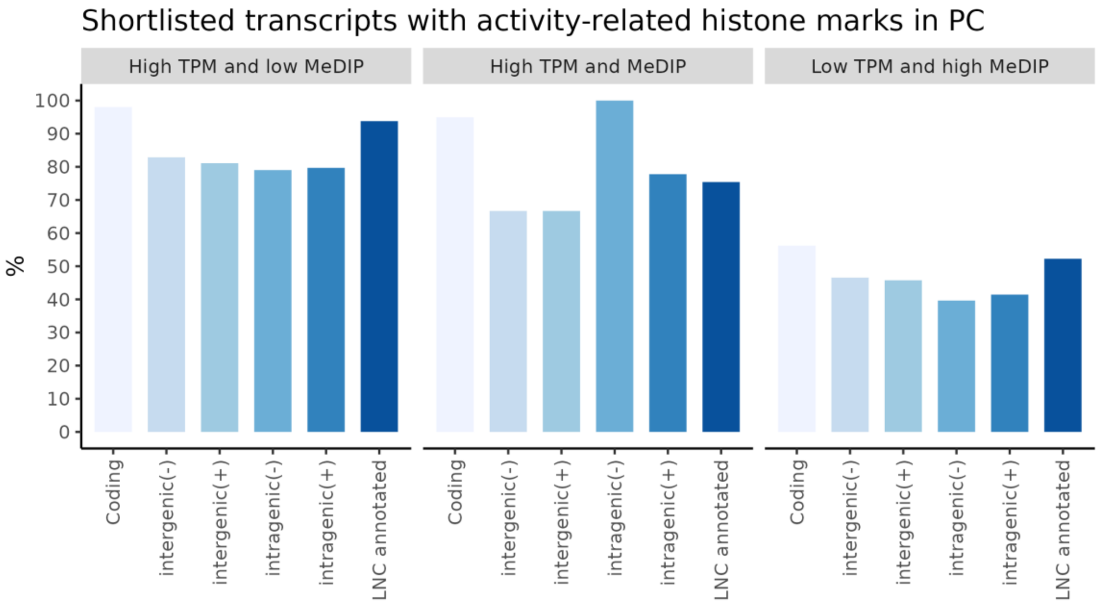


Figure 38. *Shortlisted transcripts with histone marks within their promoters in PC samples.*

Figure 39 presents the percentages of transcripts with histone marks in PC samples, whose promoters were predicted. Here, the results were very similar to the results in BPH samples. The percentages of protein-coding transcripts and LNC annotated transcripts were nearly unchanged. However, in total, less promoters of unannotated transcripts were predicted, even in the subsets with high expression and low methylation.

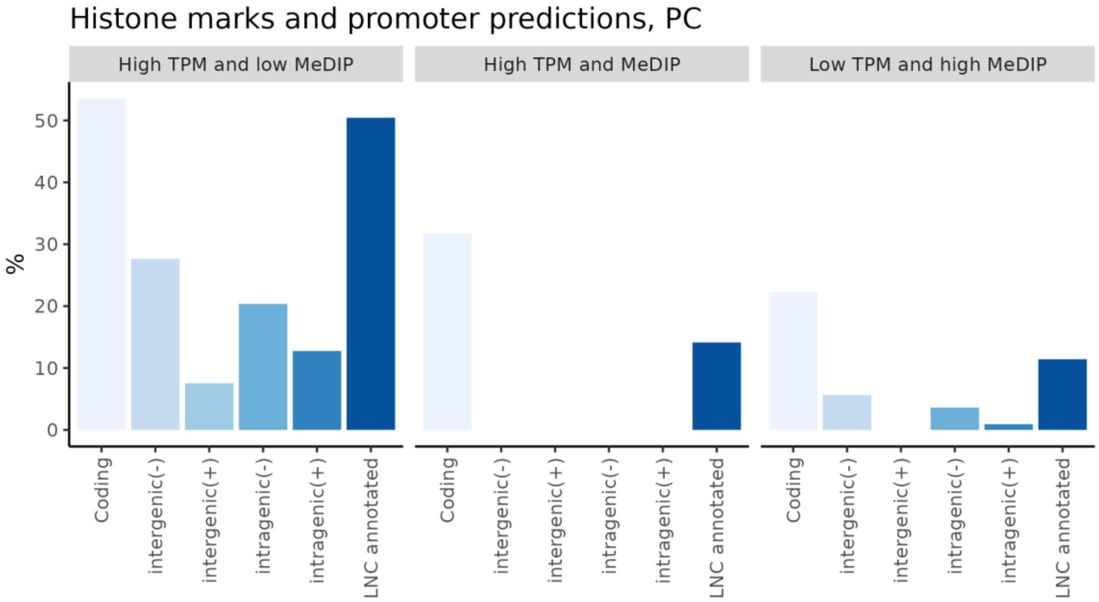


Figure 39. *The percentages of shortlisted transcripts with histone marks and promoter predictions in PC samples.*

Finally, the percentages in CRPC followed a similar patter to the ones in PC samples, however, remaining somewhat lower than in PC. The proportions of intragenic transcripts with high expression and methylation marked with histone modifications were higher than in the subset of intragenic transcripts with high expression but low methylation, and for the first time the number was higher for plus-strand assumption.

Figure 41 presents the fractions of transcripts with histone marks, whose promoters were predicted in CRPC samples. CRPC was the only condition, in which promoters of unannotated transcripts with both high expression and high methylation, and a histone mark, were predicted. Still, predictions were found only for minus-strand assumption. Surprisingly, there were even more predictions for intragenic transcripts than for LNC annotated transcripts. The percentages in other subsets were comparable with the corresponding ones in PC samples.

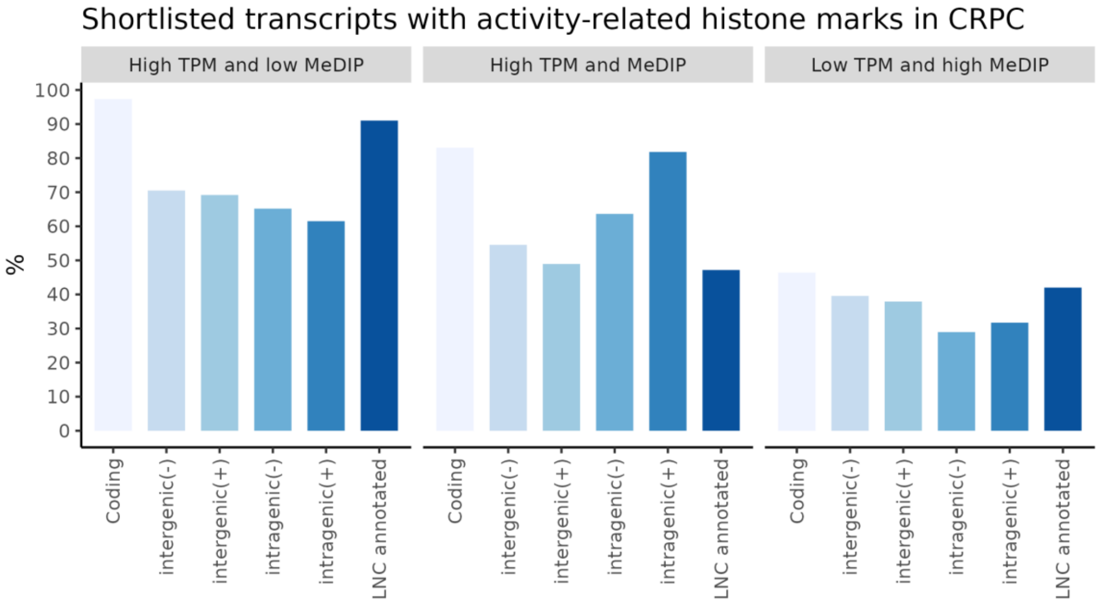


Figure 40. *Shortlisted transcripts with histone marks within their promoters in CRPC samples.*

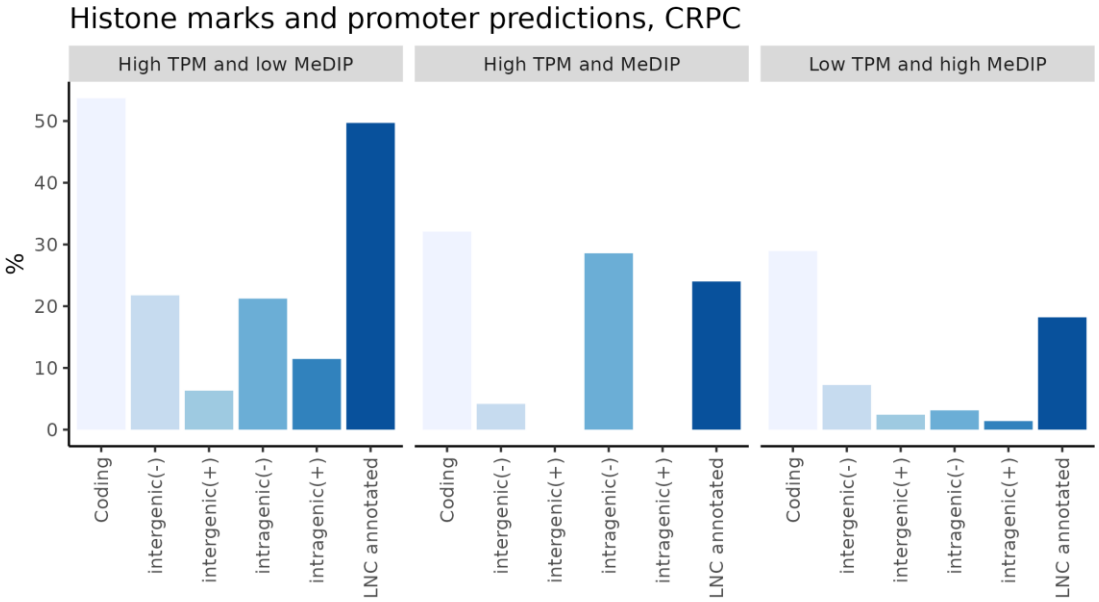


Figure 41. *The percentages of shortlisted transcripts with histone marks and predicted promoters in CRPC samples.*

5.5.4 Binding sites of transcription factor AR and shortlisted transcripts

After transcripts marked with transcriptional-activity related histone modifications were identified, the remaining question was whether they were also the targets of AR transcription factor, the main driver of prostate cancer development and progression. Figure 42 presents the findings in BPH samples. Approximately 35-50% of highly expressed transcripts with low methylation from all groups were targets of AR in BPH. Surprisingly, in the subsets with high expression, the fractions of annotated transcripts with AR binding sites were lower than those of unannotated transcripts. No intragenic transcripts with both high expression and methylation were found to bind AR.

In addition, it was checked how many of those transcripts had a predicted promoter. Figure 43 shows that overall rate of predicted promoters was low or was zero in the subsets with high methylation level in all transcript groups. Especially, for unannotated transcripts high methylation typically meant no prediction. Some predictions were found for unannotated transcripts with high expression and low methylation.

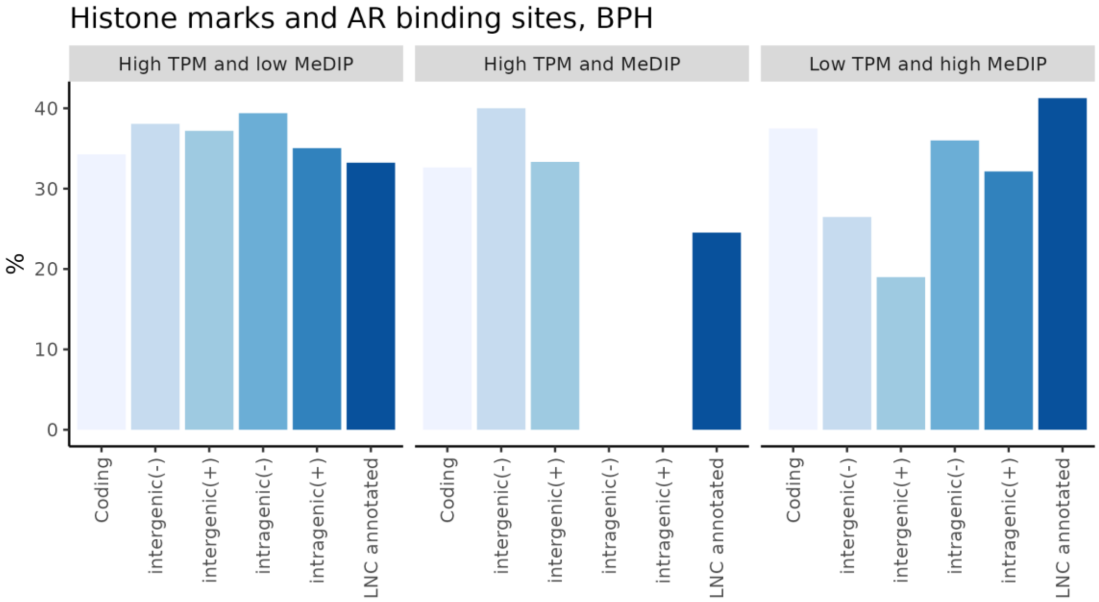


Figure 42. *Shortlisted transcriptionally active transcripts with AR binding sites within their promoters in BPH samples.*

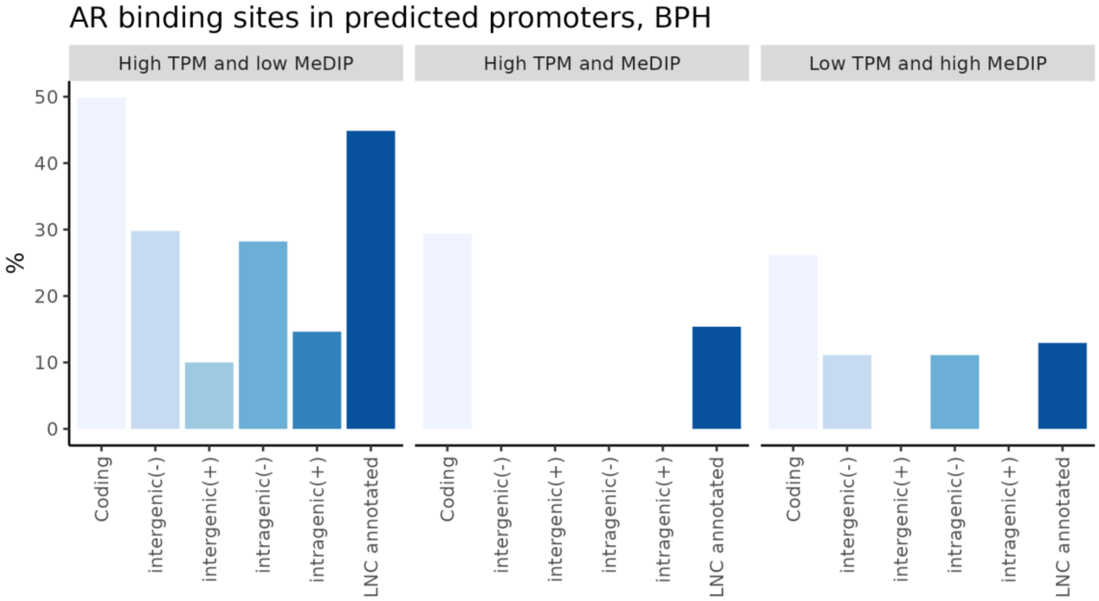


Figure 43. *Percentages of transcripts with AR binding sites and promoter predictions in BPH samples.*

As can be seen in Figure 44, in PC samples more of the transcriptionally active unannotated transcripts had AR binding sites within their promoters, while the percentages of annotated transcripts remained relatively unchanged. The biggest difference can be seen for intergenic transcripts in virtually all subsets. The fractions of intragenic transcripts with high methylation and expression were very low, like in BPH samples.

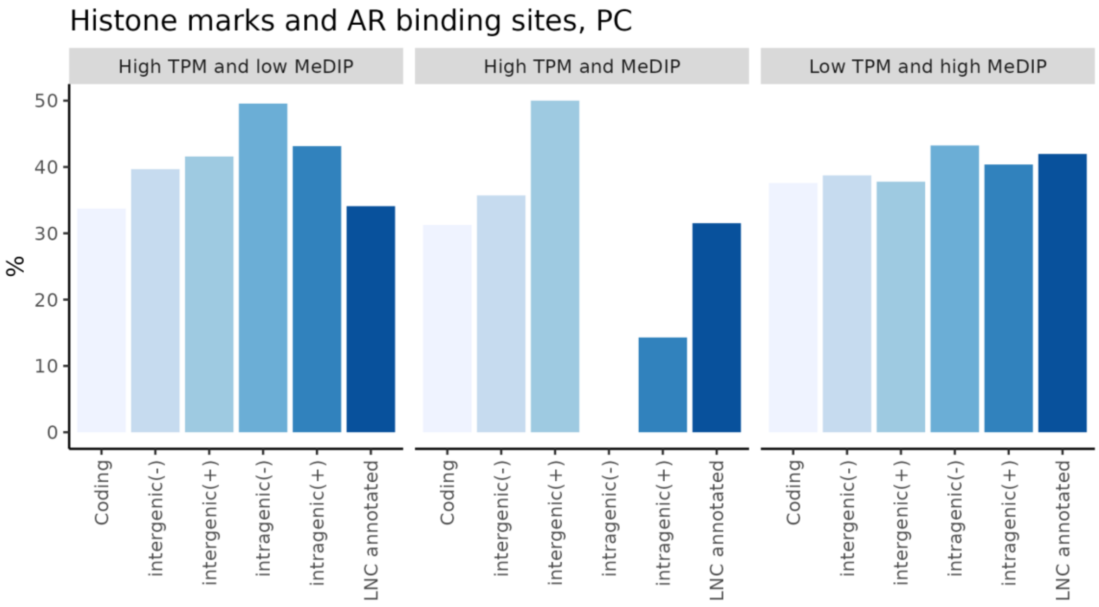


Figure 44. *Shortlisted transcriptionally active transcripts with AR binding sites with their promoters in PC samples.*

Predicted promoter among those transcripts in PC samples was rather rare, except for annotated transcripts with high expression and low methylation. The fractions can be seen in Figure 45.

Finally, Figure 46 presents the percentages of shortlisted transcripts in CRPC, whose promoters included AR binding sites. In this condition, the most notable change was 30-40% of highly expressed and methylated intragenic transcripts with AR binding sites, in contrast to none or nearly none in the other two conditions, and a drop of corresponding percentage in intergenic transcripts.

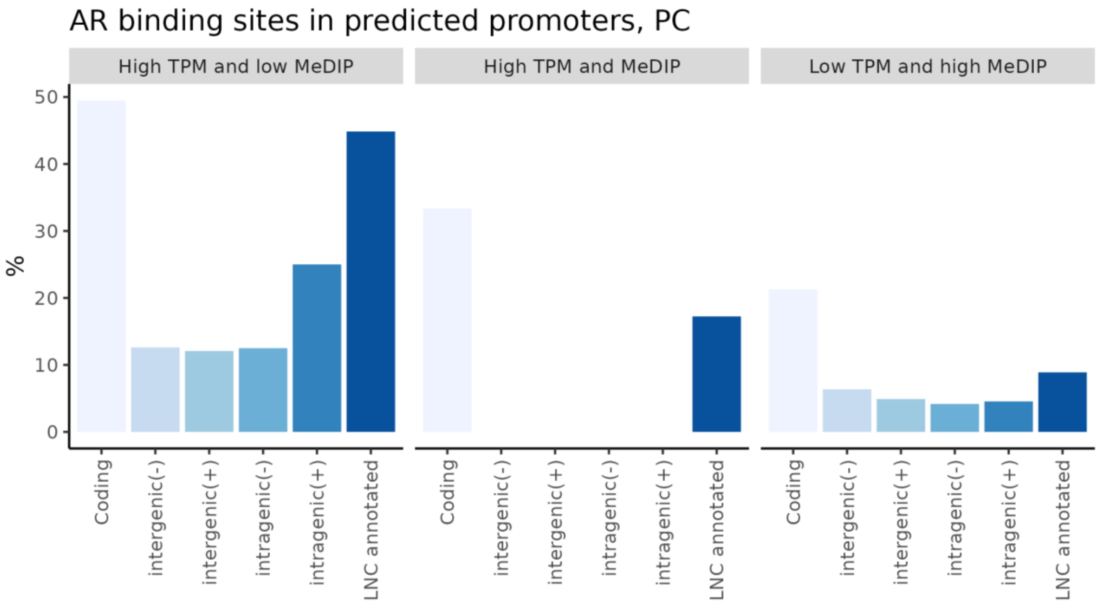


Figure 45. *Percentages of transcripts with AR binding sites, which had promoter prediction in PC samples.*

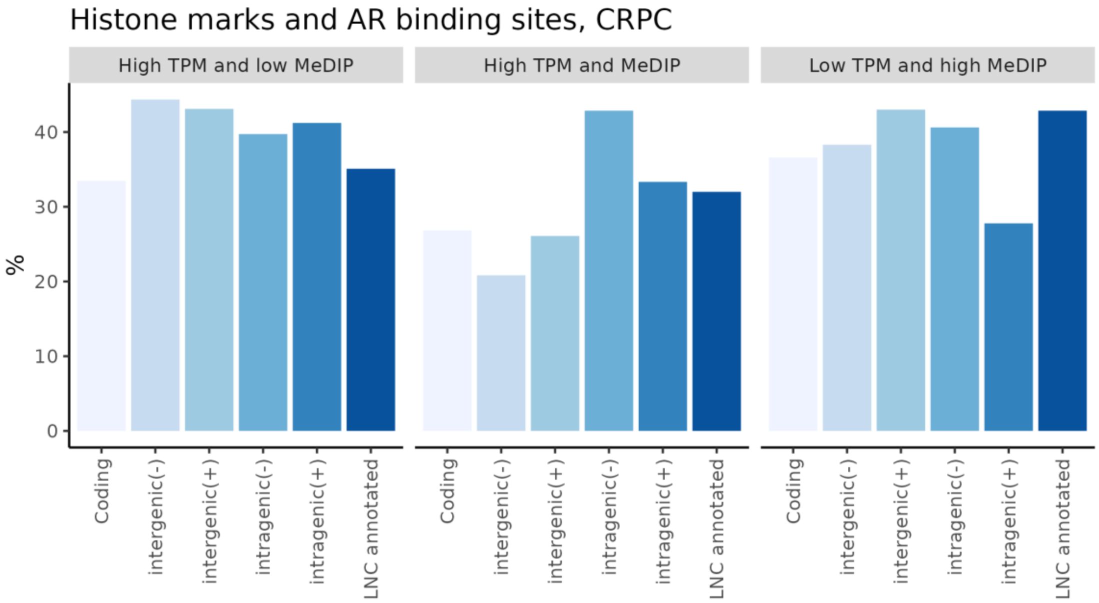


Figure 46. *Shortlisted active transcripts with AR binding sites in their promoters in CRPC samples.*

The shortlisted transcripts being targets of AR in CRPC samples had the highest rates of predicted promoters, which can be seen from Figure 47. There were predictions in the subsets of unannotated transcripts with high methylation and expression, which did not take place in the other sample groups.

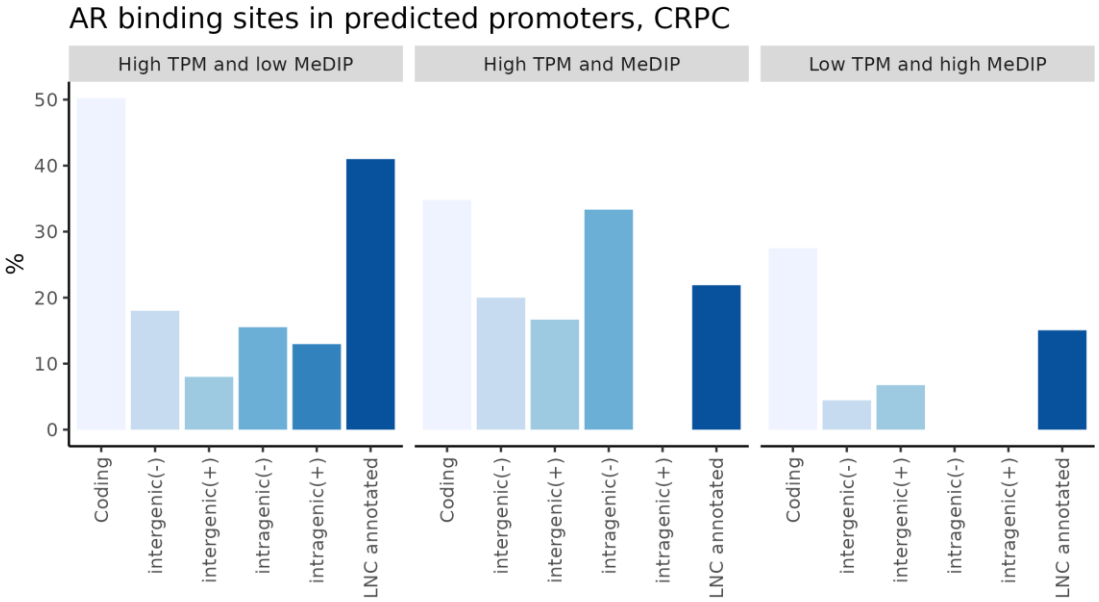


Figure 47. *Percentages of transcripts with AR binding sites, which had promoter prediction in CRPC samples.*

6. DISCUSSION

The results presented in this thesis imply that the transcription of unannotated transcripts could be regulated in a similar way to protein-coding and known LNC genes. Analysis of chromatin accessibility revealed that regions directly upstream from the TSS of a subset of the unannotated transcripts were accessible, and that the accessibility and expression of those transcripts were correlated, although not as strongly as in protein-coding genes. The correlation was stronger in PC and CRPC than in BPH, which was also true for annotated transcripts. However, in contrast to annotated transcripts, the number of expressed unannotated transcripts was not always dependent on the number of peaks in promoters. Interestingly, high expression did not always occur alongside an ATAC-seq peak in a promoter.

Integrating DNA methylation revealed higher methylation levels within promoters of unannotated transcripts than within promoters of annotated transcripts. Similar to annotated transcripts, methylation of promoters of unannotated transcripts, especially high level of methylation, typically meant no accessibility, and low expression. High methylation was not always associated with low expression. However, while expression in protein-coding transcripts with accessible promoters displayed a weak positive correlation with methylation, in the corresponding set of unannotated transcripts expression moderately anti-correlated with methylation. For annotated transcripts with inaccessible promoters, expression and methylation moderately anti-correlated, and for unannotated transcripts weakly anti-correlated, but only in BPH and PC, and there was no correlation at all in CRPC. One could hypothesize that methylation, or rather lack of it within the promoter, plays a more important role in regulation of expression of unannotated genes, at least when it comes to transcription activation and allowing chromatin opening. On the other hand, methylation is possibly more important in silencing of protein-coding genes than in activating them, thus higher and more significant correlation results for coding inaccessible transcripts. Furthermore, there is also a certain level of loss of dependency between methylation and expression in CRPC in comparison to other two conditions. Possibly, aggregation of mutations weakens the regulatory influence of methylation as the diseases progresses into castration resistance.

On average, over 50% of accessible promoters of annotated transcripts hit a promoter prediction across all conditions, but only 20-25% of accessible promoters of unannotated transcripts were predicted. The percentages of inaccessible promoters that covered a prediction were much lower in all transcript groups: approximately 12% of coding transcript promoters, 8% of LNC annotated transcript promoters, and 2-5% of unannotated transcript promoters. This was not completely surprising, since it was expected that the set of unannotated transcripts will contain only a limited number of biologically significant genes. In addition, promoter prediction programs are not able to predict all promoters, and can miss especially such promoters, whose sequence composition differs from known promoter patterns. While all above holds, the prediction rates within annotated transcript groups were also quite low. This might imply that in many cases the estimated promoter ranges do not correspond to the actual promoters and are located elsewhere. In addition, GC content analysis of predicted promoters indicated that the composition of predicted promoters with high expression and low methylation enclosed elevated percentages of GC in comparison to other subsets of transcripts, regardless of chromatin accessibility in all transcript groups. This suggests that the mentioned subsets of promoters might be actual GC-rich promoters, whose low methylation allows more efficient transcription. The transcripts regulated by these promoters might be preferably controlled by methylation, since both accessible and inaccessible promoters were found.

Histone modifications associated with transcriptionally active promoters were not as prevalent within unannotated transcript promoters as within annotated transcript promoters. They were more numerous both in CRPC and PC, than in BPH, for virtually all transcript groups. Most of the accessible promoters entailed also an activity mark, however, an activity mark in a promoter was not synonymous with accessibility, more so for unannotated transcripts than for annotated ones. While 75-80% of annotated transcripts with histone marks were expressed, the corresponding percentage of unannotated transcripts was much lower: from approximately 25% in PC, through 35-40% in BPH, to 45-50% in CRPC. While slightly more of annotated transcripts with histone marks were expressed in BPH than in other conditions, unannotated transcripts with histone marks were most often expressed in CRPC.

Studying the presence of histone marks associated with the promoters of transcriptionally active genes within subsets of transcripts with high and low expression showed that there exist some small subgroups of unannotated transcripts, whose transcription might be facilitated by the histone marks. Sometimes, their transcription might be also driven by AR. This mostly happens when the DNA methylation level is low, but not exclusively. High DNA methylation, in turn, seems to have a greater impact on the differences in expression levels and histone mark deposition between BPH, PC, and CRPC, and also between subsets of both annotated and unannotated transcripts.

Although, as discussed above, high methylation occurred with high expression, accessibility, a histone mark, and even a TF binding site, in some cases high methylation seemed to hinder expression, despite all of the abovementioned factors which in theory allow transcription. That was true in all transcript groups, and it implies that there were subsets of transcripts, whose transcription was preferentially regulated by methylation.

The rate of promoter prediction for annotated transcripts with all evidence for transcription regulation was the highest for subsets with accessible promoters, high expression and low methylation. Still, the number of predicted promoters in those subsets constituted very low fraction of all annotated transcripts. In other subsets of annotated transcripts, and in subsets of unannotated transcripts, there were either no promoter predictions, or they constituted very low fractions of those subsets.

Nevertheless, the analysis did not lead to findings of putative novel transcripts, which could be involved in PC development. The final set of transcripts with all evidence for transcription regulation was not really significantly expressed, despite passing data-derived thresholds. Even though it was expected that the unannotated transcript groups included mainly noise, lack of significant results might be caused by the many limitations of this study. First, the transcriptome assembly and functional annotation used in analyses were done some seven years ago, which is a long time in the dynamically advancing field of bioinformatics. More is known about genome every day, and new genomic annotations are published continually. It could be beneficial to perform at least the functional annotation anew. Since data was not filtered *a priori*, finding data-based thresholds which would exclude non-significant information without being too conservative was challenging. Most of the studied transcripts were barely expressed or were expressed in single samples only. Considering the genetic heterogeneity, especially as the disease progresses, finding biologically relevant transcripts in such dataset is not an easy task. Moreover, the study focused on estimations of only proximal regulatory elements upstream from the TSSs, whereas the studied transcripts could be regulated by distal enhancers. In addition, it was assumed that the promoters of the unannotated transcripts would be structurally similar to the promoters of annotated coding and non-coding genes, however, this does not need to be the case. Analyses that could be performed to explore the dataset better, would be studying the binding sites of Polymerase II, studying chromatin accessibility and methylation within gene bodies, but also interactions within transactivating domains (TADs). Furthermore, only the binding sites of transcription factor AR were investigated. The binding sites of other TFs known to play an important role in prostate cancer could be also explored within the unannotated transcripts. Finally, experimental data used to study the presence of transcriptional activity-associated histone modifications, as well as the binding sites of AR, were not specific for this sample group. Therefore, the results are not precise and some significant information might have been missed along the analysis.

7. CONCLUSIONS

The purpose of this thesis was to study the interplay between the genomic and epigenomic patterns of transcriptional activity via chromatin accessibility, DNA methylation, histone modification marks, and transcription factor AR targeting, in two subgroups of previously identified unannotated groups of transcripts and compare them with patterns observed in protein-coding (and LNC annotated) genes. The aim was to find subsets of unannotated transcripts, whose epigenomic signatures imply RNA Polymerase II transcription regulation, if such exist. Each individual layer of epigenomics, as well as multilayer data integration, provided evidence for regulation of a small subset of unannotated transcripts. However, identification of individual putative novel transcripts was not successful. The analysis is by no means exhaustive, and possibly different approaches and filtering strategies could lead to more conclusive results. The analyses performed for the needs of this thesis do not unambiguously prove that there are or that there are no biologically significant transcripts within the sets of unannotated transcripts of the studied sample cohort.

REFERENCES

Abeel, T. *et al.* (2008) 'Generic eukaryotic core promoter prediction using structural features of DNA', *Genome Research*, 18(2), p. 310. Available at: https://doi.org/10.1101/GR.6991408.

Allis, C.D. and Jenuwein, T. (2016) 'The molecular hallmarks of epigenetic control', *Nature Reviews Genetics*, 17(8), pp. 487–500. Available at : https://doi.org/10.1038/nrg.2016.59.

Armenia, J. *et al.* (2018) 'The long tail of oncogenic drivers in prostate cancer', *Nature Genetics 2018 50:5*, 50(5), pp. 645–651. Available at: https://doi.org/10.1038/s41588-018-0078-z.

Bai, L. and Morozov, A. v. (2010) 'Gene regulation by nucleosome positioning', *Trends in Genetics*, 26(11), pp. 476–483. Available at: https://doi.org/10.1016/j.tig.2010.08.003.

Baylin, S.B. and Jones, P.A. (2011) 'A decade of exploring the cancer epigenome — biological and translational implications', *Nature Reviews Cancer*, 11(10), pp. 726–734. Available at : https://doi.org/10.1038/nrc3130.

Beltran, H. *et al.* (2014) 'Aggressive variants of castration-resistant prostate cancer', *Clinical Cancer Research*, 20(11), pp. 2846–2850. Available at: https://doi.org/10.1158/1078-0432.CCR-13-3309/176285/AM/AGGRESSIVE-VARIANTS-OF-CASTRATION-RESISTANT.

Beltran, H. and Demichelis, F. (2015) 'Intrapatient heterogeneity in prostate cancer', *Nature Reviews Urology 2015 12:8*, 12(8), pp. 430–431. Available at : https://doi.org/10.1038/NRUROL.2015.182.

Buenrostro, J.D. *et al.* (2013) 'Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position', *Nature Methods*, 10(12), pp. 1213–1218. Available at : https://doi.org/10.1038/nmeth.2688.

Carioli, G. *et al.* (2020) 'European cancer mortality predictions for the year 2020 with a focus on prostate cancer', *Annals of Oncology*, 31(5), pp. 650–658. Available at: https://doi.org/10.1016/J.ANNONC.2020.02.009.

Carlberg, C. and Molnár, F. (2020) 'Genes and Chromatin', in *Mechanisms of Gene Regulation: How Science Works*. Cham: Springer International Publishing, pp. 1–17. Available at: https://doi.org/10.1007/978-3-030-52321-3_1.

Chandrasekar, T. *et al.* (2015) 'Mechanisms of resistance in castration-resistant prostate cancer (CRPC)', *Translational Andrology and Urology*, 4(3), p. 365. Available at : https://doi.org/10.3978/J.ISSN.2223-4683.2015.05.02.

Chung, S. *et al.* (2011) 'Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility', *Cancer Science*, 102(1), pp. 245–252. Available at : https://doi.org/10.1111/J.1349-7006.2010.01737.X.

Clapier, C.R. *et al.* (2017) 'Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes', *Nature Reviews Molecular Cell Biology*, 18(7), pp. 407–422. Available at: https://doi.org/10.1038/nrm.2017.26.

Cooper, S.J. *et al.* (2006) ‘Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome’, *Genome Research*, 16(1), pp. 1–10. Available at : https://doi.org/10.1101/gr.4222606.

Dalmartello, M. *et al.* (2022) ‘European cancer mortality predictions for the year 2022 with focus on ovarian cancer’, *Annals of Oncology*, 33(3), pp. 330–339. Available at : https://doi.org/10.1016/j.annonc.2021.12.007.

Danino, Y.M. *et al.* (2015) ‘The core promoter: At the heart of gene expression’, *Biochimica et Biophysica Acta (BBA) – Gene Regulatory Mechanisms*, 1849(8), pp. 1116–1131. Available at : https://doi.org/10.1016/j.bbagrm.2015.04.003.

Du, J. *et al.* (2015) ‘DNA methylation pathways and their crosstalk with histone methylation’, *Nature Reviews Molecular Cell Biology*, 16(9), pp. 519–532. Available at : https://doi.org/10.1038/nrm4043.

Du, Q. *et al.* (2015) ‘Methyl-CpG-binding domain proteins: readers of the epigenome’, *Epigenomics*, 7(6), pp. 1051–1073. Available at : https://doi.org/10.2217/epi.15.39.

Espiritu, S.M.G. *et al.* (2018) ‘The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression’, *Cell*, 173(4), pp. 1003-1013.e15. Available at : https://doi.org/10.1016/J.CELL.2018.03.029.

Gendron, J. *et al.* (2019) ‘Long non-coding RNA repertoire and open chromatin regions constitute midbrain dopaminergic neuron – specific molecular signatures’, *Scientific Reports 2019 9 :1*, 9(1), pp. 1–16. Available at : https://doi.org/10.1038/s41598-018-37872-1.

Getzenberg, R.H. and Kulkarni, P. (2014) ‘Etiology and Pathogenesis’, in S.A. Kaplan and K.T. McVary (eds) *Male Lower Urinary Tract Symptoms and Benign Prostatic Hyperplasia*. Hoboken: John Wiley & Sons, Incorporated, pp. 1–9.

Grasso, C.S. *et al.* (2012) ‘The mutational landscape of lethal castration-resistant prostate cancer’, *Nature 2012 487:7406*, 487(7406), pp. 239–243. Available at : https://doi.org/10.1038/nature11125.

Gundem, G. *et al.* (2015) ‘The evolutionary history of lethal metastatic prostate cancer’, *Nature*, 520(7547), pp. 353–357. Available at : https://doi.org/10.1038/nature14347.

Gutiérrez, J.L. *et al.* (2007) ‘Activation domains drive nucleosome eviction by SWI/SNF’, *The EMBO Journal*, 26(3), pp. 730–740. Available at: https://doi.org/10.1038/sj.emboj.7601524.

Hammerich, K.H., Ayala, G.E. and Wheeler, T.M. (2008) ‘Anatomy of the prostate gland and surgical pathology of prostate cancer’, in H. Hricak and P. Scardino (eds) *Prostate Cancer*. Cambridge: Cambridge University Press, pp. 1–14.

Handle, F. *et al.* (2019) ‘Drivers of AR indifferent anti-androgen resistance in prostate cancer cells’, *Scientific Reports 2019 9:1*, 9(1), pp. 1–11. Available at: https://doi.org/10.1038/s41598-019-50220-1.

Heinlein, C.A. and Chang, C. (2004) ‘Androgen Receptor in Prostate Cancer’, *Endocrine Reviews*, 25(2), pp. 276–308. Available at : https://doi.org/10.1210/ER.2002-0032.

Hessels, D. *et al.* (2003) ‘DD3PCA3-based Molecular Urine Analysis for the Diagnosis of Prostate Cancer’, *European Urology*, 44(1), pp. 8–16. Available at: https://doi.org/10.1016/S0302-2838(03)00201-X.

Hrdlickova, R., Toloue, M. and Tian, B. (2017) ‘<scp>RNA</scp> -Seq methods for transcriptome analysis’, *WIREs RNA*, 8(1). Available at: https://doi.org/10.1002/wrna.1364.

Huang, X., Chen, X.-G. and Armbruster, P.A. (2016) ‘Comparative performance of transcriptome assembly methods for non-model organisms’, *BMC Genomics*, 17(1), p. 523. Available at: https://doi.org/10.1186/s12864-016-2923-8.

Hutchinson, G.B. (1996) ‘The prediction of vertebrate promoter regions using differential hexamer frequency analysis’, *CABIOS*, 12(5), pp. 391–398. Available at: https://academic.oup.com/bioinformatics/article/12/5/391/210507 (Accessed: 31 October 2022).

Jacinto, F. v., Ballestar, E. and Esteller, M. (2008) ‘Methyl-DNA immunoprecipitation (MeDIP): Hunting down the DNA methylome’, *BioTechniques*, 44(1), pp. 35–43. Available at: https://doi.org/10.2144/000112708.

Jaiswal, R. and Jafa, E. (2020) ‘Epigenetics’, *Indian Journal of Medical and Paediatric Oncology*, 41(03), pp. 378–380. Available at: https://doi.org/10.4103/ijmpo.ijmpo_24_20.

JASPAR – A database of transcription factor binding profiles (no date). Available at: https://jaspar.genereg.net/ (Accessed: 17 October 2022).

Klemm, S.L., Shipony, Z. and Greenleaf, W.J. (2019) ‘Chromatin accessibility and the regulatory epigenome’, *Nature Reviews Genetics*, 20(4), pp. 207–220. Available at : https://doi.org/10.1038/s41576-018-0089-8.

Kohvakka, A. *et al.* (2020) ‘AR and ERG drive the expression of prostate cancer specific long noncoding RNAs’, *Oncogene*, 39(30), pp. 5241–5251. Available at : https://doi.org/10.1038/s41388-020-1365-6.

Kovaka, S. *et al.* (2019) ‘Transcriptome assembly from long-read RNA-seq alignments with StringTie2’, *Genome Biology*, 20(1), p. 278. Available at: https://doi.org/10.1186/s13059-019-1910-1.

Kugel, J.F. and Goodrich, J.A. (2017) ‘Finding the start site: redefining the human initiator element’, *Genes & Development*, 31(1), pp. 1–2. Available at : https://doi.org/10.1101/gad.295980.117.

Kukkonen, K. *et al.* (2021) ‘Chromatin and Epigenetic Dysregulation of Prostate Cancer Development, Progression, and Therapeutic Response’, *Cancers 2021, Vol. 13, Page 3325*, 13(13), p. 3325. Available at : https://doi.org/10.3390/CANCERS13133325.

Landolin, J.M. *et al.* (2010) ‘Sequence features that drive human promoter function and tissue specificity’, *Genome Research*, 20(7), pp. 890–898. Available at : https://doi.org/10.1101/gr.100370.109.

Li, Y. *et al.* (2021) ‘The emerging role of ISWI chromatin remodeling complexes in cancer’, *Journal of Experimental & Clinical Cancer Research*, 40(1), p. 346. Available at : https://doi.org/10.1186/s13046-021-02151-x.

Long, M.D. *et al.* (2021) ‘Dynamic patterns of DNA methylation in the normal prostate epithelial differentiation program are targets of aberrant methylation in prostate cancer’, *Scientific Reports*, 11(1), p. 11405. Available at: https://doi.org/10.1038/s41598-021-91037-1.

Luo, C., Hajkova, P. and Ecker, J.R. (2018a) ‘Dynamic DNA methylation: In the right place at the right time’, *Science*, 361(6409), pp. 1336–1340. Available at: https://doi.org/10.1126/science.aat6806.

Luo, C., Hajkova, P. and Ecker, J.R. (2018b) ‘Dynamic DNA methylation: In the right place at the right time’, *Science*, 361(6409), pp. 1336–1340. Available at : https://doi.org/10.1126/science.aat6806.

Majumdar, S. *et al.* (2011) ‘Aberrant DNA Methylation and Prostate Cancer’, *Current Genomics*, 12(7), pp. 486–505. Available at: https://doi.org/10.2174/138920211797904061.

de Medeiros Oliveira, M. *et al.* (2021) ‘TSSFinder—fast and accurate ab initio prediction of the core promoter in eukaryotic genomes’, *Briefings in Bioinformatics*, 22(6), pp. 1–12. Available at: https://doi.org/10.1093/BIB/BBAB198.

Miglani, G.S. (2014) *Gene Expression*. New Delhi: Alpha Science International Ltd.

Miller, J.L. and Grant, P.A. (2013) ‘The Role of DNA Methylation and Histone Modifications in Transcriptional Regulation in Humans’, *Sub-cellular biochemistry*, 61, p. 289. Available at: https://doi.org/10.1007/978-94-007-4525-4_13.

mitoblacklist/peaks at master · buenrostrolab/mitoblacklist · GitHub (no date). Available at: https://github.com/buenrostrolab/mitoblacklist/tree/master/peaks (Accessed: 17 October 2022).

Morey, C. *et al.* (2011) ‘DNA Free Energy-Based Promoter Prediction and Comparative Analysis of Arabidopsis and Rice Genomes’, *Plant Physiology*, 156(3), p. 1300. Available at : https://doi.org/10.1104/PP.110.167809.

Oberbeckmann, E. *et al.* (2021) ‘Genome information processing by the INO80 chromatin remodeler positions nucleosomes’, *Nature Communications*, 12(1), p. 3231. Available at : https://doi.org/10.1038/s41467-021-23016-z.

Peng, L. *et al.* (2015) ‘Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types’, *Scientific Reports 2015 5:1*, 5(1), pp. 1–18. Available at: https://doi.org/10.1038/srep13413.

Perdew, G.H., vanden Heuvel, J.P. and Peters, J.M. (eds) (2007) *Regulation of Gene Expression*. Totowa, NJ: Humana Press. Available at: https://doi.org/10.1007/978-1-59745-228-1.

Poli, J., Gasser, S.M. and Papamichos-Chronakis, M. (2017) ‘The INO80 remodeller in transcription, replication and repair’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1731), p. 20160290. Available at : https://doi.org/10.1098/rstb.2016.0290.

Prensner, J.R. *et al.* (2011) ‘Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression’, *Nature Biotechnology 2011 29:8*, 29(8), pp. 742–749. Available at : https://doi.org/10.1038/nbt.1914.

Prensner, J.R. *et al.* (2013) ‘The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex’, *Nature Genetics 2013 45:11*, 45(11), pp. 1392–1398. Available at: https://doi.org/10.1038/ng.2771.

Prensner, J.R. and Chinnaiyan, A.M. (2011) ‘The Emergence of lncRNAs in Cancer Biology’, *Cancer Discovery*, 1(5), pp. 391–407. Available at : https://doi.org/10.1158/2159-8290.CD-11-0209.

Quigley, D.A. *et al.* (2018) ‘Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer’, *Cell*, 174(3), pp. 758-769.e9. Available at: https://doi.org/10.1016/J.CELL.2018.06.039.

Rangannan, V. and Bansal, M. (2010) ‘High-quality annotation of promoter regions for 913 bacterial genomes’, *Bioinformatics*, 26(24), pp. 3043–3050. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTQ577.

Rauluseviciute, I., Drabløs, F. and Rye, M.B. (2019) 'DNA methylation data by sequencing: experimental approaches and recommendations for tools and pipelines for data analysis', *Clinical Epigenetics*, 11(1), p. 193. Available at : https://doi.org/10.1186/s13148-019-0795-x.

Rinn, J.L. *et al.* (2007) 'Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Non-Coding RNAs', *Cell*, 129(7), p. 1311. Available at : https://doi.org/10.1016/J.CELL.2007.05.022.

Robinson, D. *et al.* (2015) 'Integrative Clinical Genomics of Advanced Prostate Cancer', *Cell*, 161(5), pp. 1215–1228. Available at : https://doi.org/10.1016/J.CELL.2015.05.001.

Sanghi, A. *et al.* (2021) 'Chromatin accessibility associates with protein-RNA correlation in human cancer', *Nature Communications*, 12(1), p. 5732. Available at : https://doi.org/10.1038/s41467-021-25872-1.

Sasidharan Nair, V. *et al.* (2018) 'DNA methylation and repressive H3K9 and H3K27 trimethylation in the promoter regions of PD-1, CTLA-4, TIM-3, LAG-3, TIGIT, and PD-L1 genes in human primary breast cancer', *Clinical Epigenetics*, 10(1), p. 78. Available at : https://doi.org/10.1186/s13148-018-0512-1.

Shi, H. *et al.* (2021) 'Bias in RNA-seq Library Preparation: Current Challenges and Solutions', *BioMed Research International*, 2021, pp. 1–11. Available at: https://doi.org/10.1155/2021/6647597.

Srikantan, V. *et al.* (2000) 'PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer', *Proceedings of the National Academy of Sciences of the United States of America*, 97(22), p. 12216. Available at: https://doi.org/10.1073/PNAS.97.22.12216.

Stockslager, J.L., Cheli, R. and Haworth, K. (eds) (2002) *Lippincott Professional Guides : Anatomy & Physiology*. 2nd edn. Philadelphia: Wolters Kluwer Health.

Teves, S.S., Weber, C.M. and Henikoff, S. (2014) 'Transcribing through the nucleosome', *Trends in Biochemical Sciences*, 39(12), pp. 577–586. Available at : https://doi.org/10.1016/j.tibs.2014.10.004.

Tomlins, S.A. *et al.* (2005) 'Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer', *Science*, 310(5748), pp. 644–648. Available at : https://doi.org/10.1126/science.1117679.

Tonmoy, M.I.Q. *et al.* (2022) 'Computational epigenetic landscape analysis reveals association of CACNA1G-AS1, F11-AS1, NNT-AS1, and MSC-AS1 lncRNAs in prostate cancer progression through aberrant methylation', *Scientific Reports*, 12(1), p. 10260. Available at: https://doi.org/10.1038/s41598-022-13381-0.

Touat-Todeschini, L., Hiriart, E. and Verdel, A. (2012) 'Nucleosome positioning and transcription: fission yeast CHD remodellers make their move', *The EMBO Journal*, 31(23), pp. 4371–4372. Available at: https://doi.org/10.1038/emboj.2012.284.

Urbanucci, A. (2012) *Overexpression of Androgen Receptor in Prostate Cancer*. Tampere University.

Uusi-Mäkelä, J. *et al.* (2020) 'Chromatin accessibility analysis uncovers regulatory element landscape in prostate cancer progression', *bioRxiv*, p. 2020.09.08.287268. Available at: https://doi.org/10.1101/2020.09.08.287268.

Wang, K.C. and Chang, H.Y. (2011) 'Molecular Mechanisms of Long Noncoding RNAs', *Molecular Cell*, 43(6), pp. 904–914. Available at: https://doi.org/10.1016/j.molcel.2011.08.018.

Watson, P.A., Arora, V.K. and Sawyers, C.L. (2015) ‘Emerging mechanisms of resistance to androgen receptor inhibitors in prostate cancer’, *Nature Reviews Cancer 2015 15:12*, 15(12), pp. 701–711. Available at : https://doi.org/10.1038/nrc4016.

Weber, M. *et al.* (2005) ‘Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells’, *Nature Genetics*, 37(8), pp. 853–862. Available at : https://doi.org/10.1038/ng1598.

Woodcock, D.J. *et al.* (2020) ‘Prostate cancer evolution from multilineage primary to single lineage metastases with implications for liquid biopsy’, *Nature Communications 2020 11:1*, 11(1), pp. 1–13. Available at : https://doi.org/10.1038/s41467-020-18843-5.

Yan, F. *et al.* (2020) ‘From reads to insight: a hitchhiker’s guide to ATAC-seq data analysis’, *Genome Biology*, 21(1), p. 22. Available at: https://doi.org/10.1186/s13059-020-1929-3.

Yella, V.R., Kumar, A. and Bansal, M. (2018) ‘Identification of putative promoters in 48 eukaryotic genomes on the basis of DNA free energy’, *Scientific Reports 2018 8:1*, 8(1), pp. 1–13. Available at : https://doi.org/10.1038/s41598-018-22129-8.

Yin, Y. *et al.* (2017) ‘Impact of cytosine methylation on DNA binding specificities of human transcription factors’, *Science*, 356(6337). Available at : https://doi.org/10.1126/science.aaj2239.

Ylipää, A. *et al.* (2015) ‘Transcriptome sequencing reveals PCAT5 as a Novel ERG-Regulated long Noncoding RNA in prostate cancer’, *Cancer Research*, 75(19), pp. 4026–4031. Available at: https://doi.org/10.1158/0008-5472.CAN-15-0217/651867/AM/TRANSCRIPTOME-SEQUENCING-REVEALS-PCAT5-AS-A-NOVEL.

Zhang, M. *et al.* (2022) ‘Critical assessment of computational tools for prokaryotic and eukaryotic promoter prediction’, *Briefings in Bioinformatics*, 23(2). Available at : https://doi.org/10.1093/bib/bbab551.

APPENDIX A: GTRD EXPERIMENTS USED IN HISTONE MARK ANALYSIS

Experiment	Cell type	Treatment	Used for	Target	Associated with
HEXP001174	prostate gland	None	BPH	H3K4me3	transcriptionally active gene promoter regions
HEXP001249	prostate gland	None	BPH	H3K27ac	active gene promoters and enhancer regions
HEXP001574	prostate gland	None	BPH	H3K36me3	actively transcribing genes, gene body
HEXP002182	prostate gland	None	BPH	H3K4me1	active cis-regulatory enhancer elements
HEXP001184	22RV1 (prostate carcinoma) cell line	None	PC, CRPC	H3K27ac	active gene promoters and enhancer regions
HEXP002928	LNCaP C4-2B (prostate carcinoma) cell line	None	PC, CRPC	H3K27ac	active gene promoters and enhancer regions
HEXP014448	LNCaP (prostate carcinoma) cell line	None	PC, CRPC	H3K4me3	transcriptionally active gene promoter regions
HEXP014449	LNCaP (prostate carcinoma) cell line	None	PC, CRPC	H3K36me3	actively transcribing genes, gene body

HEXP014454	LNCaP (prostate carcinoma) cell line	None	PC, CRPC	H3K4me1	active cis-regulatory enhancer elements
HEXP014455	LNCaP (prostate carcinoma) cell line	None	PC, CRPC	H3K4me2	transcriptionally active genes; genes primed for future expression
HEXP001980	VCaP (prostate carcinoma) cell line	None	PC, CRPC	H3K27ac	active gene promoters and enhancer regions

Table 8. *GTRD experiments used in analysis of histone marks falling within promoters of the studied transcripts. Column “Used for” shows which sample group was studied using given experiment. Column “Associated with” explains the functional association of the enrichment of a given histone mark.*

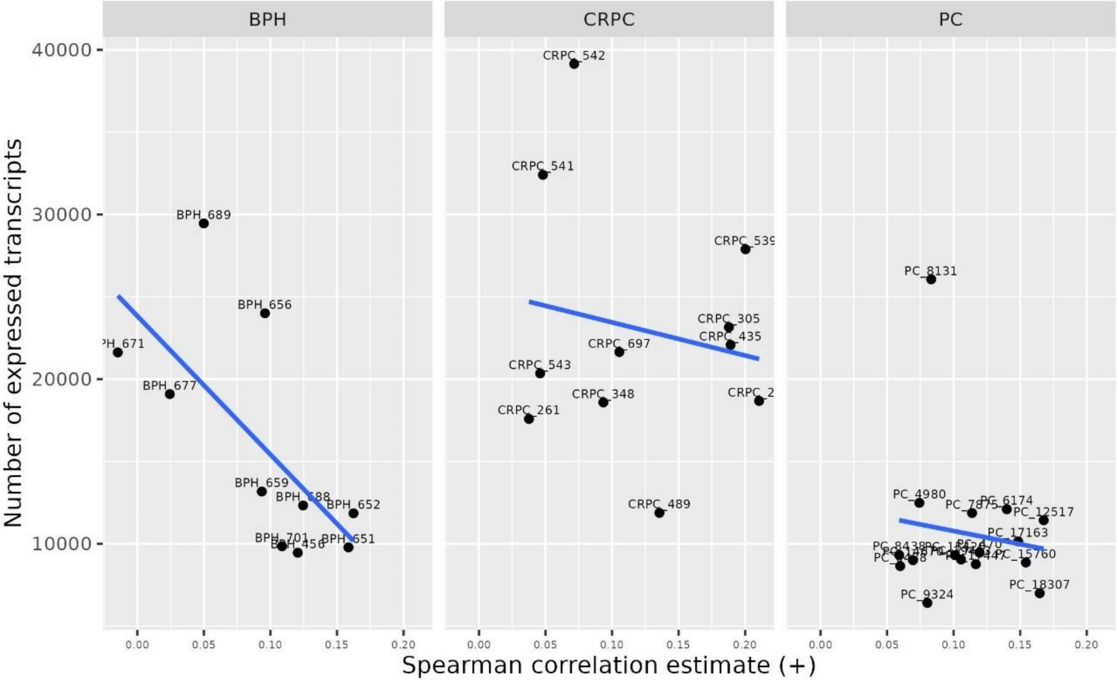
APPENDIX B: THE FRACTIONS OF “ON” AND “OFF” TRANSCRIPTS IN EACH TRANSCRIPT GROUP

Condition	Transcript type	% of all transcripts constituted by mean number of ON transcripts	% of all transcripts constituted by mean number of OFF transcripts
BPH	Intergenic (+)	2.75	97.25
	Intergenic (-)	2.82	97.18
	Intragenic (+)	2.19	97.81
	Intragenic (-)	2.31	97.69
	Coding	21.53	78.47
	LNC annotated	13.04	86.96
PC	Intergenic (+)	3.16	96.84
	Intergenic (-)	3.20	96.80
	Intragenic (+)	2.48	97.52
	Intragenic (-)	2.58	97.42
	Coding	21.78	78.22
	LNC annotated	13.28	86.72
CRPC	Intergenic (+)	3.19	96.81
	Intergenic (-)	3.25	96.75
	Intragenic (+)	2.46	97.54
	Intragenic (-)	2.57	97.43
	Coding	21.37	78.63
	LNC annotated	13.15	86.85

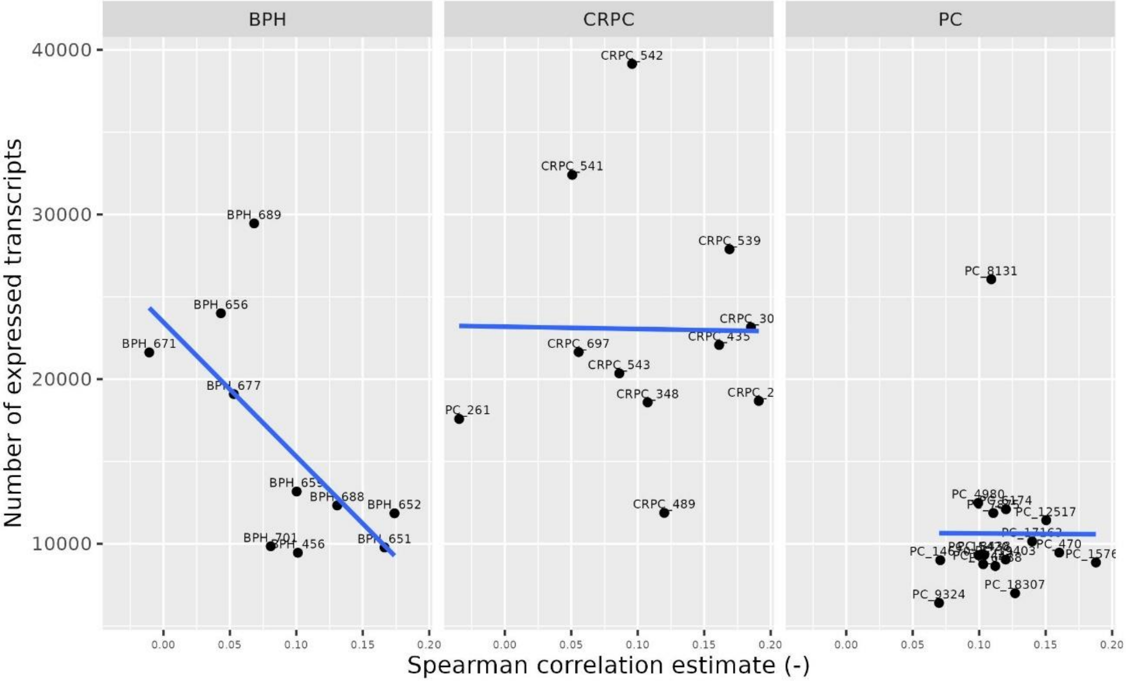
Table 9. *The percentages of ON and OFF genes of all transcripts in each transcript group, rounded to two decimal places. Plus and minus signs in the parentheses represent the sense assumption for unannotated transcripts.*

APPENDIX C: THE RELATIONSHIP BETWEEN THE NUMBER OF EXPRESSED TRANSCRIPTS AND THE STRENGTH OF CORRELATION BETWEEN CHROMATIN ACCESSIBILITY AND EXPRESSION

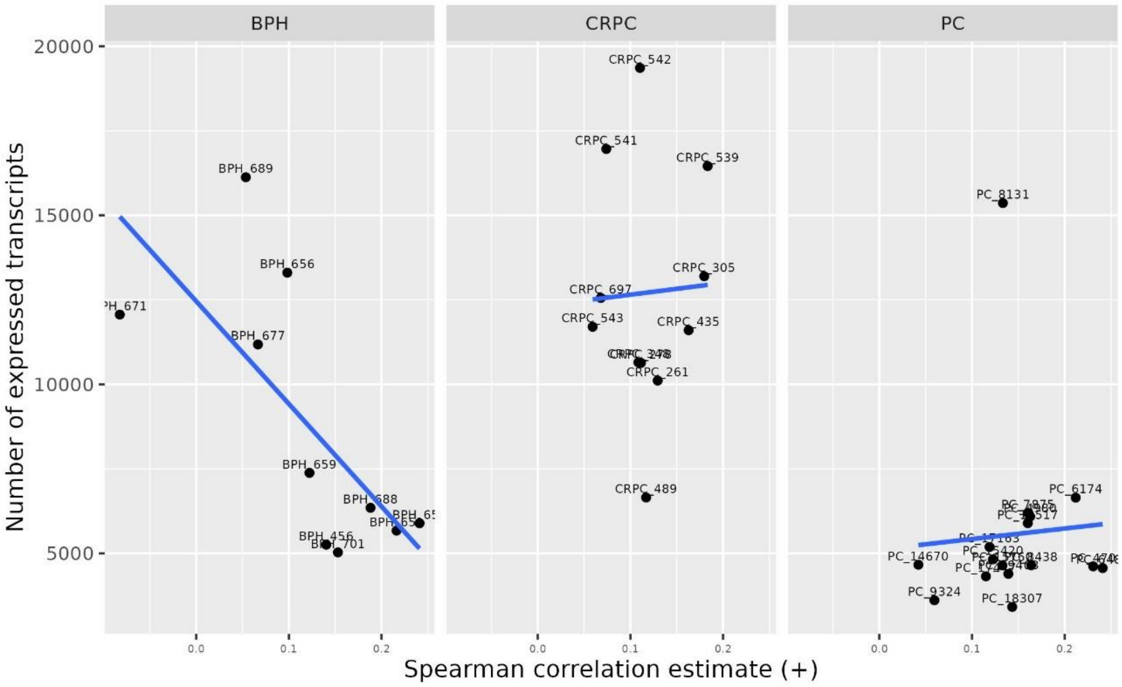
Intergenic transcripts, plus-sense assumption:



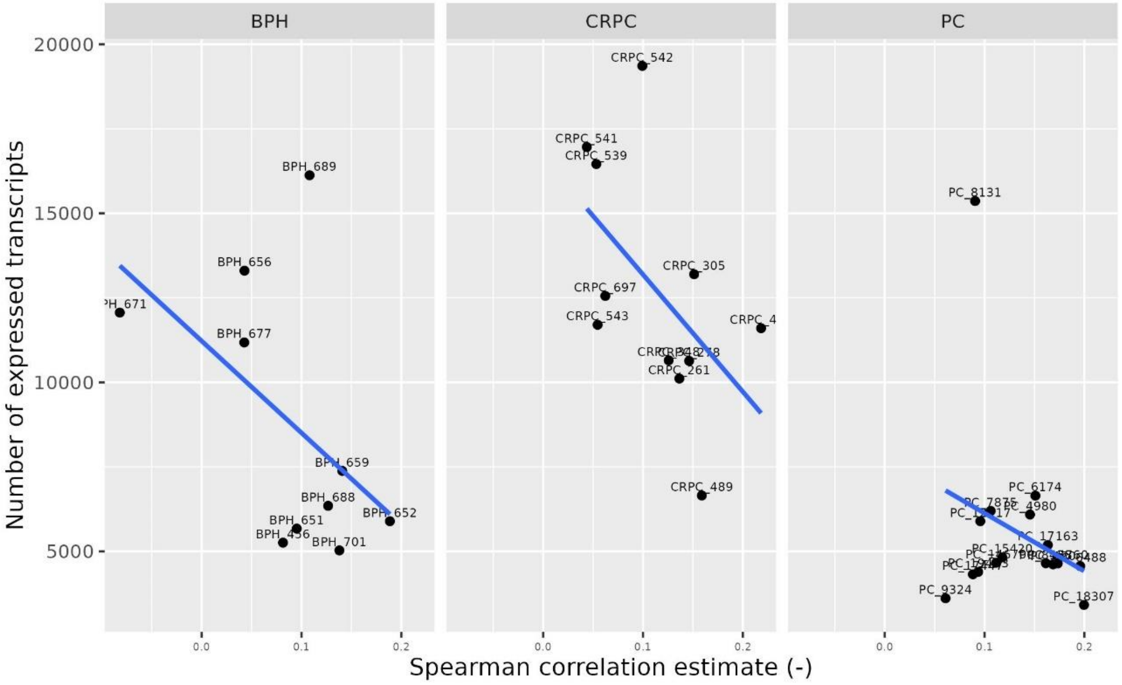
Intergenic transcripts, minus-sense assumption:



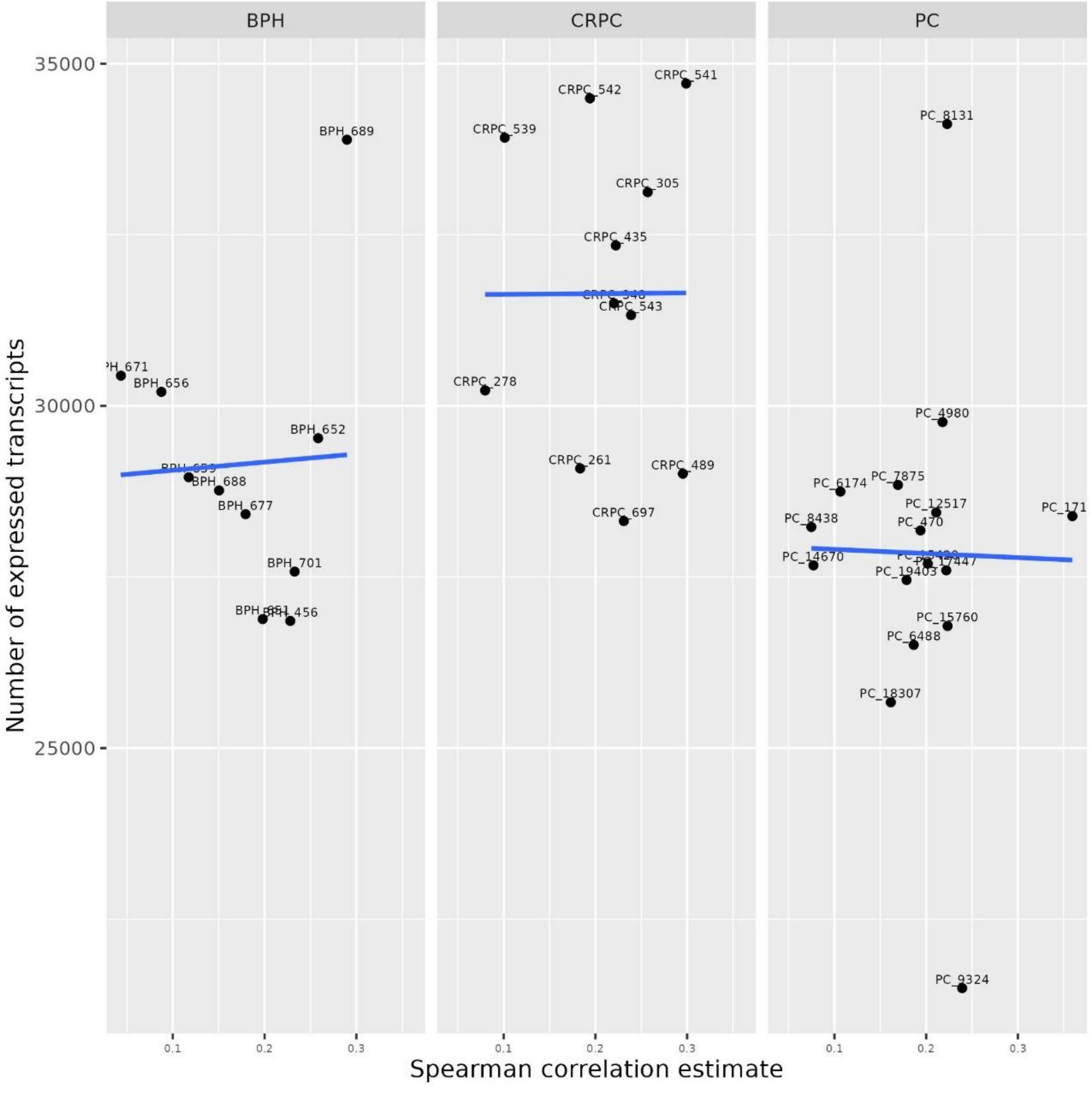
Intragenic transcripts, plus-sense assumption:



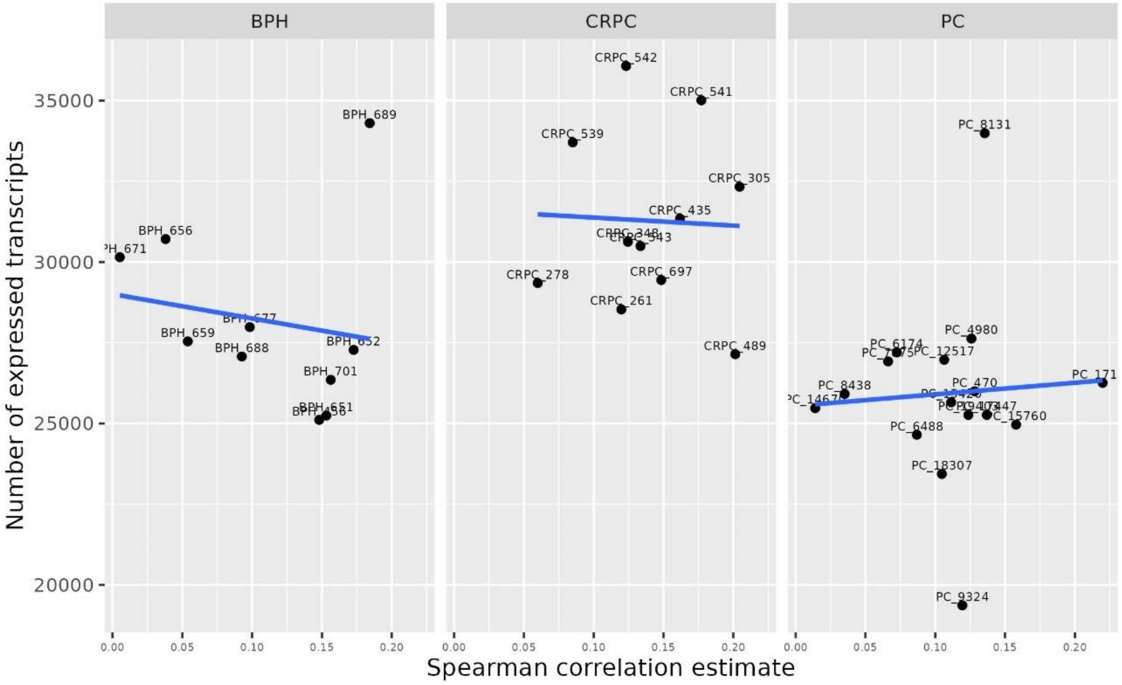
Intragenic transcripts, minus-sense assumption:



Protein-coding transcripts:

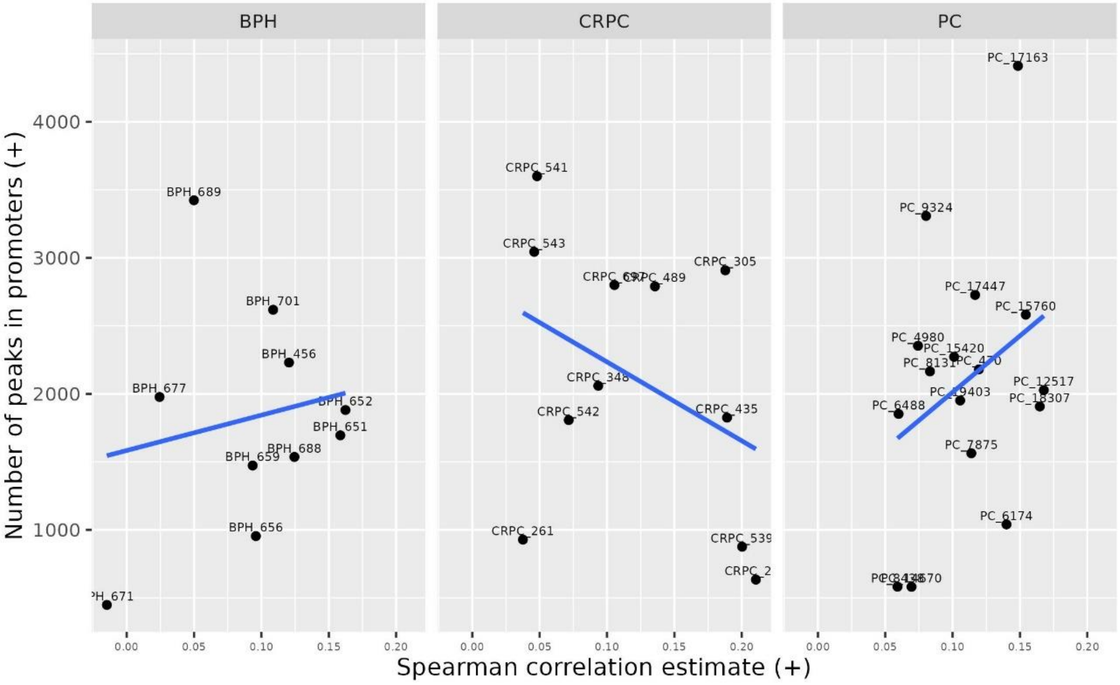


LNC annotated transcripts:

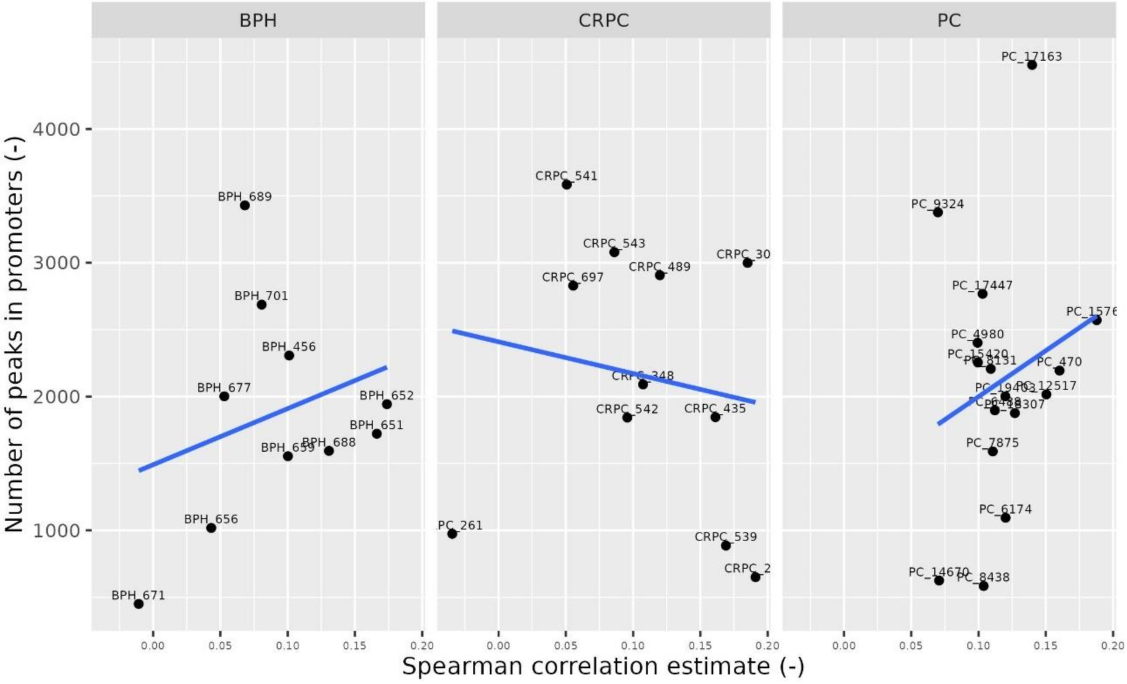


APPENDIX D: THE RELATIONSHIP BETWEEN THE NUMBER OF ATAC-SEQ PEAKS IN PROMOTERS AND THE STRENGTH OF CORRELATION BETWEEN EXPRESSION AND CHROMATIN ACCESSIBILITY

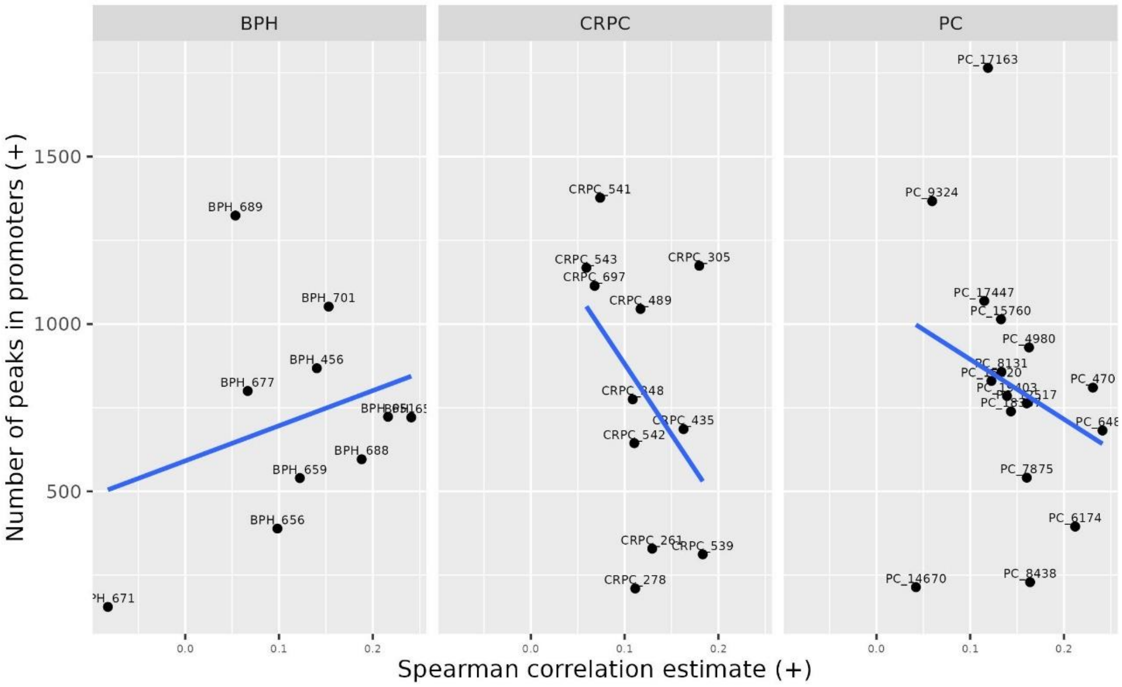
Intergenic transcripts, plus-sense assumption:



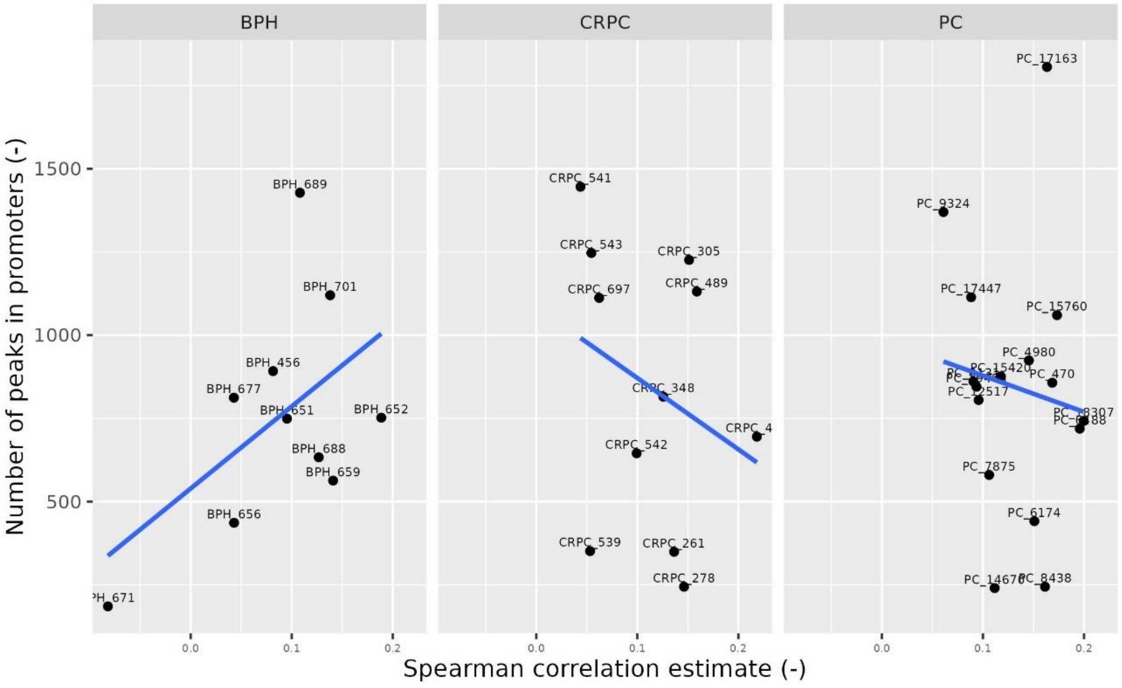
Intergenic transcripts, minus-sense assumption:



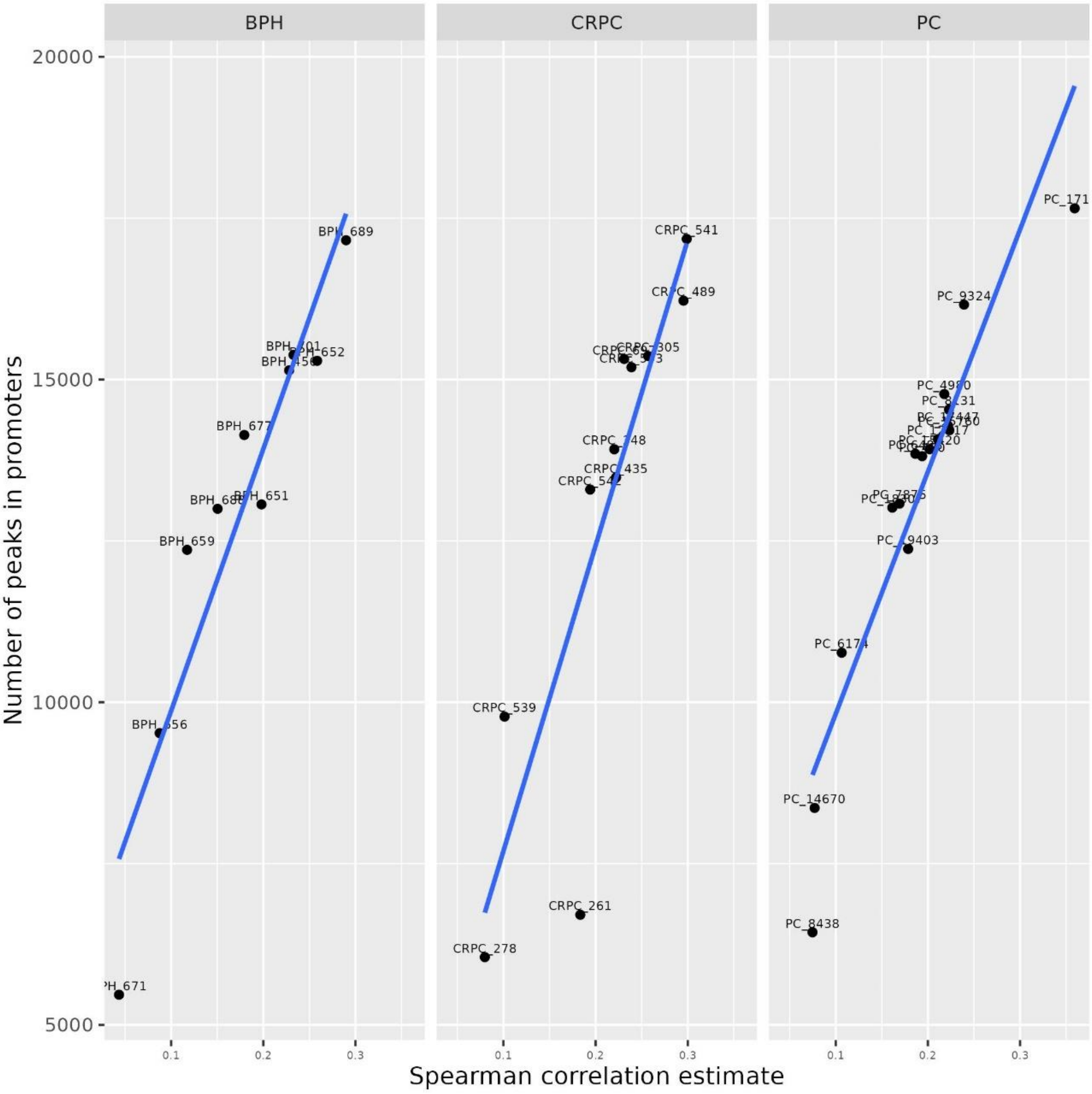
Intragenic transcripts, plus-sense assumption:



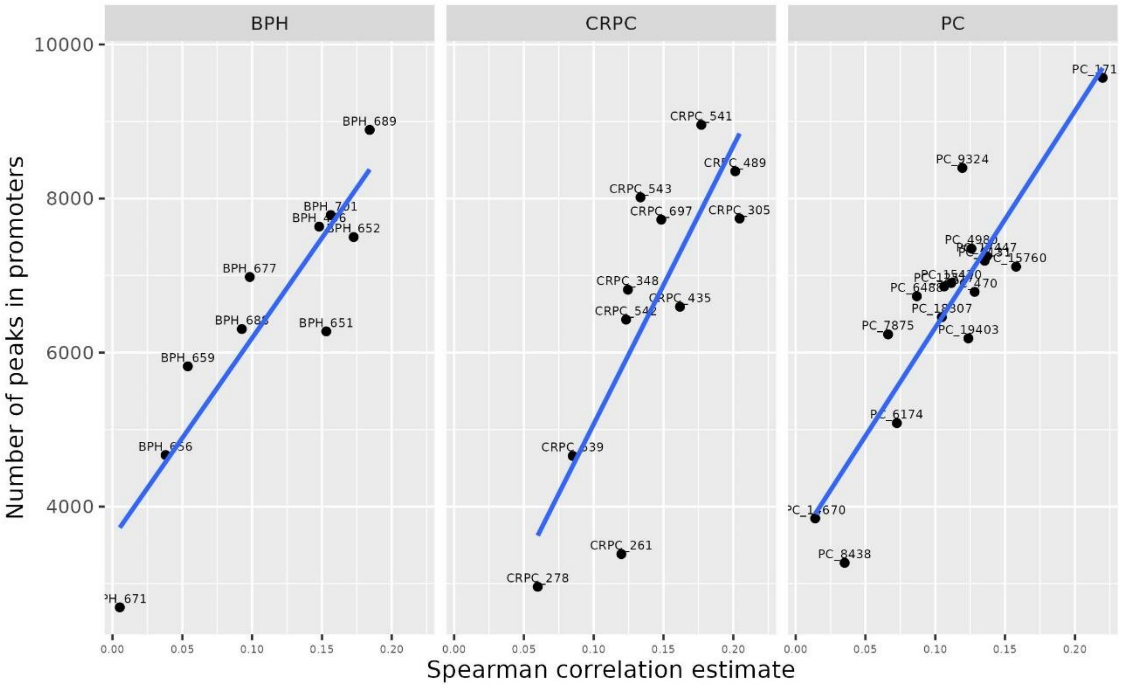
Intragenic transcripts, minus-sense assumption:



Protein-coding transcripts:



LNC annotated transcripts:



APPENDIX E: THE MOST EXPRESSED UNANNOTATED TRANSCRIPTS WITH ACCESSIBLE PROMOTERS AND LOW METHYLATION

Transcript group	Sample	Transcript	Expression (TPM)	MeDIP
Intergenic (+)	PC 15760	PCAT-4-155444754	32.29	NA
	PC 470	PCAT-20-45290869	40.95	0.22
Intergenic (-)	PC 15760	PCAT-4-155444754	32.29	NA
	PC 9324	PCAT-8-53713833	31.07	NA
Intragenic (+)	BPH 456, BPH 652, BPH 689, BPH 701, CRPC 435, CRPC 489, CRPC 541, PC 12517, PC 15760, PC 17163, PC 19403, PC 4980, PC 6488	PCAT-19-3062174	100.39, 138.40, 60.41, 154.82, 69.88, 96.01, 37.56, 104.56, 100.2, 158.74, 135.49, 174.94, 120.69	0.10, 0.10, 0.11, 0.05, 0.06, 0.14, 0.13, 0.11, 0.08, 0.10, 0.12, 0.11, 0.10
Intragenic (-)	BPH 652, BPH 689, CRPC 348, CRPC 435, CRPC 489, CRPC 541, PC 17163, PC 470, PC 4980, PC 6174, PC 8131	PCAT-19-3062174	138.4, 60.41, 39.09, 69.90, 96.01, 37.56, 158.74, 184.88, 174.94, 180.20, 56.26	0.42, 0.16, 0.24, 0.20, 0.26, 0.10, 0.26, 0.17, 0.13, 0.13, 0.17
	PC 4980	PCAT-15-71411687	38.06	0.24

Table 10. *Highly expressed ON unannotated transcripts with low methylation level, whose TPM was higher than 30.00.*

APPENDIX F: THE MOST EXPRESSED UNANNOTATED TRANSCRIPTS WITHOUT ATAC-SEQ PEAKS AND WITH HIGH METHYLATION IN PROMOTERS

Transcript group	Sample	Transcript name	TPM	MeDIP
Intergenic (+)	BPH 456, CRPC 543, PC 9324	PCAT-14-55229191	59.67, 34.26, 67.47	0.84, 0.82, 0.83
	BPH 656, CRPC 541, CRPC 697, PC 19403, PC 8131, PC 9324	PCAT-12-66057601	2448.46, 857.07, 1259.15, 68.11, 1192.40, 34980.81	0.85, 0.83, 0.85, 0.89, 0.83, 0.86
	BPH 701	PCAT-2-238910283	54.85	0.90
	CRPC 435	PCAT-4-49149406	33.36	0.98
	CRPC 543	PCAT-4-28217189	69.82	0.92
	PC 6488	PCAT-17-18689998	53.70	0.91
	PC 8131	PCAT-14-19452114	31.32	0.96
Intergenic (-)	BPH 456, BPH 651, BPH 652, BPH 659, BPH 688, BPH 701, CRPC 278, CRPC 435, CRPC 489, CRPC 541, CRPC 543, PC 12517, PC 14670, PC 15420, PC 15760, PC 17163, PC 17447, PC 18307,	PCAT-14-55229191	59.67, 34.47, 34.54, 56.11, 48.86, 47.81, 37.54, 44.31, 50.13, 31.76, 34.3, 30.83, 43.69, 69.65, 68.05, 53.21, 86.56, 44.38,	0.97, 1.00, 0.97, 1.00, 0.97, 0.94, 0.94, 0.97, 0.96, 0.95, 0.96, 0.97, 0.92, 0.97, 0.97, 0.96, 0.97, 0.96,

	PC 19403, PC 470, PC 4980, PC 6174, PC 6488, PC 7875, PC 8438, PC 9324		40.22, 58.91, 58.12, 67.31, 47.26, 67.07, 66.26, 67.47	1.00, 0.97, 1.00, 1.00, 0.97, 1.00, 0.97, 0.96
	CRPC 697	PCAT-5-29645537	33.64	0.95
	CRPC 697	PCAT-12-61471686	38.23	0.93
Intragenic (+)	BPH 671	PCAT-14-19373876	33.49	0.89
	PC 15760	PCAT-18-68277796	33.45	0.91
	PC 19403	PCAT-14-20700793	42.42	0.83
	PC 470	PCAT-16-80813353	33.56	0.93
Intragenic (-)	BPH 689, BPH 701, CRPC 489, PC 15760, PC 18307, PC 470, PC 6174, PC 7875, PC 8131	PCAT-X-133596440	92.34, 62.09, 42.27, 94.21, 49.36, 60.05, 52.57, 75.72, 38.91	0.86, 0.87, 0.85, 0.84, 0.87, 0.88, 0.84, 0.87, 0.85
	CRPC 261	PCAT-X-10730367	31.74	0.93
	PC 19403	PCAT-14-20700793	42.42	0.83

able 11. *Unannotated highly methylated OFF transcripts with particularly high expression level (TPM > 30.00).*

APPENDIX G: THE MOST EXPRESSED UNANNOTATED TRANSCRIPTS WITHOUT ATAC-SEQ PEAKS AND LOW METHYLATION IN PROMOTERS

Transcript group	Transcript name	Samples in which transcript passed filtering	Mean TPM	Samples in which TPM > 30.00	Max TPM (Sample)
Intergenic (+)	PCAT-12-66057601	CRPC: 1	623.08	CRPC: 1	899.61 (CRPC 489)
		PC: 1		PC: 1	
	PCAT-13-76592364	BPH: 3	28.37	BPH: 1	96.45 (PC 19403)
		CRPC: 7		-	
		PC: 13		PC: 7	
	PCAT-17-18689998	BPH: 7	30.40	BPH: 1	288.13 (PC 9324)
		CRPC: 8		CRPC: 1	
		PC: 12		PC: 4	
	PCAT-17-22521366	BPH: 10	141.45	BPH: 6	911.99 (BPH 688)
		CRPC: 8		CRPC: 7	
		PC: 15		PC: 9	
	PCAT-18-54406782	BPH: 6	42.85	BPH: 2	125.47 (PC 15760)
		CRPC: 5		CRPC: 2	
		PC: 7		PC: 7	
	PCAT-4-132528088	All samples	38.09	BPH: 4	144.89 (PC 9324)
				CRPC: 1	
				PC: 13	
	PCAT-9-40090155	All samples	24.46	BPH: 1	52.42 (PC 19403)
				CRPC: 1	
				PC: 7	
Intergenic (-)	PCAT-12-66057601	BPH: 3	5351.79	BPH: 3	34980.81 (PC 9324)
		CRPC: 2		CRPC: 2	
		PC: 2		PC: 2	
	PCAT-13-76592364	All samples	25.13	BPH: 1	96.45 (PC 19403)
				PC: 8	
	PCAT-17-22521366	BPH: 10	141.45	BPH: 6	911.99 (BPH 688)
		CRPC: 8		CRPC: 7	
		PC: 15		PC: 9	
	PCAT-18-54406782	BPH: 2	35.88	BPH: 1	83.39 (PC 15420)
		CRPC: 4		CRPC: 1	
		PC: 6		PC: 6	

	PCAT-2-238910283	BPH: 4 CRPC: 3 PC: 9	48.42	BPH: 4 CRPC: 1 PC: 8	145.71 (BPH 656)
	PCAT-4-132528088	All samples	38.09	BPH: 4 CRPC: 1 PC: 13	144.89 (PC 9324)
	PCAT-9-40090155	All samples	24.46	BPH: 1 CRPC: 1 PC: 7	52.42 (PC 19403)
Intragenic (+)	PCAT-1-149190384	BPH: 8 CRPC: 7 PC: 16	18.47	BPH: 4 CRPC: 2 PC: 2	69.71 (BPH 652)
	PCAT-14-82692663	BPH: 9 CRPC: 8 PC: 15	18.79	- - PC: 5	44.93 (PC 470)
	PCAT-16-35802141	All samples	419.42	All samples	901.94 (PC 17447)
	PCAT-17-53105729	All samples	27.81	BPH: 3 CRPC: 6 PC: 5	116.92 (CRPC 543)
	PCAT-19-3062174	BPH: 6 CRPC: 5 PC: 9	101.426	BPH: 6 CRPC: 4 PC: 9	213.14 (BPH 688)
	PCAT-5-70438929	BPH: 10 CRPC: 9 PC: 15	19.44	BPH: 4 CRPC: 4 -	128.19 (BPH 677)
	PCAT-5-71043785	BPH: 10 CRPC: 9 PC: 14	13.13	BPH: 3 CRPC: 3 -	52.65 (BPH 677)
	PCAT-X-133596440	All samples	62.71	BPH: 8 CRPC: 4 PC: 14	305.88 (BPH 688)
Intragenic (-)	PCAT-1-149190384	BPH: 8 CRPC: 7 PC: 16	18.47	BPH: 4 CRPC: 2 PC: 2	69.71 (BPH 652)
	PCAT-16-35802141	BPH: 1 CRPC: 4 PC: 1	311.16	BPH: 1 CRPC: 4 PC: 1	521.09 (BPH 659)
			27.81	BPH: 3	

	PCAT-17-53105729	All samples		CRPC: 5	116.92 (CRPC 543)
				PC: 5	
	PCAT-19-3062174	BPH: 8	103.83	BPH: 8	213.14 (BPH 688)
		CRPC: 4		CRPC: 3	
		PC: 10		PC: 10	
	PCAT-5-70438929	BPH: 10	19.44	BPH: 4	128.19 (BPH 677)
		CRPC: 9		CRPC: 4	
		PC: 15		-	
	PCAT-5-71043785	BPH: 10	13.13	BPH: 3	52.65 (BPH 677)
		CRPC: 9		CRPC: 3	
		PC: 14		-	

Table 12. *Unannotated OFF transcripts with high expression and low methylation, which were overexpressed (TPM > 30.00) in more than five samples or whose expression level was exceptionally high (TPM > 100.00). Column “Mean TPM” contains the mean TPM computed from values in all samples, in which given transcript passed the filtering. Column “Max TPM (Sample)” shows the largest expression of a given transcript among the samples, in which the transcript passed the filtering, and the identifier of the sample in which the expression reached this maximum value.*