Tampere University

EBRAHIM AFYOUNIAN

Computational Analysis of Multilevel High-throughput Data from Cancer Tissue

Tampere University Dissertations 931

Tampere University Dissertations 931

EBRAHIM AFYOUNIAN

Computational Analysis of Multilevel High-throughput Data from Cancer Tissue

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Medicine and Health Technology
of Tampere University,
for public discussion in the auditorium F114
of the Arvo Building, Arvo Ylpön katu 34, Tampere,
on 12 January 2024, at 12 o’clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Medicine and Health Technology
Finland

Responsible supervisor and Custos	Professor Matti Nykter Tampere University Finland	
Pre-examiners	Professor Gong-Hong Wei Fudan University China University of Oulu Finland	Associate Professor Vittorio Fortino University of Eastern Finland Finland
Opponent	Dr. Thomas Fleischer Oslo University Hospital Norway	

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Cover design: Roihu Inc.

ISBN 978-952-03-3223-5 (print)
ISBN 978-952-03-3224-2 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-3224-2

HIILINEUTRAALI PAINOTUOTE
ClimateCalc CC-000025/FI
PunaMusta Printing

Carbon dioxide emissions from printing Tampere University dissertations have been compensated.

PunaMusta Oy – Yliopistopaino
Joensuu 2023

ACKNOWLEDGMENT

This doctoral study was conducted at the Computational Biology group, Faculty of Medicine and Health Technology (MET), Tampere University. I would like to express my sincere gratitude to my supervisor, Professor Matti Nykter, for his unwavering trust in my capacity to undertake doctoral-level research. Professor Matti Nykter not only provided me with the opportunity to pursue this study but also continuously encouraged and supported me, even when I lacked confidence in my own abilities. The guidance, patience, and composure demonstrated by Professor Matti Nykter were invaluable in helping me navigate through the challenges and ultimately complete this work. Undertaking this doctoral study has undoubtedly been a formative and transformative experience for me, and I am truly grateful for Professor Matti Nykter's contributions to its fulfillment.

I extend my heartfelt gratitude to all my coauthors, who have been invaluable sources of inspiration and knowledge throughout our collaboration. Working alongside each of you has been an enriching experience, and I have learned immensely from our collective efforts. I hold each and every one of you in high regard, considering you as exemplary role models in the pursuit of scientific inquiry.

I would like to thank the members of the thesis committee, Dr. Reija Autio and Dr. Olli Lohi for their invaluable guidance and insightful feedback throughout the development of this work. Their support and supervision ensured that the project remained on track and progressed at an appropriate pace.

I am also deeply grateful to Professor Gong-Hong Wei and Professor Vittorio Fortino for their pre-examination of this work and their valuable feedback. Their expertise and constructive feedback have been instrumental in refining and strengthening the quality of this work.

I am sincerely grateful to both the MET Graduate school at Tampere University and the Computational Biology group for providing funding for this study.

I am immensely grateful to Finland for its warm embrace, remarkable hospitality, and nurturing environment. Since arriving in Finland, my sole focus has been on honing my skills as a scientific researcher and educator, thanks to the incredible support and care provided by the Finnish community. I am indebted to the wonderful people of Finland for creating an environment that has facilitated my personal and professional growth.

I would also like to extend my heartfelt gratitude to every colleague, past and present, at the Computational Biology group, as well as the members of the neighboring groups at the Kauppi Campus, Tampere University. The privilege of interacting with each of you has undeniably enriched my academic journey. Your collective expertise, diverse perspectives, and unwavering support have played an integral role in shaping my growth as a researcher. I am truly grateful for the enriching experiences and valuable interactions we have shared, which have contributed significantly to my personal and professional development.

In the end, I would like to gratefully express my deepest gratitude to every individual in my extended family. Words cannot adequately convey the immense appreciation I have for the unwavering support, patience, and love that I have received from all of you. I am particularly grateful to Tiina, my spouse, and Otso Kaave, my son who are my raisons d’être.

Helsinki, August 2023

Ebrahim Afyounian

ABSTRACT

The emergence of high-throughput measurement technologies has greatly expanded the possibilities for detecting and quantifying biomolecules involved in various subcellular and biomolecular processes on a larger scale than previously possible. These measurements, along with their accurate and robust analysis, play a crucial role in deepening our understanding and explaining a wide range of biological phenomena, including the development and progression of cancers. Consequently, this knowledge can be harnessed to develop effective interventions, particularly in the management and treatment of cancer.

Within this dissertation, we have pursued two primary aims. Firstly, we aimed to develop novel computational and statistical tools and methods for the effective and efficient analysis of high-throughput data within the context of cancer. Secondly, using the tools and methods we developed, we sought to investigate single and multilevel high-throughput data to identify key alterations that drive the development and progression of prostate cancer.

To accomplish the first aim, we devised a computational tool capable of detecting somatic copy number alterations. Additionally, we developed other computational and statistical approaches to mitigate the inherent biases present in data obtained from high-throughput sequencing technologies. As for the second aim, using the methods and tools we developed, we analyzed single and multilevel high-throughput data from a cohort of prostate cancer patients at various stages of their disease, identified multiple alterations, and presented our observations in detail.

In summary, our study demonstrates the potential of the analysis of single and multilevel high-throughput data. Through this approach, we were able to replicate previous findings and uncover alterations that impact biological processes at different levels during the development and progression of prostate cancer.

CONTENTS

List of figures

List of tables

ABBREVIATIONS

aCGH	Array-based comparative genomic hybridization
ADT	Androgen deprivation therapy
AR	Androgen receptor
ATAC-seq	Assay for transposase-accessible chromatin using sequencing
BAF	B-allele fraction
bp	Base pair
BPH	Benign prostatic hyperplasia
BWT	Burrows-Wheeler transform
cDNA	Complementary DNA
CGH	Comparative genomic hybridization
cnLOH	Copy-neutral loss of heterozygosity
CRPC	Castration-resistant prostate cancer
DAR	Differentially accessible region
ddNTP	Dideoxynucleotides triphosphate
DE	Differentially expressed
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
DNA-seq	DNA sequencing
DNase-seq	Deoxyribonuclease I hypersensitive sites sequencing
dNTP	Deoxynucleoside triphosphate
dsDNA	Double-stranded DNA
ENCODE	Encyclopedia of DNA elements
FDR	False discovery rate
FISH	Fluorescence in situ hybridization
GTRD	Gene transcription regulation database
HGP	Human Genome Project
HTS	High-throughput sequencing
Indel	Insertion/deletion
JSI	Jaccard similarity index
LC	Liquid chromatography

LC-MS/MS	Liquid chromatography coupled to tandem mass spectrometry
MACS	Model-based analysis of ChIP-Seq
MAPQ	Mapping quality
MeDIP	Methylated DNA immunoprecipitation
miRNA	MicroRNA
mRNA	Messenger RNA
MS	Mass spectrometry
NGS	Next-generation sequencing
nm	Nanometer
PC	Prostate cancer
PCA	Principal component analysis
PCR	Polymerase chain reaction
PSA	Prostate-specific antigen
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RP	Radical prostatectomy
RRBS	Reduced representation bisulfite sequencing
rRNA	Ribosomal RNA
SBS	Sequencing by synthesis
SCNA	Somatic copy number alteration
SNP	Single-nucleotide polymorphism
SNV	Single-nucleotide variant
ssDNA	Single-stranded DNA
SWATH-MS	Sequential windowed acquisition of all theoretical fragmentation - mass spectrometry
TAD	Topologically associating domain
TCGA	The cancer genome atlas
TF	Transcription factor
tRNA	Transfer RNA
t-SNE	t-distributed stochastic neighbor embedding
TSS	Transcription start site
TURP	Transurethral resection of the prostate

ORIGINAL PUBLICATIONS

Afyounian E., Annala M., Nykter M. Segmentum: a tool for copy number analysis of cancer genomes. BMC Bioinformatics. 2017 Apr 13;18(1):215.

Latonen L., Afyounian E., Jylhä A., Nättinen J., Aapola U., Annala M., Kivinummi K., Tammela T.L., Beuerman R., Uusitalo H., Nykter M., Visakorpi T. Integrative proteomics in prostate cancer uncovers robustness against genomic and transcriptomic aberrations during disease progression. Nat Commun. 2018 Mar 21;9(1):1176.

Uusi-Mäkelä J.*, Afyounian E.*, Tabaro F.*, Häkkinen T.*, Lussana A., Shcherban A., Annala M., Nurminen R., Kivinummi K., Tammela T.L., Urbanucci A., Latonen L., Kesseli J., Granberg K. G., Visakorpi T., Nykter M. Chromatin accessibility analysis uncovers regulatory element landscape in prostate cancer progression [**Manuscript**]. Available from: http://dx.doi.org/10.1101/2020.09.08.287268

* Equal contribution

Author's contributions

Publication I. Implemented the somatic copy number alteration detection method, and the sequencing read count simulator using Python 3 programming language; Conducted benchmarking and comparative analyses to evaluate the accuracy and performance of the method as compared to the state-of-the-art methods of the time; Wrote the manuscript together with other authors.

Publication II. All bioinformatics analysis in the project, except for the preprocessing of protein data. Responsibilities included analysis of genomic, epigenomic, and transcriptomic data at the individual level. In addition, development of subsequent integrative analysis steps: uncovering a list of miRNA-mRNA interactions, and comprehensive integration of proteomics data with genomic, epigenomic, and transcriptomic data. Contributed to the creation of visualizations and tables; Wrote the manuscript together with other authors.

Publication III. Analysis of differential chromatin alterations during prostate cancer progression, including: implementation of a novel ATAC-seq signal quantification approach, accounting for global and local background noise, sample collection procedure bias, and sequencing depth; Development and executing a method for detecting differentially accessible regions and differentially methylated regions with false discovery rate estimation; Created a list of reproducible, accessible chromatin loci across samples, accompanied by their quantification and characterization; Integrated the genomics and epigenomics results as well as integration with external datasets including data from multiple relevant publications as well as data from different databases to annotate, assess, and validate the findings; Contributed to the creation of multiple visualizations and tables; Wrote the manuscript together with other authors

1 INTRODUCTION

Cancer is one of the leading causes of death among humans, responsible for approximately one in five deaths [1,2]. Prostate cancer, which is the subject of study in Articles II and III, has been globally ranked as the third most commonly diagnosed cancer in 2020 [2]. The ultimate goal of cancer research is to develop effective strategies for the prevention, diagnosis, and treatment of cancer by identifying and understanding the alterations that contribute to its development and progression. Previous cancer research has identified alterations and abnormalities that enable and contribute to the transformation of normal cells into abnormal, cancerous ones [3,4].

The advent of high-throughput sequencing (HTS) technologies has enabled the detection, quantification, and sequencing of biomolecules such as DNA and RNA, and their interactions, such as DNA-protein interaction, that play a role in various biological processes on a larger scale than was previously possible [5].

It has been shown that by comparing HTS data generated from samples with different conditions (e.g., healthy versus cancer) or different stages of a particular cancer (e.g., early-stage versus advanced prostate cancer), we can potentially identify novel alterations and abnormalities that contribute to cancer development or progression. The contributions from The Cancer Genome Atlas (TCGA) are a case in point where 33 different tumor types, including ten rare cancers, have been studied by collecting seven different data types from 11000 participants [6].

While there have been notable contributions towards the understanding of cancer development and progression (some of which will be reviewed in the *literature review* section), there are still unresolved questions. For example, we would like to understand better why a fraction of prostate cancer patients progress to the advanced stage of the disease despite their initial positive response to therapy. There are also multiple challenges to be overcome with regard to HTS data analysis. Due to the decrease in the cost of HTS data generation, larger HTS datasets are being produced (e.g., the TCGA alone has produced 2.5 petabytes which is 2.5 million gigabytes of data [6]) and thus, a challenge is to develop software tools and data processing pipelines capable of robustly analyzing large datasets using a reasonable amount of time and other resources such as computing memory. Another challenge is to develop approaches and methods for the robust analysis of HTS data at single level (e.g., the genome, epigenome, transcriptome, and proteome) as well as integrating multiple layers of data. On its own, each data level affords uncovering particular alterations, such as finding genomics alterations, or changes in the gene expression. On the other hand, the integration and analysis of multilevel data can provide a more comprehensive view that may be missed when analyzing data from only one level. To illustrate, as demonstrated in Article II, alterations in gene expression at the transcriptome level do not always translate to similar changes at the proteome level.

In this work, we aim to address some of these questions and challenges by developing computational tools and methods for the analysis of cancer-related, high-throughput data either at a single level (Article I) or integrate insights from multiple levels of data to provide a more comprehensive picture (Articles II and III).

2 LITERATURE REVIEW

Cancer and the aberrant processes that contribute to its development and progression can be studied at different subcellular levels. This section introduces the levels and the level-specific biomolecules, and the processes investigated in this work, and describes the high-throughput assays used to measure the level-specific biomolecules. In addition, examples of level-specific alterations pertinent to prostate cancer will be introduced. Figure 1 provides an overview of the levels and concepts, relevant to this work, that will be presented in this section.

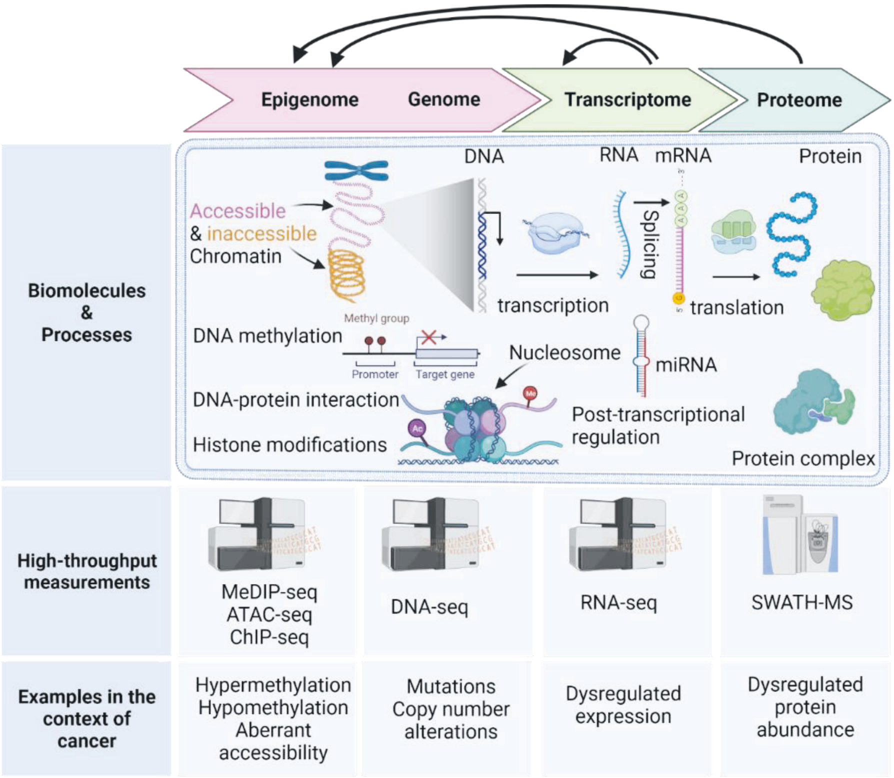


Figure 1. An overview of the subcellular levels, level-specific biomolecules and processes, and high-throughput assays used to measure them, used in this work to study cancer development and progression. Figure created with BioRender.com.

2.1 Human genome

The human genome resides inside the cell nucleus as well as by a small fraction inside the cell mitochondria in a three-dimensional organization. It is composed of a large chain of four different nucleotide biomolecules (approximately 3 x 10^9 nucleotides) packaged in 23 pairs of chromosomes (i.e., a total of 46 chromosomes). Each nucleotide is composed of three subunits: a 2' deoxyribose sugar (2' is read two-prime), a phosphate group, and a nucleobase (aka a nitrogenous base). A phosphate group connects the 5' carbon of one 2' deoxyribose to the 3' carbon of another 2' deoxyribose. Together, they form the basic building block of the sugar backbone of the deoxyribonucleic acid (DNA). There are four different nucleobases, namely adenine (A), guanine (G), cytosine (C), and thymine (T) that are attached to the 2' deoxyribose on the 1' carbon. Adenine and thymine molecules can pair together (i.e., hybridize) while cytosine and guanine can pair with each other. A DNA polynucleotide chain (or a single-stranded DNA aka ssDNA) has two ends: the end that ends with a phosphate group on the 5' carbon in the sugar backbone is called the 5' end, and the other end ending with a hydroxyl group on the 3' carbon in the sugar backbone is called the 3' end.

The human genome DNA sequence, that is the order of the nucleotides, and its composition was first drafted as a result of an international collaboration under the Human Genome Project (HGP) in 2001 [7]. The first draft of the human reference genome covered about 94% of the human genome that is around 3x10^9 nucleotides long [7]. Later efforts by the Genome Reference Consortium (GRC) and recently by the Telomere to Telomere (T2T) consortium and the Human Pangenome Reference Consortium (HPRC), have been focused on determining the remaining 6% which are inherently difficult to determine [8,9].

The human DNA sequence encodes the genetic information in the human genome. In particular, the coding regions on the DNA chain encode information or instructions on how to make different gene products such as proteins [10]. The HGP project estimated that the human genome encodes about 30000-40000 protein-coding and non-coding genes [7]. Non-coding genes are genes that do not result in proteins. MicroRNAs (aka miRNAs) are non-coding RNA with the ability to regulate the activity of other gene products. Furthermore, the DNA sequence of the non-coding regions may contain specific patterns (or motifs) allowing for DNA-protein interactions. These interactions can regulate the expression of the gene products [11].

Cancer hallmarks refer to the acquired functional capabilities that allow tumors to perform functions that contribute to their survival, proliferation, and their dissemination [4]. Genome instability and mutation is considered to be an enabling characteristic for acquiring other cancer hallmarks [3,4]. A DNA mutation is a permanent alteration in the DNA sequence that can result from errors during DNA replication, exposure to mutagens, or other genetic processes. Genome instability refers to a higher than normal rate of genomic alterations [12]. An example of genome instability is copy number alteration [4]. The human reference genome, which reveals the normal sequence of the human genome, is an invaluable resource for detecting such genomic aberrations. As it will be described later, the high-throughput DNA sequence data from cancer samples together with the human reference genome can be employed to detect genomic regions that have undergone copy number alterations such as duplication, amplification, or deletion. As an illustration, in around half of the localized prostate cancer patients, the deletion of approximately 2.7 mega base pairs on chromosome 21 has been shown to result in the *ERG* and *TMPRSS2* genes fusion resulting in the over-expression of the *ERG* gene (aka *ERG*-positive; see Figure 2) [13,14].

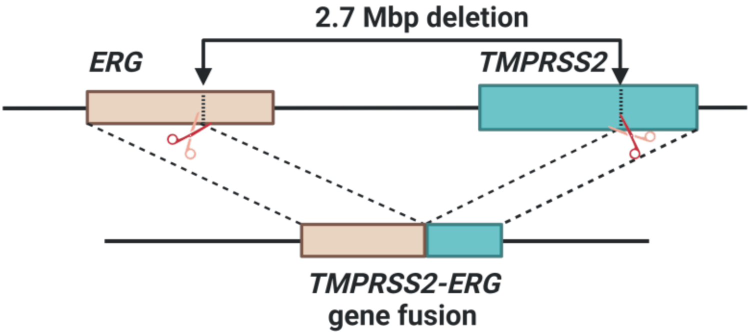


Figure 2. Schematic representation of *TMPRSS2-ERG* gene fusion as a consequence of deletion. Figure created with BioRender.com.

The ability to detect genomic alterations not only helps to understand how changes at the genomic level contribute to the development and progression of diseases such as prostate cancer, but also allows us to stratify the patients into different groups based on their genomic profiles. This stratification can be useful in predicting the patients' clinical outcomes and guiding the selection of treatment strategies [14]. For example, *ERG*-positive and *ERG*-negative prostate cancers have distinct expression signatures, morphological features, and clinical outcomes [14]. As for another example, together with others, in Article I of this work, we showed that using the copy number alterations and mutation information, grade II and III glioma samples can be divided into multiple subtypes [15].

2.2 Human epigenome

The epigenome comprises genome-wide, inheritable, and potentially reversible chemical modifications that do not change the DNA sequence itself, yet regulate gene transcription, which is the process by which RNA is synthesized from a DNA template, and as a result, regulate the cellular phenotype [16–20]. DNA methylation, posttranslational modification of DNA-associated histone proteins, and differential chromatin packaging are examples of such modifications.

DNA methylation refers to the addition of a methyl (CH_3) group especially to a cytosine nucleotide that is followed by a guanine nucleotide (i.e., CpG dinucleotides). Methylated CpGs are sparsely distributed in the human genome except for short regions with a high density of unmethylated CpGs known as CpG islands [17]. Methylated regions are transcriptionally repressed, whereas half of the genes in the human genome whose transcription start site (TSS) are covered by CpG islands are either actively expressed or ready to be transcribed [17,21]. DNA methylation was the focus of early epigenomic studies since DNA methylation withstands sample processing procedures such as DNA extraction [17].

The ability to characterize global and local DNA methylation patterns and profiles can help understand how alterations at the level of epigenome can enable a particular disease [21]. In some cancers, the genome-wide loss genomic methylation (or hypomethylation) from the repetitive genomic regions may contribute to genomic instability which is an enabling characteristic for acquiring cancer hallmarks [22,23]. In addition, it has been observed that DNA hypomethylation progresses during cancer evolution such that large regions of the non-coding and gene-poor genome of advanced cancers are hypomethylated [24,25]. As an example of local, gene-specific DNA methylation alterations, the increase in methylation (hypermethylation) of CpG islands at the *HOXB13* gene has been detected in around 30% of metastatic castration-resistant prostate cancers [26]. This hypermethylation may contribute to the lower expression of

HOXB13 gene that is involved in the regulation of androgen receptor (AR) activities and androgen-dependent prostate cancer growth [26].

Another important regulatory epigenetic mechanism is the packaging and compaction of the chromosomal DNA through a DNA and protein complex called chromatin [27]. The genomic DNA is packaged by wrapping it around a set of histone proteins (H2A, H2B, H3, and H4 known as core histones), which forms a nucleosome, the primary repeating unit of chromatin [20,27,28]. Nucleosomes are then packed further into a hierarchy of multiple loops and coils creating structures of size 30-nm, 300-nm-long, 250-nm-wide fiber respectively, and finally the chromatid of a chromosome [28,29].

When DNA is tightly packaged, it is not accessible to DNA-dependent processes such as transcription, DNA repair, replication, and recombination [30]. In contrast, when the DNA is loosely packed, the DNA sequence becomes accessible for protein interaction. One group of such proteins is transcription factors (TF). TFs possess DNA-binding domains which can recognize and bind to specific DNA sequence motifs found in regulatory regions such as promoters and enhancers [11,31].

A promoter is a DNA sequence located near the TSS of a gene. The binding of TFs to the promoter region helps to form the transcription initiation complex, which is necessary for the initiation of transcription. In the context of normal prostate development and prostate cancer, AR is a crucial TF. This protein possesses a ligand binding domain that interacts with androgens, such as testosterone. When an androgen molecule binds to AR, AR undergoes a conformational change, allowing it to translocate into the cell nucleus. Once inside the nucleus, AR utilizes its DNA binding domain to recognize and bind to specific DNA motifs. These are known as AR binding sites and together, they are referred to as the AR cistrome. Moreover, AR can modulate the transcriptional activity of its target genes through an additional domain located on its N-terminal region. This ability to regulate gene expression is essential in normal prostate development and it has implications in the progression of prostate cancer.

Enhancers are short segments of DNA that can be located far upstream or downstream from their target genes, which can regulate their transcription by recruiting TFs [11,31].

To become accessible, the chromatin structure needs to be modified via a process called chromatin remodeling. Three dynamic processes involved in chromatin remodeling are nucleosome composition alteration, nucleosome repositioning, and covalent posttranslational modification of histones [27]. Nucleosome composition alteration occurs when a canonical histone in the nucleosome such as H2A is replaced by a histone variant such as H2A.Z [27]. Nucleosome repositioning involves changing the position of the nucleosomes along with the DNA by sliding of the histone octamer or its ejection by chromatin remodelers [27]. Covalent posttranslational modification of histone proteins such as methylation and acetylation of lysine residues, and phosphorylation of serine and threonine residues is one way in which chromatin remodeling occurs [20,27,28]. Histone modifications have different effects on chromatin structure and DNA accessibility, as an example, trimethylation of lysine 27 (K27) on histone 3 (i.e., H3K27me3) has a repressive effect, whereas the trimethylation of lysine 4 (K4) on histone 3 (i.e., H3K4me3) has an activating effect and in general lysine acetylation correlates with chromatin accessibility and transcription activity [17,20,28,32].

A class of TFs, called pioneer factors, are crucial for these processes to occur [33]. They bind to DNA on the nucleosome surface of tightly-packed chromatin, loosen the chromatin, and facilitate the subsequent binding of other TFs as well as nucleosome remodeling complexes, and histone modifiers [33]. Pioneer factors can perturb the nucleosome structure and chromatin accessibility [33]. One such pioneer factor is FOXA1. In the context of normal prostate, it induces open chromatin that allows TFs such as AR to bind to specific genomic regions, and thus helps to shape AR signaling that drives the growth and survival of normal prostate [34,35]. However, FOXA1 is also a driver of prostate cancer onset and progression [34]. Together with HOXB13, it can reprogram the AR cistrome resulting in transcription of genes that contribute to oncogenesis, which is the transformation of normal cells into cancer cells [34,36]. Therefore, the ability to detect and characterize genome-wide alterations in the chromatin accessibility can provide valuable insights into how such changes may contribute to the development and progression of diseases like cancer.

2.3 Human transcriptome

Transcriptome refers to the collection of all coding and non-coding transcribed DNA sequences also known as RNA transcripts and their abundances in a cell [37,38]. In a process called transcription, different enzymes (e.g., RNA polymerase I, II, and III in non-plant eukaryotes) interact with a particular DNA template in the accessible genome and synthesize different types of RNAs. Transcription requires the binding of RNA polymerase to the promoter region of the DNA template that will undergo transcription. In addition to RNA polymerase, TFs can bind to the promoter, and they are important factors in the recruitment of RNA polymerase to the promoter and consequently in the regulation of transcription. In eukaryotes, RNA polymerase I synthesizes precursor ribosomal RNA (rRNA) 45S important for the formation of ribosomes - macromolecular complexes that are responsible for protein synthesis. RNA polymerase II synthesizes precursor messenger RNAs (mRNAs), miRNAs, and most small nuclear RNAs (snRNA) [39]. mRNAs are used by the ribosomes to synthesize proteins in a process called translation. miRNAs are non-coding molecules meaning that they do not undergo translation. However, they are functional as they are post-transcriptional regulators. They regulate the expression of their target genes by forming complementary base pairs at the 3' UTR (or the untranslated region) of their target's mRNA transcripts resulting in RNA degradation or inhibition of protein translation, which is known as RNA silencing [39–41]. Consequently, it is expected that there exists a negative correlation between miRNA expression levels and the expression levels of their target mRNAs. One miRNA can regulate several different mRNAs and conversely, each mRNA may be targeted by different miRNAs. miRNAs have been shown to be involved in multiple cellular processes such as cell proliferation, differentiation, and apoptosis [41]. Thus, their dysregulation may have an impact in diseases such as prostate cancer. In Article II of this work, we identified miRNAs that were dysregulated at different stages of prostate cancer that had an impact on the expression of their mRNA and protein targets.

In humans, miRNAs are found to be transcribed by RNA polymerase III [40]. RNA polymerase III synthesizes rRNA 5S as well as transfer RNA (tRNAs) which is responsible for transferring amino acids to the ribosomes where messenger RNAs are translated into proteins. In addition to the RNA species described here, there are other types of RNA such as small interfering RNA, long non-coding RNA, and circular RNA.

In this work, the focus is on the use of mRNA and miRNA to characterize the transcriptome of prostate cancer and its posttranscriptional regulation. Although a common expression profile characterizing each tumor stage has not yet been identified for prostate cancer, genome-wide characterization of the prostate cancer transcriptome can provide insights in understanding the development and progression of prostate cancer [41].

2.4 Human proteome

The human proteome refers to the full set of proteins that are encoded by the human genome and expressed in a particular cell [42]. A protein is made of a chain of amino acids that are bound together by peptide bonds. An amino acid is thus the building block of proteins, and is composed of an amino group, a carboxyl group, and a variable side chain. Even though there are hundreds of naturally occurring amino acids, 20 of them are encoded by the genetic code [43]. Genetic code is the set of all possible 64 permutations of 3-letter nucleotide sequences (e.g., ATG or TCC), which are called codons. It is possible that more than one codon encodes for the same amino acid. During the translation process at the ribosome, individual codons within a mature mRNA transcript are recognized by tRNA molecules. Acting as an adapter, the tRNA brings the corresponding amino acid to the ribosome, where an rRNA catalyzes the formation of a peptide bond, connecting the incoming amino acid with the preceding one in the growing polypeptide chain. This process occurs in a stepwise fashion, where each amino acid is added to the chain in an ordered sequence as dictated by the mRNA codons. The amino acid chain is also referred to as the polypeptide chain, with a peptide denoting a short chain containing more than two and up to 50 amino acids.

Proteomics is a field of study that involves the identification, quantification, and analysis of the proteome. This includes the characterization of all possible protein modifications, as well as the inference of potential interactions between different proteins [44,45]. Quantitative proteomics, as a branch of proteomics, focuses on quantifying the proteome [45]. Being able to quantify proteins enables the inference of protein expression profiles that can reveal and characterize cellular state at different conditions [45,46].

2.5 High-throughput measurement

The advent of the Sanger method used for determining the nucleotide sequences in DNA by Frederick Sanger and his colleagues in 1977 started a new era in the measurement of biological molecules [10,47,48]. Since 1977 the Sanger method has undergone multiple improvements and automation. In 1987, the first automated DNA sequencer was developed and used to analyze and determine the structure of a gene in rats [49]. All in all, it took around 25 years from the advent of the Sanger method, for the next-generation sequencing (NGS) methods to emerge, enabling the simultaneous determination of the sequence of millions of DNA templates [47,50]. NGS methods differ from the Sanger method in how they construct their sequencing libraries which improves the time required for library construction from approximately one week to approximately 2 days [50]. This and other capabilities and advances have made HTS a reality and enabled genome-wide (rather than site-specific) characterization of the genome, epigenome, and the transcriptome [50].

Currently, there are several commercially available NGS platforms/instruments. They are distinguished from one another based on factors such as their throughput (i.e., the amount of data generated per sequencing run), cost, the typical errors they make, the type of output read (e.g., single-end or paired-end reads), and the read length [51]. Each combines different methods and protocols to achieve its goal of sequencing the DNA (here, the examples explain the Illumina sequencing technology since the data used in this work are produced using Illumina sequencers. Additionally, Illumina sequencers are currently still one of the most commonly used short-read sequencing platforms [51]). These methods can be classified into four broad groups, namely, template preparation, sequencing chemistries, imaging/detection, and data analysis [51,52].

Template refers to the DNA fragment to be sequenced [51]. To prepare the template, sample DNA needs to be broken into smaller fragments. Next, DNA fragments may undergo size selection, where only DNA fragments within a certain range of length are retained. This is because different sequencing instruments work optimally with DNA fragments that are within a certain size range. The next step in template preparation is either the clonal amplification of the templates where single DNA molecules are cloned or single DNA-molecule templates that do not require clonal amplification [51,52]. In approaches that require clonal amplification, an amplification step is required so that there is a strong enough signal at the imaging/detection step required for reliable detection of the incorporated nucleotides [50,51]. In the case of the clonal amplification, a set of common adapters are ligated at each end of the DNA fragments [51]. A sequence adapter is a short chain of nucleotides that facilitates the amplification cycles as well as anchoring the ligated DNA fragment to a surface [50,51]. In addition to sequence adapters, primers are used. Primers mark the starting point for DNA synthesis and sequencing reaction by providing a free 3’ hydroxyl group to which a DNA polymerase can add a new nucleotide. Platforms using the clonal amplification, use different strategies to achieve amplification. Illumina sequencers use a solid surface to perform the amplification [51,53]. Primers are bound to the solid surface (i.e., flow cell in Illumina) in a covalent manner and the ssDNA templates can bind to the primers by hybridization. Polymerase chain reaction (PCR) is mainly used for the clonal amplification step, and it may introduce some challenges as described later [52].

In terms of sequencing, two primary approaches can be enumerated: sequencing by synthesis (SBS), used by the first sequencing platforms, that utilizes DNA polymerase enzyme, and sequencing by ligation [54], which employs DNA ligase enzymes to identify the nucleotide composition of a DNA sequence [50–52]. SOLiD and BGI sequencers are two platforms that use sequencing by ligation [51]. Illumina sequencers use SBS resembling partially to the Sanger method [53]. During a sequencing cycle, which results in the identification of one of the bases of a DNA fragment, a mixture of 4 different types of fluorophore-labeled nucleotides that lack hydroxyl group on the 3' carbon of the sugar backbone and DNA polymerases are added to the flow cell [51–53]. DNA polymerases incorporate the nucleotides in the elongating sequences. In theory, each sequence cannot be elongated more than once in each cycle due to the lack of 3' hydroxyl group. At this stage, the elongating sequences are ready to be imaged.

With regard to the imaging/detection, different methods include the optical measurement of signal intensity (e.g., two- or four-color imaging) or the non-optical measurement of changes in ionic concentration, as seen in the Ion Torrent platform [51,52,55]. Illumina sequencers use two or four laser channels (depending on the platform) to excite the fluorophores bound to the incorporated nucleotides in a given sequencing cycle and use total internal reflection fluorescence microscopy to image/detect the bioluminescence from the clusters on the solid surface [51,53]. Two advantages of two-color imaging over four-color imaging is its higher speed and lower cost since less imaging is performed and less fluorophore is used [51]. Once the imaging of a cycle is complete, in Illumina platforms, fluorophores are cleaved and washed away from the flow cell [51–53]. Additionally, the missing 3' hydroxyl groups are regenerated so that in the next cycle, the elongation can continue [51–53]. NGS methods are considered high throughput because they can simultaneously perform sequencing and detection for millions of DNA fragments/templates, eliminating the need for the electrophoresis step required in the Sanger method [50,51]. This is made possible, using Illumina technology as an example, because after the clonal amplification of DNA fragments, each colony occupies distinct sites, allowing for parallel sequencing reactions to take place [51]. Advancements in NGS technologies, including enhanced sequencing chemistries and improved detection sensitivities, have led to higher throughput, meaning higher volumes of data can be generated per sequencing run [50].

The ultimate output of NGS technologies is a large amount of sequencing reads, each of which represents the sequence of bases in a single molecule of DNA that was sequenced [51]. Sequencing a DNA template/fragment in a single direction results in the production of single-end reads. However, when a DNA fragment is sequenced from both the forward and reverse directions (i.e., from each end), it produces paired-end reads. In general, the sequence read length generated by the commonly used NGS methods is between 30 and 400 nucleotides long which is shorter than 600 to 800 base pairs of length achieved by the Sanger method [38,50,56,57]. Short sequencing reads pose a few challenges, particularly during genome assembly, alignment, and subsequent downstream analyses, as will be described later. Alternative NGS technologies with longer read length aims to overcome some of these challenges. However, they are currently more expensive and/or have lower throughput, which explains the popularity of cheaper platforms with shorter sequence read lengths [51]. Sequencing data for all the articles in this work are short read sequences.

2.5.1 Sequencing considerations and quality control

Currently, NGS technologies are not entirely error-free due to various types of errors and issues. The error rates in NGS platforms typically range from approximately 0.1% to 15% [51]. In particular, the overall accuracy of Illumina technology is estimated to be greater than 99.5% [53]. These errors and challenges can arise at various stages, including sample handling, template preparation, sequencing chemistries, and imaging/detection steps [58]. Consequently, varying error profiles can be observed across different NGS technologies and instruments [59]. These errors can be categorized as random or systematic, with the latter being referred to as bias. Illumina technology has been observed to exhibit a systematic tendency for substitution errors, where one base is mistakenly identified as another, especially after a guanine [59–61].

Some examples of errors and challenges include primer amplification bias, where in the first few nucleotides across all sequencing reads, the expected amounts of cytosines and guanines (or adenines and thymines) fluctuate; contamination with foreign bacterial or viral DNA during sample preparation before sequencing introduces foreign sequence reads [62–64]; errors or lower confidence in base-calling due to the increasing fluorescence noise as reads are elongated [52,62,65]; optical duplicate reads as a result of miscalling one cluster as two; PCR duplicate reads as a result of sequencing PCR copies of the exact same DNA fragment [70]; adapter sequence contamination where sequencing reads contain part of the adapter sequence [66]. In addition, each sequencing assay used for a specific application can introduce its own set of errors and biases (e.g., mitochondrial DNA contamination in ATAC-seq as an assay for characterizing chromatin accessibility), which can impact the accuracy and reliability of the resulting data.

To ensure the quality of NGS data, quality control tools such as `FastQC` or other assay-specific quality control tools can be employed [62]. These tools can help identify issues in the data and guide the proper preprocessing before downstream analyses. To illustrate, in case of adapter sequence contamination, tools such as Cutadap can be used to remove the adapter sequences and improve the accuracy of the sequencing results [66].

2.5.2 Sequence alignment

A crucial step in preparing sequencing data for downstream analyses is the accurate mapping of the sequencing reads to their corresponding locations in a reference genome. This process involves identifying the genomic coordinates from which a DNA fragment originated. This process is commonly referred to as sequencing read alignment, or simply alignment, and it is performed by specialized computer software known as sequence aligners or aligner for short [67]. One algorithm that has become widely used in short-read alignment is the Burrows-Wheeler Transform (BWT) [68]. Tools such as `BWA` [67] and `Bowtie2` [69] use the BWT algorithm.

The task of sequencing read alignment assumes the availability of a reference genome. If a reference is not available, a reference genome or transcriptome can be constructed by assembling the sequencing reads, a process known as de novo assembly [70]. Thanks to the HGP and similar efforts [7,71], the human reference genome is already available.

Accurate sequence alignment is crucial for many genomic analyses, but several issues can result in inaccuracies referred to as artifacts. One example of a challenge in the sequence alignment is the presence of multi-mapped reads, where a read can be aligned equally well to multiple regions in the genome. These reads can confound downstream analyses and interpretation of results. This situation can arise when there are regions of sequence redundancy in the genome, such as repeats or duplicated regions [72]. According to some estimates around 50% to 60% of the human genome is composed of repetitive DNA [73]. Sequence reads obtained from genomic regions with repetitive sequences are thus called unmappable reads and these regions are called unmappable regions [72]. Certain aligners discard unmappable reads, which may result in lower coverage in unmappable regions compared to other regions and introducing mappability bias [74]. Another approach used by certain aligners to handle multi-mapped reads is randomly selecting one of the potential locations and assigning the sequence read to that location while setting the mapping quality score (MAPQ) to 0 [56,75]. MAPQ score is calculated as $-10\log_{10}(p)$ where *p* indicates the likelihood of a read being incorrectly mapped to the reference genome. It is assigned to each aligned read, and it is a measure of how confident the aligner was about that particular alignment (e.g., a MAPQ of 30 means that there is 1 in 1000 chance that the read was wrongly aligned) [75]. The MAPQ score can be used to remove reads with low scores from the downstream analysis. For example, discarding reads with MAPQ=0 results in the removal of multi-mapped reads. An additional step in dealing with unmappable regions and regions with low mappability is to discard them before starting the downstream analysis [76,77].

2.6 Measurement of genome and its applications

Whole-genome sequencing (WGS) and whole-exome sequencing are two commonly used high-throughput methods for the genomic DNA sequence detection and measurement. The first human whole-genome was sequenced using next-generation technologies in 2008 [78].

Utilizing WGS data allows for comprehensive detection and measurement of various types of genomic alterations, including single-nucleotide variations (SNVs), insertions and deletions (indels), copy number alterations, and structural rearrangements, across the entire genome. SNVs or point mutations arise when a single base is found to differ from the corresponding base in the reference genome, such as a change from cytosine to thymine. Indels manifest when a small group of nucleotides is inserted into or deleted from a specific region of the genome, respectively. In addition, there exist larger genomic alterations, also known as structural variants, which include duplications, amplification, deletions, and copy-neutral loss of heterozygosity (cnLOH), resulting in changes to the copy number. Copy number refers to the number of copies of a genomic locus per cell. Humans, being diploid organisms, possess two copies of each autosomal DNA. Copy number alterations that occur in the somatic cells are called somatic copy number alterations (SCNA), while changes larger than 50 nucleotides that occur in germline cells are called copy number variations [79]. SCNAs can impact crucial genes involved in the oncogenesis process [80].

In genomic terms, duplication refers to an event where a larger segment of the genome is duplicated. Amplification, on the other hand, involves the duplication of a genomic segment multiple times, leading to the presence of multiple copies of a specific locus. Deletion occurs when one or both copies of a larger genomic segment is deleted, resulting in hemizygous and homozygous deletions, respectively. Loss of heterozygosity occurs when one of the two copies of the at a heterozygous locus is lost. cnLOH occurs when the lost copy is replaced with a duplicated copy of the surviving copy, resulting in an unchanged copy number at that locus.

In the last few decades, multiple methods have been developed to detect SCNAs, providing a way to identify potential disease-associated genes, including those involved in cancer [81]. These methods include fluorescence in situ hybridization (FISH), comparative genomic hybridization (CGH) [82], microarray-based CGH (aCGH) [83], and single nucleotide polymorphism (SNP) arrays [84]. However, these methods generally exhibit lower resolution. In this context, resolution refers to the level of detail in accurately detecting and distinguishing genomic alterations. As an example, FISH has a resolution on the scale of several megabases, while aCGH provides a resolution of approximately one megabase [81].

As the resolution improves, the ability to detect smaller events, such as deletions, and the accuracy of localizing the event increase [81,85]. Methods using HTS data offer enhanced accuracy and higher resolution compared to earlier mentioned methods, with WGS having the potential to detect SCNAs with single-nucleotide accuracy [86].

HTS-based SCNA detection typically follows a series of steps. Initially, the number of sequencing reads or DNA fragments within overlapping or non-overlapping genome-wide windows is tallied, with the window size determined manually or algorithmically. Assuming no biases like mappability or GC content, the estimated read count reflects the underlying copy number [86]. Here, GC content bias refers to a correlation between the proportion of G and C bases in a specific genomic region and the count of mapped DNA fragments associated with that region [87]. If a paired normal sample is accessible for a tumor sample, it becomes feasible to calculate the ratio of read counts between the tumor and normal samples during this step, which potentially helps in canceling out some of the biases.

The subsequent step involves identifying the breakpoints that delineate genomic regions into segments exhibiting significantly distinct copy numbers. Once the segments are identified, using the read count ratio between the tumor and normal samples, the copy number within each segment is estimated. In certain methods, this step is subsequently accompanied by the categorization of the identified segments [86].

Tools that solely rely on read count information have limitations in detecting certain genomic events like cnLOH. However, when the depth of coverage is sufficiently high, the detection of cnLOH events becomes possible by utilizing and calculating the B-allele fraction (BAF). BAF represents the proportion of the alternate allele at heterozygous SNP positions. SNPs are single nucleotides that vary among individuals and are present in at least 1% of the population. Heterozygous SNPs indicate the presence of two different alleles at a specific SNP locus, while homozygous SNPs indicate the presence of the same allele on both copies of a chromosome pair. In normal diploid cells, the BAF for heterozygous SNPs is expected to be 0.5 since half of the reads contain the alternate allele. Alterations in the genome lead to deviations from the expected BAF value of 0.5. In the case of cnLOH in a sample from a pure tumor population (with no normal cell contamination), the BAF values are typically 0 or 1, and these values shift towards 0.5 as the purity of the sample decreases.

There are other types of structural alterations or chromosomal rearrangements, such as translocations and inversions. Translocation occurs when two nonhomologous chromosomes exchange chromosomal segments, while inversion refers to the reversal of a stretch of DNA sequence within a chromosome. Figure 3 provides a schematic representation of four classes of genomic alterations.

SCNAs have been implicated as key drivers in various cancer types, with specific gene amplifications such as *MCL1* and *BCL2L1* found to be crucial for the survival of cancer cells [79,88]. In the context of prostate cancer, the short arm (p arm) of chromosome 8 is commonly deleted in 55.7% of localized cases and 90.5% of advanced cases [89]. This region harbors the *NKX3.1* gene, a prostate-specific tumor suppressor gene and the deletion of one copy of this gene may predispose to prostate carcinogenesis [90].

In Article I, we developed a tool that utilizes WGS data to detect copy number alterations. Additionally, the tool is also capable of detecting cnLOHs.

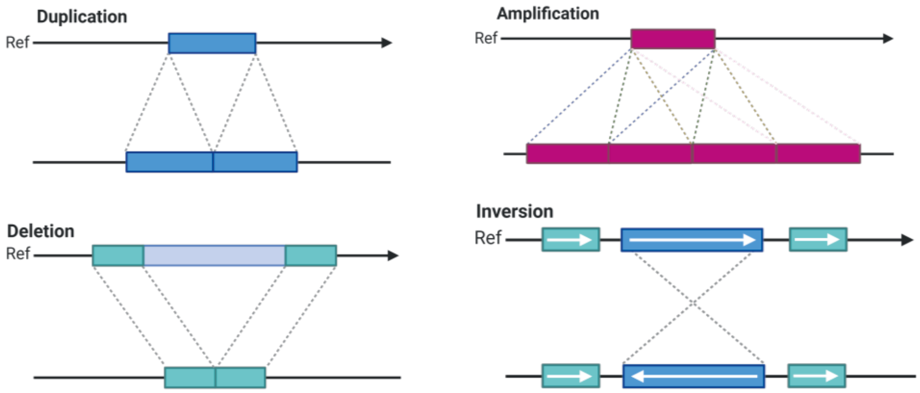


Figure 3. Schematic representation of four classes of genomic aberration. Figure created with BioRender.com.

2.7 Measurement of DNA methylation and its applications

Several approaches have been developed to detect and measure genome-wide DNA methylation sites. They are categorized into three groups: (1) enrichment-based methods that enrich for methylated DNA fragments, (2) digestion-based methods that employ methylation-sensitive restriction enzymes, and (3) sequencing-based methods that determine the sequence of bisulfite-converted DNA [18,21].

Methylated DNA immunoprecipitation (MeDIP), used in Articles II and III, is an enrichment-based method that employs an antibody specific for methylated cytosine to immunocapture denatured, methylated genomic DNA fragments followed by either sequencing or DNA microarray detection [18,21,91]. While enrichment-based methods allow for rapid and efficient genome-wide assessment of DNA methylation, they do not provide information on individual CpG dinucleotides and exhibit a non-linear relationship between enrichment level and DNA methylation level [21,91]. This enrichment bias leads to the preferential enrichment of CpG-rich DNA fragments and potential underrepresentation of CpG-poor regions with less than 1.5% CpG dinucleotides [18,21,92]. This may lead to the misinterpretation of methylated CpG-poor regions as unmethylated in the absence of appropriate corrections [92]. The microarray-based MeDIP method has a resolution of 80 kilobase pairs, whereas MeDIP-seq can achieve a resolution of 100-300 base pairs [91,93]. MeDIP-seq may suffer from biases such as copy number alteration bias, GC content bias, and CpG density bias [21]. GC content bias leads to lower numbers of DNA fragments map to genomic regions with high or low GC content than regions with medium GC content due to PCR inefficiency in those regions [94,95].

Sequencing-based methods of bisulfite-converted DNA, on the other hand, offer genome-wide base-resolution and unbiased information, but they involve a more laborious procedure compared to MeDIP and are generally more expensive [18,21,91]. Reduced representation bisulfite sequencing (RRBS), a hybrid approach combining digestion-based and sequencing-based methods, is a high-resolution method that uses a restriction enzyme to digest genomic DNA, enriching for DNA fragments with high CpG content regions. Subsequently, bisulfite conversion is performed, and DNA fragments are sequenced [18,96].

The sequencing reads obtained from either of these assays are quality controlled and then aligned to a reference genome. After addressing potential biases, the aligned sequencing reads can be used to quantify methylation levels. Proper normalization enables the comparison of methylation patterns between samples, facilitating the identification of differentially methylated regions (DMRs) across conditions, such as healthy vs. disease. Additionally, DNA methylation data can be used to investigate epigenetic silencing of tumor suppressor genes [97]. Notably, DNA methylation information has been effectively used in the classification of central nervous system tumors [98].

MEDIPS [99] and QSEA [99,100] are two computational tools that can be utilized for the analysis of MeDIP-seq data. These tools offer the capability to quantify DNA methylation levels in a genome-wide manner while addressing potential biases. They also enable the detection of regions exhibiting differential methylation across different samples.

2.8 Measurement of chromatin accessibility and its applications

Chromatin accessibility refers to the accessibility of a genomic locus, indicating its level of packaging and compaction. Regions with accessible chromatin are characterized by a more open structure. Several assays are available for the genome-wide measurement of the chromatin accessibility. Direct methods include Assay for transposase-accessible chromatin using sequencing (ATAC-seq), DNase-seq, and Formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq), while indirect methods include Micrococcal nuclease sequencing (MNase-seq) [30].

ATAC-seq, as used in Article III, is an HTS assay that maps open chromatin regions, TF and nucleosome occupancy, enabling the inference of nucleosome packing, positioning, nucleosome-TF spacing patterns, and TF occupancy and footprints [30]. It uses hyperactive Tn5 transposase enzymes to cleave dsDNA and tag them with sequencing adaptors. This is followed by library construction via PCR and paired-end NGS [30,101]. The resulting sequencing reads undergo quality control, alignment to a reference genome, and peak detection using tools like Model-based analysis of ChIP-Seq (MACS) [102], which identifies regions with enrichment of aligned reads compared to inaccessible loci. These regions, known as peaks, indicate accessible chromatin loci. Peak detection is influenced by two factors: sequencing read depth and TSS enrichment score [103]. A higher sequencing read depth provides greater statistical power to detect weaker sites, resulting in the identification of more peaks [103,104]. As a quality metric, a higher TSS enrichment score indicates an overall better sample quality, and in general enables the detection of a greater number of peaks [103,104]. However, the TSS enrichment score relies on TSS annotation, and as a result, it may not directly correlate with the characteristics of the detected peaks like those detected in intronic and intergenic regions. Another metric used to measure sample and peak call quality is the fraction of reads in peaks (FRiP). This metric quantifies the proportion of sequencing reads that align to peaks. A low value suggests suboptimal enrichment and may indicate issues with the experiment's quality. These metrics, together with scores or adjusted P-values generated by peak callers like MACS, collectively provide insights into the reliability of the detected peaks. As different samples may have different sequencing read depths and TTS enrichment scores, these factors should be considered during the analysis of data from multiple samples. With potential biases addressed, the aligned sequencing reads can also be used to quantify chromatin accessibility. Normalization allows for the comparison of different samples, facilitating the detection of differentially accessible regions (DARs) across samples with distinct conditions, such as healthy and disease states.

ATAC-seq offers several advantages compared to other chromatin accessibility assays. One key advantage is its requirement for a lower number of starting cells (500-50,000), making it suitable for analyzing samples with limited starting materials, including clinical samples. In contrast, other assays typically require a larger number of cells (100,000-

1,000,000) [30]. Additionally, ATAC-seq exhibits comparable sensitivity and specificity to DNase-seq, which is widely regarded as the gold standard for studying chromatin accessibility [30,101]. However, ATAC-seq does have some limitations. As the distance from the accessible loci increases, its ability to map nucleosomes diminishes. Another limitation is its susceptibility to mitochondrial DNA contamination [30,101]. This is because mitochondria lack the chromatin packaging, making their DNA more accessible [105]. Overall, ATAC-seq is gaining popularity in the field of epigenomic research.

2.9 Measurement of transcriptome and its applications

The transcriptome is the complete collection of all small and large, coding, and non-coding RNA molecules transcribed from the DNA sequences in a cell [37,38]. It includes mRNA, tRNA, rRNA, and other non-coding RNA molecules such as miRNA [106]. Depending on the cell type and its physiological state only 1 to 2% of the total RNA is mRNA [106].

Transcriptomics or transcriptome profiling refers to the mapping and quantifying the composition and the structure of the transcriptome in different cell types and conditions [16,38,107]. Elucidating the similarities and differences in transcriptomes between different cell types, tissues, or conditions is valuable for gaining insight into the underlying molecular mechanisms that govern biological processes [38,107]. Several approaches have been developed for transcriptome profiling, employing methods such as hybridization-based microarray technology or sequencing.

Sequencing-based approaches can be divided into two categories. The first category includes low-throughput methods like cDNA sequencing and expressed sequence tag (EST) sequencing. These methods rely on Sanger sequencing. The second category consists of HTS approaches [38,107].

RNA sequencing (RNA-seq) is an HTS-based approach that enables the profiling of the entire transcriptome [107]. RNA-seq enables the detection of transcripts, including novel ones and alternative splicing events, even in the absence of known genomic sequences. Splicing refers to the process of removing introns and joining exons to form mature mRNA transcripts. Additionally, RNA-seq offers the advantage of single base resolution in localizing transcription boundaries, enabling precise mapping of transcription start and end sites. Furthermore, through the comparison of aligned reads with a reference genome or transcriptome, RNA-seq facilitates the detection of sequence alterations, including SNVs, within transcribed regions. Additionally, RNA-seq overcomes the issue of high background signals often associated with hybridization-based approaches. Finally, RNA-seq exhibits high accuracy and a low error rate, making it a reliable technique for comprehensive transcriptome analysis [38,107].

In RNA-seq, the total RNA or a specific fraction, such as poly(A)+ mRNA, is reverse transcribed into cDNA, followed by the addition of adaptors. The cDNA is then subjected to single-end or paired-end sequencing using various high-throughput sequencing technologies [38,107].

After sequencing, sequencing reads typically undergo quality control, including the assessment of sequence quality, GC content, and the presence of adaptor sequences. The reads are typically aligned or mapped to a reference genome or transcriptome to determine their source and abundance [106]. The mapping process enables the quantification of gene expression levels and the detection of alternative splicing events and novel transcripts. In cases where the identification of novel transcripts is desired, de novo assembly methods can be applied [106].

After quantification, addressing biases, and applying proper normalization, the resulting data can be used to determine the transcript composition and the abundances of each transcript (or gene expression) in a given sample [106]. Proper normalization is crucial for comparing different samples. Gene expression data can be used to identify differentially expressed (DE) genes across various cell types or conditions. Statistical tools such as `DESeq` [108,109] or `edgeR` [110] employ different approaches to detect DE genes, taking into account various biases and performing appropriate normalizations. Furthermore, gene expression data can be integrated with pathway information to identify dysregulated pathways, particularly in diseases like cancer, where a pathway refers to a sequential set of molecular interactions and signaling events that collectively regulate a specific biological process or cellular function [97].

The detection and quantification of miRNAs can be achieved through small RNA sequencing, which follows a standardized procedure. Initially, total RNA undergoes size fractionation to isolate RNA molecules within a specific length range (typically 18 to 30 nucleotides). Following size selection, a series of laboratory steps are performed before sequencing on high-throughput platforms, such as the Illumina sequencers.

2.10 Measurement of the proteome and its application

Initially, proteomics relied on the separation of proteins using two-dimensional gel electrophoresis, followed by mass spectrometric identification. However, gel-free approaches have gained popularity due to their higher throughput capabilities [42,44].

Mass spectrometry (MS) is a technique used to measure the mass-to-charge ratio of ions in a sample, enabling protein identification and quantification [45]. A typical mass spectrometer consists of three main components: an ion source, a mass analyzer, and a detector [45]. Different types of ion sources, mass analyzers, and detectors offer distinct advantages and drawbacks, allowing for the customization of instrument configurations to suit specific applications and experimental requirements [45].

During proteomics analysis, several factors must be considered, including the complexity of the sample, the detection of low abundance proteins, and accurate quantification. A popular approach for analyzing complex peptide mixtures is liquid-chromatography (LC) coupled with tandem mass spectrometry (MS/MS) or LC-MS/MS [45,111]. In this method, proteins are enzymatically digested, typically using trypsin enzymes, to generate peptides. The resulting peptides are then separated using liquid chromatography to reduce sample complexity [46,111]. Subsequently, the separated peptides undergo two rounds of mass spectrometry analysis.

The mass spectra obtained from MS are searched against a database, and each spectrum is assigned to a specific peptide using pattern-matching algorithms [45]. Peptide identification is often preferred over whole protein identification in proteomics, as peptide identification methods are more sensitive [45].

The LC-MS/MS method can be used to quantify the relative abundances of proteins in a sample when the proteins are labeled with stable isotopes [45]. To achieve comprehensive protein quantification, it is necessary to synthesize various reagents capable of labeling different groups of proteins [45]. However, the use of isotopic labeling can increase sample preparation time and costs, which may limit its applicability to larger samples [46]. As a result, the throughput of LC-MS method is typically limited to a few hundred peptides per analysis, which can be considered relatively low [46,112].

In the realm of LC-MS/MS, two primary strategies can be distinguished: shotgun proteomics and targeted proteomics [111]. These strategies differ in their mass spectrometric methods [111]. Shotgun proteomics is effective for protein discovery, allowing for the identification of a maximal number of proteins. However, it may not be suitable for high-throughput quantification when numerous samples need to be analyzed [46,111]. In contrast, targeted proteomics is a more suitable approach for reproducible detection and quantification of a predefined set of proteins across many samples, although it may leave a substantial portion of the proteome undetected and unquantified [111,113].

In Article II, we employed a quantitative proteomics technique known as sequential windowed acquisition of all theoretical fragmentation - mass spectrometry (SWATH-MS) [113]. This method aims to address the limitations associated with the shotgun proteomics and targeted proteomics by combining the comprehensive nature of shotgun proteomics with the reproducibility of targeted proteomics [111,112,114]. By doing so, SWATH-MS increases the throughput, allowing for the identification and quantification of a larger number of proteins across multiple samples, while maintaining the consistency and reproducibility achieved through targeted mass spectrometry using the selected reaction monitoring method, which is regarded as the gold standard for quantitative proteomics [111,112].

Furthermore, SWATH-MS is a label-free method, making it a cost-effective option with simplified sample preparation [46,113]. Instead of relying on a database search like other methods, SWATH-MS utilizes spectral libraries that contain information about previously identified peptides to identify peptides of interest [46,111]. Thus, the availability of comprehensive proteome-wide spectral libraries becomes crucial for successful peptide identification [111,114,115]. Despite being less sensitive than the selected reaction monitoring method, SWATH-MS is well-suited for large-scale, high-throughput, and high-quality quantitative proteomics [113,114].

2.11 Integration of data from different levels

High-throughput data obtained from each of the aforementioned levels offers valuable insights into the state of a biological system and its components [16,116]. When comparing samples with a specific disease against control samples, using single-level data can yield a list of intergroup differences, including biological pathways and processes associated with the disease [16]. However, it is important to acknowledge that multiple levels may contribute to and regulate a given phenotype. Relying solely on single-level data may overlook these interactions, despite each level being crucial for constructing our understanding of a biological system and enabling the exploration of diseases from diverse perspectives [117,118]. Moreover, a sole reliance on single-level data may undermine confidence in identifying causal effects due to the higher prevalence of reactive effects compared to causative effects [16]. Consequently, integrating multilevel data can foster a more comprehensive understanding of a biological system and its underlying mechanisms.

To achieve a comprehensive understanding and enhance confidence in elucidating causal changes, it is imperative to integrate and analyze data from various levels in a multilevel manner [16,44,116–121]. This approach enables us to explore the flow of information from the primary cause of a specific disease to its functional consequences, discover molecular mechanisms of disease, cluster samples, and make predictions about outcomes such as survival [16,44,116–121].

The omics data from multiple levels, encompassing the genome, epigenome, transcriptome, and proteome, can be analyzed in multiple ways: sequentially or jointly [16,116]. In the sequential approach, the results from one level are refined and made more specific by incorporating data from other levels, assuming a causal link, such as from genomics to transcriptomics [16,116].

According to Hasin and colleagues, the sequential approach can be further divided into three approaches: genome-first, phenotype-first, and environment-first [16]. In the genome-first approach, the identified genomic alterations are further characterized with other omics layers to uncover downstream interactions and pathways [16]. To illustrate, genes that exhibit copy number alterations such as amplifications or deletions, at the genome level, can be assessed for differential expression at the transcriptome level [97]. Consequently, the genome-first approach often concentrates on a specific locus or a few specific loci.

In the phenotype-first approach, various levels of omics data are collected for a particular phenotype of interest, such as primary prostate cancer. Data from each level is analyzed to identify factors that correlate with and may explain the phenotype of interest. These identified factors are then combined to further elucidate their roles and the affected pathways associated with the phenotype under study [16].

The environment-first approach involves analyzing an environmental factor, such as diet, as its primary focus. This analysis employs multilevel omics data analysis to uncover potential links to diseases [16].

In this work, Articles II and III can be categorized under the phenotype-first approach. This is because they aim to gather, analyze, and integrate different levels of omics data, focusing specifically on prostate cancer as the phenotype of interest.

Data integration and analysis pose several challenges, including technological limitations, data quality issues, high dimensionality arising from a multitude of biological variables, and the relatively limited number of available biological samples [116,120]. Consider the case where the current technologies enable the characterization and quantification of the majority of a biological system's transcriptome, while at the time of writing, only a fraction of the proteome can be detected and quantified by the existing technologies [117]. Moreover, in many omics studies, the number of biological samples is significantly smaller than the number of biological variables, such as genes. For instance, the TCGA project encompasses approximately 500 prostate adenocarcinomas, while more than 50,000 transcripts are quantified for each sample [122]. These challenges demonstrate the complexities involved in integrating and analyzing omics data, highlighting the need for careful consideration and appropriate methodologies to overcome such limitations.

2.12 High-throughput data analysis techniques and considerations

2.12.1 High-throughput data normalization

When comparing two groups of samples under different conditions, our main objective is to uncover biological differences. However, the process of sample preparation and measurement can introduce technical effects that affect the accuracy of our measurements. This means that the observed measurements may not accurately reflect the true biological differences between the groups.

One such technical effect is the variability in the total number of molecules sequenced across different samples, which can result in variations in library sizes or sequencing depths. For instance, in RNA-seq analysis, observing a large difference in the read count of a gene between two different conditions may simply be the result of differences in sequencing depth rather than the gene's actual differential expression [123].

Another technical effect is the variability in library composition due to differences in signal-to-noise ratio. The signal-to-noise ratio refers to the ratio of the actual biological signal to the background noise present in the sequencing data. In the context of ATAC-seq data, the term signal refers to the sequencing reads originating from accessible regions of the genome, while noise represents sequencing reads that do not originate from accessible regions. Such reads can arise from various sources, including sequencing errors.

To overcome these challenges, normalization techniques are employed as correction methods to adjust the quantifications, such as gene expressions. The goal is to ensure accurate comparisons between samples with different sequencing read depths and composition.

It is important to note that each normalization method operates based on a specific set of assumptions. Therefore, the selection of an appropriate normalization method requires careful consideration of whether these assumptions can be met [123]. Failure to choose a proper normalization method can have adverse consequences in downstream analysis and the resulting outcomes, such as an increased number of false positives [123].

Normalization methods can be classified into various groups. These include normalization by library size (e.g., total count normalization), normalization by the distribution of read counts (e.g., median of ratio normalization), and normalization by controls (e.g., using housekeeping genes or spike-ins expressions for normalization) [123,124].

Normalization by library size relies on the assumption of comparable total expression levels across samples, irrespective of the experimental conditions. This assumption is based on the idea that the majority of genes exhibit consistent expression across conditions, with only a minority of genes displaying differential expression. Based on this assumption, this normalization approach aims to mitigate differences in sequencing depth between samples. It achieves this by dividing the values in each sample by the total number of sequencing reads in that sample [123,124]. On the other hand, normalization by read count distribution adjustment assumes that the distributions of read counts are similar across the samples [123,124].

Median of ratio normalization falls under the category of normalization by distribution of read counts [109]. In this technique, a sample scaling factor (or size factor) is calculated through the following process: for each gene, the geometric mean of its read counts across all samples is determined, serving as a pseudo-reference or consensus sample [109]. Subsequently, for each gene, the ratio of its read count to the read count of the same gene in the pseudo-reference sample is computed. The median of these ratios within each sample yields the sample scaling factor [109,123,124]. The normalized values are obtained by dividing the values in each sample by the corresponding sample scaling factor.

Median of ratio normalization technique operates on the assumption that the majority of genes are not DE and, therefore, they should exhibit similar read counts across different conditions. Additionally, it assumes that technical effects impact both non-DE genes and DE genes in a similar manner. Consequently, it is expected to see a ratio of 1 for most of the genes [124].

Normalization techniques based on read counts distribution are effective when the number of dysregulated genes is similar across conditions, regardless of differences in the total quantity of a specific molecule per cell [123]. However, it is important to note that these normalization methods may be ineffective when there is a global shift in one condition, which violates the assumption of symmetry [123].

Assuming that the underlying assumptions of median of ratio normalization hold for count data other than RNA-seq data, it is reasonable to consider applying this method to other data types based on count data. When comparing samples from the same tissue but under different conditions, it can be reasonably assumed that the majority of regions in the genome are not differentially methylated or accessible. As a result, median of ratio normalization can be employed on MeDIP-seq and ATAC-seq data.

2.12.2 Dimensionality reduction

High-throughput techniques generate high-dimensional data, meaning they produce data with a large number of variables or features. To illustrate, in an ATAC-seq experiment, tens of thousands of genomic regions with accessible chromatin can be identified from a single sample. Similarly, in an RNA-seq experiment the expression of tens of thousands of genes can be quantified for every sample. However, many statistical methods struggle with high-dimensional data, which reduces their power [125].

Moreover, the visual exploration of the data is also affected by high dimensionality. To overcome this issue, dimensionality reduction can be employed to simplify the data by accounting for uninformative and redundant variables, enabling the extraction of hidden structures and patterns in the data and their visualization [125]. Dimensionality reduction can be applied to ATAC-seq accessible features obtained from an experiment. The purpose is to map the data to a lower-dimensional space in order to assess whether it can effectively distinguish between different sample groups, such as healthy and cancer samples. Various methods exist for dimensionality reduction, including principal component analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE).

PCA serves as a technique for dimensionality reduction. It operates by generating a set of new orthogonal variables known as principal components, which are linear combinations of the original variables [126]. Together, these principal components capture all the variance within the data. In many cases, for visualization purposes, researchers opt to employ the first two or three principal components. However, it is important to note that this choice comes at the cost of information loss since the initial principal components may not capture the entirety of the variance in data. Additionally, a drawback of using PCA in computational biology is its limited ability to optimize the separation of different groups of samples. Moreover, PCA is insensitive to the source of variation in the data. Thus, if some of the variation is caused by systematic experimental artifacts, PCA would still incorporate it into the principal components [126]. Finally, PCA is susceptible to outliers and cannot effectively capture the local structure in the data [127].

t-SNE is an additional method used for dimensionality reduction and visualization [128]. In addition to revealing the global structure in the data, t-SNE preserves the inherent clustering of the data points in the high-dimensional space, when it projects the data to a lower-dimensional space. This is achieved by preserving the pairwise data point distances. This allows for the capture of local structures in the data, addressing one of the limitations of PCA [127,128]. To accomplish this, t-SNE converts the pairwise data point Euclidean distances in the high-dimensional space into conditional probabilities determined by a Gaussian distribution. Consequently, data points that are closer receive higher probabilities (or scores), while those that are farther receive smaller scores. To determine the similarity between two points in the low-dimensional space, t-SNE uses the Student's t-distribution instead of a Gaussian distribution to address optimization problems and the crowding problem observed in the earlier stochastic neighbor embedding method [128]. Finally, t-SNE repositions the data points in the low-dimensional space such that their similarity score profile resembles the similarity score profile calculated for the high-dimensional space. One practical issue with t-SNE is that the results are not deterministic due to the random positioning of the data points in the low-dimensional space. In the case of reducing dimensionality to two dimensions, the resultant plot in a two-dimensional space might change with each execution of the algorithm, unless the random state parameter is explicitly set with a similar value for each execution.

2.12.3 Hierarchical clustering

Clustering methods are commonly used to group similar data points, such as biological samples or their features, together based on their characteristics. They are useful in gaining insights into complex, high-dimensional data and identifying hidden patterns or structures, such as co-expressed genes in an RNA-seq experiment.

One major class of clustering methods is hierarchical clustering [129,130]. Hierarchical clustering aims to create a tree-shaped data structure, or dendrogram, composed of a hierarchy of clusters showing the relationships between the clusters [129,131]. This can be done using either a top-down approach, called divisive clustering, or a bottom-up approach, called agglomerative clustering. In divisive clustering, one starts with one cluster containing all *n* observations and divides it into smaller clusters until there are *n* clusters. In agglomerative clustering, one starts with *n* observations each considered as one cluster and clusters them together until there is one cluster of *n* observations. [132].

To form a cluster, a similarity metric or distance metric is used. Different distance metrics are available, such as Euclidean, Manhattan distance, and Pearson correlation coefficient. Additionally, linkage methods, such as single-linkage, average-linkage, and complete-linkage define the inter-cluster distance used to cluster two clusters or a cluster and an element, such as a gene, together. The selection of the appropriate distance metric and linkage method depends on the task at hand [129].

Hierarchical clustering results are often displayed with a heatmap accompanied by dendrograms. This allows for the visual inspection of samples or genes that are more similar to each other than other samples or genes. Dendrograms represent the similarity between two elements, such as a gene or a cluster of genes, as well as the cluster formation order determined by the length of the branch in the dendrogram.

However, there are a few practical issues to consider when using clustering methods like hierarchical clustering. Each clustering algorithm aims to optimize a particular cluster property. For instance, in complete linkage clustering, the aim is to merge clusters in a way that minimizes the maximum pairwise distance within the merged clusters [129, 130]. Put differently, the distance between two clusters in the complete linkage method corresponds to the maximum distance between any pair of points, with one point from each of the two clusters. Additionally, clustering methods may cluster the data even in the absence of an actual cluster [129,130].

2.12.4 Comparing sample sets through statistical hypothesis testing

In the context of high-throughput data analysis, it is often desirable to identify features like genes that exhibit differences when comparing sample sets under different conditions. Consider the case of comparing ATAC-seq experiment data from localized versus advanced prostate cancer samples where we might be interested in detecting genomic regions with differential accessibility between these two sets. Statistical significance tests provide a means to assess whether there is sufficient evidence in the data to support a significant difference. Typically, a statistical test yields a P-value, which represents the probability of observing a difference at least as extreme as the one observed, assuming that there is no difference between the underlying populations [133]. In other words, the P-value represents the probability that the observed difference is solely due to random chance [134]. When the P-value falls below a predefined significance threshold, we conclude that the difference is significant. Historically, in the 1930s when computing resources were limited, the significance threshold was set at 0.05 or 0.01, with the latter being more common in genomic studies [135]. However, it is important to note that a statistically significant test outcome does not necessarily equate to a biologically significant difference, although it suggests the possibility [134,135].

Statistical tests can be categorized into two types: parametric and nonparametric. Parametric tests rely on assumptions regarding the underlying distribution of the data, and the reliability of the test results can be affected if these assumptions are not met. For example, the t-test assumes that samples are drawn from populations with a normal distribution and have similar variances [136,137]. In contrast, nonparametric tests, like the Wilcoxon rank-sum test, do not make distributional assumptions and can be employed even with skewed data or outliers [137,138]. For instance, the Wilcoxon rank-sum test (aka Mann–Whitney U test) serves as a nonparametric alternative to the unpaired t-test. It does not rely on the assumption of normality, which is a prerequisite for the t-test [137,139]. Instead, it assumes that the distributions of the two groups being compared are identical, with the potential difference being a shift in location. While nonparametric tests are well-suited for analyzing real-world data, they may be less sensitive than their corresponding parametric tests when dealing with small sample sizes [137].

2.12.5 Multiple-testing correction

During high-throughput data analysis, it is common to perform a large number of statistical hypothesis tests to identify differential features, such as genes that are differentially expressed across two conditions. However, this can lead to a substantial number of false positives, which are results that appear statistically significant by chance alone.

Let us consider a scenario where we aim to identify differentially accessible regions between two conditions, dividing the genome into non-overlapping windows of size 500 base pairs. In this case, approximately 6.5 million statistical tests would be conducted. If we set the significance threshold at 0.01, without any actual significant differences, around 1% of the windows would be called statistically significant purely by chance. This would result in 65,000 false positives when none of them are genuinely differentially accessible.

Let us consider, in a hypothetical manner, that 1% of all 6.5 million genomic regions are actually differentially accessible, and the statistical power to detect them is 70%. In this scenario, we would be able to detect 45,500 of the genomic regions that are truly differentially accessible. However, alongside these true positives, we would also detect 64,545 false positives. Consequently, the false discovery fraction would be 0.59, indicating that roughly 1 out of 2 discoveries are false. This example highlights the necessity for multiple-testing correction methods.

To address this issue, multiple-testing correction methods are employed [133]. These methods can be broadly classified into two families: family-wise error rate (FWER) methods, such as the Bonferroni method, and false discovery rate (FDR)-based methods, such as the Benjamini-Hochberg method [133].

The Bonferroni method adjusts the P-values by multiplying them by the number of tests conducted in an analysis [140]. However, this method is highly conservative and tends to yield low statistical power [140]. In the aforementioned example, the P-value for a truly differentially accessible genomic region should be smaller than 1.538*10^-9 to retain statistical significance after correction.

In contrast, the Benjamini-Hochberg correction offers better statistical power compared to the Bonferroni method. The Benjamini-Hochberg correction arranges the P-values in order, ranks them, and adjusts the P-values inversely proportion to their rank [133].

Reducing the number of tests is another strategy to enhance detection power. One approach is to filter out genes with low expression levels prior to conducting differential expression analysis, thereby reducing the number of tests performed [106,141].

2.12.6 False discovery rate calculations

When analyzing high-throughput data, it is a common practice to set a stringent significance level (α) in order to minimize false positives, but this approach often leads to an increase in false negatives, leading to a reduction in statistical power [142]. Moreover, studies utilizing data simulation experiments have shown that the application of multiple-testing correction methods tends to decrease statistical power, particularly when a large number of tests are conducted and only a small fraction of these tests exhibit a genuine effect, indicating a true difference between the two sets of samples under investigation [133]. As a consequence, many genomic regions that are truly differentially accessible or expressed across the two conditions may remain undetected after the application of multiple-testing correction. Additionally, the use of nonparametric significance tests can further decrease the statistical power [137].

In certain scenarios, particularly during the discovery phase when the cost of false positives is deemed acceptable compared to false negatives, it may be justifiable to omit the multiple-testing correction. However, this can be compensated for by incorporating additional criteria, such as considering the effect size. Nonetheless, it remains valuable to calculate the FDR, as it provides information about the proportion of discoveries that are likely to be false positives. One approach to estimate the FDR is through the use of permutation tests. In this procedure, the labels of the samples are randomly shuffled for each genomic region. In particular, a localized prostate cancer sample label may be assigned to an advanced prostate cancer sample, or vice versa. Subsequently, a statistical significance test is performed. This process is repeated multiple times for each genomic region and all other regions. The number of tests that yield significance (i.e., P-value < α, where α is the significance threshold) under this permutation setup can provide an estimation of the number of false positives. Consequently, the false positive rate and the FDR can be estimated, enabling the establishment of criteria to ensure the FDR remains below a certain acceptable threshold [138].

2.12.7 Enrichment analysis

In computational biology, it is often of interest to assess whether a set of results from a previous analysis, such as a set of DE genes exhibit an association with a specific functionality, such as aerobic respiration. The objective is to determine whether the data suggests there is an enrichment of certain functionalities within the list of DE genes [143]. To conduct such a test, methods like Fisher's exact test or the hypergeometric test, which rely on the hypergeometric distribution, can be employed [144,145].

Let us consider an example to illustrate this concept. Suppose we have collected gene expression data for 2000 genes across two conditions, and from this dataset, we have identified 60 genes as DE genes. Furthermore, let us assume that among the 2000 genes, there are 75 genes known to be involved in aerobic respiration. If seven out of the 60 DE genes are associated with aerobic respiration, we would like to determine if aerobic respiration is enriched in our list of DE genes. These values can be represented in a tabular format known as a contingency table as shown in Table 1.

Fisher's exact test calculates the probability of observing 7 out of 60 DE genes being involved in aerobic respiration or other events that are at least as extreme, under the assumption that aerobic respiration is not enriched in the DE gene list. If this probability is below a certain significance threshold (e.g., 0.01), we can conclude that it is unlikely to observe such a result purely by chance, indicating a significant enrichment. In this example, the calculated P-value is 0.0061. Since the P-value is significantly low, it is highly improbable to observe a table as extreme as this if aerobic respiration was not enriched in the DE gene list. Thus, we can conclude that aerobic respiration is indeed enriched in our list of DE genes.

The hypergeometric distribution can also be utilized to calculate the expected value and variance for this example. The expected value is determined as 60 * (75/2000) = 2.25, and the variance is approximately 2.1. Another way to frame the question is to evaluate the likelihood of observing values as extreme as 7 purely by chance when the expected value and variance are 2.25 and 2.1, respectively.

Table 1. Contingency table for the DE gene list enrichment example.

	Involved in Aerobic respiration	Not involved in Aerobic respiration	Marginal row totals
DE genes	7	53	60
Not DE genes	68	1872	1940
Marginal column totals	75	1925	2000

It is important to acknowledge that Fisher's exact test can be computationally demanding as it involves exhaustive enumeration of all possibilities [143,145]. Alternatively, the P-value can be approximated using the chi-squared test, provided that the expected value is greater than 5 [145].

In the example mentioned earlier, we focused on testing the enrichment of a single functionality (aerobic respiration). However, in practical scenarios, multiple tests may be conducted to investigate the enrichment of various functionalities and categories. Therefore, it becomes necessary to address the issue of multiple testing and adjust the obtained P-values accordingly. Pathway analysis, also referred to as functional enrichment analysis, follows a similar approach to identify biological pathways that demonstrate enrichment among a given list of identified genes, such as DE genes.

2.13 Cancer

Cancer is a disease characterized by a diverse set of alterations in the genome and epigenome acquired via a multi-step process [3,4]. For example, tumor suppressor genes play a critical role in constraining cells from uncontrolled growth. Alterations to a particular tumor suppressor gene may occur through a variety of mechanisms, such as deletion, mutations that impair its function, or epigenetic modifications that repress its expression, including aberrant promoter hypermethylation [4,91].

Cells that have acquired such changes may become capable of uncontrolled growth and replication, as well as evading mechanisms that are in place to protect an organism, such as humans, from such behavior [3,4]. One single cell in a person with such abilities along with its progeny may potentially survive and reproduce over time to create a mass of cells or a tumor containing up to 10^12 cells. If left untreated, this can ultimately result in the death of the individual [146].

Metastasis, which is the spread of tumors to distant organs, is estimated to be the cause of 90% of deaths from solid tumors [147]. Despite the existence of multiple defense mechanisms against uncontrolled growth and replication, cancer remains a common disease, and it is estimated that one in five humans will die from cancer [3,146].

Over the years, multiple techniques with increasing resolution have been developed to study cancer-related alterations in DNA, such as chromosomal metaphase analysis, fluorescence in situ hybridization (FISH), microarray-based comparative genome hybridization, and PCR amplification followed by Sanger sequencing [148]. However, these techniques had certain limitations, such as scalability issues due to the need for a large amount of tumor tissue. High-throughput technologies have improved on these techniques, allowing for the study of the cancer genome, epigenome, transcriptome, and proteome with increased resolution and sensitivity, and facilitating the discovery of alterations that enable cells to survive and reproduce limitlessly [148].

In 2008, Ley and colleagues were among the first scientists to use high-throughput WGS technology to study the cancer genome of acute myeloid leukemia [149]. Since then, high-throughput technologies have been used to study and characterize different cancers at various omics levels. Measurements from the genome, epigenome, transcriptome, and proteome have been utilized in the context of prostate cancer (see the following section for some examples). As an example, the loss of DNA methylation as well as hypermethylation of CpG islands have been implicated in cancer [17,32] or it has been used to classify tumors [98].

The exploration of alterations that contribute to the onset and progression of cancer, especially through integrated analysis of various omics data, holds immense potential. Not only does it pave the way for the development of better treatments, but it also enhances our understanding of the fundamental principles of cell biology [146,148].

2.14 Prostate cancer

Prostate cancer ranks globally as the third most commonly diagnosed cancers in 2020 [2]. In the same year, it was estimated to be the sixth leading cause of cancer death in men [150].

Prostate cancer is typically diagnosed through a combination of prostate-specific antigen (PSA) blood tests and digital rectal examinations (DREs), followed by a biopsy of the prostate gland [35]. The PSA blood test measures the levels of PSA, which is synthesized by the prostate gland and is encoded by the *KLK3* gene in humans. Elevated PSA levels can be a sign of prostate cancer, although other factors such as age and inflammation can also affect PSA levels. During a biopsy, small samples of tissue are taken from the prostate gland using a thin needle and examined under a microscope to check for the presence of cancer cells.

Prostate cancer can be broadly classified into two main categories: primary prostate cancer (PC) and castration-resistant prostate cancer (CRPC). The latter is considered a more aggressive form of prostate cancer. During the early stages of prostate cancer, the tumor typically remains confined to the prostate gland without spreading to other areas of the body. This stage is referred to as localized prostate cancer. Various treatment options are available for localized prostate cancer, including active surveillance (close monitoring of cancer progression), radical prostatectomy (surgical removal of the prostate gland), and ablative radiotherapy (targeted destruction of cancer cells) [35]. If the localized prostate cancer diagnosed at an early stage, the life expectancy for 99% of men with this disease is over ten years [35].

If prostate cancer recurs or relapses after prostatectomy, treatment options may involve salvage radiotherapy and/or androgen deprivation therapy (ADT) for local relapse. For systemic relapse, ADT may be combined with chemotherapy or novel androgen signaling-targeted agents [35]. ADT serves as the cornerstone in the treatment of prostate cancer. However, in some cases, prostate cancer becomes resistant to ADT, leading to the progression of advanced prostate cancer. At this stage, prostate cancer is classified as castration-resistant and considered incurable [35]. Figure 4 provides a visual summary of the described steps.

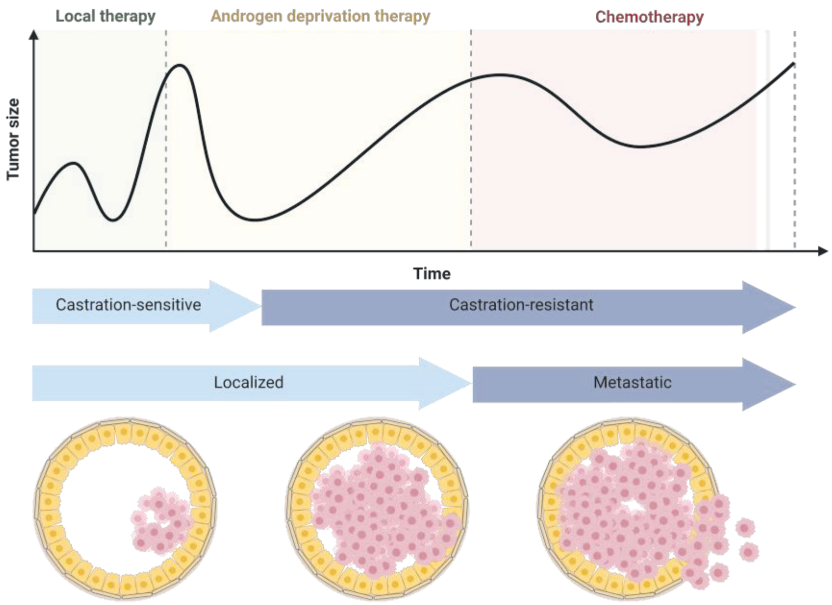


Figure 4. Prostate cancer progression across different stages. Figure created with BioRender.com.

The analysis of data across various omics levels has not only confirmed previous findings obtained by non-HTS methods but has revealed novel factors that play a role in the development and progression of prostate cancer. These include genetic susceptibility factors [151–153], alterations in the genome [154–164], epigenome [36,103,157,160], transcriptome [157,160,161,163,165–168], and proteome [157,169–171]. Here, we provide some examples of these discoveries from different levels of omics analysis.

Genomic alterations are commonly found in prostate cancer, encompassing genomic deletions, amplifications, and mutations. Notable examples of these alterations include the deletion of *NKX3.1* (8p copy number loss), *PTEN*, and *TP53*, as well as the amplification of the *AR* and *MYC* (8q copy number gain). Additionally, ETS family gene rearrangements, such as TMPRSS2-ERG gene fusion resulting from a deletion on chromosome 21q22, are also observed. More than 200 risk regions associated with prostate cancer have been identified through genome-wide association studies (GWASs). These regions contain germline causal variants that disrupt the normal regulation of nearby target genes in tissues relevant to the prostate [172]. Several somatic mutations have been identified in genes such as *AR*, *SPOP*, *FOXA1*, *TP53*, *IDH1*, and various chromatin- and histone-modifying genes, including *MLL2* [154–156,173–176].

According to the Pan-Cancer Analysis of Whole Genomes (PCAWG) study, *FOXA1*, *TP53*, and *SPOP* are the three most common driver genes in prostate cancer [177]. Driver gene mutations are recognized for their critical role in the initiation and progression of cancerous cells.

Epigenome analysis has also enhanced our understanding of prostate cancer. Notably, through the examination of DNA methylation in primary prostate cancer, alongside other omics data types, the TCGA consortium has unveiled distinctive molecular subsets within this disease. This analysis has revealed the existence of various subtypes based on DNA methylation levels, particularly in the *ERG* fusion-positive subgroup [157].

Another notable example involves the TCGA consortium's utilization of ATAC-seq to examine samples from 23 different cancer types, including primary prostate cancer. This analysis yielded a comprehensive catalog of loci exhibiting accessible chromatin, which are considered as DNA regulatory elements [103].

Prostate cancer is dependent on AR, a master transcription factor in the prostate. AR is essential for promoting the growth and survival of prostate cancer cells. Due to this, researchers consider prostate cancer as an epigenetic disease [36].

Transcriptome analysis has contributed to the characterization of prostate cancer subtypes and the discovery of novel transcripts associated with the disease. For example, Glinsky and colleagues employed microarray-based gene expression data to classify prostate cancer patients into different subgroups [167]. Building upon the work of Glinsky et al. and other similar studies, Tomlins and colleagues identified another prostate cancer subgroup, which is driven by *SPINK1* outlier expression [168]. Furthermore, Ylipää and colleagues utilized transcriptome analysis to establish the oncogenic role of *PCAT5*, a long non-coding RNA [165]. Annala and colleagues leveraged the same data to identify *SKIL* as another oncogene [166].

The application of proteome analysis has enhanced our understanding of prostate cancer by uncovering deregulated pathways and identifying proteins specifically overexpressed in tumors [178]. Notably, Iglesias-Gato and colleagues generated the first large-scale proteome dataset comprising over 9000 proteins from 28 primary, localized prostate cancer tumor samples and 8 normal samples [170]. Through the analysis of this dataset, they identify tumor-specific overexpressed proteins involved in anabolic processes such as fatty acid synthesis [170]. In another study, Iglesias-Gato et al.

expanded their investigation to include a dataset of over 5000 proteins from 22 bone metastatic prostate cancer samples [171]. Their analysis revealed overexpression of proteins involved in cell-cycle progression and DNA damage response, along with the underexpression of proteins involved in cell adhesion [171].

Studies that have integrated data from different omics levels have shown that the transcriptome is a poor predictor of the proteome due to a lack of strong correlation between these two data types [169]. To achieve their findings, several studies have adopted an integrative approach, combining data from multiple omics levels [154,157,160,161,163,169,171,176]. As an example, the TCGA consortium integrated DNA methylation data with mRNA expression data, leading to the identification of 164 epigenetically silenced genes. Among these genes were those known to be downregulated in metastatic prostate cancer, as well as genes involved in prostate organ development [157]. In another example, Sinha and colleagues utilized proteomic data to classify primary prostate cancer into five distinct subtypes, independent of subtypes inferred from other omics data [169,178].

Despite substantial progress in identifying alterations and understanding the mechanisms underlying prostate cancer, there remain knowledge gaps, particularly concerning the progression of the disease towards the castration-resistant state. This advanced stage is currently deemed incurable, necessitating further research to gain a comprehensive understanding of its underlying mechanisms. The development of effective therapeutic interventions hinges on bridging these gaps. Therefore, the aim of this work is to contribute towards closing these gaps by uncovering novel alterations that play a role in the disease. This will be achieved through the development and utilization of computational and statistical tools and methods, which will enable the proper analysis of multilevel high-throughput data on prostate cancer.

3 AIMS OF THE STUDY

As previously mentioned, a subset of prostate cancer patients experiences disease progression to an advanced stage known as CRPC, despite their initial positive response to therapy. Unfortunately, this stage of the disease is currently deemed incurable. Hence, the purpose of this study is to enhance our understanding of the alterations and molecular mechanisms that drive this progression, with the hope that the insights gained from this study can contribute to the development of effective treatments.

We hypothesize that through a comprehensive and robust analysis and integration of high-throughput omics data, collected from a unique cohort of prostate cancer patients, which represent different stages of the disease, we can identify additional key alterations that drive the progression of prostate cancer. To achieve this goal, this study focuses on addressing the following two specific aims:

Aim 1: Develop novel computational and statistical tools and methods for the effective and efficient analysis of high-throughput, omics data.

Aim 2: By utilizing the developed tools and methods from Aim 1, investigate both single and multilevel high-throughput omics data in order to identify key alterations that drive the progression of prostate cancer.

All of the Articles have contributed to the accomplishment of Aim 1 in this study. Additionally, Articles II and III have contributed to the achievement of Aim 2 in this study.

4 MATERIAL AND METHODS

4.1 Material

4.1.1 Tampere prostate cancer cohort

The Tampere prostate cancer cohort includes fresh-frozen tissue specimens from three different groups of patients. Untreated primary prostate cancer (PC), locally recurrent castration-resistant prostate cancer (CRPC), and benign prostatic hyperplasia (BPH) groups. BPH is used as the control group and as a model for the normal prostate. This is because BPH is similar to PC in terms of histological, pathological, and genetic characteristics [179]. Thus, these three groups can be used to track prostate cancer development and progression to castration-resistance.

The samples were collected at the Tampere University Hospital (Tampere, Finland) either via radical prostatectomy (RP) or transurethral resection of the prostate (TURP). All samples were examined microscopically and found to contain at least 70% cancerous or hyperplastic cells. These samples were used to produce genomics (low-coverage DNA-seq, 4X), epigenomics (MeDIP-seq and ATAC-seq), transcriptomics (RNA-seq and miRNA-seq), and proteomics (SWATH-MS) high-throughput data. Figure 5 illustrates the multilevel dataset used in this work.

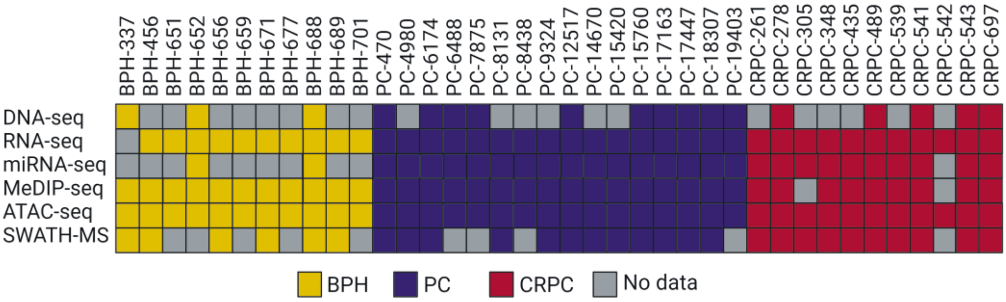


Figure 5. Multilevel dataset used in Article III. Figure created with BioRender.com.

In Articles II and III, high-throughput data from PC and BPH samples were compared to identify early alterations that occur during prostate cancer development. Additionally, the comparison between CRPC and PC samples aimed to identify alterations associated with cancer progression and castration resistance.

4.1.2 The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) program was launched in 2006 as a collaborative initiative between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) in the United States. The program conducted molecular characterization of over 20,000 cancer samples, along with their corresponding matched normal from 33 different cancer types. The TCGA program generated an extensive collection of data across multiple molecular levels, including genomics, epigenomics, transcriptomics, and proteomics [180].

In Article I, a subset of genomic data from the TCGA was used, focusing on prostate adenocarcinoma and low-grade glioma. The WGS data was used to evaluate the performance of our developed tool, `Segmentum`, in identifying SCNAs using actual WGS data (the list of sample barcode names used for the analysis can be found in Appendix 1).

Moreover, the TCGA program has published numerous peer-reviewed scientific articles that document their comprehensive analyses of the TCGA dataset. These publications provide a valuable resource for conducting integrative analysis. In Article III, we compared our findings derived from the analysis of chromatin accessibility data obtained from the Tampere prostate cancer cohort with the findings from TCGA-generated chromatin accessibility data [103].

4.1.3 miRWalk 2.0 database

miRWalk2.0 is a comprehensive and freely available database that contains experimentally validated and predicted miRNA-target interactions [181]. In Article II, we used miRWalk2.0 database to generate a list of predicted mRNA targets of miRNAs.

4.1.4 Gene Transcription Regulation Database

The Gene Transcription Regulation Database (GTRD) is a collection of uniformly processed ChIP-seq data used to identify TF binding sites within the human and mouse genomes. As of the time of writing, this database contained data from thousands of ChIP-seq experiments, encompassing information on hundreds of unique TFs [182]. In Article III, we used information provided by the GTRD database to annotate genomic intervals that were identified during the analysis of ATAC-seq data derived from the Tampere PC cohort.

4.1.5 GeneHancer

In Article III, we utilized GeneHancer, a database of human enhancers and their corresponding target genes. GeneHancer integrates data from various sources including the ENCODE project, the Ensembl regulatory build, the functional annotation of the mammalian genome (FANTOM) project, and the VISTA Enhancer Browser [183]. This resource allowed us to annotate the genomic loci identified through the analysis of ATAC-seq and MeDIP-seq data obtained from the Tampere PC cohort.

4.1.6 Other datasets

In Article III, we used data obtained from multiple external publications to guide our analysis and enhance the annotation and assessment of our findings. Specifically, we incorporated the following information:

Massie et al. (2011) provided a list of topologically associating domain (TAD) boundaries inferred from the LNCaP cell line as well as a list of AR binding sites obtained from ChIP-seq experiments conducted in LNCaP and VCaP cell lines [184].

Pomerantz et al. (2015) provided a list of AR binding sites derived from ChIP-seq experiments performed on prostate cancer samples [185].

Pomerantz et al. (2020) provided AR binding sites and histone modifications lists obtained from ChIP-seq experiments, as well as a list of accessible chromatin regions derived from ATAC-seq experiments conducted on samples across different stages of prostate cancer [36].

Stelloo et al. (2018) for the list AR binding sites and histone modifications obtained from ChIP-seq experiments conducted on primary prostate cancer samples [186].

4.2 Methods

4.2.1 Somatic copy number alteration detection

In Article I, we introduced `Segmentum`, a tool we developed for the detection of SCNAs in cancer samples using WGS data. `Segmentum` was implemented using Python version 3. The overall workflow of `Segmentum` is shown in Figure 6.

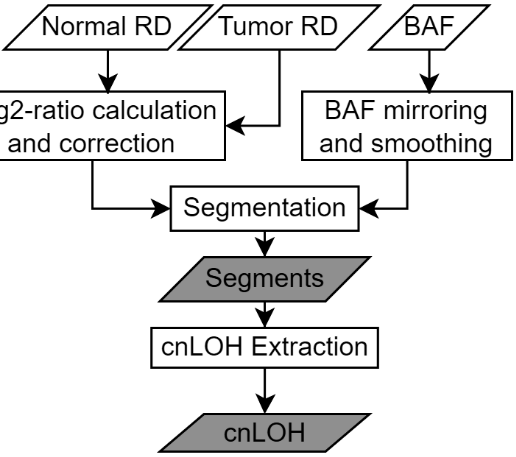

Normal RD
Tumor RD
BAF
Log2-ratio calculation and correction
BAF mirroring and smoothing
Segmentation
Segments
cnLOH Extraction
cnLOH

Figure 6. Overall workflow of `Segmentum`. RD, read depth; BAF, B-allele fraction; cnLOH, copy-neutral loss of heterozygosity. Figure created with app.diagrams.net.

In `Segmentum`, the ratio of read counts between tumor and its matched normal samples is calculated for windows of a specified size determined by the user across the genome. This ratio then undergoes a logarithm base 2 (log2) transformation using:

$$logr_i \ = \ \log_2(\frac{tC_i}{nC_i}),$$

where *logr_i* is the log2-ratio of the *i*th genomic window and *tC_i* and *nC_i* are the read counts from the *i*th window for the tumor and normal samples, respectively.

In addition to the log2-ratios, `Segmentum` also calculates genome-wide BAFs with:

$$BAF_i \ = (\frac{A_i}{A_i + B_i}),$$

where *BAF*$_i$ is the BAF value for the *i*th heterozygous SNP, and *B*$_i$ and *A*$_i$ refer to the alternative and reference allele respectively, of the *i*th heterozygous SNP.

The calculated BAFs undergo simultaneous mirroring and smoothing. Mirroring the BAFs around the 0.5 axis simplifies BAF data analysis by assigning a value of 0 to heterozygous SNPs. Deviations from the expected heterozygous SNP BAF value of 0.5 are assigned values larger than 0, rather than a value between 0 and 1 (excluding 0.5). Smoothing is performed using a median filter, which helps mitigate the impact of outlier values on the BAF signal profile and effectively reduces noise. The simultaneous mirroring and smoothing are done using:

$$cBAF_i = (H \times |0.5 - \mathrm{M}_9(BAF_i)|) + ((1-H) \times \mathrm{M}_9(|0.5 - BAF_i|)), \text{ and}$$

$$H = 1 - 2 \times |0.5 - BAF_i|,$$

where *cBAF*$_i$ is the simultaneously mirrored and smoothed *BAFi* and M$_9$() is a median filter function that calculates the median value by considering nine estimated *BAF*s in the vicinity of and including the *i*th SNP.

To detect breakpoints, a double sliding window is then applied to the estimated *logr*s and *cBAF*s. The user specifies the size used for both windows within the double sliding window. The location of a breakpoint is determined by calculating a value *S* using the following formula. If the calculated value of *S* exceeds 1, a breakpoint is placed at the center of the double sliding window.

$$S = \frac{|\overline{logr_{win_i}} - \overline{logr_{win_{i+1}}}|^2}{\tau_{logr}} + \frac{|\overline{cBAF_{win_i}} - \overline{cBAF_{win_{i+1}}}|^2}{\tau_{BAF}} \equiv \frac{(\Delta logr)^2}{\tau_{logr}} + \frac{(\Delta BAF)^2}{\tau_{BAF}},$$

where $\overline{logr_{win_i}}$ and $\overline{cBAF_{win_i}}$ are the mean of the log2-ratios and *cBAFs* falling in the left window within the double sliding window at the *i*th position respectively. $\overline{logr_{win_{i+1}}}$ and $\overline{cBAF_{win_{i+1}}}$ are the mean of the log2-ratios and *cBAFs* falling in the right window within the double sliding window at the *i+1*th position respectively. User-provided thresholds, τ_{logr} and τ_{BAF}, are the thresholds for the absolute mean difference in the log2-ratios and the *BAFs* between the two adjacent windows, respectively. Figure 7 depicts the decision boundary for breakpoint detection, as implemented by the formula described above.

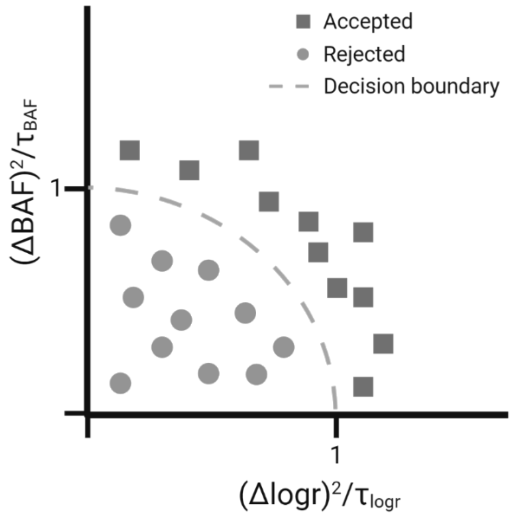


Figure 7. Decision boundary using two criteria for detecting breakpoints.

A segment is defined by the genomic region bounded by two breakpoints. To characterize each segment, the average log2-ratios and average *cBAFs* within that segment are calculated and reported. The average log2-ratio of a segment represents the relative copy number of that specific genomic region. Segments with log2-ratios close to 0 and *cBAFs* close to 0.5 are considered as cnLOH.

4.2.2 Benchmarking Segmentum

To evaluate the accuracy of `Segmentum` in detecting SCNAs, we conducted a series of benchmark experiments. One such experiment involved the analysis of WGS data obtained from one glioma sample, a prevalent type of brain tumor. The WGS data had an average coverage depth of approximately 46X. For this particular benchmark experiment, we generated five distinct subsamples of the original sample using `samtools`, a widely used genomics research tool [187]. These subsamples represented different fractions of the original sample, namely 75%, 50%, 25%, 10%, and 5%. Subsequently, we applied `Segmentum` to analyze each of these subsamples. To validate the results, we compared them against a ground truth derived from the TCGA SNP array data, specifically the Affymetrix Genome-wide Human SNP array, which was obtained from the same sample. To quantify the similarity between the results obtained by `Segmentum` and the ground truth, we calculated the Jaccard similarity index (JSI). The JSI is a measure of the similarity between two sets, and it can be used to compare the overlap or similarity between two datasets such as the results from `Segmentum` against the ground truth. The JSI ranges from 0 to 1, where 0 indicates no similarity between the sets, while 1 signifies complete similarity or identical sets.

We further benchmarked `Segmentum`, comparing it with other state-of-the-art tools available at the time of publication, including `Patchwork`, `CLImAT`, and `control-FREEC` [188–190]. The benchmarking process involved applying these tools to real data with varying depth of coverages. We used ten high-coverage TCGA glioma samples (30X to 100X average depth of coverage), and ten low-coverage TCGA prostate adenocarcinoma samples (6X average depth of coverage). To assess the performance, we compared the results of each tool to a ground truth derived from TCGA SNP array data. The comparison was done by calculating the JSI.

In order to further evaluate `Segmentum`'s ability to detect SCNAs in the presence of normal cell contamination, we developed a sequencing read count simulator. This simulator generates whole-genome read counts for both normal and tumor samples, as well as BAF based on deletion, amplification, and cnLOH events. We then applied `Segmentum` to the simulated data, varying the fractions of normal cell contamination.

To assess `Segmentum`'s performance, we calculated metrics such as precision, recall (sensitivity), and F-measure (F1 score). Precision measures the accuracy by calculating the proportion of correctly predicted positive instances out of all instances predicted as positive. Recall, on the other hand, assesses the ability to correctly identify positive instances by calculating the proportion of correctly predicted positive instances out of all actual positive instances. The F measure combines precision and recall into a single metric using their harmonic mean, providing a balanced assessment of the tool's performance.

4.2.3 Enrichment and pathway analysis

In Article II, we utilized Fisher's exact test to examine whether the evidence supports the hypothesis that a specific DE miRNA negatively regulates a greater number of genes within a set of DE genes than would be expected by chance alone. Furthermore, in Article II, we conducted pathway analysis by utilizing Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City, USA) in order to identify proteins associated with relevant biological pathways of interest.

4.2.4 ATAC-seq data processing and peak calling

After quality control using `FastQC` and trimming the ATAC-seq sequencing reads with `TrimGalor`, the reads were aligned to the GRCh38 reference genome using `Bowtie2` aligner [69]. The resulting aligned reads underwent preprocessing using `samtools` to filter out reads with a MAPQ below 20, and `Picard Markduplicates` to mark the duplicate reads [67,191]. Subsequently, peaks were called using `MACS2` [102,192]. A second round of quality control was performed using the `ataqv` tool, to assess the quality of the called peaks and other relevant quality metrics [193]. Additionally, regions on autosomal and sex chromosomes affected by the alignment of mitochondrial DNA were identified, and peaks overlapping these regions were filtered out.

4.2.5 ATAC-seq signal quantification

To robustly quantify the ATAC-seq signal across the genome, we developed the following approach. Initially, we divided the genome into overlapping bins of size 500 bp with steps of 250 bp. To count the number of overlapping sequencing reads in each bin, we utilized the `bedtools coverage -counts` subcommand [194].

Subsequently, we implemented several correction steps to address various biases and normalize the data. The initial step was background correction, which addressed local background biases such as copy number alterations [102]. This correction accounted for situations in which accessibility could not be detected in deleted genomic regions due to the absence of sequencing reads, or where amplification influenced the measurement in amplified genomic regions. By decoupling these factors, the background correction enabled us to focus on the accessibility itself. To apply the background correction, we devised the following formula:

$$\mathrm{c}(x) = \max(0, (\mathrm{R}(x) - \max(\mathrm{Q}_1(\mathrm{P}_{10}(x)), \mathrm{Q}_1(\mathrm{P}_{100}(x)), (\mathrm{P}_{\mathrm{chr}}(x))))),$$

where R(x) represents the read count for the bin located at position x. The terms $\mathrm{P}_{10}(x)$ and $\mathrm{P}_{100}(x)$ denote the lists of read counts encompassing all bins within a range of +/-5 kilobases and +/-50 kilobases around position x, respectively. Furthermore, $\mathrm{P}_{\mathrm{chr}}(x)$ denotes the list of read counts for all bins within a specific chromosome arm. $\mathrm{P}_{10}(x)$, $\mathrm{P}_{100}(x)$, and $\mathrm{P}_{\mathrm{chr}}(x)$ exclude the bin at position x. Additionally, Q_1 denotes the value corresponding to the first quartile.

After applying the background correction, we utilized the median of ratio normalization method to obtain normalized read counts [109]. This step ensures that samples with varying sequencing depths can be effectively compared.

Additionally, we performed sample collection procedure bias correction. This correction accounted for potential biases introduced by collecting samples using two different procedures: radical prostatectomy (RP) and transurethral resection of the prostate (TURP).

To correct for sample collection procedure bias, we divided the samples into two groups: RP (consisting of four BPH and four PC samples) and TURP (comprising four BPH and four CRPC samples). For each bin, we conducted a two-sided Wilcoxon rank-sum test and recorded the resulting P-value. This procedure was repeated 100 times, with the samples assigned to each group shuffled each time. Bins with a P-value of ≤0.01 in 5% or more iterations were identified. For the identified bins, we subtracted the difference between the medians of the normalized read counts of all the TURP samples and all the RP samples from the read counts of all the TURP samples.

The quantified ATAC-seq signal was utilized for various purposes in Article III. These include the detection of DARs, performing correlation analysis with gene expression and protein abundance data, and visualization of the results.

4.2.6 ATAC-seq consensus peak set compilation and quantification

In order to construct and quantify a consensus ATAC-seq peak set from the peaks identified across multiple samples using `MACS2`, we made modifications to the approach initially proposed by Corces et al., enabling the integration of our correction procedure [103]. The following steps were employed:

1. The peak summits in each sample were extended by ±250 bp, resulting in 500 bp wide peak windows.
2. The quantified ATAC signal, as described earlier, was used to quantify the peaks within these windows.
3. In instances where overlapping peaks were present within a sample, the peak with the highest signal was selected.

4. To ensure comparability between samples, the peak signals in each sample were scaled using a scaling factor. This factor was determined by calculating the sum of all peak signals in a sample and dividing it by 10^6.
5. All peaks from the individual samples were pooled together, and overlapping peaks were removed, retaining the peaks with the highest scaled signal value.
6. Finally, peaks that were supported by only one sample were filtered out, resulting in the generation of a consensus peak set.

For each individual sample, the quantification of the consensus peak set was carried out using a similar approach as described in the *ATAC-seq signal quantification* section. This quantification was utilized to discard peaks if all samples had a signal below a data-driven threshold of 5, which was calculated using the elbow method.

The peak set was subsequently divided into seven groups, categorized based on the sample types that exhibited a peak at the location of each consensus peak. This categorization, along with the ATAC-seq signal, was utilized to study the accessibility in prostate cancer and to create visualizations to effectively represent and illustrate the data.

4.2.7 MeDIP-seq signal quantification

Following quality control using `FastQC`, the sequencing reads from MeDIP-seq were aligned to the GRCh38 reference genome using the `Bowtie2` aligner [69]. The resulting aligned reads were then preprocessed using `samtools` and `Picard Markduplicates` to eliminate duplicate reads [67,191].

To robustly quantify the MeDIP-seq signal, a similar procedure as described for ATAC-seq signal quantification was employed. This involved partitioning of the genome into bins, performing background correction to address local biases, and applying median of ratio normalization.

4.2.8 Detection of differential genomic features between two conditions

In this study, we employed a specific strategy to identify genomic features that exhibit differential expression, accessibility, or methylation between two groups under different conditions. Our strategy incorporates three criteria to select such features.

The first criterion involves utilizing the nonparametric Wilcoxon rank-sum test to determine if the data from the two groups originate from distinct distributions. Essentially, this test assesses the significance of the observed difference between the medians of the two samples [138]. In addition to the statistical hypothesis test, we take into consideration the effect size, which denotes the magnitude of the difference between the medians of the two groups [135,138]. Consequently, the second criterion examines whether the magnitude of the median difference exceeds a predetermined threshold.

Lastly, the third criterion examines whether the absolute log2-ratio of the medians between the two groups surpasses a specified threshold. This information allows for determining both the ratio and direction of the change. In the case of ATAC-seq data, if the log2-ratio of the median from group 2 to group 1 for a genomic region is +1, it suggests that the accessibility of this genomic region is twice as high in group 2 compared to group 1. To ensure symmetry of positive and negative ratios around zero, we apply a log2 transformation to the data.

The second and third criteria are complementary to each other. To illustrate, if the medians of the two compared groups are close to 0 (say 0.016 and 0.004), this might result in a considerable log2-ratio (log2(0.016/0.004) = 2) which could satisfy the threshold set for the third criterion. However, the second criterion (|0.016 - 0.004| = 0.012) would not be satisfied. These two criteria work together to reduce the false discovery rate in the absence of multiple-testing correction [195]. Although these additional criteria do not guarantee a specific level of false discovery rate theoretically, our calculations in Article III demonstrate that by appropriately setting the thresholds, the false discovery rate can be maintained below 10%. Figure 8 illustrates the effect of three criteria, namely P-value, absolute median difference, and log2-ratio, on the estimation of FDR through permutation testing. The figure focuses on the detection of DARs from ATAC-seq data obtained from two distinct prostate cancer groups, namely BPH and PC, from the Tampere PC cohort.

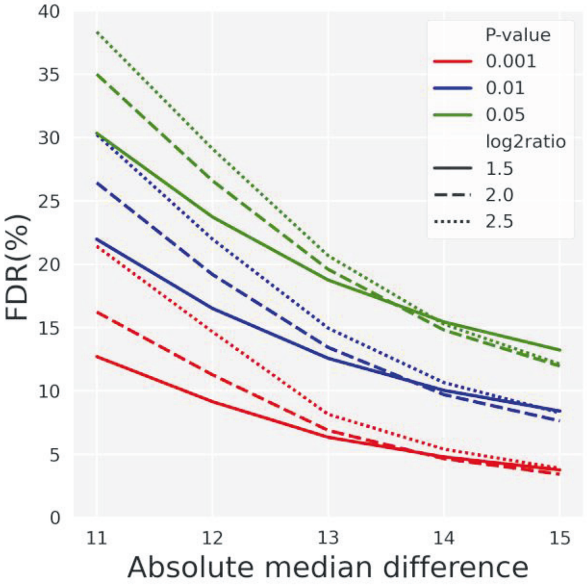


Figure 8. Impact of P-value, absolute median difference, and log2-ratio on FDR in detecting DARs from ATAC-seq data between BPH and PC prostate cancer groups.

In Articles II and III, we applied this strategy to identify DE genes at the transcriptome and proteome levels, as well as differentially accessible regions (DARs), and differentially methylated regions (DMRs). In Article III, when identifying DARs and DMRs, we determined and set the threshold values for the aforementioned criteria (through performing permutation tests, as explained earlier) to maintain the FDR within reasonable limits.

DARs were identified using the following criteria: |log2-ratio| > 2, P-value < 0.01, absolute-median-difference > 14. These criteria corresponded to FDR of 9.7% for the BPH to PC comparison and 9.14% for the PC to CRPC comparison. DARs with positive log2-ratio were classified as opening DARs, while DARs with negative log2-ratio were classified as closing DARs.

DMRs were identified using the following criteria: |log2-ratio| > 2, P-value < 0.01, absolute-median-difference > 10. These criteria corresponded to FDR of 4.61% for the BPH to PC comparison and 7.90% for the PC to CRPC comparison. DMRs with positive log2-ratio were classified as hypermethylated DMRs, while DMRs with negative log2-ratio were classified as hypomethylated DMRs.

In Article II, we initially identified the common genes in both the transcriptome and proteome datasets, resulting in a total of 3310 genes. Subsequently, employed specific criteria to identify DE genes, utilizing a threshold of |log2-ratio| > 1.5 and an adjusted P-value < 0.05 obtained from a non-parametric Wilcoxon test. Similar criteria values were used to find DE miRNAs.

It should be noted that several other methods have been proposed for detecting DE genes between two conditions. Notably, DESeq [108,109] and edgeR [110] are two widely used methods in this context. The principles introduced and utilized by these methods have also been adapted to detect differential genomic regions, such as in terms of methylation [196] or protein-DNA binding [197,198]. In our approach described above, we incorporated the normalization step proposed in DESeq, specifically the median of ratio normalization, to ensure comparability between samples [109].

4.2.9 Data integration

Our work, particularly in Articles II and III, aligns with the sequential, phenotype-first approach, as explained in the literature review. We focus on prostate cancer progression as the phenotype of interest and apply data integration techniques on multilevel omics data to characterize this phenotype.

In a previous study, conducted on the Tampere PC cohort, RNA-seq was used to identify somatic point mutations in PC and CRPC samples, as well as germline point mutations in BPH samples within transcribed loci. These variants were further validated through targeted sequencing [166]. Building upon this research, in Article II, we investigated the impact of these point mutations on gene and protein expression of the genes with these mutations. To assess the impact, we employed an impact score calculation method as described by Zhang et al. [199]. The impact score was determined using the following formula:

$$Score = (EXP - Median_{non-SNV}) \,/\, MAD_{non-SNV},$$

where *EXP* represents the expression level of the gene or protein in the sample containing the specific SNV. *$Median_{non-SNV}$* and *$MAD_{non-SNV}$* refer to the median and median absolute deviation of the expression levels of the gene or protein across all samples without the SNV, respectively.

In Article II, we investigated the impact of alterations in copy number on gene/protein expressions by utilizing copy number estimates in terms of log2-ratios obtained from a previous study [165]. The analysis involved a dataset comprising 3185 common genes and 23 common PC and CRPC samples. For each sample, we computed the Spearman correlation between the copy numbers and gene/protein expressions. The resulting correlations were then visualized to provide a clear representation of the findings.

Similarly, we examined the impact of alterations in methylation on gene/protein expressions by utilizing DMRs obtained from a previous study [165]. Specifically, we focused on genes that had a DMR within a 10 kilobases vicinity. The analysis involved a dataset consisting of 751 common genes and 25 common PC and CRPC samples. For each sample, we computed the Spearman correlation between the DMR log2-ratios and gene/protein expressions. These correlations were subsequently visualized to provide a clear representation of the findings.

In Article III, we utilized DNA-seq to find copy number alterations in the genome, using `Segmentum`, the tool we developed in Article I. Additionally, we employed MeDIP-seq and ATAC-seq data to detect epigenetic alterations such as DARs and DMRs across different groups, that corresponded to the observed changes in gene expression. Furthermore, we investigated how these alterations manifested in the proteome using correlation analysis and other techniques.

To complement our analysis, we incorporated external datasets into our study. To illustrate, in Article III, we identified a few thousands of DARs across BPH, PC, and CRPRC groups. To gain insights into the functionality of these loci, we leveraged external resources such as GeneHancer to identify potential enhancer regions. We also employed data and results from other relevant prostate cancer studies to annotate and assess our findings. Moreover, we annotated these sites with experimentally validated transcription factor binding sites (TFBS) from the GTRD database. To accomplish this, we employed various tools such as `bedtools` and Hypergeometric Optimization of Motif EnRichment (HOMER) [194,200].

In Article II, we employed the miRWalk2.0 database to compile a comprehensive list of experimentally verified and predicted mRNA targets of miRNAs. Subsequently, we integrated multiple datasets including RNA-seq, miRNA-seq, MeDIP-seq, and proteomics data from the Tampere prostate cancer cohort with this list. Our objective was to identify miRNAs that hold particular relevance in the context of prostate cancer. Specifically, we used the information on mRNA and miRNA expression levels, protein abundances, their DE status, and the list of miRNA targets to guide our correlation analysis. We initiated the analysis by considering DE miRNAs and their corresponding target list. We then calculated the Spearman correlation between the expression levels of DE miRNAs and their target genes, either in terms of mRNA or protein levels, and focused on those exhibiting a negative correlation greater than -0.50. Furthermore, we performed an enrichment analysis using a hypergeometric test (with a significance threshold of P-value < 0.05). The purpose was to identify DE miRNAs that demonstrated a significant enrichment in the number of negatively correlating mRNAs when compared to the total number of the targets associated with the miRNA under study.

Additionally, we integrated information on the position of TADs obtained from cell line studies. This integration allowed us to limit the number of correlation calculations between DARs and gene expressions, thus mitigating the issue of multiple testing.

4.2.10 Data visualization

In this study, we employed various visualization techniques to effectively convey the information. Multiple visualizations, including PCA, t-SNE, hierarchical clustering plots, and boxplot showing coverage at peaks, DARs, and DMRs were generated to provide clear visual representations of the data. To accomplish this, we primarily utilized the Python programming language and widely used visualization modules such as `matplotlib` to create the visualizations and present our results [201].

5 RESULTS

In this chapter, we present the findings of this work. The findings are arranged in the same sequence as we introduced the omics levels in the literature review.

5.1 Alterations in the genome

5.1.1 Somatic copy number alterations

SCNAs contribute to genome instability, which is an enabling characteristic for acquiring other cancer hallmarks [3,4]. This phenomenon is particularly notable in prostate cancer, where a subset of patients with advanced disease undergoing ADT display large copy number amplifications within the *AR* gene locus and its enhancer region [202]. Consequently, multiple tools and techniques have been developed to identify and characterize these alterations, each varying in accuracy, usability, speed, and integrability with other tools. Thus, our aim was to develop a tool that is accurate, user-friendly, efficient, and capable of seamless integration with other existing tools when creating computational pipelines.

To fulfill this aim, we developed `Segmentum` in Article I, a computational tool designed to detect SCNAs, including cnLOH events, by analyzing WGS data from tumor and matched normal samples. Additionally, `Segmentum` can detect recurrent alterations across multiple samples within a cohort. The output generated by `Segmentum` is formatted as a SEG file, which can be parsed by other tools or programmatically and visualized using tools like the Integrative Genomics Viewer (IGV) [203,204].

The benchmark results from a subsampling experiment demonstrated that `Segmentum` can accurately detect SCNAs. It achieved a JSI exceeding 0.93, even with an average depth of coverage as low as approximately 4X, indicating a high level of concordance with the ground truth.

We further benchmarked `Segmentum` against other state-of-the-art tools available at the time of publication. The evaluation was based on the accuracy of the results and the analysis time requirement, using different datasets with varying depths of coverage.

For the high depth of coverage sequencing data, we utilized 10 TCGA glioma samples with an average depth of coverage ranging from 30X to 100X. The ground truths were established using TCGA SNP array data. Our results demonstrated that `Segmentum` generated highly similar results to the ground truth, with a JSI of 0.90. The next best tool achieved a JSI score of 0.86.

In the case of low depth of coverage sequencing data, we employed 10 TCGA prostate adenocarcinoma samples with an average depth of coverage of 6X. The ground truths were also derived from TCGA SNP array data. `Segmentum` ranked second in generating results most similar to the ground truth, with a JSI score of 0.88, while the best tool achieved a JSI score of 0.93.

In terms of analysis time requirement, `Segmentum` outperformed other tools. However, when considering both the average preparation time and analysis time, `Segmentum` ranked second. Nonetheless, it was twice as fast as the second-best performing tool in terms of accuracy. These findings suggest that `Segmentum` offers comparable accuracy and speed to other state-of-the-art tools available at the time of publication, while providing a simpler approach, and easier to use compared to the other tools included in the benchmark.

It is important to note that `Segmentum` does not infer the ploidy or tumor content. However, benchmarking `Segmentum` with simulated data demonstrated that `Segmentum` accurately detects SCNAs when the normal cell contamination is below 50% (or in other words the tumor purity exceeds 50%). Beyond this threshold, the ability to detect changes drops significantly.

It is worth mentioning that `Segmentum`'s output is compatible as an input to other tools such as `ABSOLUTE` [192], which is capable of inferring ploidy and tumor purity. This compatibility makes `Segmentum` suitable for integration into tumor analysis pipelines.

5.2 Alterations in the epigenome

5.2.1 Chromatin accessibility alterations

In Article III, we utilized ATAC-seq data to characterize the alterations in the chromatin accessibility in the context of prostate cancer progression. In order to accurately analyze the data and account for any systematic errors and biases, we developed a robust method for quantifying the ATAC-seq signal. This approach effectively mitigated the impact of both global background noise levels and local background factors, such as the influence of SCNAs on ATAC-seq readouts and the bias introduced by the sample collection procedure. Figure 9 illustrates the efficacy of our correction approach in mitigating the impact of SCNAs on chromatin accessibility readouts. Specifically, regions that exhibit copy number gains or amplifications in CRPC samples are appropriately corrected, resulting in a more comparable profile to that of normal regions.

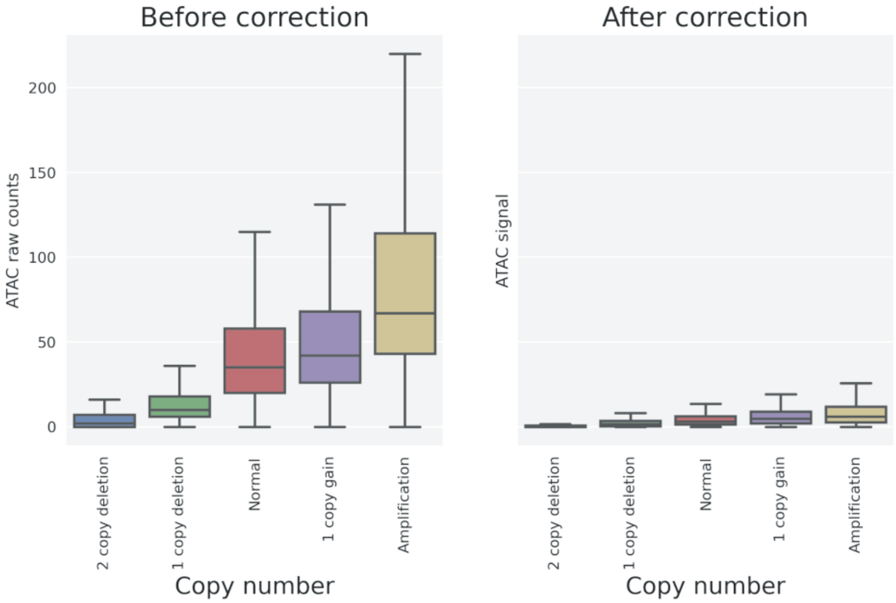


Figure 9. The correction approach developed in Article III effectively addresses the copy number alterations as intended. We utilized seven CRPC samples with available WGS data. We calculated the ATAC-signal and the raw counts at the ATAC-seq peaks set. Segmenum was used to calculate the log2 ratios. The log2 ratios were categorized into five groups assuming an 80% tumor purity.

Furthermore, we identified a range of 23,840 to 138,942 open chromatin loci or peaks per sample (refer to Figure 10A). From these, we constructed a consensus ATAC-seq peak set comprising 178,206 reproducible open chromatin regions.

When annotating these open chromatin loci using external data, we discovered a significant overlap of 79% with regulatory regions identified in normal tissues or TCGA data [103,205] (refer to Figure 10B). Furthermore, we found that 66% of the prostate cancer-specific peaks identified in the TCGA study overlapped with our accessible chromatin loci [103]. This substantial agreement with external data underscores the consistency of our findings with previous investigations. Moreover, our analysis uncovered 38,157 novel, putative, regulatory open chromatin loci relevant to prostate cancer.

Further annotation revealed that the loci commonly observed in a substantial number of samples within the cohort were predominantly located in promoter regions. Notably, 60% of the loci that displayed accessibility across all samples were located in promoters. Conversely, the loci common to only a few samples were mainly situated in intronic or exonic regions indicating their potential role in trans-regulatory functions (refer to Figure 10C). Additionally, it was observed that in all three groups, peaks within the promoter region displayed a stronger signal intensity compared to peaks in other regions (refer to Figure 10D). These observations indicate that promoters remain accessible, and their accessibility remains unchanged throughout the development and progression of prostate cancer.

Additionally, we classified the peak set into seven groups based on the sample types that had a peak at the location of the consensus peak. This analysis revealed that only a minor fraction of these loci was specific to a particular group, indicating a considerable overlap in the accessible chromatin regions across different sample types (refer to Figure 10E).

We employed the quantified signal obtained from ATAC-seq analysis and using our approach, we identified 1,727 DARs during the transition from BPH to PC, with an FDR of 9.7%. Similarly, we identified 3,498 DARs during the transition from PC to CRPC, with an FDR of 9.14%. Notably, a large proportion of these DARs were found in the intronic and intergenic regions (refer to Figure 11A).

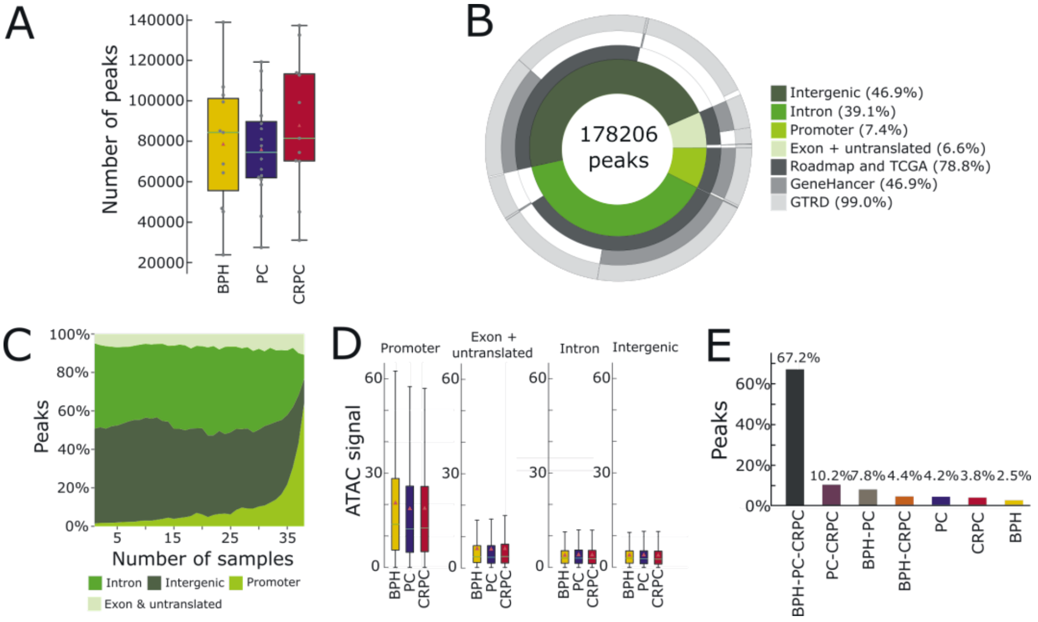


Figure 10. The landscape of prostate cancer open chromatin. **A**. The distribution of identified peak counts reveals variations in the number of peaks across different samples, both within and across groups. **B**. The annotation of peak sets not only confirms consistent results with earlier studies but also sheds light on the potential functionality of the identified open chromatin regions. **C**. Majority of the open chromatin loci common to a few samples may have trans-regulatory functions as they primarily coincide in the intergenic and intronic regions. **D**. Peaks in the promoter region exhibited stronger signal intensity compared to peaks in other regions **E**. The majority of peaks in the peak set are shared across all three prostate cancer groups.

Interestingly, we observed minimal overlap between the DARs identified during the BPH to PC and PC to CRPC transitions, comprising only 113 DARs, which accounts for approximately 2% of the total DARs. This finding suggests distinctive alterations in chromatin accessibility during the development and progression of prostate cancer (refer to Figure 11B).

To further investigate the patterns within the ATAC-seq signal across all DARs, we employed t-SNE dimension reduction technique. This enabled the separation of samples into their respective groups (refer to figure 11C).

Lastly, hierarchical clustering of the ATAC-seq signal at DARs, resulted in the division of these DARs into four distinct groups. One subgroup was specifically associated with disease progression (refer to figure 11D).

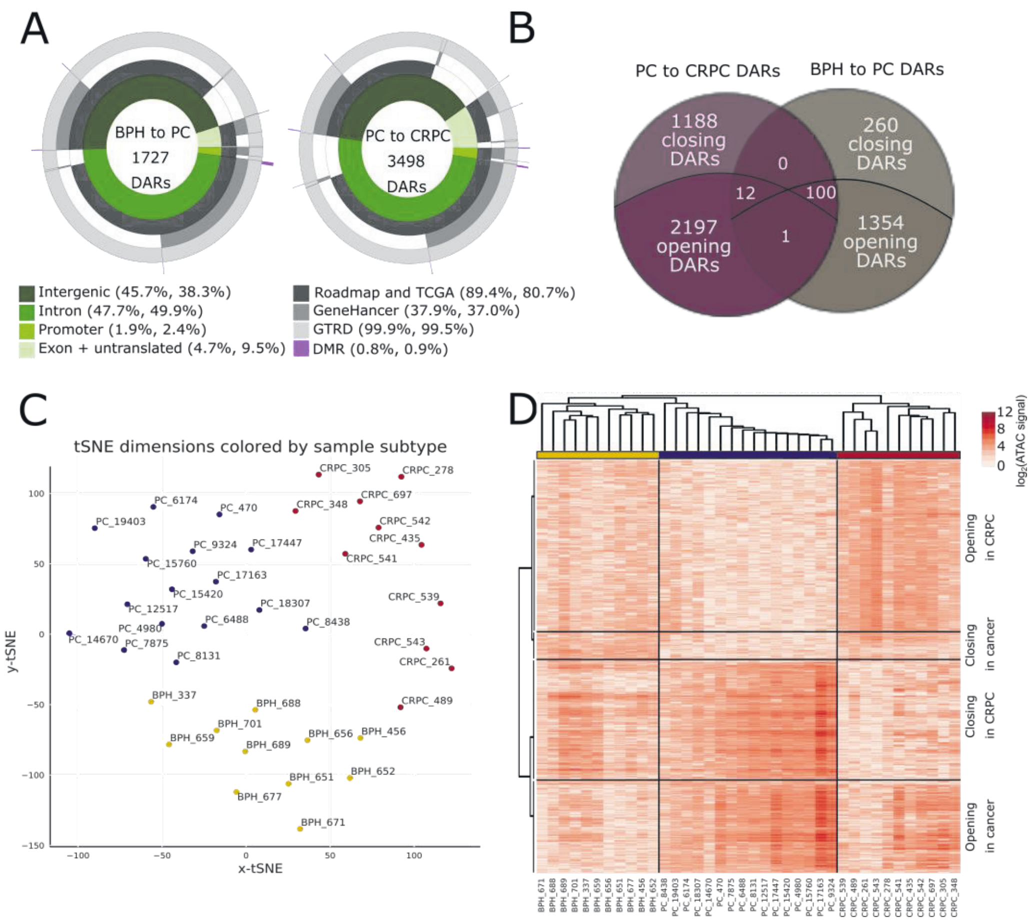


Figure 11. DARs provide insights on the development and progression of prostate cancer. DAR, differentially accessible region. **A**. DAR annotation reveals that the majority of the DARs are located in intergenic and intronic regions. **B**. Diagram reveals a limited overlap between the DARs from the two contrasts. **C**. The DARs can be used to classify samples into their respective groups using t-SNE dimension reduction. **D**. Hierarchical clustering of the DARs divides the samples into their corresponding groups and identifies prostate cancer progression-related chromatin accessibility patterns (metric: Pearson correlation; linkage: Weighted Pair Group Method with Arithmetic Mean).

5.2.2 DNA methylation alterations

Using the approaches we developed, we quantified the signal from MeDIP-seq data and identified 2,061 DMRs during the transition from BPH to PC, with an FDR of 4.61%. Similarly, we identified 2,723 DMRs during the transition from PC to CRPC with an FDR of 7.90%. These findings highlight significant DNA methylation alterations during the onset and progression of prostate cancer.

Similar to the DAR results, a large proportion of the DMRs were also found in the intronic and intergenic regions, as illustrated in Figure 12. Furthermore, the majority of these alterations are hypermethylation, particularly nearby CpG islands at promoter regions (refer to Figure 12).

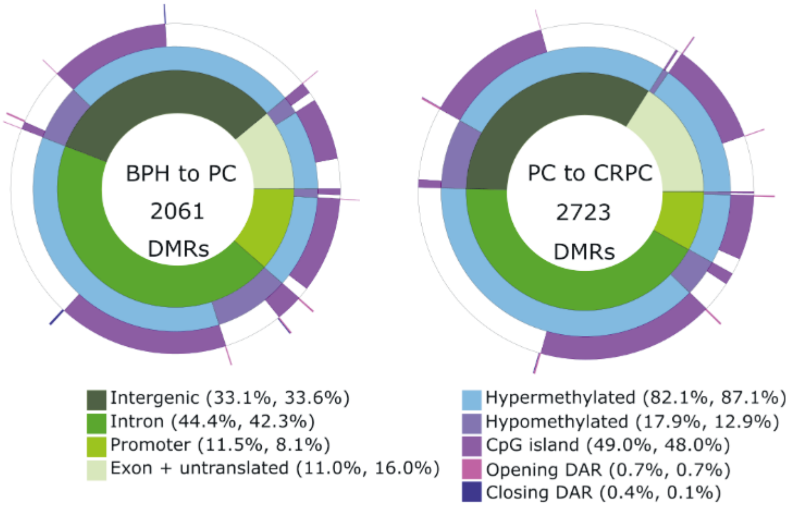


Figure 12. Hypermethylation of CpG islands during both prostate cancer onset and progression is evident. DMR, differentially methylation region.

5.3 Alterations in the transcriptome

In Article II, a total of 3310 common genes were identified in both the transcriptome and proteome datasets. Our approach in differential expression analysis revealed 425 DE genes during the transition from BPH to PC, and 203 DE genes during the transition from PC to CRPC.

Additionally, in a previous study, the Tampere PC cohort samples underwent miRNA expression profiling using small RNA sequencing [165]. Leveraging this data, our approach identified a total of 95 DE miRNAs during the transition from PC to CRPC.

5.4 Alterations in the proteome

In Article II, we employed the SWATH-MS method to detect and quantify a total of 3394 proteins with high-confidence in the Tampere prostate cancer cohort. We identified 728 DE proteins during the transition from BPH to PC. Additionally, in the progression from PC to CRPC, we detected 382 DE proteins. Intriguingly, only 153 DE proteins were found to be common between these two transitions, highlighting the involvement of distinct protein dysregulation events at different stages of prostate cancer.

Our pathway analysis of DE proteins identified alterations in the regulatory pathways. Notably, we uncovered changes in key pathways such as the tricarboxylic acid cycle metabolic pathway, which remained undetectable in our transcriptomic data.

5.5 Multilevel observations

Our study integrated multilevel data, incorporating genomics (DNA-seq), epigenomics (MeDIP-seq and ATAC-seq), transcriptomics (mRNA-seq and miRNA-seq), and proteomics (SWATH-MS). The integration revealed several noteworthy observations as described in the following sections.

5.5.1 Point mutation, copy number, and methylation impact on gene and protein expression

The analysis of point mutations impact on expression, encompassing both somatic and germline SNVs, revealed a statistically significant association between somatic mutations and mRNA levels when compared to germline variants (Fisher's exact test, P-value=0.0055). However, no statistically significant impact on protein abundance levels was observed.

In Article II, we observed that alterations in copy number and methylation had an impact on the mRNA expressions of CRPC samples. However, these changes had comparatively lower impact at the proteome level (refer to Figure 13).

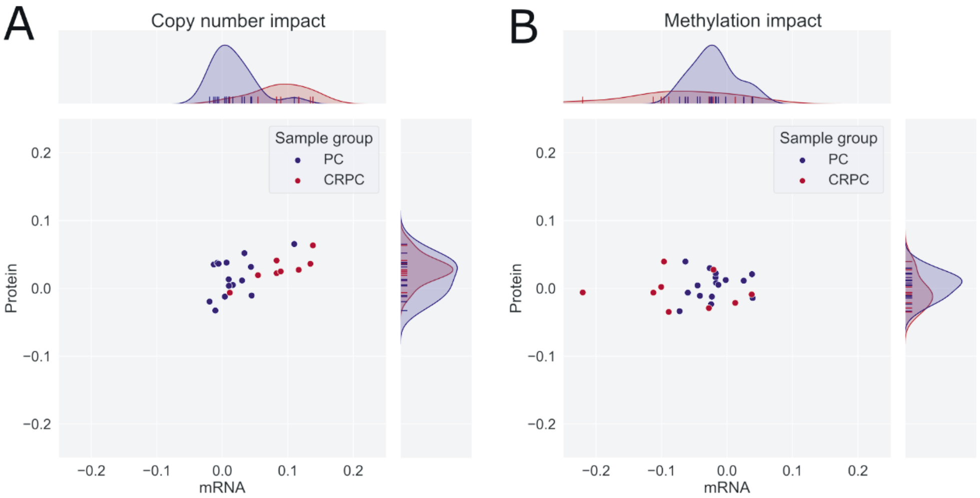


Figure 13. The impact of alterations in copy number and methylation on CRPC samples is more pronounced at the transcriptome than at the proteome level. **A**. Spearman correlations between copy number and gene/protein expressions reveal that copy number alterations exert a relatively greater influence on CRPC samples at the transcriptome level. **B**. Spearman correlations between genes' nearby DMRs and their gene/protein expressions show that methylation alterations have a relatively stronger impact on CRPC samples at the transcriptome level. PC, prostate cancer; CRPC, Castration-resistant prostate cancer; DMR, differentially methylation region.

In Article II, our data analysis uncovered a total of 140 DE genes at either the transcriptome or proteome level, which had a DMR within a 10 kilobases vicinity. Notably, we observed that hypermethylation can influence the expression of these DE genes, resulting in either a decrease or increase in their expression levels.

5.5.2 Interactions at the epigenome

Upon examining the accessible chromatin loci within the consensus peak set and their vicinity, we observed a clear enrichment of chromatin accessibility signals and a concurrent depletion of DNA methylation across all three prostate cancer groups (refer to Figure 14A). Furthermore, the extent of overlap between DARs and DMR was found to be modest (refer to Figure 14B, also Figure 11A and Figure 12). This finding underscores the distinctive regulatory roles played by chromatin accessibility and DNA methylation throughout the course of prostate cancer development and progression.

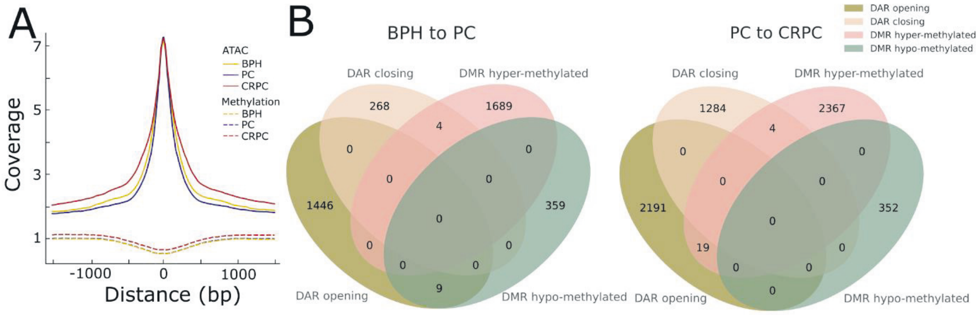


Figure 14. The interplay between chromatin accessibility and methylation reveals interesting insights. **A.** Depletion of DNA methylation signal at the accessible chromatin loci within the peak set indicates a potential regulatory relationship. **B**. Only a modest overlap between DAR and DMRs across different transitions were observed. DAR, differentially accessible region; DMR, differentially methylated region.

5.5.3 miRNA alterations impact the transcriptome and proteome

In Article II, integrative analysis of miRNA expression with mRNA and protein abundance data resulted in a few observations. Specifically, during the transition from PC to CRPC, we identified 474 genes that exhibited a negative correlation with 95 DE miRNAs at the transcriptome level. Similarly, at the proteome level, 482 genes displayed a negative correlation with the same set of 95 DE miRNAs. Interestingly, only 122 genes were found to be common between the transcriptome and proteome levels (refer to Figure 15).

Upon further examination of the targets of these 95 DE miRNAs, we discovered 115 DE genes at the transcriptome level and 218 DE genes at the proteome level in the PC to CRPC transition. Surprisingly, there were only 24 DE genes that overlapped between the transcriptome and proteome levels (refer to Figure 15).

This outcome highlights the efficacy of our integrative analysis in generating a comprehensive list of relevant miRNAs and their target genes in the progression of prostate cancer. Notably, a few of the miRNA-target interactions observed solely at the proteome level underwent successful validation in the laboratory using prostate cancer cell lines.

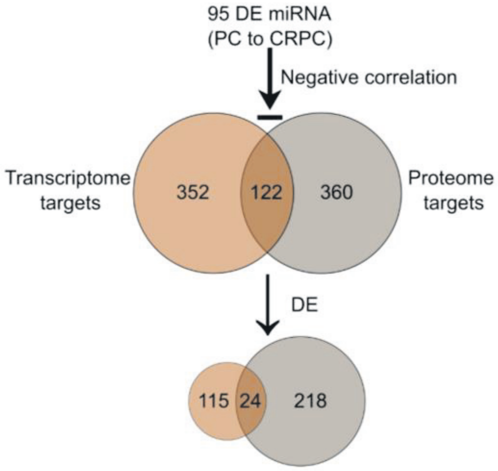


Figure 15. The Venn diagrams illustrate the number of genes exhibiting negative correlation with a targeting DE miRNA and the number of a subset of genes that are DE themselves, categorized by their expression at either the mRNA or protein level. Notably, only a small fraction of the miRNA targets is identified simultaneously at both the mRNA and protein level. DE, differentially expressed.

5.5.4 mRNA expression data alone cannot reliably predict protein abundances

In Article II, our observations revealed that a mere 73% of the genes quantified at the proteomics level exhibited a positive correlation with mRNA expression. Notably, the average correlation among these genes was a modest 0.15 (refer to Figure 16A). Additionally, it was noted that CRPC samples exhibited a comparatively lower correlation between mRNA and protein expression when compared to the other two groups (refer to Figure 16B). Furthermore, by analyzing the overlap of DE genes between mRNA and protein abundance datasets, we observed that only a small subset of DE genes was common between the two datasets (refer to Figure 16C). These findings offer valuable perspectives on the constraints of relying solely on mRNA expression data to predict protein levels accurately. Furthermore, each dataset reveals largely distinct events in the development and progression of prostate cancer.

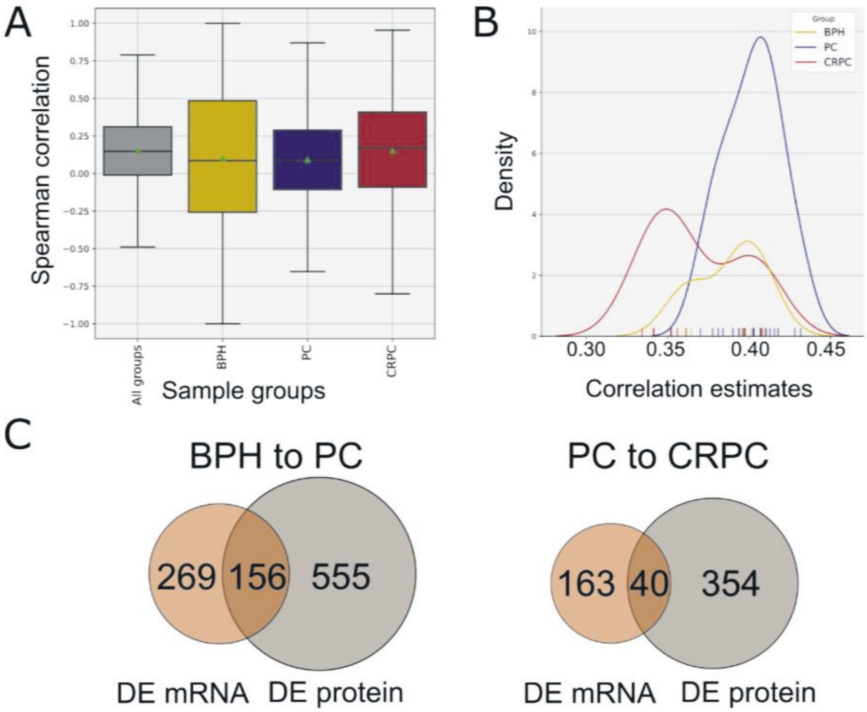


Figure 16. mRNA expression data alone is insufficient for accurate prediction of protein abundances. **A**. On average, there is a moderate correlation between mRNA and protein expression when Spearman correlation is calculated at the gene level. The mean is represented by a green triangle. **B**. In comparison to the BPH and PC groups, CRPC samples demonstrate a relatively weaker correlation at the sample level between mRNA and protein expression. **C**. The Venn diagrams illustrate a relatively small overlap between the numbers of DE genes and proteins across two transitions. DE, differentially expressed.

6 DISCUSSION

6.1 Comprehensive and robust analysis of high-throughput omics data

High-throughput measurements at the subcellular and molecular level not only provide valuable data on cellular processes but also offer insights into alterations associated with specific conditions, like cancer. In this study, we employed multiple high-throughput measurement techniques, including DNA-seq, MeDIP-seq, ATAC-seq, RNA-seq, small-RNA-seq (miRNA-seq), and SWATH-MS, to comprehensively investigate the genome, epigenome, transcriptome, and proteome of prostate cancer. Our objective was to elucidate the alterations contributing to disease progression. To ensure reliable findings, we addressed technical variability, biases inherent to high-throughput data, and project-specific issues.

In Article I, we benchmarked our developed method for detecting SCNAs, demonstrating its robustness to normal cell contamination under practical conditions. The use of matched normal samples minimized biases such as mappability and GC content biases in HTS data.

In Article II, we partly used preprocessed and analyzed data from earlier studies conducted on the Tampere PC cohort. These studies had already addressed various issues and biases, including the bias resulting from the use of different RNA extraction reagents. We applied appropriate normalization techniques to ensure reliable comparison between sample groups. Moreover, we applied the Benjamini-Hochberg P-value adjustment method to account for multiple hypothesis testing.

In Article III, we developed an ATAC-seq data analysis method that effectively mitigated global background noise levels and local background factors, such as the influence of SCNAs. We established criteria thresholds to maintain a reasonable FDR and corrected for a project-specific bias rustling from sample collection procedure. Moreover, we normalized the data to enable reliable comparison between sample groups.

Ensuring the robustness of tools and methods designed for the analysis of HTS data is of paramount importance. We believe that, taken together, our considerations allowed us to develop robust HTS data analysis tools and methods.

6.1.1 Other considerations

When detecting DARs, our approach employed a genome-wide detection of DARs rather than solely focusing on the detection of differentially accessible peaks within the ATAC-seq consensus peak set. This decision was made because the genome-wide DAR detection method allows for the unbiased identification of a larger number of actual DARs enhancing the comprehensiveness of our analysis [206]. While this approach may potentially result in more false positives, the chosen criteria threshold values have been selected to maintain the FDR below a reasonable level.

6.2 Exploring prostate cancer progression through single and multilevel high-throughput data analysis

By conducting single-level analyses, we successfully identified multiple distinct alterations occurring at various stages of prostate cancer development and progression across different omics levels. Our findings encompassed common and unique DARs, DMRs, DE genes and miRNAs, as well as DE proteins. The intriguing aspect of these observations lies in the presence of a small overlap between these alterations across different transition stages of prostate cancer. For example, out of 1110 DE proteins, only 153 (approximately 14%) were found to be common between the two transition stages, indicating the existence of discrete dysregulation events at each stage of prostate cancer.

Through multilevel analyses, we uncovered intricate interactions of different processes occurring across various stages and levels of prostate cancer. At the epigenome level, we examined the interplay between chromatin accessibility and DNA methylation. Throughout the genome, an increase in chromatin accessibility coincided with a decrease in DNA methylation. Furthermore, we observed a minimal overlap between DARs and DMRs, suggesting that chromatin accessibility and DNA methylation play distinctive regulatory roles throughout the course of prostate cancer development and progression.

In another multilevel example, we explored the impact of alterations at the genome or epigenome level on the transcriptome and proteome. Intriguingly, we observed that certain alterations, such as copy number changes, exert a stronger impact on the transcriptome compared to the proteome. Furthermore, we observed that transcriptome expression levels alone are insufficient for accurately predicting proteome expression levels. By integrating data encompassing miRNA, mRNA, and proteome abundances, we uncovered miRNAs that primarily regulate a subset of the proteome through inhibition of translation, rather than mRNA degradation. Collectively, these insights underscore the robustness of the proteome in the face of genomic and epigenomic alterations as compared to the transcriptome. It should be noted that these insights would remain hidden without the integrative analysis of multilevel data.

In summary, this study has compiled a comprehensive catalog of alterations across different levels and has elucidated their intricate relationships and their impacts on each other. These results not only corroborated existing knowledge regarding prostate cancer development and progression but also offer novel observations that warrant further investigation and validation in other patient cohorts.

6.3 Challenges and limitations of the study

Throughout this work, we encountered a number of challenges that we made efforts to overcome. Furthermore, it is important to acknowledge that our study had certain limitations, which we aim to elucidate in the following sections.

6.3.1 Key considerations in using Segmentum

In Article I, we demonstrated the capability of `Segmentum` to accurately detect SCNAs, even when the average depth of coverage was as low as approximately 4X. However, at such coverage levels, the reliable identification of heterozygous SNPs is limited. Consequently, the detection of cnLOH events becomes infeasible, restricting the analysis solely to the identification of deletions and gains. It is worth noting that this limitation is not unique to `Segmentum` but affects other tools that rely on heterozygous SNPs for cnLOH detection.

Furthermore, our investigation revealed that when normal cell contamination exceeds 50%, `Segmentum`'s ability to detect changes diminishes significantly. However, considering that the typical range of normal cell contamination in real tumors is usually around 30-40%, we can expect `Segmentum` to perform well under practical conditions.

While `Segmentum` does not provide estimates of absolute copy number for each SCNA event, the ability to do so would be valuable for various applications, such as stratifying cancer patients based on the copy number of relevant SCNAs to determine appropriate treatment strategies [207]. Fortunately, there are existing tools like `ABSOLUTE`, `ACE`, and `Rascal` specifically designed to estimate absolute copy number [208–210]. `Segmentum`'s output is compatible as input for tools like `ABSOLUTE`, and it can be utilized to estimate absolute copy number. This compatibility makes `Segmentum` well-suited for integration into tumor analysis pipelines.

6.3.2 Tampere prostate cancer cohort

In comparison to some recent studies in prostate cancer, the Tampere prostate cancer cohort may be considered relatively small [36]. It is often the case that cohorts with smaller sample sizes are criticized for their limited statistical power and the potential increase in false positive rates during discovery. These concerns are indeed valid and warrant attention.

Addressing the issue of limited statistical power poses a considerable challenge, and one of the most effective approaches to overcome it is by utilizing larger sample sizes. However, an indirect assessment of the power of a cohort or approach in discovery can be made by integrating and annotating the discovery results with available external data. In Article III, we demonstrated that 66% of the prostate cancer-specific peaks found in the TCGA study overlapped with the accessible chromatin loci detected in our study [103].

On the other hand, addressing the concern regarding the increase in false positive rates during discovery, as we have shown (e.g., in Article III), is a less daunting task. It is possible to calculate the FDR of a discovery analysis by employing techniques such as permutation tests and setting appropriate criteria thresholds to ensure that the FDR remains below a certain predetermined value. When detecting DARs and DMRs in Article III, we established criteria thresholds that maintained the FDR below 10%, which is an acceptable threshold for discovery purposes.

It is important to note that a smaller cohort may possess unique merits that larger cohorts may not offer. For instance, smaller cohorts may exhibit distinctive sample group compositions that provide a novel perspective for research or have comprehensive characterizations across various omics levels. Hence, the size of a cohort alone should not warrant its outright dismissal, as it can still provide valuable insights and perspectives.

During the course of our research for Article III, we observed that the sample collection procedure for prostate cancer (i.e., RP vs. TURP) had introduced a bias in the measurements obtained from the ATAC-seq assay. To address this issue, we refined our correction approach to effectively account for this bias and minimize its impact on the results. It is important to note that this specific bias was not observed in other data types analyzed during our study.

It is important to highlight that collecting normal prostate tissues for research purposes raises ethical concerns and is not always feasible. In the absence of such samples, one approach employed is to utilize the normal tissue inadvertently removed during the surgical extraction of a prostate tumor. In the Tampere prostate cancer cohort, BPH samples have been used as the control group and as a surrogate for the normal prostate tissue. This choice is justified by the fact that BPH exhibits similarities to prostate cancer in terms of histological, pathological, and genetic characteristics [179].

6.3.3 Considerations in measurements and the measurement assays

In Article II, we presented findings indicating a lack of concordance between mRNA expression and protein abundances, a phenomenon also observed in other studies [169]. This lack of concordance can be attributed to various processes, including post-transcriptional and post-translational regulation, protein degradation, local availability of resources for protein biosynthesis, and buffering of mRNA fluctuations [118,211]. While these factors help explain the observed discrepancy, two additional reasons may also contribute to this phenomenon.

Firstly, in our study, we employed SWATH-MS to quantify 3394 proteins, which was an impressive number at the time of publication. However, it is important to acknowledge that a significant portion of proteins remained undetected and unquantified.

Secondly, it is crucial to note that the omics data used in this study were generated from the same set of samples, with different sections of each specimen utilized for each omics measurement, which may contribute to variations observed across different omics levels. It is worth noting that this is not unique to our study but affects other studies as well. To address these concerns, emerging approaches that enable the simultaneous measurement of multiple levels of molecular information can be employed in future studies, helping to mitigate some of these limitations.

Nonetheless, the work presented in Article II stands as a pioneering integrative study in prostate cancer. This is thanks to its comprehensive multilevel analysis of genomic, epigenomic, transcriptomic, and proteomic data.

In Article III, our analysis revealed a limited overlap between the detected DMRs and DARs. This finding suggests that chromatin accessibility and DNA methylation may represent distinct epigenetic regulatory mechanisms in prostate cancer development and progression. However, it is important to consider that the observed low overlap could be partially attributed to the inherent limitations of the MeDIP-seq approach in detecting DMRs when compared to alternative technologies for DNA methylation measurement, such as MethylCap-seq or RRBS [212].

Studies conducted by Bock and colleagues have demonstrated that MethylCap-seq outperforms MeDIP-seq in detecting DMRs, detecting approximately twice as many regions at comparable depth of coverage [212]. Therefore, the relatively low overlap observed between DMRs and DARs in our analysis may be influenced, at least in part, by the limitations of the MeDIP-seq method. Nevertheless, it remains crucial to further investigate and validate this observation in additional datasets.

The data utilized in Articles II and III are considered bulk data, where each measurement represents an average level across a large population of input cells. It is important to recognize that such bulk data can encompass a heterogeneous group of cells, including distinct cell types or diverging subpopulations originating from similar cells within the context of cancer. Consequently, relying solely on average measurements may not accurately reflect the true state of each specific cell group.

Furthermore, bulk measurement methods typically require a substantial number of cells, often ranging in the hundreds of thousands. To address the limitations associated with bulk data, the field of single-cell sequencing and measurement approaches has made significant progress since the emergence of techniques like single-cell mRNA-seq in

2009 [213]. Single-cell mRNA-seq, for example, enables the characterization of transcriptome at a single-cell resolution, providing insights into cellular heterogeneity and capturing individual cell states more faithfully.

The prospect of acquiring and analyzing multilevel single-cell data from cohorts, like the Tampere prostate cancer cohort utilized in Article II and III, is not only intriguing but also holds tremendous potential in unraveling additional mechanisms underlying the development and progression of prostate cancer. Meanwhile, as the acquisition of such resources is underway, it is important to acknowledge the existence of tools like `CibersortX`, which have been designed to deconvolve bulk data by utilizing single-cell data [214]. These tools enable the extraction of cell-type-specific information from bulk measurements, leveraging single-cell information.

6.4 Future work

As of the writing of this dissertation, our ongoing work involves further characterization of the Tampere prostate cancer cohort by incorporating and conducting integrative analysis of Hi-C measurements [215]. Hi-C data provides valuable insights into the 3D architecture of genomes, allowing us to study the spatial organization of chromatin within the nucleus. This spatial organization plays a crucial role in gene transcription regulation [216].

By integrating Hi-C data, we aim to complement and provide further explanations for the findings obtained in our earlier studies. In Article III, we observed correlations between the chromatin accessibility of certain loci, as determined by ATAC-seq data, and the abundance of corresponding mRNA or protein. Hi-C data can provide additional evidence to support some of these observed correlations. As an example, it can reveal that a distantly accessible chromatin region functions as an enhancer that spatially interacts with a specific gene promoter, potentially influencing its expression. Moreover, Hi-C data can also be utilized to detect sequence variations and SCNAs, thereby complementing the WGS data available in our cohort [217,218]. The analysis of higher-order chromatin interactions and their influence on gene regulation holds the potential to unveil novel insights into the development and progression of prostate cancer.

7 CONCLUSIONS

As highlighted in the literature review, despite initial positive response to therapy, a subset of prostate cancer patients experiences disease progression to an advanced stage. Unfortunately, this advanced stage is currently considered incurable, despite numerous efforts to understand the underlying mechanisms and devise suitable therapeutic interventions. Given these challenges, the specific aims of this study were twofold. Firstly, we aimed to develop computational and statistical tools and methods for the effective and efficient analysis of high-throughput, omics data within the context of cancer. Secondly, we sought to use the developed tools and methods from our first aim, to investigate single and multilevel high-throughput, omics data to uncover additional key alterations that drive the prostate cancer development and progression, with the aspiration that these newfound insights could contribute to the development of effective treatments. While acknowledging the scope and limitations of this study, we believe that we have achieved these goals by presenting the findings of this study in three original articles. All of the Articles contributed to the accomplishment of Aim 1 in this study. Additionally, Articles II and III contributed to the achievement of Aim 2 in this study.

SCNAs play a significant role in genome instability, a critical factor in acquiring other cancer hallmarks. Therefore, the identification of SCNAs is important. In Article I, we developed `Segmentum`, a novel tool capable of accurately identifying SCNAs using WGS data. Through benchmarking analysis, we demonstrated that `Segmentum` performed equally well as other state-of-the-art tools, while also exhibiting a notable speed advantage of at least twofold. Additionally, we also examined the limitations of `Segmentum` in accurately identifying SCNAs under certain scenarios, such as the presence of normal cell contamination.

Shifting our focus to multilevel analysis of high-throughput omics data, Article II presents our findings from the investigation of proteomic data obtained from the Tampere prostate cancer cohort, encompassing various stages of prostate cancer onset and progression. By developing integrative approaches and integration of proteomics data with other sources, we identified alterations occurring at different levels and examined their interactions across multiple layers. Notably, we observed the robustness of the proteome in the presence of alterations at other levels, and we concluded that transcriptomic gene expression information alone is insufficient to predict the proteome abundances.

Article III expands upon our multilevel analysis by incorporating chromatin accessibility data derived from the ATAC-seq assay into the characterization of the Tampere prostate cancer cohort. To ensure the reliability of downstream analyses, we developed a correction approach to effectively mitigate both global background noise levels and local background factors as well as the influence of SCNAs on ATAC-seq readouts, the biases introduced by the sample collection procedure, and the variability in sequencing depths. Consequently, we confidently identified alterations in chromatin accessibility and integrated them with other datasets from multiple omics levels for a more comprehensive understanding.

In summary, our study demonstrates the potential of single and multilevel analysis of high-throughput omics data, which allows us to reproduce previous findings and uncover the alterations that affect biological processes at different levels during prostate cancer development and progression. We hope that these findings enhance our understanding of the disease and provide valuable prospects for the development of future therapeutic interventions.

8 APPENDIX 1

Tables 1-3 present the list of TCGA sample barcode names utilized in Article I to benchmark Segmentum's performance on real data as well as the classification of grade II and III gliomas based on genomic alterations.

Table 1. TCGA low-grade glioma sample barcode names employed for the evaluation of Segmentum's performance on high coverage actual data, ranging from 30X to 100X.

Sample barcode names
TCGA-HT-7689-01A-11D-2253-08
TCGA-DB-5278-01A-01D-1468-08
TCGA-DU-7301-01A-11D-2086-08
TCGA-DU-5872-01A-11D-A465-08
TCGA-DU-5874-01A-11D-1705-08
TCGA-CS-5395-01A-01D-1468-08.1
TCGA-DU-6401-01A-11D-1705-08.2
TCGA-DU-7013-01A-11D-A461-08.1
TCGA-DU-7304-01A-12D-A461-08.4
TCGA-FG-8182-01A-11D-2253-08.3

Table 2. TCGA prostate adenocarcinoma sample barcode names employed for the evaluation of Segmentum's performance on low-coverage (6X) actual data.

Sample barcode names
TCGA-G9-6332-01A-11D-1785-01
TCGA-G9-6338-01A-12D-1959-01
TCGA-G9-6342-01A-11D-1959-01
TCGA-G9-6362-01A-11D-1785-01
TCGA-G9-6364-01A-21D-1785-01
TCGA-G9-6373-01A-11D-1785-01
TCGA-G9-6494-01A-11D-1785-01
TCGA-HI-7171-01A-12D-2112-01
TCGA-HC-7211-01A-11D-2112-01
TCGA-EJ-7784-01A-11D-2112-01

Table 3. TCGA low-grade glioma sample barcode names employed for the classification of grade II and III gliomas based on their genomic alterations.

Sample barcode names
TCGA-CS-6668-01A-11D-1893-08
TCGA-DB-5278-01A-01D-1468-08
TCGA-DH-A669-01A-12D-A31L-08
TCGA-DU-5870-01A-11D-A461-08
TCGA-DU-5874-01A-11D-1705-08
TCGA-DU-6397-01A-11D-A461-08
TCGA-DU-7009-01A-11D-2024-08
TCGA-E1-5318-01A-01D-1468-08
TCGA-E1-5319-01A-01D-1893-08
TCGA-EZ-7264-01A-11D-2024-08
TCGA-FG-5964-01A-11D-1705-08
TCGA-HT-7695-01A-11D-2253-08
TCGA-HW-7486-01A-11D-2024-08
TCGA-HW-7487-01A-11D-2024-08
TCGA-CS-6665-01A-11D-1893-08
TCGA-DU-5872-01A-11D-A465-08
TCGA-DU-6407-01A-13D-1705-08
TCGA-DU-7301-01A-11D-2086-08
TCGA-FG-5965-01B-11D-1893-08
TCGA-HT-7689-01A-11D-2253-08
TCGA-HT-A5R7-01A-11D-A461-08
TCGA-HT-A61B-01A-11D-A461-08
TCGA-IK-7675-01A-11D-2086-08
TCGA-TQ-A7RK-01A-11D-A33T-08
TCGA-TQ-A7RV-01A-21D-A34A-08
TCGA-TQ-A8XE-01A-11D-A36O-08
TCGA-DU-6401-01A-11D-1705-08
TCGA-DU-7304-01A-12D-A461-08
TCGA-FG-8182-01A-11D-2253-08
TCGA-FG-A4MT-01A-11D-A461-08
TCGA-HT-7602-01A-21D-2086-08
TCGA-TM-A7CF-01A-11D-A32B-08
TCGA-DU-7013-01A-11D-A461-08
TCGA-CS-5395-01A-01D-1468-08
TCGA-CS-6669-01A-11D-1893-08
TCGA-DU-6404-01A-11D-A461-08
TCGA-FG-7643-01A-11D-A461-08
TCGA-HT-8104-01A-11D-A461-08

9 REFERENCES

1. Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, et al. Molecular Biology of the Cell. Garland Science; 2014. 1464 p.

2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May;71(3):209–49.

3. Hanahan D, Weinberg RA. The Hallmarks of Cancer [Internet]. Vol. 100, Cell. 2000. p. 57–70. Available from: http://dx.doi.org/10.1016/s0092-8674(00)81683-9

4. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 Mar 4;144(5):646–74.

5. Metzker ML. Emerging technologies in DNA sequencing. Genome Res. 2005 Dec;15(12):1767–76.

6. Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell. 2018 Apr 5;173(2):283–5.

7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.

8. Eisenstein M. Closing in on a complete human genome [Internet]. Vol. 590, Nature. 2021. p. 679–81. Available from: r

9. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome [Internet]. bioRxiv. 2021 [cited 2021 May 28]. p. 2021.05.26.445798. Available from: https://www.biorxiv.org/content/10.1101/2021.05.26.445798v1.abstract

10. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016 Jan;107(1):1–8.

11. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012 Sep;13(9):613–26.

12. Tubbs A, Nussenzweig A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. Cell. 2017 Feb 9;168(4):644–56.

13. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005 Oct 28;310(5748):644–8.

14. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. Recurrent gene fusions in prostate cancer. Nat Rev Cancer. 2008 Jun 19;8(7):497–511.

15. Suzuki H, Aoki K, Chiba K, Sato Y, Shiozawa Y, Shiraishi Y, et al. Mutational landscape and clonal architecture in grade II and III gliomas. Nat Genet. 2015 May;47(5):458–68.

16. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017 May 5;18(1):83.

17. Shen H, Laird PW. Interplay between the cancer genome and epigenome. Cell. 2013 Mar 28;153(1):38–55.

18. Pelizzola M, Ecker JR. The DNA methylome. FEBS Lett. 2011 Jul 7;585(13):1994–2000.

19. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012 May 29;13(7):484–92.

20. Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. Cell. 2007 Feb 23;128(4):669–81.

21. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet. 2010 Mar 1;11(3):191–203.

22. Robertson KD. DNA methylation and human disease. Nat Rev Genet. 2005 Aug;6(8):597–610.

23. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discov. 2022 Jan;12(1):31–46.

24. Johnstone SE, Gladyshev VN, Aryee MJ, Bernstein BE. Epigenetic clocks, aging, and cancer. Science. 2022 Dec 23;378(6626):1276–7.

25. Guo H, Vuille JA, Wittner BS, Lachtara EM, Hou Y, Lin M, et al. DNA hypomethylation silences anti-tumor immune genes in early prostate cancer and CTCs. Cell. 2023 Jun 22;186(13):2765–82.e28.

26. Lu X, Fong KW, Gritsina G, Wang F, Baca SC, Brea LT, et al. HOXB13 suppresses de novo lipogenesis through HDAC3-mediated epigenetic reprogramming in prostate cancer. Nat Genet. 2022 Apr 25;54(5):670–83.

27. Saha A, Wittmeyer J, Cairns BR. Chromatin remodelling: the industrial revolution of DNA around histones. Nat Rev Mol Cell Biol. 2006 Jun;7(6):437–47.

28. Fischle W, Wang Y, Allis CD. Histone and chromatin cross-talk. Curr Opin Cell Biol. 2003 Apr;15(2):172–83.

29. Annunziato A. DNA packaging: nucleosomes and chromatin. Nature Education. 2008;1(1):26.

30. Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. Epigenetics Chromatin. 2014 Nov 20;7(1):33.

31. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018 Feb 8;172(4):650–65.

32. Baylin SB, Jones PA. A decade of exploring the cancer epigenome - biological and translational implications. Nat Rev Cancer. 2011 Sep 23;11(10):726–34.

33. Zaret KS. Pioneer Transcription Factors Initiating Gene Network Changes. Annu Rev Genet. 2020 Nov 23;54:367–85.

34. Teng M, Zhou S, Cai C, Lupien M, He HH. Pioneer of prostate cancer: past, present and the future of FOXA1. Protein Cell. 2021 Jan;12(1):29–38.

35. Rebello RJ, Oing C, Knudsen KE, Loeb S, Johnson DC, Reiter RE, et al. Prostate cancer. Nat Rev Dis Primers. 2021 Feb 4;7(1):9.

36. Pomerantz MM, Qiu X, Zhu Y, Takeda DY, Pan W, Baca SC, et al. Prostate cancer reactivates developmental epigenomic programs during metastatic progression. Nat Genet. 2020 Aug;52(8):790–9.

37. Rizzo JM, Buck MJ. Key principles and clinical applications of "next-generation" DNA sequencing. Cancer Prev Res [Internet]. 2012; Available from: https://cancerpreventionresearch.aacrjournals.org/content/5/7/887.short

38. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan 1;10(1):57–63.

39. Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, et al. MicroRNA genes are transcribed by RNA polymerase II. EMBO J. 2004 Oct 13;23(20):4051–60.

40. Borchert GM, Lanier W, Davidson BL. RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol. 2006 Dec;13(12):1097–101.

41. Fabris L, Ceder Y, Chinnaiyan AM, Jenster GW, Sorensen KD, Tomlins S, et al. The Potential of MicroRNAs as Prostate Cancer Biomarkers. Eur Urol. 2016 Aug;70(2):312–22.

42. Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, et al. From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Arnino Acid Analysis. Biotechnology. 1996 Jan;14(1):61–5.

43. Wagner I, Musso H. New naturally occurring amino acids. Angew Chem Int Ed Engl. 1983 Nov;22(11):816–28.

44. Tyers M, Mann M. From genomics to proteomics. Nature. 2003 Mar 13;422(6928):193–7.

45. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003 Mar 13;422(6928):198–207.

46. Anjo SI, Santa C, Manadas B. SWATH-MS as a tool for biomarker discovery: From basic research to clinical applications. Proteomics. 2017 Feb;17(3-4):1600278.

47. McCombie WR, McPherson JD, Mardis ER. Next-Generation Sequencing Technologies. Cold Spring Harb Perspect Med [Internet]. 2019 Nov 1;9(11). Available from: http://dx.doi.org/10.1101/cshperspect.a036798

48. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463–7.

49. Gocayne J, Robinson DA, FitzGerald MG, Chung FZ, Kerlavage AR, Lentes KU, et al. Primary structure of rat cardiac beta-adrenergic and muscarinic cholinergic receptors

obtained by automated DNA sequence analysis: further evidence for a multigene family. Proc Natl Acad Sci U S A. 1987 Dec;84(23):8296–300.

50. Mardis ER. DNA sequencing technologies: 2006–2016. Nat Protoc. 2017 Feb 1;12(2):213–8.

51. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016 May 17;17(6):333–51.

52. Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet. 2010;11(1):31–46.

53. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008 Nov 6;456(7218):53–9.

54. Landegren U, Kaiser R, Sanders J, Hood L. A ligase-mediated gene detection technique. Science. 1988 Aug 26;241(4869):1077–80.

55. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011 Jul 20;475(7356):348–52.

56. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet. 2008 Mar 1;24(3):142–9.

57. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012 Jul 5;2012:251364.

58. Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, et al. Analysis of error profiles in deep next-generation sequencing data. Genome Biol. 2019 Mar 14;20(1):50.

59. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. 2011 Nov 8;12(11):R112.

60. Kircher M, Stenzel U, Kelso J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol. 2009 Aug 14;10(8):R83.

61. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 2015 Mar 31;43(6):e37.

62. Andrews S, Others. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.

63. Trivedi UH, Cézard T, Bridgett S, Montazam A, Nichols J, Blaxter M, et al. Quality control of next-generation sequencing data without a reference. Front Genet. 2014 May 6;5:111.

64. Xi W, Gao Y, Cheng Z, Chen C, Han M, Yang P, et al. Using QC-Blind for Quality Control and Contamination Screening of Bacteria DNA Sequencing Data Without Reference Genome. Front Microbiol. 2019 Jul 9;10:1560.

65. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, et al. The challenges of sequencing by synthesis. Nat Biotechnol. 2009 Nov;27(11):1013–23.

66. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011 May 2;17(1):10–2.

67. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

68. Burrows M, Wheeler DJ. A Block-sorting Lossless Data Compression Algorithm. 1994. 18 p.

69. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009 Mar 4;10(3):R25.

70. Chaisson MJ, Brinza D, Pevzner PA. De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Res. 2009 Feb;19(2):336–46.

71. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001 Feb 16;291(5507):1304–51.

72. Li W, Freudenberg J. Mappability and read length. Front Genet. 2014 Nov 10;5:381.

73. de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 2011 Dec;7(12):e1002384.

74. Taub MA, Corrada Bravo H, Irizarry RA. Overcoming bias and systematic errors in next generation sequencing data. Genome Med. 2010 Dec 10;2(12):87.

75. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008 Nov;18(11):1851–8.

76. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep. 2019 Jun 27;9(1):9354.

77. Pickrell JK, Gaffney DJ, Gilad Y, Pritchard JK. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. Bioinformatics. 2011 Aug 1;27(15):2144–6.

78. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008 Apr 17;452(7189):872–6.

79. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb;463(7283):899–905.

80. Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS. A census of amplified and overexpressed human cancer genes. Nat Rev Cancer. 2010 Jan;10(1):59–64.

81. Pinkel D, Albertson DG. Comparative genomic hybridization. Annu Rev Genomics Hum Genet. 2005;6:331–54.

82. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 1992 Oct 30;258(5083):818–21.

83. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet. 1998 Oct;20(2):207–11.

84. Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, et al. High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Res. 2004 Feb;14(2):287–95.

85. Chiang DY, Getz G, Jaffe DB, O'Kelly MJT, Zhao X, Carter SL, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods. 2008 Nov 30;6(1):99–103.

86. Alkodsi A, Louhimo R, Hautaniemi S. Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. Brief Bioinform. 2015 Mar;16(2):242–54.

87. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012 May;40(10):e72.

88. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013 Oct;45(10):1134–40.

89. Williams JL, Greer PA, Squire JA. Recurrent copy number alterations in prostate cancer: an in silico meta-analysis of publicly available genomic data. Cancer Genet. 2014 Sep 16;207(10-12):474–88.

90. Bhatia-Gaur R, Donjacour AA, Sciavolino PJ, Kim M, Desai N, Young P, et al. Roles for Nkx3.1 in prostate development and cancer. Genes Dev. 1999 Apr 15;13(8):966–77.

91. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet. 2005 Aug;37(8):853–62.

92. Rauluseviciute I, Drabløs F, Rye MB. DNA methylation data by sequencing: experimental approaches and recommendations for tools and pipelines for data analysis. Clin Epigenetics. 2019 Dec 12;11(1):193.

93. Taiwo O, Wilson GA, Morris T, Seisenberger S, Reik W, Pearce D, et al. Methylome analysis using MeDIP-seq with low DNA concentrations. Nat Protoc. 2012 Mar 8;7(4):617–36.

94. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011 Feb 21;12(2):R18.

95. Liu B, Morrison CD, Johnson CS, Trump DL, Qin M, Conroy JC, et al. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. Oncotarget. 2013 Nov;4(11):1868–81.

96. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 2005 Oct 13;33(18):5868–77.

97. Zhao EY, Jones M, Jones SJM. Whole-Genome Sequencing in Cancer. Cold Spring Harb Perspect Med [Internet]. 2019 Mar 1;9(3). Available from: http://dx.doi.org/10.1101/cshperspect.a034579

98. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. DNA methylation-based classification of central nervous system tumours. Nature. 2018 Mar 14;555(7697):469–74.

99. Lienhard M, Grimm C, Morkel M, Herwig R, Chavez L. MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. Bioinformatics. 2014 Jan 15;30(2):284–6.

100. Lienhard M, Grasse S, Rolff J, Frese S, Schirmer U, Becker M, et al. QSEA—modelling of genome-wide DNA methylation from sequencing enrichment experiments. Nucleic Acids Res. 2016 Nov 29;45(6):e44–e44.

101. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013 Dec;10(12):1213–8.

102. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008 Sep 17;9(9):R137.

103. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science [Internet]. 2018 Oct 26;362(6413). Available from: http://dx.doi.org/10.1126/science.aav1898

104. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012 Sep;22(9):1813–31.

105. Bogenhagen DF. Mitochondrial DNA nucleoid structure. Biochim Biophys Acta. 2012 Sep-Oct;1819(9-10):914–20.

106. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016 Jan 26;17:13.

107. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008 Jul;5(7):621–8.

108. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

109. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010 Oct 27;11(10):R106.

110. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.

111. Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics. 2012 Jun;11(6):O111.016717.

112. Selevsek N, Chang CY, Gillet LC, Navarro P, Bernhardt OM, Reiter L, et al. Reproducible and consistent quantification of the Saccharomyces cerevisiae proteome by SWATH-mass spectrometry. Mol Cell Proteomics. 2015 Mar;14(3):739–49.

113. Shi T, Song E, Nie S, Rodland KD, Liu T, Qian WJ, et al. Advances in targeted proteomics and applications to biomedical research. Proteomics. 2016 Aug;16(15-16):2160–82.

114. Röst HL, Aebersold R, Schubert OT. Automated SWATH Data Analysis Using Targeted Extraction of Ion Chromatograms. Methods Mol Biol. 2017;1550:289–307.

115. Schubert OT, Gillet LC, Collins BC, Navarro P, Rosenberger G, Wolski WE, et al. Building high-quality assay libraries for targeted analysis of SWATH MS data. Nat Protoc. 2015 Mar;10(3):426–41.

116. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016 Jan 20;17 Suppl 2:15.

117. Qin G, Liu Z, Xie L. Multiple Omics Data Integration [Internet]. Reference Module in Biomedical Sciences. 2019. Available from: http://dx.doi.org/10.1016/b978-0-12-801238-3.11508-9

118. Liu Y, Beyer A, Aebersold R. On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell. 2016 Apr 21;165(3):535–50.

119. Liu Y, Devescovi V, Chen S, Nardini C. Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. BMC Syst Biol. 2013 Feb 19;7:14.

120. Joyce AR, Palsson BØ. The model organism as a system: integrating'omics' data sets. Nat Rev Mol Cell Biol. 2006;7(3):198–210.

121. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science. 2001 May 4;292(5518):929–34.

122. GDC [Internet]. [cited 2020 Apr 24]. Available from: https://portal.gdc.cancer.gov/

123. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform. 2018 Sep 28;19(5):776–92.

124. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013 Nov;14(6):671–83.

125. Nguyen LH, Holmes S. Ten quick tips for effective dimensionality reduction. PLoS Comput Biol. 2019 Jun;15(6):e1006907.

126. Ringnér M. What is principal component analysis? Nat Biotechnol. 2008 Mar;26(3):303–4.

127. Li W, Cerise JE, Yang Y, Han H. Application of t-SNE to human genetic data. J Bioinform Comput Biol. 2017 Aug;15(4):1750017.

128. Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9(Nov):2579–605.

129. D'haeseleer P. How does gene expression clustering work? Nat Biotechnol. 2005 Dec;23(12):1499–501.

130. Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. Bioinformatics. 2005 Aug 1;21(15):3201–12.

131. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998 Dec 8;95(25):14863–8.

132. Nugent R, Meila M. An overview of clustering applied to molecular biology. Methods Mol Biol. 2010;620:369–404.

133. Krzywinski M, Altman N. Comparing samples--part II: when a large number of tests are performed, P values must be interpreted differently. Nat Methods. 2014;11(4):355–7.

134. Krzywinski M, Altman N. Significance, P values and t-tests. Nat Methods. 2013 Nov;10(11):1041–2.

135. Hoffman JIE. Hypothesis Testing: The Null Hypothesis, Significance and Type I Error [Internet]. Basic Biostatistics for Medical and Biomedical Practitioners. 2019. p. 159–71. Available from: http://dx.doi.org/10.1016/b978-0-12-817084-7.00010-3

136. Krzywinski M, Altman N. Comparing samples—part I [Internet]. Vol. 11, Nature Methods. 2014. p. 215–6. Available from: http://dx.doi.org/10.1038/nmeth.2858

137. Krzywinski M, Altman N. Nonparametric tests [Internet]. Vol. 11, Nature Methods. 2014. p. 467–8. Available from: http://dx.doi.org/10.1038/nmeth.2937

138. Smalheiser NR. Chapter 12-Nonparametric Tests. Data Literacy. 2017;157–67.

139. Hoffman JIE. Comparison of Two Groups: t-Tests and Nonparametric Tests [Internet]. Basic Biostatistics for Medical and Biomedical Practitioners. 2019. p. 341–66. Available from: http://dx.doi.org/10.1016/b978-0-12-817084-7.00022-x

140. Hoffman JIE. Multiple Comparisons [Internet]. Basic Biostatistics for Medical and Biomedical Practitioners. 2019. p. 375–90. Available from: http://dx.doi.org/10.1016/b978-0-12-817084-7.00024-3

141. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. Proc Natl Acad Sci U S A. 2010 May 25;107(21):9546–51.

142. Krzywinski M, Altman N. Power and sample size: the ability to detect experimental effects is undermined in studies that lack power. Nat Methods. 2013;10(12):1139–41.

143. Holmes S, Huber W. Modern Statistics for Modern Biology. Cambridge University Press; 2018. 400 p.

144. Hoffman JIE. Hypergeometric Distribution [Internet]. Biostatistics for Medical and Biomedical Practitioners. 2015. p. 179–82. Available from: http://dx.doi.org/10.1016/b978-0-12-802387-7.00013-5

145. Altman N, Krzywinski M. Tabular data [Internet]. Vol. 14, Nature Methods. 2017. p. 329–30. Available from: http://dx.doi.org/10.1038/nmeth.4239

146. Alberts B. Molecular Biology of the Cell. Garland Science, Taylor and Francis Group; 2015. 1464 p.

147. Gupta GP, Massagué J. Cancer metastasis: building a framework. Cell. 2006 Nov 17;127(4):679–95.

148. Mardis ER. The Impact of Next-Generation Sequencing on Cancer Genomics: From Discovery to Clinic. Cold Spring Harb Perspect Med [Internet]. 2019 Sep 3;9(9). Available from: http://dx.doi.org/10.1101/cshperspect.a036269

149. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature. 2008 Nov 6;456(7218):66–72.

150. Culp MB, Soerjomataram I, Efstathiou JA, Bray F, Jemal A. Recent Global Patterns in Prostate Cancer Incidence and Mortality Rates. Eur Urol. 2020 Jan;77(1):38–52.

151. Al Olama AA, Kote-Jarai Z, Berndt SI, Conti DV, Schumacher F, Han Y, et al. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat Genet. 2014 Oct;46(10):1103–9.

152. Farashi S, Kryza T, Clements J, Batra J. Post-GWAS in prostate cancer: from genetic association to biological contribution. Nat Rev Cancer. 2019 Jan;19(1):46–59.

153. Eeles R, Goh C, Castro E, Bancroft E, Guy M, Olama AAA, et al. The genetic epidemiology of prostate cancer and its clinical implications. Nat Rev Urol. 2013 Dec 3;11(1):18–31.

154. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell. 2010 Jul 13;18(1):11–22.

155. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012 May 20;44(6):685–9.

156. Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, et al. The mutational landscape of lethal castration-resistant prostate cancer. Nature. 2012 May 20;487(7406):239–43.

157. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015 Nov 5;163(4):1011–25.

158. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, et al. Punctuated evolution of prostate cancer genomes. Cell. 2013 Apr 25;153(3):666–77.

159. Lindberg J, Klevebring D, Liu W, Neiman M, Xu J, Wiklund P, et al. Exome sequencing of prostate cancer supports the hypothesis of independent tumour origins. Eur Urol. 2013 Feb;63(2):347–53.

160. Weischenfeldt J, Simon R, Feuerbach L, Schlangen K, Weichenhan D, Minner S, et al. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. Cancer Cell. 2013 Feb 11;23(2):159–70.

161. Bova GS, Steven Bova G, Kallio HML, Annala M, Kivinummi K, Högnäs G, et al. Integrated clinical, whole-genome, and transcriptome analysis of multisampled lethal

metastatic prostate cancer [Internet]. Vol. 2, Molecular Case Studies. 2016. p. a000752. Available from: http://dx.doi.org/10.1101/mcs.a000752

162. Gundem G, Van Loo P, Kremeyer B, Alexandrov LB, Tubio JMC, Papaemmanuil E, et al. The evolutionary history of lethal metastatic prostate cancer. Nature. 2015 Apr 1;520(7547):353–7.

163. Robinson D, Van Allen EM, Wu YM, Schultz N, Lonigro RJ, Mosquera JM, et al. Integrative Clinical Genomics of Advanced Prostate Cancer. Cell. 2015 Jul 16;162(2):454.

164. Armenia J, Wankowicz SAM, Liu D, Gao J, Kundra R, Reznik E, et al. The long tail of oncogenic drivers in prostate cancer. Nat Genet. 2018 Apr 2;50(5):645–51.

165. Ylipää A, Kivinummi K, Kohvakka A, Annala M, Latonen L, Scaravilli M, et al. Transcriptome Sequencing Reveals PCAT5 as a Novel ERG-Regulated Long Noncoding RNA in Prostate Cancer. Cancer Res. 2015 Oct 1;75(19):4026–31.

166. Annala M, Kivinummi K, Tuominen J, Karakurt S, Granberg K, Latonen L, et al. Recurrent SKIL-activating rearrangements in ETS-negative prostate cancer. Oncotarget. 2015 Mar 20;6(8):6235–50.

167. Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL. Gene expression profiling predicts clinical outcome of prostate cancer. J Clin Invest. 2004 Mar;113(6):913–23.

168. Tomlins SA, Rhodes DR, Yu J, Varambally S, Mehra R, Perner S, et al. The role of SPINK1 in ETS rearrangement-negative prostate cancers. Cancer Cell. 2008 Jun;13(6):519–28.

169. Sinha A, Huang V, Livingstone J, Wang J, Fox NS, Kurganovs N, et al. The Proteogenomic Landscape of Curable Prostate Cancer. Cancer Cell. 2019 Mar 18;35(3):414–27.e6.

170. Iglesias-Gato D, Wikström P, Tyanova S, Lavallee C, Thysell E, Carlsson J, et al. The Proteome of Primary Prostate Cancer. Eur Urol. 2016 May;69(5):942–52.

171. Iglesias-Gato D, Thysell E, Tyanova S, Crnalic S, Santos A, Lima TS, et al. The Proteome of Prostate Cancer Bone Metastasis Reveals Heterogeneity with Prognostic Implications. Clin Cancer Res. 2018 Nov 1;24(21):5433–44.

172. Giambartolomei C, Seo JH, Schwarz T, Freund MK, Johnson RD, Spisak S, et al. H3K27ac HiChIP in prostate cell lines identifies risk genes for prostate cancer susceptibility. Am J Hum Genet. 2021 Dec 2;108(12):2284–300.

173. Shen MM, Abate-Shen C. Molecular genetics of prostate cancer: new prospects for old challenges. Genes Dev. 2010 Sep 15;24(18):1967–2000.

174. Rubin MA, Demichelis F. The Genomics of Prostate Cancer: A Historic Perspective. Cold Spring Harb Perspect Med [Internet]. 2019 Mar 1;9(3). Available from: http://dx.doi.org/10.1101/cshperspect.a034942

175. Taplin ME, Bubley GJ, Shuster TD, Frantz ME, Spooner AE, Ogata GK, et al. Mutation of the androgen-receptor gene in metastatic androgen-independent prostate cancer. N Engl J Med. 1995 May 25;332(21):1393–8.

176. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, et al. The genomic complexity of primary human prostate cancer. Nature. 2011 Feb 10;470(7333):214–20.

177. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature. 2020 Feb;578(7793):102–11.

178. Scaravilli M, Afyounian E, Nykter M, Visakorpi T, Latonen L. Integrative proteomics of prostate cancer. Current Opinion in Endocrine and Metabolic Research. 2020 Feb 1;10:43–9.

179. Cao D, Sun R, Peng L, Li J, Huang Y, Chen Z, et al. Immune Cell Proinflammatory Microenvironment and Androgen-Related Metabolic Regulation During Benign Prostatic Hyperplasia in Aging. Front Immunol. 2022 Mar 21;13:842008.

180. The Cancer Genome Atlas Program [Internet]. 2018 [cited 2021 Aug 27]. Available from: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

181. Dweep H, Gretz N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. Nat Methods. 2015 Aug;12(8):697.

182. Yevshin I, Sharipov R, Valeev T, Kel A, Kolpakov F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. Nucleic Acids Res. 2017 Jan 4;45(D1):D61–7.

183. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database [Internet]. 2017 Jan 1;2017. Available from: http://dx.doi.org/10.1093/database/bax028

184. Massie CE, Lynch A, Ramos-Montoya A, Boren J, Stark R, Fazli L, et al. The androgen receptor fuels prostate cancer by regulating central metabolism and biosynthesis. EMBO J. 2011 May 20;30(13):2719–33.

185. Pomerantz MM, Li F, Takeda DY, Lenci R, Chonkar A, Chabot M, et al. The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. Nat Genet. 2015 Nov;47(11):1346–51.

186. Stelloo S, Nevedomskaya E, Kim Y, Schuurman K, Valle-Encinas E, Lobo J, et al. Integrative epigenetic taxonomy of primary prostate cancer. Nat Commun. 2018 Nov 21;9(1):4900.

187. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

188. Mayrhofer M, DiLorenzo S, Isaksson A. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. Genome Biol. 2013 Mar 25;14(3):R24.

189. Yu Z, Liu Y, Shen Y, Wang M, Li A. CLImAT: accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data. Bioinformatics. 2014 Sep 15;30(18):2576–83.

190. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012 Feb 1;28(3):423–5.

191. Picard [Internet]. [cited 2021 Jun 4]. Available from: http://broadinstitute.github.io/picard/

192. Gaspar JM. Improved peak-calling with MACS2 [Internet]. Available from: http://dx.doi.org/10.1101/496521

193. Orchard P, Kyono Y, Hensley J, Kitzman JO, Parker SCJ. Quantification, Dynamic Visualization, and Validation of Bias in ATAC-Seq Data with ataqv. Cell Syst. 2020 Mar 25;10(3):298–306.e4.

194. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.

195. Pascovici D, Handler DCL, Wu JX, Haynes PA. Multiple testing corrections in quantitative proteomics: A useful but blunt tool. Proteomics. 2016 Sep;16(18):2448–53.

196. Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. 2014 Apr;42(8):e69.

197. Stark R, Brown G, Others. DiffBind: differential binding analysis of ChIP-Seq peak data. R package version [Internet]. 2011;100(4.3). Available from: http://129.217.206.11/packages/2.13/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf

198. Lun ATL, Smyth GK. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic Acids Res. 2016 Mar 18;44(5):e45.

199. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014 Jul 20;513(7518):382–7.

200. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010 May 28;38(4):576–89.

201. Hunter. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007 May 1;9:90–5.

202. Koivisto P, Hyytinen E, Palmberg C, Tammela T, Visakorpi T, Isola J, et al. Analysis of genetic changes underlying local recurrence of prostate carcinoma during androgen deprivation therapy. Am J Pathol. 1995 Dec;147(6):1608–14.

203. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan 1;29(1):24–6.

204. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013 Mar;14(2):178–92.

205. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb 19;518(7539):317–30.

206. Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol. 2020 Feb 3;21(1):22.

207. Pladsen AV, Nilsen G, Rueda OM, Aure MR, Borgan Ø, Liestøl K, et al. DNA copy number motifs are strong and independent predictors of survival in breast cancer. Commun Biol. 2020 Apr 2;3(1):153.

208. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012 Apr 29;30(5):413–21.

209. Poell JB, Mendeville M, Sie D, Brink A, Brakenhoff RH, Ylstra B. ACE: absolute copy number estimation from low-coverage whole-genome sequencing data. Bioinformatics. 2019 Aug 15;35(16):2847–9.

210. Sauer CM, Eldridge MD, Vias M, Hall JA, Boyle S. Absolute copy number fitting from shallow whole genome sequencing data. bioRxiv [Internet]. 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.07.19.452658.abstract

211. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. 2012 Mar 13;13(4):227–32.

212. Bock C, Tomazou EM, Brinkman AB, Müller F, Simmer F, Gu H, et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol. 2010 Oct;28(10):1106–14.

213. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009 May;6(5):377–82.

214. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015 May;12(5):453–7.

215. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009 Oct 9;326(5950):289–93.

216. Cremer T, Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet. 2001 Apr;2(4):292–301.

217. Chakraborty A, Ay F. Identification of copy number variations and translocations in cancer cells from Hi-C data. Bioinformatics. 2018 Jan 15;34(2):338–45.

218. Harewood L, Kishore K, Eldridge MD, Wingett S, Pearson D, Schoenfelder S, et al. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. Genome Biol. 2017 Jun 27;18(1):125.

PUBLICATION
I

Segmentum: a tool for copy number analysis of cancer genomes

Afyounian E., Annala M., Nykter M.

BMC Bioinformatics. 2017 Apr 13;18(1):215.

DOI: 10.1186/s12859-017-1626-8

Afyounian *et al. BMC Bioinformatics* (2017) 18:215
DOI 10.1186/s12859-017-1626-8

BMC Bioinformatics

SOFTWARE

Open Access

Segmentum: a tool for copy number analysis of cancer genomes

Ebrahim Afyounian, Matti Annala and Matti Nykter*

Abstract

Background: Somatic alterations, including loss of heterozygosity, can affect the expression of oncogenes and tumor suppressor genes. Whole genome sequencing enables detailed characterization of such aberrations. However, due to the limitations of current high throughput sequencing technologies, this task remains challenging. Hence, accurate and reliable detection of such events is crucial for the identification of cancer-related alterations.

Results: We introduce a new tool called Segmentum for determining somatic copy numbers using whole genome sequencing from paired tumor/normal samples. In our approach, read depth and B-allele fraction signals are smoothed, and double sliding windows are used to detect breakpoints, which makes our approach fast and straightforward. Because the breakpoint detection is performed simultaneously at different scales, it allows accurate detection as suggested by the evaluation results from simulated and real data. We applied Segmentum to paired tumor/normal whole genome sequencing samples from 38 patients with low-grade glioma from the TCGA dataset and were able to confirm the recurrence of copy-neutral loss of heterozygosity in chromosome 17p in low-grade astrocytoma characterized by *IDH1/2* mutation and lack of 1p/19q co-deletion, which was previously reported using SNP array data.

Conclusions: Segmentum is an accurate, user-friendly tool for somatic copy number analysis of tumor samples. We demonstrate that this tool is suitable for the analysis of large cohorts, such as the TCGA dataset.

Keywords: Somatic copy number analysis, Loss of heterozygosity, Segmentation, Whole-genome sequencing, Cancer

Background

Somatic copy number alterations (SCNA) are a group of genomic aberrations commonly observed in many cancers [1]. Copy number is the number of copies per cell of a particular gene or DNA sequence. Somatically acquired chromosomal rearrangements such as deletions and duplications may change the copy number of a gene. Consequently, the expression level of a gene is often correlated with its copy number [2] - a phenomenon known as the gene dosage effect. Loss of heterozygosity (LOH) is an event in which one of the two alleles at a heterozygous locus is lost due to segmental aneuploidy, gene conversion, mitotic recombination, or mitotic nondisjunction [3]. LOH events involving tumor suppressor genes such as *PTEN*, *RB1*, and *TP53* have been observed in many cancer. LOH may alter gene expression. For example, monoallelic expression (MAE), which is the expression of a gene from only one of two alleles in a diploid organism, is associated with LOH [3]. By analyzing a cohort of 23 triple-negative breast cancer patients, Ha et al. [3] have shown that LOH is a prominent aberration in this type of cancer, and modulates a significant portion of the transcriptome in the form of MAE. Copy-neutral LOH (cnLOH) is a specific type of LOH that occurs when the lost allele is replaced with a duplicated copy of the surviving allele, resulting in the copy number remaining unchanged. Suzuki et al. have shown recurring cnLOH at chromosome 17p (harboring *TP53* gene) in low-grade astrocytoma [4]. The altered expression of genes with allelic imbalance due to LOH events may bring about selective advantages for tumorigenesis and tumor progression. Additionally, regions with cnLOH may harbor genes with driver mutations [5]. Hence, accurate and reliable detection and characterization of events, such as SCNAs and LOH,

* Correspondence: matti.nykter@uta.fi
Faculty of Medicine and Life Sciences and BioMediTech institute, University of Tampere, Tampere, Finland

are crucial for the identification of prospective cancer-related genes, such as tumor suppressor genes and oncogenes, and eventually for informing new approaches to treat cancer [6].

High throughput sequencing (HTS)-based SCNA detection approaches (including both whole exome sequencing (WES) and whole genome sequencing (WGS)) have become popular due to their potential for accurate copy number estimation and breakpoint detection with single nucleotide accuracy. However, the short read length of current HTS technologies makes it difficult to map some reads to unique locations in the genome. Furthermore, due to GC-content bias, GC-content-rich regions in the genome will have increased number of reads. These ambiguities make accurate estimation of coverage and consequently copy number a challenge [7]. Additionally, tumor ploidy and normal cell contamination introduce further challenges in SCNA detection [8].

HTS-based copy number analysis is, in most cases, based on read depth (RD) estimations at each genomic location and further segmentation and quantification of the RD profiles into segments of consistent copy number (Additional file 1: Table S1 for a list of SCNA tools) [9, 10]. However, such tools are only capable of detecting deletions and duplications. Recently, RD-based analysis has been augmented to identify cnLOH events by incorporating information from an alternate allele's fraction at heterozygous single nucleotide polymorphism (SNP) positions (or B-allele fraction (BAF)). The BAF of a heterozygous SNP has an expected value of 0.5 in normal diploid cells. Deviation from 0.5 in the heterozygous SNP BAF points to an aberration. In the case of cnLOH, BAF values are expected to be either 0 or 1 in a pure tumor population. Tools such as Control-FREEC [11], Patchwork [12], and CLImAT [13] incorporate BAF data to extend SCNA detection. Control-FREEC determines the breakpoints using a least absolute shrinkage estimator (LASSO) regression. Sample ploidy is provided by the user to Control-FREEC. It also evaluates and corrects for normal cell contamination, GC-content, and mapability biases while inferring the copy number profile of a tumor genome. Patchwork performs GC and positional normalization and segments the genome using a circular binary segmentation (CBS) algorithm. It also estimates normal cell contamination and tumor ploidy. CLImAT implements corrections for GC-content and mapability bias and models the RD and BAF data with a hidden Markov model (HMM) to infer the somatic copy number variation, normal cell contamination and tumor ploidy (Additional file 1: Overview of Tools section for more details on these tools). While the above tools are well-suited for SCNA detection, their use has some limitations. Control-FREEC and Patchwork utilize computationally costly models, which leads to long analysis times. The main motivation of our study was to develop an accurate and user-friendly tool that could be used to analyze large WGS datasets, such as the cancer genome atlas (TCGA) datasets. In our approach, the RD and BAF signals are smoothed, and double sliding windows subsequently are used to detect breakpoints, which makes our approach fast and straightforward. Because the breakpoint detection is performed simultaneously at different scales, it allows accurate detection. Our tool, Segmentum, is freely available under MIT license at: https://github.com/eafyounian/Segmentum (Additional file 2 contains the software code. For the lates version of the software code please visit the project's online repository).

Implementation

Pipeline

Segmentum was developed and written in the Python programming language (version 3) and requires the SciPy library to be installed (If the user wishes to use the 'plot' sub-command to inform parameter value selection, matplotlib library is also required). Segmentum employs SAMtools to extract RD and heterozygous SNPs BAF data from BAM files containing WGS data. These constitute the inputs required by Segmentum to perform copy number analysis. Figure 1 illustrates the

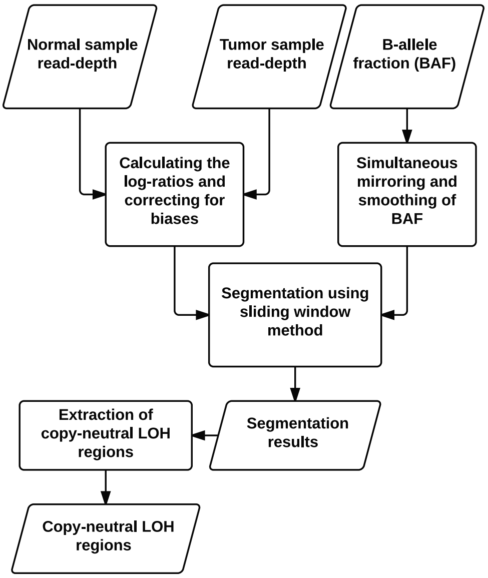

Normal sample read-depth
Tumor sample read-depth
B-allele fraction (BAF)
Calculating the log-ratios and correcting for biases
Simultaneous mirroring and smoothing of BAF
Segmentation using sliding window method
Extraction of copy-neutral LOH regions
Segmentation results
Copy-neutral LOH regions

Fig. 1 Segmentum pipeline. Normal and tumor RDs are used to calculate RD log-ratios. RD log-ratios are then corrected for biases. BAF data are simultaneously mirrored and smoothed. Using RD log-ratios and BAF, the genome is segmented with a double sliding window method. Segmentation results are used to identify cnLOH regions in the genome (see the following sections for more details on each step)

Segmentum pipeline. Each step is explained in more detail in the following sections.

RD extraction and BAF calculation for heterozygous SNPs

To extract the RD from the BAM files, the genome is divided into bins of user-defined length (2 kbp by default) and the number of reads overlapping each bin is counted to determine the RD at each bin. To calculate the BAF values, heterozygous SNPs in the normal sample are identified at known SNP sites in the human genome (based on SNP annotations such as those produced by the 1000 Genomes project). Next, the number of reference and alternative alleles at each heterozygous SNP position is extracted from the tumor sample and the BAF for the i^{th} heterozygous SNP is calculated using the following equation:

$$BAF_i = \frac{alt_i}{alt_i + ref_i}$$

where alt_i and ref_i refer to the alternative and reference allele, respectively, of the i^{th} heterozygous SNP.

It should be noted that by default reads with mapping quality score 10 are filtered out before RD extraction and BAF calculation in order to address the challenges raised by reads not mapping to a unique region in the genome (the read filtration criterion based on the mapping quality score is a parameter to Segmentum and can be set by the user).

Log-ratio calculation

The RD log-ratio is calculated using the following equation:

$$logr_i = log_2\left(\frac{tRD_i}{nRD_i}\right)$$

where $logr_i$ is the log-ratio of the i^{th} genomic window and tRD_i and nRD_i are RDs extracted from the i^{th} genomic window of a specific size (determined by user; default is 2 kbp) for the tumor and normal samples, respectively.

Differences in the total number of aligned reads in the normal and tumor samples may bias the estimation of the RD log-ratios. The correction was performed by finding the mode of log-ratio values for each chromosome and subtracting the median of all of the modes from each log-ratio value. It should be noted that median, in the correction step, is robust to the changes in one mode. For instance, one chromosomal arm having a copy number change has no effect on the correction since it only affects one of the chromosomal modes.

Mirroring and smoothing of the BAF values

The BAF of a heterozygous SNP has an expected value of 0.5 in normal diploid cells. In the presence of somatic copy number alterations, the BAF can diverge from 0.5 if the relative abundance of the two alleles changes. To make smoothing and segmentation of BAF data possible, the BAF values must be mirrored about the 0.5 axis so that the B allele fraction always represents the allele fraction of the dominant allele. Without this mirroring step, the BAF values will be symmetric about the BAF = 0.5 axis and smoothing will underestimate the absolute divergence from 0.5 [14]. In this study, a median filter is used for smoothing the BAF data. Simultaneous mirroring and smoothing is implemented using the following equation:

$$cBAF_i = H * |0.5 – M_9(BAF_i)| + (1 – H) * M_9(|0.5 – BAF_i|).$$

where BAF_i is the BAF value for the i^{th} heterozygous SNP, $cBAF_i$ is the simultaneously mirrored and smoothed BAF_i, H is a heterozygosity measurement calculated with the following equation: $H = 1 - 2 * |0.5 - x|$, and M_9 refers to applying a median filter to 9 SNPs in the vicinity of and including the i^{th} SNP.

Segmentation using a double sliding window approach

To detect changes in the RD log-ratio and BAF signals, two non-overlapping, fixed-sized windows (determined by the user) are slid over the RD log-ratio and BAF values and a compound score (S) is calculated for each of the adjacent two windows. If the compound score is greater than 1, a change is detected and a breakpoint is placed at the place where the two windows touch each other. The compound score is calculated using the following equation:

$$S = \frac{|\,\overline{logr_{win_i}} - \overline{logr_{win_{i+1}}}\,|^2}{\tau_{logr}} + \frac{|\overline{cBAF_{win_i}} - \overline{cBAF_{win_{i+1}}}\,|^2}{\tau_{BAF}}$$

where $\overline{logr_{win_i}}$ is the mean of the RD log-ratio values in the i^{th} window, $\overline{cBAF_{win_i}}$ is the mean of the mirrored and smoothed BAF values in the i^{th} window, τ_{logr} and τ_{BAF} are thresholds for the absolute mean difference in the RD log-ratios and the absolute mean difference in the BAF values in the two adjacent windows, respectively.

It is possible that some breakpoints will not be detected by a single pass of a double sliding window due to a given window size. Thus, to increase the sensitivity, Segmentum analyzes the signals for the detection of breakpoints multiple times with different window-sizes and thresholds. Each new window is 1.5 times larger than the previous one. The increase in the window size decreases the detection thresholds. This is due to the fact that increasing the window size increases the sample size (assuming sampling from normal distribution with $N(\mu, \sigma^2)$) and consequently decreases the standard deviation of the mean (mean having probability distribution of $N\left(\mu, \sigma^2\big/_{n}\right)$). The new standard deviation of the mean when window size is increased 1.5 times is $\frac{1}{\sqrt[2]{1.5}}$ times the old standard deviation. Let $\tau = \alpha\sigma$ where τ is the threshold and α is a scalar and σ is the standard deviation. It follows that:

$$\tau_{new} = \alpha.\sigma_{new} = \frac{1}{\sqrt[2]{1.5}}.\,\alpha.\sigma_{old} = \frac{1}{\sqrt[2]{1.5}}.\,\tau_{old}$$

Thus both the τ_{logr} and τ_{BAF} thresholds are updated using the following equation:

$$\tau_{new} = \frac{1}{\sqrt[2]{1.5}} * \tau_{old}$$

The process of increasing the window-size is continued as long as the updated thresholds are greater than the thresholds for the merging two consecutive segments (see below). After detecting all the breakpoints, a consensus list of breakpoints is created by accepting all of the breakpoints detected by the first pass of the double sliding window and adding the breakpoints detected from the larger windows to the list only if the breakpoint is not in the vicinity of an existing breakpoint in the list (i.e., $|cp_{current} - cp_{existing}| > window\ size$, where cp is a detected breakpoint). Consensus breakpoints are used to create the segments. Two consecutive breakpoints constitute a segment. For each segment, the average RD log-ratio and average mirrored and smoothed BAF is calculated. Two consecutive segments are merged if the following conditions are met:

$$\left|\, \overline{logr_{seg_i}} - \overline{logr_{seg_{i+1}}} \,\right| < \tau_{merge_{logr}}$$

$$and \quad \left|\, \overline{cBAF_{seg_i}} - \overline{cBAF_{seg_{i+1}}} \,\right| < \tau_{merge_{BAF}}$$

where $\overline{logr_{seg_i}}$ is the mean RD log-ratio of the i^{th} segment, $\overline{cBAF_{seg_i}}$ is the mean mirrored and smoothed BAF of the i^{th} segment, and $\tau_{merge_{logr}}$ and $\tau_{merge_{BAF}}$ (determined by user) are the RD log-ratio and BAF merging thresholds, respectively.

Detection of cnLOH events within a single sample

A segment is considered to be a cnLOH segment if the following conditions are met:

$$\left|\, \overline{logr_{seg_i}} \,\right| < \tau_{cnLOH_{logr}} \quad and \quad \left(0.5 - \overline{cBAF_{seg_i}}\right) < \tau_{cnLOH_{BAF}}$$

where $\overline{logr_{seg_i}}$ is the mean RD log-ratio of the i^{th} segment, $\overline{cBAF_{seg_i}}$ is the mean mirrored and smoothed BAF of the i^{th} segment, $\tau_{cnLOH_{logr}}$ and $\tau_{cnLOH_{BAF}}$ (determined by the user) are thresholds for calling a cnLOH segment.

Detection of recurrent cnLOH regions across multiple samples

To find genomic regions with recurrent cnLOH events, all cnLOH regions for individual samples are identified following the procedure described earlier. Then, the number of occurrences of a cnLOH event for a specific region across multiple samples is counted using an interval tree data structure (Additional file 1: Figure S1).

Simulator

To evaluate Segmentum in terms of segmentation accuracy, a simulator capable of simulating whole-genome RD for both normal and tumor samples and BAF based on events such as deletions, amplifications and cnLOH was developed. The simulator receives a normal sample RD data and outputs 4 sets of data including the simulated normal and tumor RD, BAF data and a ground truth. First, the simulator learns the distribution of the RD data from the provided normal sample by simply counting the number of times two consecutive RD values (e.g., 368 and 299) occur together throughout the genome (Additional file 1: Figure S2). The learned distribution also accounts for the inherent noise in the RD data. Next, inverse transform sampling (Smirnov transform) is used to generate RD values for each position in the genome based on the learned distribution. Then, noise is removed using a median filter. A normal RD is constructed by adding independent Poisson noise to the simulated RD data. To construct the tumor RD, two copy number tracks (because autosomal chromosomes come in maternal and paternal pairs) harboring random SCNAs are constructed. The tumor sample RD is calculated using the copy number tracks, the simulated normal sample RD and the normal sample contamination (i.e., a parameter determined by user). To construct the BAF data, heterozygous SNPs are initially randomly distributed across the genome (1 heterozygous SNP per 1.5 Kbp). The number of B-alleles at a heterozygous SNP is calculated using a binomial distribution with the parameters n (total number of reads at heterozygous SNP position) and p (probability that a read is coming from the B-allele). n is extracted from the simulated normal RD at heterozygous SNP positions. p is calculated using the two constructed copy number tracks and the normal sample contamination. Once the number of B-alleles is calculated, it is used to calculate the BAF values (Additional file 1: Figures S3 and S4 for the simulator pipeline and the simulated data visualized in the integrative genomics viewer (IGV) [15], respectively).

Results

Segmentum segmentation accuracy for the simulated data

Using the simulator (see the 'Simulator' section for more details), RD data for both normal and tumor samples and BAF values for heterozygous SNPs from the tumor sample as well as a ground truth were simulated with different percentages of normal contamination (an example set of simulated data is available at Segmentum's online repository. See the 'Availability and requirements' section for the link to the repository). The simulated data were analyzed by Segmentum. The segmentation results were evaluated against the ground truth. The precision, recall, and the F-measure values were calculated based on this evaluation (Fig. 2 and Additional file 1 for the definitions of precision, recall, and F-measure).

Segmentum segmentation accuracy for real data compared to other tools

To assess segmentation accuracy of Segmentum for real data, paired tumor/normal whole genome sequencing samples (30x < coverage < 100x) from 10 individuals diagnosed with low-grade glioma (LGG) were downloaded from the TCGA dataset and used as is. Furthermore, segmentation results from SNP-array data (level 3 data) (completed by TCGA using an Affymetrix Genome-wide human SNP array 6.0) was used as ground truth (Additional file 1: Table S3). Segmentum's results were evaluated against Control-FREEC, Patchwork, and CLImAT as competing tools. To evaluate the segmentation accuracy, the genome was broken into 100 bp. blocks (excluding all blocks in centromeres and sex chromosomes). Using block annotations from different tools, genome-wide proportions of the blocks annotated as SCNA by different combinations of tools were calculated and the results were illustrated by a Venn diagram (Fig. 3).

Additionally, to measure the pairwise degree of similarity of the segmentation results between two tools, the *Jaccard similarity index* (JSI) was calculated for all of the pairs using the following equation:

$$JSI = \frac{|\cap pair|}{|\cup pair|}$$

where $|\cap pair|$ and $|\cup pair|$ are the cardinalities of intersection and union, respectively. Intersection and union values were extracted from the Venn diagrams. Figure 4 represents a heat map of the JSI values for each pair of tools averaged over 10 TCGA LGG samples. According to the heat map, on average, Segmentum produces the most similar results to the SNP array segmentation results with a JSI score of 0.9, followed by Patchwork with a JSI score of 0.86.

Similar evaluations using low coverage data (6x average coverage) are shown in Additional file 1: Figures S5 and S6. The low coverage data is comprised of the paired tumor/normal whole genome sequencing samples of 10 individuals diagnosed with prostate adenocarcinoma (PRAD). With regard to the low coverage data, Patchwork produces the most similar results to the SNP array segmentation results with a JSI score of 0.93, followed by Segmentum with a JSI score of 0.88. Additional file 1: Table S4 contains the names of the 10 TCGA PRAD samples (Additional file 1: Tables S6-S10 represent the parameter values used for running the

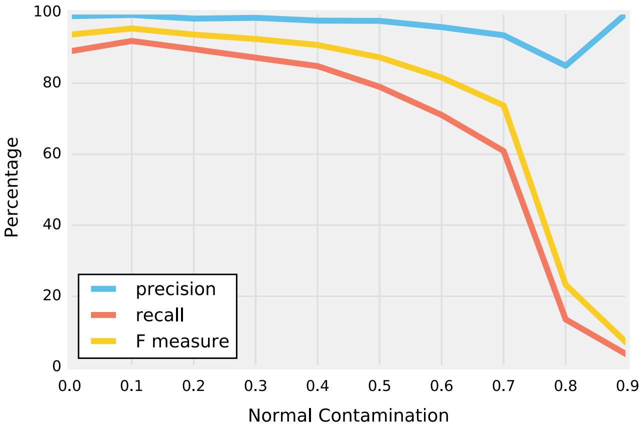


Fig. 2 Segmentation accuracy of Segmentum for simulated data with different degrees of normal contamination. Estimated precision, recall, and F-measure values for simulated data at different normal contamination levels (Additional file 1, Derivation of the precision, recall, and F-measure of the simulated data)

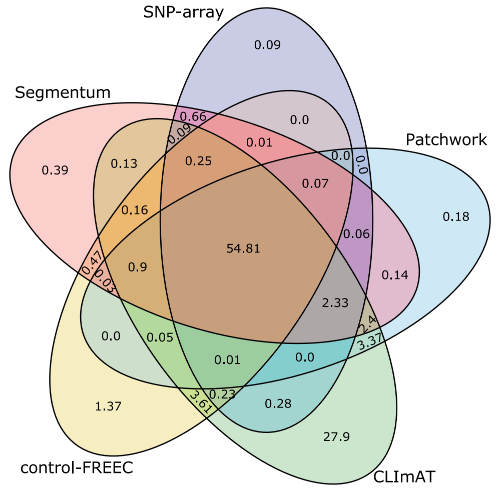


Fig. 3 Comparison of the SCNA results with different tools and the SNP array (ground truth). Venn diagram values (averaged over ten TCGA LGG samples) represent the percentage of overlap among the SCNA calls

competing tools. Additional file 1: Segmentum's parameter value selection section provides guidance on selecting parameter values for Segmentum. Additional file 1: Figure S9 represents an example plot made by Segmentum's 'plot' sub-command that can be used to guide the parameter value selection).

Segmentum segmentation accuracy for the subsampled real data

To assess the segmentation accuracy of Segmentum for real data with respect to sample's coverage, we subsampled one of the LGG samples (i.e. TCGA-CS-5395) at different subsampling fractions (i.e. 75%, 50%, 25%, 10%, and 5%) using Samtools (version 1.3.1). We analyzed each subsample by Segmentum and benchmarked it against ground truth in the same manner as explained earlier. Figure 5 represents the JSI scores for each subsample (Additional file 1: Figure S7 shows the average coverage of the subsamples for normal and tumor pairs). It can be seen that Segmentum reaches high accuracies even with low coverage data. For instance, the accuracy for the 10%-fraction subsample was 93.4% (where the average coverage for tumor and normal subsamples were 3 and 4 respectively).

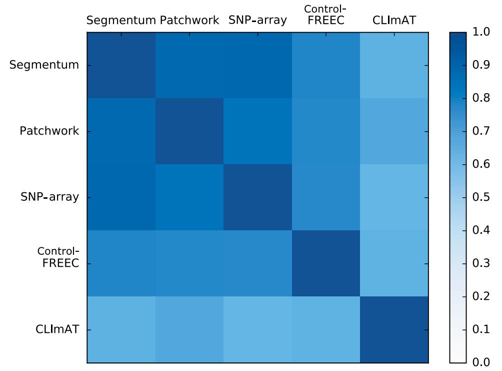


Fig. 4 Pairwise JSI scores averaged over ten TCGA LGG samples. JSI scores range between 0 and 1, where 0 means no similarity and 1 represents identical results between two tools

It should be noted that as the coverage decreases the number of identified heterozygous SNPs decreases (Additional file 1: Figure S8). For instance, for the 10%-fraction subsample only 1997 heterozygous SNPs were identified from the entire genome (in contrast to the original sample where the number of identified heterozygous SNPs was more than 3 million SNPs). Even though Segmentum is shown to work with low coverage data, one should note the implications of low amounts of detected heterozygous SNPs on the reliable detection of cnLOH events.

Time usage evaluation

All of the computations were completed on the same UNIX server. Table 1 shows the average time required by each tool to perform the analysis for 10 TCGA LGG samples (30x < coverage < 100x). Based on the results, on average, CLImAT appears to be the fastest, followed by Segmentum, Patchwork, and Control-FREEC. It should be noted that to assign the allele-specific copy number to genomic segments, Patchwork requires users to determine some parameter values by interpreting plots produced by the tool, and this interpretation time is not included here. Additionally, the time required to create the pileup files used by Patchwork and Control-FREEC is different due to the use of different parameter values in SAMtools. It should be noted that time required for making pileup files can be decreased by parallelizing the process on machines with multiple cores or on computer clusters (e.g. by assigning one core to each chromosome). Similarly, BAF calculation for Segmentum can be parallelized. However, since this is not a core feature of the benchmarked tools and not all tools support parallelization, to be fair, only the required linear time is reported here. A similar time usage evaluation, using low coverage data (average coverage 6x), is shown in Additional file 1: Table S2. With regard to the low coverage data, Segmentum comes second after CLImAT in terms of analysis time, which is consistent with the results from the high coverage data.

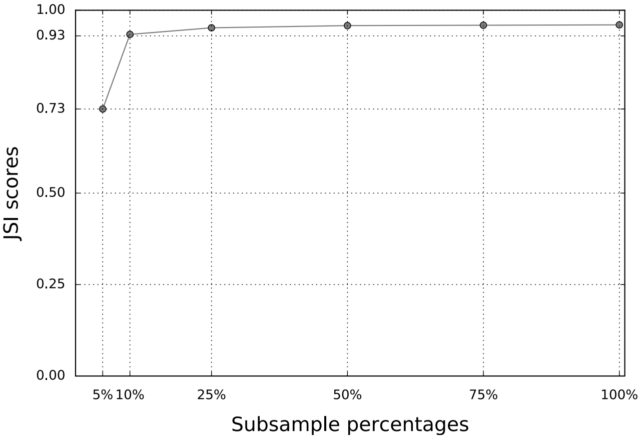


Fig. 5 Pairwise JSI scores (Segmentum vs. SNP array as ground truth) for different subsamples. JSI scores range between 0 and 1, where 0 means no similarity and 1 represents identical results between two tools

Recurrent cnLOH detection case study

In a study of lower grade gliomas (LGGs), i.e., grade II and III gliomas, Suzuki et al. [4] characterized the mutational landscape of these glioma types by dividing them into 3 distinct subtypes based on their distinct sets of mutations and clinical behaviors. These subtypes are distinguished with the following criteria: (1) mutation in *IDH1/2* accompanied by co-deletion of chromosomes 1p and 19q (subtype I), (2) mutation in *IDH1/2* without co-deletion of chromosomes 1p and 19q (subtype II), and (3) *IDH1/2* wild type (subtype III). Of interest to our study was the recurrence of cnLOH events in chromosome 17p in subtype II [4]. To show the ability of Segmentum to detect such aberrations from large datasets, 38 paired-end WGS samples from the TCGA dataset (30x < coverage < 100x)) for patients diagnosed with LGG were downloaded and analyzed by Segmentum. We were able to distinguish all three subtypes as characterized in [4], including the recurrence of cnLOH in subtype II at chromosome 17p. We also identified a fourth subtype with a mutation in *IDH1/2* without co-deletion of chromosomes 1p and 19q and no cnLOH at 17p (Fig. 6).

Table 1 Average tool analysis time for high coverage data (30x < coverage < 100x)

Tool	Average preparation time	Average analysis time
Segmentum	- 10 h 34 min for extracting RD from normal or tumor BAM file - 4 h 25 min for calculating BAF values	- 1 min 45 s
Patchwork	- 29 h 37 min for creating pileups from normal or tumor BAM file	- 3 h 56 min
Control-FREEC	- 33 h 28 min for creating pileups from normal or tumor BAM file	- 7 h 11 min
CLImAT	- 2 h 12 min for extracting RD	- 29 min

Discussion

By comparing the simulated (Fig. 2) and real data (Figs. 3, 4 and 5 and Additional file 1: Figures S5 and S6), we can conclude that Segmentum can recover true copy number aberrations with high accuracy even when the coverage is as low as ~4 reads (Fig. 5, Additional file 1: Figure S7). On average, Segmentum produces results that are the most concordant with the copy number aberrations identified from the SNP array data (i.e. ~90% of concordance) (Fig. 4). As shown in Table 1, our tool is more than twice as fast as the second best performing tool in terms of accuracy. Segmentum is also the second fastest tool after CLImAT compared to the other tools evaluated in this study (Table 1). However, CLImAT ranks last in terms of accuracy (Fig. 4). One explanation for the speed of CLImAT is that it computes the BAF values for a subset of known SNPs (~13.7 million SNPs that are retrieved from the dbSNP database [16]). In contrast, Segmentum, computes the BAF values for heterozygous SNPs determined from the 1000 Genomes project's SNP list (~85 million SNPs) [17]. The other reason for the speed of CLImAT might be that it does not require a normal sample for analysis.

As the normal contamination in the simulated data increases, the number of false negatives increases and the recall rate decreases (Fig. 2). However, within the ranges of realistic amounts of normal contamination (i.e. ~30% to 40%), Segmentum performs consistently well.

Segmentum is able to report recurrent cnLOH regions across multiple cancer genome samples; a characteristic

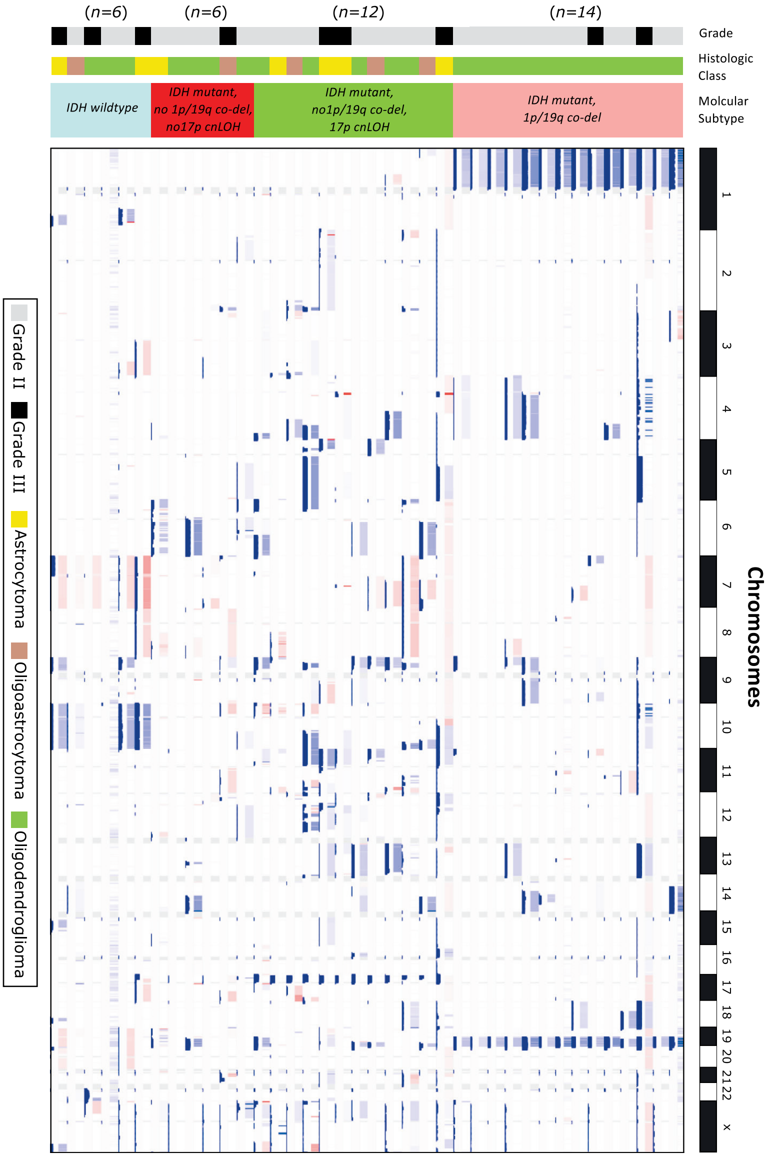


Fig. 6 SCNA landscape in grade II and III gliomas. WHO-grade, histological class, and molecular subtype classification are shown by color as indicated. The thirty-eight samples are divided into 4 distinct subtypes based on the occurrence of a mutation in *IDH1/2*, co-deletion of chromosomes 1p and 19q and the presence of 17p cnLOH. Deletions and amplifications are visualized by boxes with different shades of blue and red, respectively. White regions are either normal or cnLOH regions. The bar charts below each box represent the mirrored and smoothed BAF values. Large mirrored and smoothed BAF values (close to 0.5) point to heterozygous SNP allelic imbalance. In the second subtype (from the top), at chromosome 17p, recurring cnLOH is apparent where the bar charts point to large mirrored and smoothed BAF values, though no deletion or amplification is detected at that region (Additional file 1: Table S5 for TCGA LGG sample barcode names)

of cancer genomes that has been neglected until recently [4]. By applying Segmentum to TCGA data, we were able to recover recurrent cnLOH events from low-grade glioma samples that were reported earlier by SNP array-based data analysis. It is worth mentioning that Segmentum can work in two modes, i.e., with or without BAF value. In the case where BAF values are not used, Segmentum cannot detect regions with cnLOH. Furthermore, Segmentum is capable of reliably segmenting the cancer genome using both high (Figs. 3 and 4) and low (Fig. 5 and Additional file 1: Figures S5 and S6) sequence coverage data. However, with the low sequence coverage data, the estimated BAF values for the heterozygous SNPs will be less reliable. This is also reflected in Additional file 1: Figure S8, where it is shown that the number of detected heterozygous SNPs drop as the average coverage decreases. The implications of low amounts of detected heterozygous SNPs on the reliable detection of cnLOH events should not be overlooked.

Even though we have shown that Segmentum is highly accurate at recovering the true copy number, other tools in this study do more than just segmenting the genome. For instance, CLImAT and Patchwork are capable of estimating tumor ploidy and tumor purity and consequently, reporting the integral copy numbers for each segment. Patchwork and Control-FREEC are also capable of reporting the genotype of each segment and CLImAT reports the genotype for each SNP within each segment. This is in contrast to Segmentum that only reports the mean RD log-ratio and BAF value of each segment. However, tools such as ABSOLUTE [18] or THetA [19] can be used to estimate tumor impurity and ploidy from Segmentum's segmentation result, meaning that Segmentum can be used as part of a larger tumor evolution analysis pipeline. Finally, a strength of our tool is its minimum dependence on third party tools, with the exception of SAMtools, for calculating the RD and BAF.

Conclusions

We have developed Segmentum as a tool for the identification of SCNAs, including cnLOH in tumor samples, using WGS data. We have shown that Segmentum is accurate and fast with regards to other state-of-the-art tools, making it suitable for analyzing cohorts with a large number of samples, such as TCGA cohorts.

Availability and requirements

Project name: Segmentum
Project homepage: https://github.com/eafyounian/Segmentum
Operating system(s): Linux
Programming language: Python
Other requirements: SciPy, Samtools, and matplotlib if the 'plot' sub-command is used.
License: MIT license
Any restrictions to use by non-academics: None

Additional files

Additional file 1: This file contains supplementary information, tables and figures supporting the manuscript. **Figure S1.** Detection of regions harboring recurrent cnLOH across multiple samples. **Figure S2.** Read depth spatial correlation. **Figure S3.** Simulator pipeline. **Figure S4.** Simulated data visualized in Integrative Genomics Viewer (IGV). **Figure S5.** Comparison of SCNA results from different tools and SNP array (ground truth) for low sequence coverage data. **Figure S6.** Pairwise JSI scores for low sequence coverage data (averaged of 10 TCGA PRAD samples) **Figure S7.** Subsample average coverages in the subsampling evaluation. **Figure S8** Detected number of heterozygous SNPs in different subsamples **Figure S9.** Copy number – B-allele fraction clusters. **Table S1.** List of SCNA tools using WGS data. **Table S2.** Average tool analysis time for low sequence coverage data (average coverage 6x). **Table S3** TCGA LGG sample barcode names and the estimated sample purity by ABSOLUTE. **Table S4.** TCGA PRAD sample barcode names and the estimated sample purity by ABSOLUTE. **Table S5.** TCGA LGG sample barcode names categorized based on inferred subtype. **Table S6.** Parameter values for running DFExtract. **Table S7.** Parameter values for running CLImAT. **Table S8.** Parameter values for running Patchwork for 10 TCGA LGG samples. **Table S9.** Parameter values for running Patchwork for 10 TCGA PRAD samples. **Table S10.** Parameter values for running Control-FREEC. (DOCX 857 kb)

Additional file 2: Software code. This compressed file contains the software code (for the latest version of the software code please visit the project's online repository). (ZIP 32 kb)

Abbreviations

BAF: B-Allele fraction; BAM: Binary alignment map; CBS: circular binary segmentation; CGH: Array comparative genomic hybridization; cnLOH: Copy-neutral loss of heterozygosity; CNV: Copy number variation; dbSNP: Single nucleotide polymorphism database; DNA: Deoxyribonucleic acid; FISH: Fluorescent in situ hybridization; HMM: Hidden Markov model; HTS: High throughput sequencing; IGV: Integrative genomics viewer; JSI: Jaccard similarity index; LASSO: Least absolute shrinkage eStimatOr; LGG: Low grade glioma; LOH: Loss of heterozygosity; MAE: Monoallelic expression; PRAD: PRostate ADenocarcinoma; RD: Read-depth; SAM: Sequence alignment/map; SCNA: Somatic copy number alteration; SNP: Single nucleotide polymorphism; TCGA: The cancer genome atlas; WES: Whole exome sequencing; WGS: Whole genome sequencing

Acknowledgments

The results published here are in part based upon data generated by The Cancer Genome Atlas project (dbGaP Study Accession: phs000178.v9.p8) established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov. The authors would also like to acknowledge CSC — IT Center for Science Ltd. (https://www.csc.fi/csc) for providing the computational resources.

Funding

This work was funded by the Academy of Finland (project no. 269474), Cancer Society of Finland, and Sigrid Juselius Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

EA, MA implemented the method. EA performed the data analysis and drafted the manuscript. MA initiated the study and designed the method. MN supervised the study. All authors edited and approved the manuscript.

Authors' information

EA is a Ph.D. student at Faculty of Medicine and Life Sciences and BioMediTech institute, University of Tampere, Tampere, Finland.
MA is a Ph.D. student at Faculty of Medicine and Life Sciences and BioMediTech institute, University of Tampere, Tampere, Finland.
MN is a professor of bioinformatics at Faculty of Medicine and Life Sciences and BioMediTech institute, University of Tampere, Tampere, Finland.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable

Ethics approval and consent to participate

Not applicable

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 27 May 2016 Accepted: 6 April 2017
Published online: 13 April 2017

References

1. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, Ligon AH, Cho YJ, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye FJ, Sasaki H, Tepper JE, Fletcher JA, Tabernero J, Baselga J, Tsao MS, Demichelis F, Rubin MA, Janne PA, Daly MJ, Nucera C, Levine RL, Ebert BL, Gabriel S, Rustgi AK, Antonescu CR, Ladanyi M, Letai A, Garraway LA, Loda M, Beer DG, True LD, Okamoto A, Pomeroy SL, Singer S, Golub TR, Lander ES, Getz G, Sellers WR, Meyerson M. The landscape of somatic copy-number alteration across human cancers. Nature. 2010; 463(7283):899–905.
2. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Graf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, METABRIC Group, Langerod A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Borresen-Dale AL, Brenton JD, Tavare S, Caldas C, Aparicio S. The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. Nature. 2012; 486(7403):346–52.
3. Ha G, Roth A, Lai D, Bashashati A, Ding J, Goya R, Giuliany R, Rosner J, Oloumi A, Shumansky K, Chin SF, Turashvili G, Hirst M, Caldas C, Marra MA, Aparicio S, Shah SP. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. Genome Res. 2012; 22(10):1995–2007.
4. Suzuki H, Aoki K, Chiba K, Sato Y, Shiozawa Y, Shiraishi Y, Shimamura T, Niida A, Motomura K, Ohka F, Yamamoto T, Tanahashi K, Ranjit M, Wakabayashi T, Yoshizato T, Kataoka K, Yoshida K, Nagata Y, Sato-Otsubo A, Tanaka H, Sanada M, Kondo Y, Nakamura H, Mizoguchi M, Abe T, Muragaki Y, Watanabe R, Ito I, Miyano S, Natsume A, Ogawa S. Mutational landscape and clonal architecture in grade II and III gliomas. Nat Genet. 2015;47(5): 458–68.
5. Barresi V, Romano A, Musso N, Capizzi C, Consoli C, Martelli MP, Palumbo G, Di Raimondo F, Condorelli DF. Broad copy neutral-loss of heterozygosity regions and rare recurring copy number abnormalities in normal karyotype-acute myeloid leukemia genomes. Genes Chromosomes Cancer. 2010; 49(11):1014–23.
6. Stuart D, Sellers WR. Linking somatic genetic alterations in cancer to therapeutics. Curr Opin Cell Biol. 2009;21(2):304–10.
7. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 2013, 14(Suppl 11);S1-2105-14-S11-S1. Epub 2013 Sep 13.
8. Liu B, Morrison CD, Johnson CS, Trump DL, Qin M, Conroy JC, Wang J, Liu S. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. Oncotarget. 2013;4(11):1868–81.
9. Alkodsi A, Louhimo R, Hautaniemi S. Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. Brief Bioinform. 2015;16(2):242–54.
10. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12(5):363–76.
11. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012;28(3):423–5.
12. Mayrhofer M, DiLorenzo S, Isaksson A. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. Genome Biol. 2013;14(3):R24. -2013-14-3-r24.
13. Yu Z, Liu Y, Shen Y, Wang M, Li A. CLImAT: accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data. Bioinformatics. 2014; 30(18):1-8.
14. Staaf J, Lindgren D, Vallon-Christersson J, Isaksson A, Goransson H, Juliusson G, Rosenquist R, Hoglund M, Borg A, Ringner M. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. Genome Biol. 2008;9(9):R136. -2008-9-9-r136 . Epub 2008 Sep 16.
15. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.
16. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29(1):308–11.
17. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74.
18. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30(5):413–21.
19. Oesper L, Mahmoody A, Raphael BJ. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biol. 2013;14(7):R80. -2013-14-7-r80.

PUBLICATION
II

Integrative proteomics in prostate cancer uncovers robustness against genomic and transcriptomic aberrations during disease progression.

Latonen L., Afyounian E., Jylhä A., Nättinen J., Aapola U., Annala M., Kivinummi K., Tammela T.L., Beuerman R., Uusitalo H., Nykter M., Visakorpi T.

Nat Commun. 2018 Mar 21;9(1):1176.

DOI: 10.1038/s41467-018-03573-6

nature
COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-03573-6 OPEN

Integrative proteomics in prostate cancer uncovers robustness against genomic and transcriptomic aberrations during disease progression

Leena Latonen[1,2], Ebrahim Afyounian[1], Antti Jylhä[3], Janika Nättinen[3], Ulla Aapola[3], Matti Annala[1], Kati K. Kivinummi[1], Teuvo T.L. Tammela[4], Roger W. Beuerman[3,5,6,7,8], Hannu Uusitalo[3,9], Matti Nykter[1,10] & Tapio Visakorpi[1,2]

To understand functional consequences of genetic and transcriptional aberrations in prostate cancer, the proteomic changes during disease formation and progression need to be revealed. Here we report high-throughput mass spectrometry on clinical tissue samples of benign prostatic hyperplasia (BPH), untreated primary prostate cancer (PC) and castration resistant prostate cancer (CRPC). Each sample group shows a distinct protein profile. By integrative analysis we show that, especially in CRPC, gene copy number, DNA methylation, and RNA expression levels do not reliably predict proteomic changes. Instead, we uncover previously unrecognized molecular and pathway events, for example, several miRNA target correlations present at protein but not at mRNA level. Notably, we identify two metabolic shifts in the citric acid cycle (TCA cycle) during prostate cancer development and progression. Our proteogenomic analysis uncovers robustness against genomic and transcriptomic aberrations during prostate cancer progression, and significantly extends understanding of prostate cancer disease mechanisms.

[1] Prostate Cancer Research Center, Faculty of Medicine and Life Sciences and BioMediTech Institute, University of Tampere, Tampere 33014, Finland. [2] FimLab Laboratories, Tampere University Hospital, Tampere 33101, Finland. [3] Department of Ophthalmology, Faculty of Medicine and Life Sciences, University of Tampere, Tampere 33014, Finland. [4] Department of Urology, University of Tampere and Tampere University Hospital, Tampere 33521, Finland. [5] Singapore Eye Research Institute, Singapore 169856, Singapore. [6] Duke-NUS Neuroscience, Singapore 169857, Singapore. [7] Duke-NUS Medical School Ophthalmology and Visual Sciences Academic Clinical Program, Singapore 169857, Singapore. [8] Ophthalmology, Yong Loo Lin Medical School, National University of Singapore, Singapore 119228, Singapore. [9] Tays Eye Centre, Tampere University Hospital, Tampere 33521, Finland. [10] Science Center, Tampere University Hospital, Tampere 33521, Finland. These authors contributed equally: Ebrahim Afyounian, Antti Jylhä, Janika Nättinen. Correspondence and requests for materials should be addressed to M.N. (email: matti.nykter@uta.fi) or to T.V. (email: tapio.visakorpi@uta.fi)

Prostate cancer is the most common male malignancy in Western countries, and the second most common cancer among men overall[1]. Currently, no curative treatment exists for castration resistant prostate cancer (CRPC)[2]. To understand the etiology of the disease and to find more specific drug targets, the driver mutations and expressional changes in prostate cancer have been examined through extensive genomic and transcriptomic characterization[3–7]. Although significant insight has been gained through these efforts, it is clear that not all molecular alterations influencing the tumor outcome can be captured through these approaches.

Proteins are regulated at multiple levels, and their expression is not always reflecting the levels of mRNA[8,9]. Thus, a comprehensive understanding of the molecular events in cancer require thorough investigation of the proteome[10]. Recent developments in mass spectrometric methods[11–13] have enabled high throughput analysis of clinical patient samples, and the first integrative studies involving large scale, mass spectrometry-based proteomics of human cancer have recently been published[14–16]. For prostate cancer, recent proteomic advancements have included high scale, mass spectrometry-based studies performed in diagnostic body fluids[17,18], as well as primary tumors[19] and the tumor microenvironment[20]. So far, the only integrative proteogenomic analysis of clinical prostate cancer involved genomic and transcriptomic data of CRPC combined with phosphoproteomic analysis[21]. Despite the merits of this study in interrogating the active signaling pathways in CRPC, the large-scale proteomic view of PC and CRPC, and reflections of them to the disease progression are still lacking.

Here, we provide the first integrative view on human prostate cancer with the proteome of clinical patient samples of benign prostatic hyperplasia (BPH), untreated primary prostate cancer (PC) and locally recurrent CRPC. Our analysis adds a new level to the current knowledge of prostate cancer development and

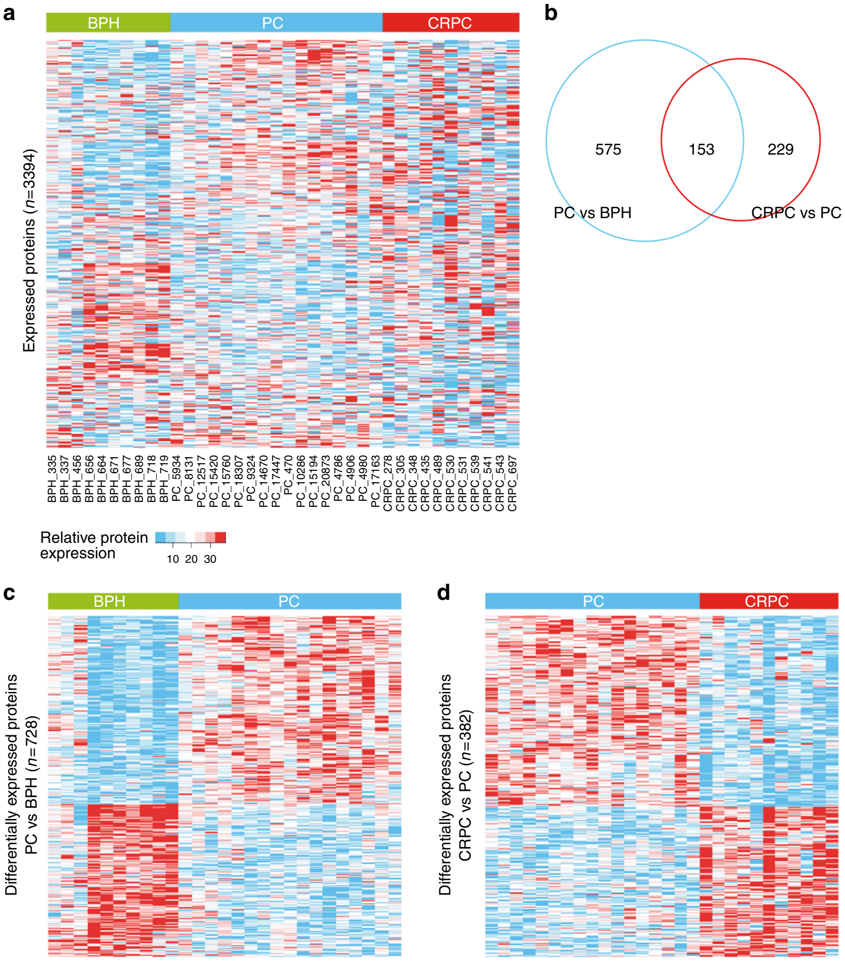


Fig. 1 Proteomic analysis reveals distinct protein expression patterns in PC and CRPC. **a** Heat map of all protein expressions identified and quantified by mass spectrometry in the proteomic analysis of BPH and prostate cancer samples (PC and CRPC). Each column of heat map represents a patient sample and each row represents a specific protein ($n = 3394$). **b** Venn diagram showing the numbers of differentially expressed proteins in PC vs BPH and CRPC vs PC comparisons. Only a minority of the differentially expressed proteins overlap between the comparisons. **c**, **d** Heat maps of the differentially expressed proteins in **b** show clearly distinctive patterns of protein expression between disease groups. PC compared to BPH samples ($n = 728$) is shown in **c**, and CRPC compared to PC samples ($n = 382$) is shown in **d**. Color key of relative expression in **a** applies also to **c** and **d**

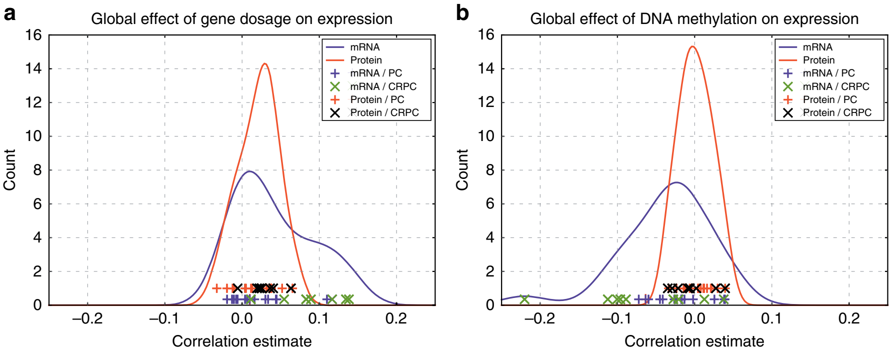


Fig. 2 Global expression changes associated with gene copy number and DNA methylation are visible at the transcriptomic but not at proteomic level. **a** Correlation distributions of mRNA and protein expression with gene copy number. Lines represent effects in all analyzed genes in all samples, and show that gene dosage has higher positive correlation with mRNA expression than protein expression in prostate cancer on a global scale. Symbols on the bottom of the graph represent individual samples, and show how most of the CRPC samples have a higher positive correlation compared to PC samples at the mRNA level, as at the protein level no such difference between the disease groups is observed. **b** Correlation distributions of mRNA and protein expression with DNA methylation. Lines represent effects in all analyzed genes in all samples, and show that DNA methylation has higher negative correlation with mRNA expression than protein expression in prostate cancer on a global scale. Symbols on the bottom of the graph represent individual samples, and show how most of the CRPC samples have a decreased correlation compared to PC samples at the mRNA level, as at the protein level no such difference between the disease groups is observed

progression by identifying several molecular and pathway events not previously described based on transcriptomic data.

Results

Mass spectrometric analysis of proteomic profiles. Samples of 10 BPH, 17 untreated PC (Supplementary Table 1), and 11 CRPC (Supplementary Table 2) were analyzed. The CRPC samples came from patients that had been treated either by castration and/or antiandrogens and experienced urethral obstruction (ie. local recurrence) during the treatment. With sequential window acquisition of all theoretical fragment ion spectra mass spectrometry (SWATH-MS), we identified a total of 213,979 peptides, corresponding to 1,753,161 identified spectra in an assembly of 4601 protein groups using false discovery rate of 1%. Protein and peptide quantification data can be found from Supplementary Data 1. From this library, 3394 proteins had distinct peptides sequences with matching spectras to SWATH-MS analysis and were quantified in all samples (Supplementary Data 2). The SWATH-MS data was reproducible with mean intraclass correlation (ICC) coefficient of 0.98 between technical replicate MS analyses. Permutation tests (Spearman correlation) showed that 98.6% of the technical replicate MS analyses had a p-value < 0.05, demonstrating excellent quality. The represented protein classes (PANTHER protein class) and gene ontology groups (GO; molecular functions, cellular components, and biological processes) are shown in Supplementary Fig. 1a. The distribution of the proteins into different protein classes was largely according to expected as compared to *Homo sapiens* reference list (Supplementary Fig. 1a,b). The major overrepresented groups included the highly abundant nucleic acid binding (mainly RNA binding) and ribosomal proteins, oxidoreductases, and hydrolases. The major underrepresented groups were transcription factors and receptors, including immunoglobulins, consistent with the cell type-dependent expression of especially the latter group.

Expression profiles of the identified proteins in the prostate tissue samples are shown in Fig. 1a. We wanted to assess changes occurring at the protein level during prostate cancer development and progression. As a model for benign tissue, we used BPH samples, against which primary PC samples were compared to identify early cancerous events. To identify events related to cancer progression and castration resistance, CRPC samples were compared to PC samples. We identified 728 proteins in PC vs BPH and 382 proteins in CRPC vs PC to be differentially expressed (Wilcoxon rank sum test with Benjamini & Hochberg adjustment p-value < 0.05 and median ratio (fold change) >1.5) between the comparison groups (Fig. 1b). While the overall protein classes of the differentially expressed proteins and their distribution to groups of molecular function, cellular component, and biological process were similar between PC vs BPH and CRPC vs PC comparisons (assessed by Panther analysis; data not shown), only a subset ($n = 153$) of the differentially expressed proteins were common between the comparison groups (Fig. 1b). The expression profiles of the differentially expressed proteins clearly distinguished between the patient sample groups, as shown in Fig. 1c (PC compared to BPH) and Fig. 1d (CRPC compared to PC). These results show that the proteomic profile of prostate cancer is significantly altered during the course of the disease.

Correlations of copy number and methylation with proteomics. We have previously performed whole genome sequencing for copy number analysis, DNA methylation sequencing, and whole transcriptome sequencing to majority of the samples used in the proteomic analysis described here (Supplementary Table 3)[7,22]. We compared the correlation between gene copy number, and mRNA or protein expression levels between the common samples. While at the transcriptome level, the mRNA expression and copy number have an increased overall correlation in the CRPC samples compared to PC samples (Fig. 2a, Supplementary Fig. 2a), a similar global correlation change with gene copy number is not present at the proteomic level. Next, we compared the correlation between DNA methylation at differentially methylated regions (DMRs), and mRNA or protein expression levels in the same samples. Similarly as with the copy number data, the increased negative correlation between DNA methylation and mRNA expression at a global level in the CRPC samples compared to PC samples is not detected at the level of the proteome (Fig. 2b, Supplementary Fig. 2b). These results suggest that,

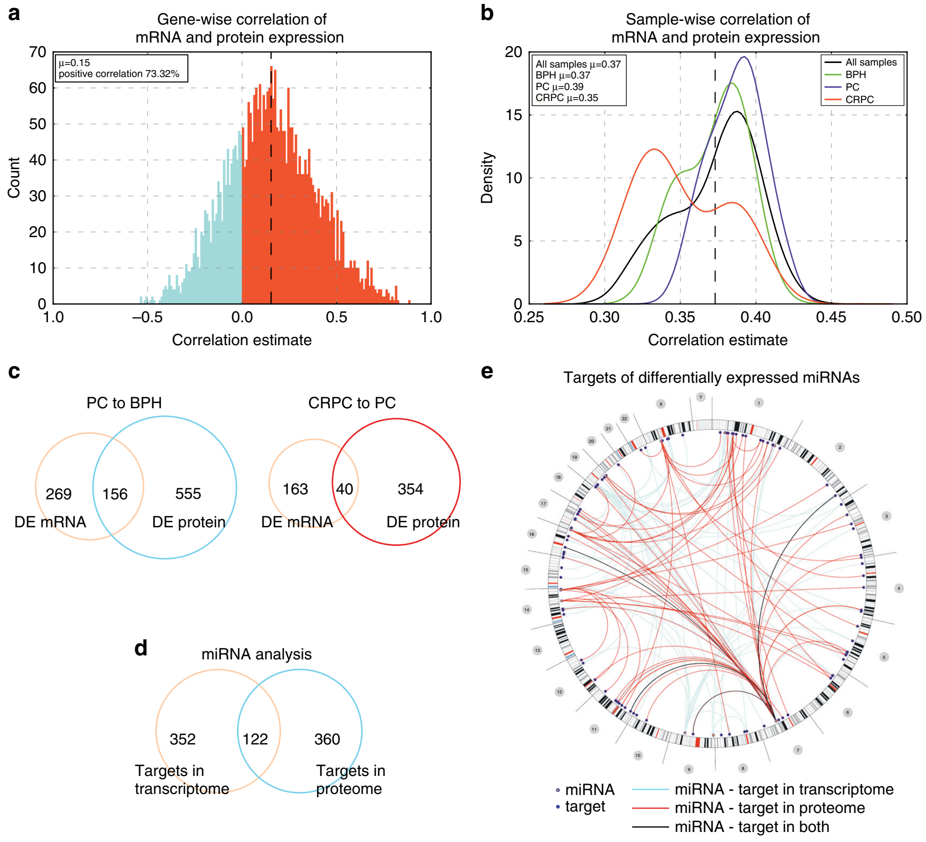


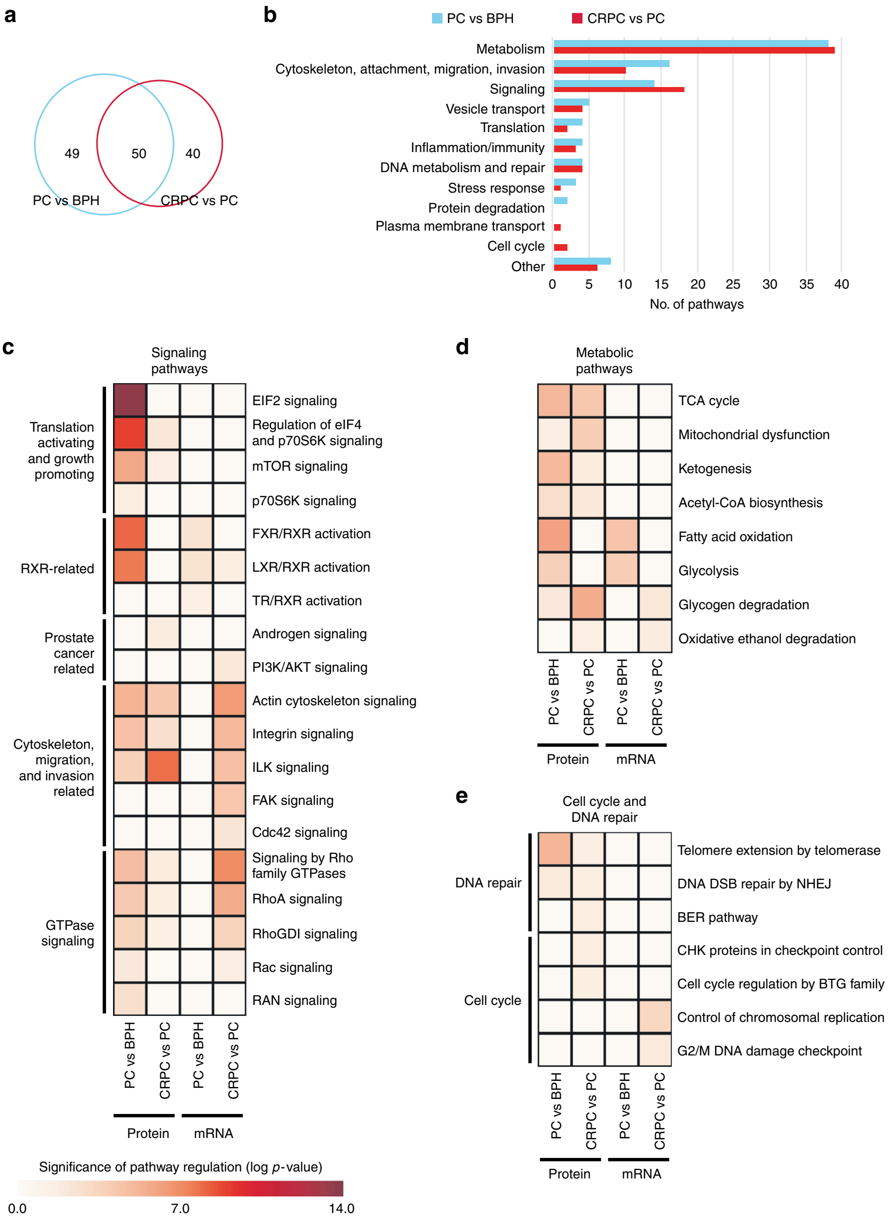
Fig. 3 Transcriptomic and proteomic data show distinct patterns of expression at the RNA and protein level in prostate cancer. **a** Correlation between mRNA and protein expression of individual genes. The graph shows correlations of all genes identified in proteomic analysis in all samples used in this study. Most of the genes (>73%) show positive correlation between expression of their mRNA and protein. μ is the mean of the correlations. **b** Disease group-wise correlation between mRNA and protein expression of the genes identified in the proteomic analysis shows that in CRPC there is a decreased correlation between mRNA–protein expression pairs compared to primary PC. Compared to all samples (black line) and BPH (green line), the PC samples (blue line) have a higher correlation between their mRNA-protein expression pairs, while CRPC samples (red line) have a lower correlation. μ is the mean of the correlations. **c** Venn diagram showing the numbers differentially expressed (DE) genes in PC vs BPH and CRPC vs PC comparisons identified based on mRNA or protein expression. The numbers of overlapping genes show that only a minority of the differentially expressed genes show expression changes in both mRNA and protein levels. **d** Venn diagram showing the numbers of genes that are negatively correlating with a targeting miRNA based on their expression at the mRNA or protein level. Only a minority of the miRNA targets are identified both at the mRNA and protein level, indicating that correlations at the protein level help to identify mostly a different pool of miRNA targets than correlations at the mRNA level. **e** Circos plot depicting genomic locations of miRNAs and their targets that are both negatively correlating at expression, as well as differentially expressed during prostate cancer progression (CRPC vs PC samples). Outer ring indicates chromosomes and cytobands, with chromosome numbers in the gray circles. Each line in the center maps a prostate cancer-related miRNA-target pair indicated through transcriptomic (blue lines), proteomic (red lines), or both (black lines) analyses. The blue circles mark the genomic location of the miRNAs, and the solid blue dots mark the targets

on a global level, the genomic and epigenomic events that influence mRNA levels are not directly translated to protein expression in prostate cancer.

The effect of altered methylation in prostate cancer on selected genes is, on the other hand, evident also at the proteomics data. There were 140 genes, which were differentially expressed either at mRNA or protein level, with a DMR close by (<10 kb). Within this group, there were several examples of methylation correlating with, and thus likely affecting, mRNA and protein expression. For example, the previously described increased DMR methylation in prostate cancer on genes *ALDH1A2*, *GSTP1*, *GPX3*, and *CYB5R2* correlate with decreased expression of their mRNA and protein according to our data (Supplementary Fig. 3). We further identified increased promoter DMR methylation in prostate cancer correlating with decreased expression of mRNA and protein expression also on *FBXO2*, *TGFB1I1*, and *TNS1* (Supplementary Fig. 4). Increased gene body methylation in prostate cancer correlating with decreased expression of mRNA and protein expression was identified on *GNAO1*, *LGALS1*, *TNS1*, and *PPAP2B* (Supplementary Fig. 5). Decreased methylation significantly correlating with increased expression was identified for *ENO1*, *SOAT1*, *RPS2*, and *TACSTD2* (Supplementary Fig. 6). Altered DMR methylation found in prostate cancer samples identifies also genes that are less likely to affect directly the outcome of the cancer cells. This is due to either their expression primarily in stromal cells (e.g., *CSRP1*, *CA3*) or the fact that, despite mRNA expression being affected, the expression level of the protein is not being affected by the differential methylation of the gene (e.g., *CLU, CNTN1*) (Supplementary Fig. 7). Interestingly, we also identified genes whose differentially increased methylation significantly correlated with increased expression in mRNA and/or protein level (*GMDS, MCCC2, MIA3*, and *PYCR1*) (Supplementary Fig. 8).

Impact of mutations on protein expression. We identified amino acid altering mutations in expressed genes from the RNA-sequencing data of the samples used in this study, and validated these from the DNA using targeted sequencing (Supplementary Table 4). For all somatic and germline variants, we evaluated the impact of the variant to mRNA and protein expression as described earlier[14]. While somatic mutations had a statistically significant impact to mRNA levels in relation to germline variants (Fisher's exact test, p-value = 0.0055, Supplementary Fig. 9), we observed no impact on protein expression levels between somatic and germline mutations or in relation to null distribution estimated from unmutated genes (Supplementary Fig. 9).

To screen for proteins with potential involvement in mutation accrual during prostate cancer development and progression, we assessed correlations of protein expression in relation to mutation



burden of the samples, including the somatic point mutations, copy number alterations, and genomic rearrangements. The two proteins, expression of which correlated best with point mutation burden, were mitochondrial antioxidant regulator PRDX3 (peroxiredoxin 3) and CAD (carbamoyl-phosphate synthetase 2) functioning in de novo synthesis of pyrimidine nucleotides (Supplementary Table 5). For copy number alterations and genomic rearrangements, the best correlating proteins had functions mostly in mitochondria and cytoskeleton. Notably, the strongest negative correlations with the number of rearrangements were with expression of two Talin proteins (TLN2 and TLN1)(Supplementary Table 5).

Comparison of expression profiles at RNA and protein levels. The expression levels of most of the proteins identified in our dataset were positively correlated with the expression level of their mRNA, as expected (Fig. 3a). However, when comparing the sample groups, we found that in CRPC, the correlation between individual mRNA-protein pairs was lower in general than in BPH or PC samples (Fig. 3b). We next tested whether similar genes are identified as differentially expressed based on both mRNA and protein expression data. In both PC vs BPH and CRPC vs PC comparisons, only a fraction of the differentially expressed genes were common between the identifications based on transcriptomic and proteomic data, the difference being larger in CRPC vs PC comparison (Fig. 3c). Of the commonly identified genes, 97 and 95% of the differential expressions detected were oriented to the same direction (up or downregulated) in both PC vs BPH and CRPC vs PC comparisons based on mRNA and protein expression data, respectively. According to these results, proteomic and transcriptional data help identify largely different events during prostate cancer development and progression.

Next, we integrated small RNA sequencing data for PC and CRPC samples common between the proteomics and mRNA expression data. MicroRNAs regulate gene expression by binding to mRNA molecules and preventing translation, which leads to decreased target protein expression. miRNA binding to the target can induce degradation of the mRNA, however, also stabilization of the target mRNA has been reported[23]. To study how much of the observed gene expression in prostate cancer is potentially connected to regulation by miRNAs, we studied the pool of differentially expressed genes and their correlating miRNAs. As one miRNA can have several target mRNAs, and one mRNA can be targeted by several miRNAs, we considered individual miRNA-target pairs based on both transcriptome and proteome data, and the predicted or verified miRNA target annotations. Negative correlations between miRNA and differentially expressed targeted mRNAs in CRPC vs PC samples revealed 30 miRNAs and 205 individual miRNA-target pairs (Supplementary Table 6). Of these, 9 miRNAs were also differentially expressed (Supplementary Table 6). For 34 of the miRNA-target pairs, negative correlation was also found between miRNA and protein expression of the target, indicating a functional impact of miRNA regulation for these particular targets (Supplementary Table 6). To look for the miRNA targets for which the miRNA does not induce mRNA degradation, but effect primarily through inhibition of translation, we searched for negative correlations between miRNA and differentially expressed targeted proteins in the proteome of CRPC vs PC samples. This analysis identified additional 49 miRNAs and 268 individual miRNA-target pairs (Supplementary Table 7). Of these, 8 miRNAs were also differentially expressed (Supplementary Table 7). This pool of miRNA-target pairs represents a resource of novel associations in prostate cancer that have not been visible through previous transcriptome analyses.

To understand the capacity that miRNAs have in regulating prostate cancer progression, we assessed the number of differentially expressed miRNAs and the fraction of the proteome they are collectively able to regulate. There were 95 miRNAs that were differentially expressed between CRPC and PC samples. Assuming negative correlation between a miRNA and its database-predicted or verified target either at the mRNA or protein expression level, the differentially expressed miRNAs in our dataset had the potential to target 16% of the genes in the study. There were 474 and 482 genes according to mRNA and protein expression, respectively, targeted by and negatively correlating with at least one regulating miRNA (Fig. 3d, Supplementary Data 3-4). Of these, only 122 genes were commonly identified (Supplementary Table 8). To look for the miRNA targets which most likely affect prostate cancer progression, we assessed the fraction of the miRNA-regulated genes that were differentially expressed. Of the above miRNA-regulated targets identified based on mRNA expression, 24% ($n = 115$) were differentially expressed between CRPC and PC samples at the mRNA level (Supplementary Fig. 10a, Supplementary Table 9). Similarly, of the regulatory targets identified based on the proteomics data, 45% ($n = 218$) were differentially expressed at the protein level (Supplementary Fig. 10b, Supplementary Table 10). There were 24 genes common between these groups (21% or 11% of the genes identified based on mRNA and protein expression, respectively). A genomic map of the differentially expressed miRNAs and their differentially expressed targets in CRPC vs PC samples based on transcriptomics and proteomics is shown in Fig. 3e. Collectively, these data indicate that by studying the miRNA-target correlations at the protein expression level we were able to identify a significant number of potential regulatory events, which were not identified based on mRNA expression data of clinical prostate cancer samples.

Fig. 4 Proteomic analysis identifies novel pathways as regulated in PC and CRPC. **a** Venn diagram showing numbers of differentially regulated pathways according to Ingenuity Pathway Analysis in PC vs BPH and CRPC vs PC comparisons. Despite partial overlap, the different disease states have a significant number of pathways specifically regulated. **b** Differentially regulated pathways in **a** according to pathway types. Metabolism is the largest group in both comparisons, with roughly a similar number of pathways differentially regulated. Numbers of most of the other pathway types that are differentially regulated between the disease states vary. **c–e** Examples of signaling pathways found to be differentially regulated according to proteomics (protein) or transcriptomics (mRNA) data in PC vs BPH and CRPC vs PC comparisons. **c** Examples of signaling pathways groups identified as regulated according to proteomic data. Especially translation activating, growth promoting pathways are identified as regulated solely based on proteomic data. RXR-related pathways are identified better by proteomics than transcriptomics to be regulated in PC. Pathways related to cytoskeleton, migration, and invasion, as well as GTPase signaling pathways are identified to be regulated in PC solely by proteomics, although in CRPC they are better identified as regulated by transcriptomics. **d** Metabolic pathways differentially identified as regulated based on proteomic and transcriptomic data include pathways identified as regulated in both PC and CRPC solely based on proteomics (TCA cycle, mitochondrial dysfunction, ketogenesis, acetyl-CoA biosynthesis), and pathways that are equally identified by proteomics and transcriptomics, but are specific for PC (fatty acid oxidation, glycolysis) or CRPC (glycogen degradation, oxidative ethanol degradation). **e** While DNA repair pathways regulated in PC and CRPC were identified based on proteomics only, the regulated cell cycle pathways were altered in CRPC and identified based on either proteomic or transcriptomic data. The color key below panel **c** applies to panels **c**, **d**, and **e**

Table 1 TCA cycle proteins with altered expression levels in prostate cancer

Symbol	Entrez gene name	PC vs BPH	CRPC vs PC
ACO2	aconitase 2	3.141	0.472
CS	citrate synthase	1.705	n.s.
FH	fumarate hydratase	1.598	n.s.
IDH3A	isocitrate dehydrogenase 3 (NAD (+)) alpha	n.s.	0.653
MDH2	malate dehydrogenase 2	2.167	1.912
OGDH	oxoglutarate dehydrogenase	1.653	0.608
SUCLA2	succinate-CoA ligase ADP-forming beta subunit	1.909	n.s.
SUCLG1	succinate-CoA ligase alpha subunit	2.091	0.469

Fold changes in protein expression are shown. n.s., not significantly altered

To validate our analysis for miRNA targets detectable both at the mRNA or the protein level, we transfected PC-3 prostate cancer cells with pre-miRNA constructs and assessed the mRNA and protein levels of predicted targets. We selected two representative miRNAs that were differentially expressed to opposite directions during prostate cancer progression, namely miR-22 as downregulated and miR-493 as upregulated in CRPC compared to PC, and verified their successful transfection by TaqMan RT-qPCR (Supplementary Fig. 11a). As positive controls for miRNA targeting at the mRNA level, we performed RT-pPCR on two predicted targets of miR-493 that were identified as negatively correlated based on our analysis at the target transcript level (Supplementary Data 3). Supplementary Fig. 11b shows that, as expected, the mRNA levels of ENDOD1 and GOLM1 are significantly decreased by miR-493 expression. Further, the negatively correlating miRNA-target pairs identified only in the proteomic analysis show decreased protein expression in MS/MS quantification, but no decrease in mRNA levels in RT-qPCR assay, as shown for miRNA-target pairs miR-22—KHRSP1 and miR-493—DNML1 (Supplementary Fig. 11c and d, respectively). These results confirm that our miRNA-target analyses based on the proteomics data have identified miRNA targets that are not identified at the mRNA level.

Proteomic analysis reveals novel regulated pathways. To test whether proteomics reveal pathway alterations in prostate cancer that have not previously been found by interrogation of mRNA expression changes, we next performed pathway analysis comparison between mRNA and protein expression data from the same samples. Supplementary Fig. 12a shows that roughly similar numbers of pathways were found significantly regulated based on proteomics and RNA expression data when comparing PC to BPH, and slightly more based on proteomics in CRPC to PC comparison. However, only a minority (16–26%) of the pathways found in each comparison category were common between RNA and proteomics data. These results show that proteomic data is able to reveal pathway regulations not visible at the RNA expression level, especially when comparing CRPC to PC.

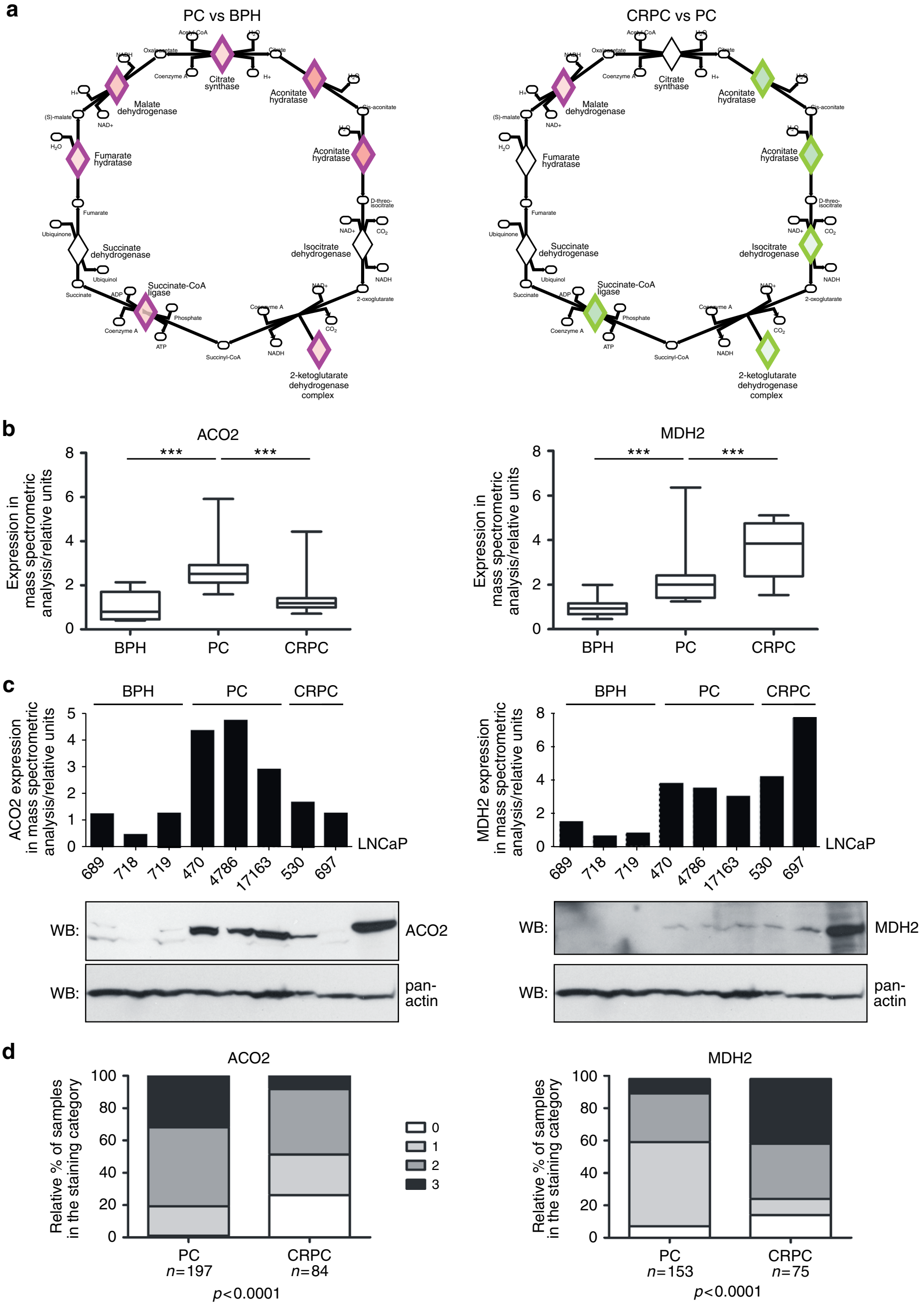
We further analyzed which signaling pathways were deregulated during prostate cancer development and progression at the proteomic level. Comparing PC samples to BPH, 99 pathways were found regulated according to Ingenuity Pathway Analysis, while 90 pathways were regulated in CRPC vs PC (Fig. 4a, Supplementary Table 11). Fifty pathways were common between these comparisons. The pathway categories were similar in both comparisons, with metabolic pathways being the most prominent (Fig. 4b). In PC vs BPH, cytoskeleton, attachment, and motility-related pathways were the second largest group, while in CRPC vs PC it was the signaling pathways. Exclusively in PC vs BPH, there were protein degradation pathways found significantly regulated, while in CRPC vs PC, certain cell cycle pathways were significantly regulated. The pathways common between the PC vs BPH and CRPC vs PC comparisons (Supplementary Table 11) included mostly metabolic pathways, as well as cytoskeleton, attachment and motility-related pathways (62% of the common pathways). It is noteworthy that all significantly regulated DNA metabolism and repair pathways, and most of the vesicle transport pathways, were common between the comparison groups. In contrast, only a few of the regulated signaling pathways were common between the comparison groups, all of which represented Rho GTPase signaling pathways (Supplementary Table 11).

In PC vs BPH, the top significantly regulated pathways included EIF2, eIF4 and p70S6K signaling, as well as FXR/RXR and LXR/RXR activation (Supplementary Table 11, Fig. 4c). While the former pathways promote growth- and survival through alterations in levels of several translation initiation factors and ribosomal proteins, the latter signal to metabolic pathways regulated by farnesoid X receptor (FXR), liver X receptor (LXR), and retinoid X receptor (RXR). When comparing CRPC to PC samples, the most significantly altered pathways during progression of prostate cancer include ILK signaling and glucose metabolism-related pathways (Fig. 4c, d, Supplementary Table 11).

Next, we wanted to further understand the differences in pathway regulation at mRNA and protein levels. Despite being largely different pathways, the biological functions of the pathways most often found by either RNA expression or proteomics were similar, with metabolic, signaling, and cytoskeleton and cell movement-related pathways being the most common (Fig. 4c, d, Supplementary Fig. 12b, Supplementary Fig. 13a,b). Examples of differentially identified signaling pathways are shown in Fig. 4c. Most prominently, the translation-activating and growth-promoting EIF, p70S6, and mTOR signaling, and cytoskeleton-related signaling in PC vs BPH were found solely based on proteomics data. RXR-related signaling in PC vs BPH, as well as several cytoskeleton-related signaling pathways in CRPC vs PC, were found by both transcriptomics and proteomics similarly. Interestingly, GTPase signaling was significantly regulated in PC vs BPH by proteomics and in CRPC vs PC by transcriptomics.

The group of metabolic pathways that was regulated in both PC vs BPH and CRPC vs PC comparisons was extensive (Supplementary Fig. 12b). Despite the relatively low overlap in individual pathways between the comparisons (between disease groups, and between proteomic and transcriptomic data), all the analyses identified pathways from the major groups of energy, amino acid, and lipid metabolism (Supplementary Table 11; examples shown in Fig. 4d, Supplementary Fig. 13b). It is noteworthy that the mitochondria-related metabolic pathways and ketogenesis were identified as differentially regulated only by proteomics, while e.g., glycolytic and glycogen degradation-related pathways were identified by both transcriptomics and proteomics (Fig. 4d). One of the most prominent group of metabolic pathways in prostate cancer were amino acid metabolic pathways (Supplementary Fig. 13b). Interestingly, while different cell cycle regulatory pathways were found regulated by transcriptomic and proteomic data, DNA repair pathways were found solely by proteomics analysis (Fig. 4e). Other interesting groups of differentially identified pathways based on RNA and protein expression were vesicular traffic-related and protein degradation pathways (Supplementary Fig. 13c,d).

Changes in TCA during prostate cancer evolution. Based on our analysis, metabolic changes are prominent during both development and progression of prostate cancer. One of the most interesting pathways identified by our proteomic data was the tricarboxylic acid cycle (TCA; also referred to as the citric acid cycle, or the Krebs cycle), which was altered in both PC vs BPH and CRPC vs PC comparisons. This pathway was not found regulated by RNA expression data, suggesting changes taking place primarily at the protein level. Furthermore, although alterations in certain enzyme activities in TCA have previously been shown to occur during prostate cancer development[24], our proteomics results indicated a previously undescribed, two-step modulation of the TCA cycle. The TCA pathway proteins that were considered regulated by the pathway analysis were mostly



altered to opposite directions in PC vs BPH and CRPC vs PC comparisons: upregulated in PC vs BPH, and downregulated in CRPC vs PC (Table 1, Fig. 5a). An exception was malate dehydrogenase 2 (MDH2), levels of which continued to increase in CRPC (Table 1, Fig. 5a). Comparison of protein and RNA expression of TCA genes[8,23] in all three groups of samples revealed that the TCA proteins are divided into three classes: (1) proteins whose mRNA and protein expression go hand in hand indicating primary regulation by gene expression (CS, FH, IDH3A, IDH2, and SUCLG2), (2) proteins, whose protein levels are not changed (IDH3B, IDH3G), and (3) proteins, that exhibit regulation at the protein level not correlating with mRNA (ACO2, MDH2, OGDH, SUCLA2, and SUCLG1) (Supplementary Fig. 14). From the latter group of proteins, ACO2, OGDH, SUCLA2, and SUCLG1 were all upregulated at the protein level, but not at the mRNA level, in PC vs BPH, while being downregulated in CRPC vs PC either at the mRNA or protein level. Increase in MDH2 protein expression in PC vs BPH did correlate with an increase in mRNA levels, but the increase in CRPC vs PC did not, suggesting posttranslational regulation (Supplementary Fig. 14).

To study more closely the events identified in the TCA cycle, and to validate the results of the proteomics data, we selected two TCA proteins showing significant but different alterations at their protein expression between the prostate cancer sample groups to study further. As a representative of the most common alteration pattern we chose aconitase 2 (ACO2) which showed statistically highly significant ($p < 0.001$, Mann–Whitney test) upregulation of the protein in PC vs BPH, as well as statistically highly significant ($p < 0.001$, Mann–Whitney test) downregulation in CRPC vs PC (Fig. 5b). As a second protein we chose MDH2 exhibiting the deviant behavior amongst the TCA proteins, as it was upregulated statistically significantly both in PC vs BPH ($p < 0.001$; Mann–Whitney test) and further upregulated in CRPC vs PC ($p < 0.05$; Mann–Whitney test) (Fig. 5b). We performed western blotting on these proteins with representative samples of BPH, PC, and CRPC used in the proteomic analysis, and found similar changes than by mass spectrometry (Fig. 5c, Supplementary Fig. 15 and 16), validating the mass spectrometry detection and analysis results.

We performed further validation on the differential regulation of these proteins during prostate cancer progression by immunohistochemical stainings on larger sample sets of clinical PC and CRPC. Grading of the immunohistochemical staining intensity (example staining intensities of grades 0–3 displayed in Supplementary Fig. 17) showed that relative percentage of samples with no or low staining intensities (0–1) of ACO2 increased in CRPC vs PC (Fig. 5d), indicating that the relative levels of ACO2 decreased in CRPC. On the other hand, the relative percentage of samples with higher staining intensities (2–3) of MDH2 increased in CRPC vs PC (Fig. 5d), indicating that the relative levels of MDH2 increased in CRPC. These results confirm the mass spectrometry results and show that the TCA cycle proteins ACO2 and MDH2 are differentially regulated at the protein level during prostate cancer progression.

We further assessed potential mechanisms that could explain the distinct regulation of MDH2. We found that two miRNAs predicted to target MDH2, namely miR-22 and miR-205, were identified as differentially expressed in our analysis and were negatively correlating with MDH2 protein (Supplementary Data 4) but not mRNA (Supplementary Data 3) levels in the large scale datasets. We transfected PC-3 prostate cancer cells with these miRNAs, and verified the transfection efficiency with TaqMan RT-qPCR analysis (Supplementary Fig. 18a). We detected no significant alterations at MDH2 mRNA levels in RT-qPCR analysis upon elevated expression of the miRNAs (Supplementary Fig. 18b). In contrast, luciferase assay showed statistically significant decrease in reporter production from a MDH2 3′-UTR construct by both miR-22 and miR-205 overexpression (Supplementary Fig. 18c), indicating that these miRNAs are able to directly target MDH2 mRNA. Furthermore, MS/MS quantification showed a substantial decrease in MDH2 protein levels by both miR-22 and miR-205 expression (Supplementary Fig. 18c). These results validated the predictions of miR-22 and miR-205 to directly target MDH2, and identified these miRNAs as prostate cancer-relevant, differentially expressed regulators of the TCA.

Discussion

We have provided the first extensive proteomic view of prostate cancer development and progression. With over 3000 individual proteins quantified in each of the BPH, PC and CRPC samples analyzed, we described the protein level alterations occurring in clinical prostate cancer, and found several previously undescribed biological events with important implications and potential for future studies. In addition, we provided novel views on the relationship of proteomic, genomic, and transcriptomic changes occurring during castration resistance. The comprehensive view obtained by our integrative analysis underlines the importance of protein level dissection of the molecular mechanisms supporting cancer growth and progression.

Our results showed that neither the altered gene dosages, nor the global methylation changes were translated to the level of the proteome to the same extent as they influence the global RNA expression in CRPC. This suggests that, in the progressed stage, a large proportion of changes in gene copy number and differential DMR methylation are side products of the catastrophic state of cancer cell regulatory systems which are untranslated and thus, subsequently, left without a functional effect at the protein level. Yet, our data confirmed several previously identified regulatory DNA methylation events with associated expression changes occurring in prostate cancer. We also identified several previously

Fig. 5 TCA cycle is differentially regulated during prostate cancer progression. **a** A schematic view of the TCA cycle protein expression changes in PC vs BPH and CRPC vs PC comparisons according to the Ingenuity Pathway Analysis. Differential expression of TCA enzymes (diamonds) are highlighted in green (downregulation) and red (upregulation). As mostly the same enzymes are involved in both PC and CRPC, the primary mode of expression change is upregulation in PC and downregulation in CRPC. **b** Examples of a typical (ACO2) and a unique (MDH2) TCA protein expression patterns as identified by mass spectrometry proteomics. ACO2 is upregulated in PC compared to BPH, and gets downregulated in CRPC compared to PC. MDH2 protein expression levels increase in PC compared to BPH, and continue to increase in CRPC. Boxplots show interquartiles with mean values, whiskers represent minimum and maximum values. ***p-value < 0.001 (Mann–Whitney test). **c** ACO2 and MDH2 protein expression patterns verified in a subset of BPH, PC, and CRPC samples by western blotting. ACO2 and MDH2 protein expression according to the proteomic mass spectrometry analysis (upper panel bar graph) and in corresponding samples according to western blotting (WB; lower panels). Pan-actin is used as a loading control. **d** Change in ACO2 and MDH2 protein expression patterns during progression of prostate cancer verified by immunohistochemistry. Immunohistochemical analysis in clinical tumor samples of PC and CRPC show statistically significantly decreased ACO2 and increased MDH2 staining intensity in CRPC compared to PC and (Chi squared test; 0 = no staining, 1 = weak staining, 2 = intermediate staining, 3 = strong staining)

undescribed protein expression alterations in PC and CRPC associated with differential methylation of DMRs.

We showed that the proteomic profile of prostate cancer is significantly altered during the course of the disease. We identified differentially expressed proteins, potential miRNA regulatory effects, and significantly altered pathway events. The key notion is that these have not been identified through transcriptomic analyses. This supports the view that not all proteins apply to changes at the mRNA level, and underlines the importance of mechanistic studies at the protein level.

Especially intriguing is the group of predicted miRNA-target pairs that we found to have negative correlations between miRNA expression and target protein expression without alterations detected at the target mRNA level. These target mRNAs may be bound by the miRNAs without induced degradation of the target. For each miRNA-target pair, the targeting and relevance for prostate cancer needs to be verified by follow-up experiments. Here, we verified several targets for three example miRNAs that are differentially expressed in CRPC vs PC, and thus may play regulatory roles during prostate cancer progression. Our proteomic pathway analysis identified especially translation-related growth pathways as significantly altered in primary PC compared to BPH samples. In addition, changes in protein degradation pathways were better detected by proteomics than transcriptomics. Thus, protein homeostasis in prostate cancer seems to be regulated primarily at the protein level. In CRPC, the proteome-specific pathway alterations were concentrated on mitochondria-related metabolism and DNA repair. While the glycolytic and long-term energy storage utilization pathways were significantly regulated in prostate cancer at both proteomic and transcriptomic levels, the changes in the core TCA and mitochondrial pathways are evident solely based on the proteomic data. This indicates that posttranscriptional events are taking place in the mitochondria during castration resistance, in order for the cancer cells to ensure survival and propagation under the altered conditions.

As a key finding, we detected two metabolic shifts involving the TCA during prostate cancer development and progression. The changes in TCA enzyme activities during prostate cancer development have been studied earlier, but the second shift occurring during progression to CRPC is previously undescribed. In primary prostate cancer, it is well-established that the normally high tissue citrate levels decrease[24,25]. Costello and Franklin[24] suggested that normal citrate-producing prostate epithelial cells become citrate-oxidizing when they turn malignant. Under this bioenergetic hypothesis, mitochondrial aconitase ACO2 is a key enzyme for the bioenergy transformation[26]. Subsequently, Juang[27] showed that downregulation of mitochondrial aconitase in cultured prostate cancer cells decreases cell proliferation rate. Mitochondrial aconitase gene expression was earlier shown to be regulated by testosterone in prostate epithelial cells in vitro[28], suggesting that in high AR activity tumors ACO2 gene expression could be upregulated. In our gene expression data, ACO2 mRNA levels increase in PC compared to the levels in BPH. However, in CRPC compared to PC, reflecting events during formation of castration resistance and involving increased AR expression, ACO2 mRNA levels are not increased further, and the protein levels decrease. Thus, while our results support previous evidence of upregulation of mitochondrial aconitase levels during development of prostate cancer, progression to CRPC seems to involve primarily posttranslational regulation of the enzyme, reflecting the differences between the first and the second metabolic shift during the course of prostate cancer evolution.

Most of the TCA enzymes are upregulated during the first metabolic shift in prostate cancer, and then either stay upregulated (CS, FH) or are downregulated (*e.g.* ACO2, OGDH, and SUCLG1) during the second shift. The exception is MDH2, protein levels of which continue to increase in the second shift during prostate cancer progression. As a mechanism explaining the continued increase in MDH2 protein levels in CRPC, we identified decreased expression of miR-22 and miR-205, miRNAs which were both confirmed to decrease MDH2 protein levels without decreasing the MDH2 mRNA expression. MDH2 is mitochondrial malate dehydrogenase, which is an enzyme that catalyzes the NAD/NADH-dependent, reversible oxidation of malate to oxaloacetate. It has been reported previously that patients with MDH2 overexpression have a significantly shorter period of relapse-free survival after undergoing neoadjuvant combination chemotherapy followed by surgery[29]. Further, stable knockdown of MDH2 via shRNA in prostate cancer cell lines decreased cell proliferation and increased docetaxel sensitivity[29]. Together with our data, these results collectively suggest MDH2 inhibition as a mechanism to target castration resistant tumors. MDH2 druggability has been studied in the context of doxorubicin-induced cardiomyopathy, where the non-specific MDH2 inhibitors mebendazole, thyroxine, and iodine have been found promising[30]. Thus, development of MDH2-specific chemical inhibitors could be of great benefit against progressed prostate cancer, as well as for prevention of cardiotoxicity during chemotherapy.

In conclusion, we identified here several key aspects of prostate cancer biology with the most comprehensive proteomics on primary and progressed prostate cancer samples so far. In addition to increasing our understanding of prostate cancer biology, our study identified several important aspects of prostate cancer signaling and metabolism for future studies.

Methods

Samples. Fresh-frozen tissue specimens from 10 BPH, 17 untreated PC, and 11 CRPC samples were acquired from Tampere University Hospital (Tampere, Finland). PC samples (Supplementary Table 1) were obtained by radical prostatectomy. Mean age at diagnosis was 62.0 years (range: 47.4–71.8) and mean PSA at diagnosis was 9.8 ng/ml (range: 3.5–19.8). Locally recurrent CRPC samples (Supplementary Table 2) were obtained by transurethral resection of the prostate. Samples were snap-frozen and stored in liquid nitrogen. Histological evaluation and Gleason grading were performed by a pathologist based on hematoxylin/eosin-stained slides. All samples contained a minimum of 70% cancerous or hyperplastic cells. The use of clinical material was approved by the ethical committee of the Tampere University Hospital and the National Authority for Medicolegal Affairs. Written informed consent was obtained from the subjects.

Chemicals and materials. Acetonitrile (ACN), formic acid (FA), water (UHPLC-MS grade), triethyl ammonium bicarbonate buffer (TEAB), sodium dodecyl sulfate (SDS), iodoacetamide (IAA), trifluoro acetic acid (TFA), ammonium bicarbonate (ABC), tris-(2-carboxyethyl)phosphine (TCEP), urea and pellet pestles were all purchased from Sigma Aldrich (St. Louis, MO, USA). RIPA lysis buffer, protease inhibitor cocktail (Halt™) and sample clean up tips (C18) were from Thermo Fisher Scientific (San Jose, CA, USA). Bio-Rad DC™ protein assay kit and bovine serum albumin standard were purchased from Bio-Rad (Hercules, CA, USA) and 30 kDa MWCO centrifugal devices from PALL (Port Washington, NY, USA). TPCK-treated trypsin was from AB Sciex (Framingham, MA, USA). HRM Calibration Kit was purchased from Biognosys AG (Zurich, Switzerland).

Protein extraction from tissue samples and enzymatic digestion. Five 5 μm slices were cut from fresh-frozen tissue samples. Tissues were homogenized with polypropylene pellet pestle in ice-cold RIPA lysis buffer containing Halt protease inhibitor. The disrupted tissues were subjected to sonication for 5 min followed by a 30 min incubation on ice. After incubation, lysates were centrifuged to remove any remaining cell debris (16,000 xg, 20 min, +4 °C). Total protein concentration of the samples was measured with Bio-Rad DC protein assay. Mean amount of protein recovered from frozen tissues was 91.5 ± 67.3 μg (SD). From 9 to 50 μg of protein was precipitated with acetone (−20 °C) overnight. The protein amounts were selected based on our previous testing of suitable injection volume of 5 μg total protein in 2 μl volume in SWATH. Precipitated proteins were centrifuged, supernatant was decanted, and samples were allowed to dry for 5 min. Proteins were dissolved in 0.05 M ABC with 2% SDS and reduced by 0.05 M TCEP. After 60 min of incubation at + 60 °C, samples were transferred into 30 kDa molecular weight cut-off centrifugal filters and flushed twice with 8 M urea in 0.05 M Tris-HCl. Cysteine residue blocking was carried out by 0.05 M IAA in 0.5 M Tris-HCl at room temperature in the dark. Samples were repeatedly flushed with 8 M urea and

0.05 M ABC to remove urea prior to digestion with trypsin for 16 h at + 37 °C at a trypsin-to-protein ratio of 1:25. Digests were collected by rinsing the centrifugal devices with 0.1 M TEAB followed by 0.5 M NaCl and dried in a speed vacuum concentrator. Samples were dissolved in 0.1% TFA and desalted with C18 tips. Sample clean-up and desalting was performed with Pierce C18 tips according to manufacturer's instructions. Samples were dried in speed vacuum concentrator and stored at −20 °C until reconstituted in loading solution (5% ACN, 0.1% FA) at equal concentrations. HRM peptide mix was added to each sample before NanoRPLC-MSTOF SWATH analysis.

NanoRPLC-MSTOF for discovery proteomics. Digested peptides were analyzed by Nano-RPLC-MSTOF instrumentation using Eksigent 425 NanoLC coupled to high speed TripleTOF™ 5600 + mass spectrometer (Ab Sciex, Concord, Canada). A capillary RP-LC column (cHiPLC® ChromXP C18-CL, 3 μm particle size, 120 Å, 75 μm i.d × 15 cm, Eksigent Concord, Canada) was used for LC separation of peptides. Samples were first loaded into trap column (cHiPLC® ChromXP C18-CL, 3 μm particle size, 120 Å, 75 μm i.d × 5 mm) from autosampler and flushed for 10 min at 2 μl/min (2% ACN, 0.1% FA). The flush system was then switched to line with analytical column and gradient alution. All samples were analyzed with 120 min 6 step gradient using eluent A: 0.1% FA in 1% ACN and eluent B: 0.1% FA in ACN (eluent B from 5 to 7% over 2 min, 7 to 24% over 55 min, 24 to 40% over 29 min, 40 to 60% over 6 min, 60 to 90% over 2 min and kept at 90% for 15 min, 90 to 5% over 0.1 min and kept at 5% for 13 min) at 300 nl/min.

In order to perform SWATH-MS quantification, we first generated a spectral identification library with 57 different samples (prostate tissue and cancer cell line samples). Key parameters for MSTOF mass spectrometer in SWATH ID library analysis were: ion spray voltage floating (ISVF) 2300 V, curtain gas (CUR) 30, interface heater temperature (IHT)+125 °C, ion source gas 1 13, declustering potential (DP) 100 V. All methods were run by Analyst TF 1.5 software (Ab Sciex, USA). For IDA parameters, 0.25 s MS survey scan in the mass range 350–1250 mz were followed by 60 MS/MS scans in the mass range of 100–1500 Da (total cycle time 3.302 s). Switching criteria were set to ions greater than mass to charge ratio (m/z) 350 and smaller than 1250 (m/z) with charge state 2–5 and an abundance threshold of more than 120 counts. Former target ions were excluded for 12 s. Information dependent acquisition (IDA) rolling collision energy (CE) parameters script was used for automatically controlling CE. SWATH quantification analysis parameters were the same as for spectral identification library analyses, with the following exceptions: cycle time 3.332 s and MS parameters set to 15 Da windows with 1 Da overlap between mass range 350–1250 Da followed by 40 MS/MS scans in the mass range of 350–1250 Da.

Mass spectrometric data analysis. SWATH library analysis were performed with Protein pilot software version 4.7 (Ab Sciex, Canada) which was used to analyze MS/MS data and searched against the UniprotKB/Swiss-prot database for protein identification. Settings in the Paragon search algorithm in Protein pilot were configured as follows. Sample type: identification, Cys-alkylation: MMTS, Digestion: Trypsin, Instrument: TripleTOF 5600 + , Search effort: thorough ID. False discovery rate (FDR) analysis was performed in the Protein pilot and FDR < 1% was set for protein identification. The data from all the identification runs were combined as a batch and used for library creation for SWATH relative quantification.

For quantification we used PeakView® software 2.0 with SWATH-plug in to assign the correct peaks to correct peptides in the library. Two replicate MS analyses were done from each sample. iRT peptides (Biognosys, Switzerland) was used for retention time calibration with PeakView. 1–15 specified peptides per protein were selected to be used in SWATH quantification. Peptide peak areas were extracted and filtered to remove all peptides, which do not have a single measurement with an FDR <1% across all measurements. The SWATH-MS data exhibited excellent quality and reliability with p-value < 0.05 in 98.6% of replicate MS analyses (permutation tests, Spearman's rank correlation) and mean interclass correlation (ICC) coefficient of 0.98.

Statistical analysis of proteomics data. Data processing included $\log_2$-transformation and quantile normalization. The quality of the replicate MS analyses was analyzed by calculating the intraclass correlation (ICC) and Spearman's rank correlation was used to generate p-values in permutation tests ($n = 1000$ permutations/replicate MS analyses). Further analysis was performed on the mean values of the replicate MS analyses. Wilcoxon rank sum test was implemented to analyze the differences between sample types. Benjamini and Hochberg adjustment were applied to all initial p-values, where applicable, to account for the multiple testing issues. R software version 3.2.3 (R Core Team. Foundation for Statistical Computing, Vienna, Austria) was used to analyze data. Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City, USA) was used to conduct pathway analysis and identify proteins connected to pathways of interest. Protein grouping and classification was performed by using PANTHER Classification System[31].

Analysis of differentially expressed mRNA and protein. Common samples (Supplementary Table 3) between the proteomic analysis performed here and previously described mRNA expression (RNA sequencing) data[7] were used to extract common genes ($n = 3310$) between the protein and mRNA data. mRNAs and proteins were considered differentially expressed across different comparisons (BPH vs PC, PC vs CRPC) if absolute median ratio of two conditions was greater than 1.5 with a p-value < 0.05 (Benjamini–Hochberg adjusted p-value of a non-parametric Wilcoxon test).

Association between protein expression and gene copy number. The Spearman's rank correlation was calculated for each sample between gene level DNA copy numbers[7] and mRNA expressions, or copy numbers and protein expressions (Supplementary Table 3). Using the correlation values probability density functions (PDFs) for each correlation value sets were estimated using kernel density estimation with Gaussian kernels and Scott's rule for bandwidth determination. The estimated PDFs were then plotted and supplemented by rug plots of the exact correlation values. To estimate background distributions, the Spearman's rank correlations of each sample with all other samples were calculated between copy numbers and mRNA expressions or copy numbers and protein expressions. Using the correlation values PDFs for each correlation value set were estimated as detailed above.

Association between protein expression and DNA methylation. Based on MeDIP-sequencing data[7] we identified 751 differentially methylated regions (DMRs) within 10 kb from TSS of 557 unique genes with available expression values for RNA expression and protein expression (Supplementary Table 3). Subsequently, the Spearman's rank correlations were calculated for each sample between DMR normalized fragment counts and mRNA expressions or DMR normalized fragment counts and protein expressions. Kernel density estimation was used for visualization of the correlation values as described above. Background distributions were calculated in the same manner as explained earlier. Furthermore, we identified 2773 genes common between mRNA and protein expression datasets where their absolute distance to a nearby DMR was <250 kb. Of these, 745 genes were showing absolute correlation >0.3 between their gene expression and nearby DMRs. 140 out of 745 genes were differentially expressed both at mRNA and protein level. Finally only 79 of these genes had absolute distance ≤10 kb from 117 DMRs (these were used for scatter plots).

Structural variation analysis. To identify rearrangements whole genome sequencing reads were aligned against the GRCh37 reference genome using Bowtie-2.0.0-beta7[32]. An in-house structural variant calling software called Breakfast (https://github.com/annalam/pypette) was then used to identify paired end reads where the mates aligned discordantly. A paired alignment was considered discordant if both mates aligned to the genome but aligned to separate chromosomes or >100 kb apart. Mates with an alignment quality phred value < 20 were discarded from analysis. Next, individual mates that did not initially align to the reference genome were split into 25 bp anchors. The 25 bp anchor pairs were then realigned and searched for discordant alignments using the same criteria as with paired end reads. The full 90 bp sequences corresponding to discordant anchor pairs were compared against the reference genome to identify exact breakpoints and to analyze for sequence homologies. A discordant anchor pair was discarded if the sequence homology between the read and one of the breakpoint flanking sequences was above 70% for the nucleotides matching with the discordant anchor. The exact breakpoint was determined by selecting the breakpoint associated with the lowest amount of nucleotide mismatches. After identifying discordant pairs from paired end and split reads, the discordant pairs were reoriented so as to always have the pair with the lower chromosome or coordinate first. Discordant pairs were then clustered using a sliding window approach. A cluster of discordant pairs was accepted as a putative structural variant if it contained at least one paired end read and one split read indicating the structural variant. To filter out false positives, structural variants were also called in BPH samples, and all genomic regions within 1 kb of a breakpoint identified in a BPH sample were blacklisted.

Point mutation impact analysis. Somatic and germline point mutations[22] in each sample were used to find their impact on the expression level of the genes harboring the mutations as described earlier[14]. Null distribution was generated by comparing expression of randomly selected unmutated sample to other unmutated samples.

Association between protein expression and mutation burden. Number of somatic point mutations[22], rearrangements (as described above), and chromosomal instability (CIN) in each sample across common samples between protein and mRNA expression data (Supplementary Table 3) were used to find their association (Spearman's rank correlation) with individual genes in the protein dataset. CIN was calculated as the mean of integer copy numbers assigned to non-overlapping blocks of size 500 bp spanning across the entire genome.

Association between protein and miRNA expression. miRNA expression data (small RNA sequencing)[7] were used to extract miRNAs with negative correlation with their targets using the common samples (Supplementary Table 3). Predicted targets of miRNAs were downloaded from miRWalk 2.0 database[33] using the following parameter values: Input parameters Promoter 2 kb, 3′ UTR, minimum seed length 7 and/or p-value 0.05. We considered mRNA to be a target for miRNA if targeting was predicted by 2/3 of the databases miRanda, PICTAR2, and Targetscan[34–36]. Differentially expressed miRNAs were defined as having an absolute median ratio between two conditions >1.5, and the Benjamini-Hochberg adjusted p-value of a non-parametric Wilcoxon test <0.05. miRNAs were considered unexpressed if all samples had read count below 8 and they were excluded from differential expression analysis. Spearman's rank correlations were calculated between the miRNA expression and the expression of its predicted targets with a threshold for negative correlation $< = -0.50$. For enrichment analysis, hypergeometric test was used to test statistical significance (p-value < 0.05) of the number of negatively correlating predicted targets of a miRNA. miRNA—target associations were visualized as circos plot using POMO[37].

Transfections of pre-miRNA. PC-3 cells (ATCC, Rockville, MD, USA) were cultured under the recommended conditions and reverse transfected with 10 nM non-targeting control (miR-control) or pre-micro-RNA constructs (Applied Biosystems/Ambion, Austin, TX, USA) using INTERFERin transfection reagent (Polyplus Transfection SA, Illkirch, France) according to manufacturer's instructions. Cells were incubated for 48 or 72 h before collection for RNA or protein samples, respectively.

RNA extraction and RT-qPCR. RNA was extracted using TriReagent® (Sigma-Aldrich) according to manufacturer's instructions. Quantitative RT-PCR for miRNAs was performed using TaqMan microRNA Assay (Applied Biosystems, Foster City, CA, USA) according to the manufacturer's recommendations. RNU6B was used as a reference gene. Quantitative RT-PCR for mRNAs was performed using Maxima SYBR Green (Fermentas Inc., Burlington, Ontario, Canada) from cDNA made using Maxima RT reverse transcriptase (ThermoFischer Scientific Inc.). *TBP* was used as a reference gene. qPCR reactions were performed with the CFX96 q-RT-PCR detection system (Bio-Rad Laboratories Inc., Hercules, CA, USA).

Luciferase reporter assay. PC-3 cells were reverse transfected with 10 nM non-targeting control (miR-control) or pre-micro-RNA constructs (Applied Biosystems/Ambion, Austin, TX, U.S.A.), and MDH2–3′-UTR in pEZX-MT05-GLuc-SEAP luciferase reporter plasmid (GeneCopoeia, Rockville, MD, USA; 10 ng/well) in 96 well plates using jetPRIME transfection reagent (Polyplus Transfection SA, Illkirch, France) according to manufacturer's instructions. Cells were incubated for 24 h before the medium was collected for analysis of secreted Gaussia luciferase (GLuc) and secreted alkaline phosphatase (SEAP) activities with Secrete-Pair™ Dual Luminescence Assay Kit (GeneCopoeia) according to manufacturer's instructions.

MicroLC-MSTRAP for targeted protein validation analysis. Proteins for targeted MS/MS analysis were selected based on their expression in discovery analysis. Peptides for each protein were selected based on their specificity, intensity (based on SWATH-MS analysis), amino acid composition, and water solubility in the tissue samples. All peptides with methionine or modifications or missing cleavage sites were disqualified. For each selected peptide, an isotopically labeled standard peptide (AQUA-peptides, Sigma-Aldrich) was used to confirm the identification. For each protein in the analysis, two peptides for targeted MS analysis were selected, and each peptide analysis was confirmed using 3 fragment ions. The peptides, fragment ions, and corresponding isotopical standards for each protein are represented in Supplementary Table 12.

Cell lysis, protein measurements, and tryptic digestion were performed as before. TEAB-solution supplemented with 20 fg of each targeted peptide isotope per 1 μg of total protein in the sample was used to flush the digested peptides of the membrane. Sample cleanup was performed as before. 1 μg of cleaned samplewas used for MicroLC-MSTrap analysis. Analysis was performed with Sciex 6500 + MSTrap coupled with Eksigent NanoLC 425 with 1–10 μl/min microLC flow cell. MicroLC utilized a 42 min 6 step gradient using eluent A: 0.1% FA in MQ and eluent B: 0.1% FA in ACN (eluent B from 10 to 30% over 22 min, 30–50% over 8 min, 50–80% over 2 min, kept at 80% for 5 min, 80–10% over 0.2 min and kept at 10% for 5 min, at 5 μl/min. MSTrap settings were as follows; Curtain gas: 30, Spray voltage: 5300, Collision gas: medium, Temperature: 150 °C, Ion source gas 1: 20, Ion source gas 2: 20, were set the same for all peptides. Collision energy was specifically set to 40 for OLA1 peptide IPAFLNVVDIAGLVK and to its respective isotope standard and to 30 for all others. Results were normalized against their representative isotopically labeled standard peptide and then compared between samples. Standard deviation for each peptide in the analysis method was calculated using isotope labeled peptide standards. Relative standard deviation for all the peptides was under 10%.

Western blotting. LNCaP cells (ATCC) were cultured under the recommended conditions. Cells and sections of frozen tissue were lysed in Triton-X lysis buffer containing 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 0,5% Triton x-100, 1 mM PMSF, 1 mM DTT and 1× complete protease inhibitor cocktail (Roche Inc., Mannheim, Germany), after which the lysates were sonicated four times for 30 s at medium power with Bioruptor equipment (Diagenode Inc., Liège, Belgium), and cellular debris was removed by centrifugation. Proteins were separated by polyacrylamide gel electrophoresis (SDS-PAGE) and transferred to PVDF membrane (Immobilon-P; Millipore Inc., Billerica, Massachusetts, USA). Primary antibodies against ACO2 (HPA001097; Sigma-Aldrich, St. Louis, MO, U.S.A.; dilution 1:1000), MDH2 (HPA019714; Sigma-Aldrich; 1:1000), and pan-actin (ACTN05; NeoMarkers, Portsmouth, NH, USA; 1:1000) were used and detected using anti-rabbit HRP-conjugated antibody produced in swine (1:5000, DAKO Inc., Denmark) or by anti-mouse HRP-conjugated antibody produced in rabbit (1:5000, DAKO Inc., Denmark) and Western blotting luminol reagent (Santa Cruz Inc., Santa Cruz, California, USA) with autoradiography. Original scans including molecular weight information for the western blots are presented in Supplementary Fig. 19.

Immunohistochemistry. Formalin-fixed, paraffin-embedded tumor microarrays of PC and CRPC samples[38] were used. Sections were deparaffinized and antigen retrieval was performed by using Tris-EDTA buffer 0.05% Tween-20 (pH 9) at + 98 °C for 15 min. The staining was performed by Lab Vision Autostainer (ThermoFischer Scientific Inc., Waltham, MA, USA). Primary antibodies (as above) and secondary antibody (N-Histofine® Simple Stain MAX PO; Nichirei, Tokyo, Japan) were used. ImmPACT DAB (Vector Laboratories, Burlingame, CA, USA) was used as a chromogen. The sections were counterstained with hematoxylin and mounted with DPX mounting medium (Sigma-Aldrich). Scoring of staining intensity on tumor areas was performed on a 0–3 scale (Supplementary Fig. 17), and the difference in score distributions between PC and CRPC groups was statistically assessed with Chi squared test.

Data availability. Mass Spectrometry data has been deposited to Peptide Atlas repository under dataset identifier PASS01126. Deep sequencing data has been deposited to European Genome-Phenome Archive under accession number EGAS00001000526.

Received: 17 May 2017 Accepted: 21 February 2018
Published online: 21 March 2018

References

1. Jemal, A. et al. Global cancer statistics. *CA Cancer J. Clin.* **61**, 69–90 (2011).
2. Wong, Y. N. S., Ferraldeschi, R., Attard, G. & de Bono, J. Evolution of androgen receptor targeted therapy for advanced prostate cancer. *Nat. Rev. Clin. Oncol.* **11**, 365–376 (2014).
3. Taylor, B. S. et al. Integrative genomic profiling of human prostate cancer. *Cancer Cell.* **18**, 11–22 (2010).
4. Barbieri, C. E. et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.* **44**, 685–689 (2012).
5. Grasso, C. S. et al. The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
6. Robinson, D. et al. Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–1228 (2015).
7. Ylipää, A. et al. Transcriptome sequencing reveals PCAT5 as a novel ERG-regulated long noncoding RNA in prostate cancer. *Cancer Res.* **75**, 4026–4031 (2015).
8. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).
9. Liu, Y., Beyer, A. & Aebersold, R. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165**, 535–550 (2016).
10. Boja, E. S. & Rodriguez, H. Proteogenomic convergence for understanding cancer pathways and networks. *Clin. Proteom.* **11**, 22 (2014).
11. Megger, D. A., Bracht, T., Meyer, H. E. & Sitek, B. Label-free quantification in clinical proteomics. *Biochim. Biophys. Acta* **1834**, 1581–1590 (2013).
12. Chapman, J. D., Goodlett, D. R. & Masselon, C. D. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass. Spectrom. Rev.* **33**, 452–470 (2014).
13. Rosenberger, G. et al. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* **1**, 140031 (2014).
14. Zhang, B. et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387 (2014).
15. Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62 (2016).
16. Zhang, H. et al. Integrated proteogenomic characterization of human high-grade serous ovarian. *Cancer Cell.* **166**, 755–765 (2016).
17. Jia, X. et al. Detection of aggressive prostate cancer associated glycoproteins in urine using glycoproteomics and mass spectrometry. *Proteomics* **16**, 2989–2996 (2016).

18. Larkin, S. E. et al. Detection of candidate biomarkers of prostate cancer progression in serum: a depletion-free 3D LC/MS quantitative proteomics pilot study. *Br. J. Cancer* **115**, 1078–1086 (2016).
19. Iglesias-Gato, D. et al. The proteome of primary prostate cancer. *Eur. Urol.* **69**, 942–952 (2016).
20. Staunton, L. et al. Pathology-driven comprehensive proteomic profiling of the prostate cancer tumor microenvironment. *Mol. Cancer Res.* **15**, 281–293 (2017).
21. Drake, J. M. et al. Phosphoproteome integration reveals patient-specific networks in prostate. *Cancer Cell.* **166**, 1041–1054 (2016).
22. Annala, M. et al. Recurrent SKIL-activating rearrangements in ETS-negative prostate cancer. *Oncotarget* **6**, 6235–6250 (2015).
23. He, L. & Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.* **5**, 522–531 (2004).
24. Costello, L. C., Franklin, R. B. & Feng, P. Mitochondrial function, zinc, and intermediary metabolism relationships in normal prostate and prostate cancer. *Mitochondrion* **5**, 143–153 (2005).
25. Mycielska, M. E. et al. Citrate transport and metabolism in mammalian cells: prostate epithelial cells and prostate cancer. *Bioessays* **31**, 10–20 (2009).
26. Costello, L. C. & Franklin, R. B. Bioenergetic theory of prostate malignancy. *Prostate* **25**, 162–166 (1994).
27. Juang, H. H. Modulation of mitochondrial aconitase on the bioenergy of human prostate carcinoma cells. *Mol. Genet. Metab.* **81**, 244–252 (2004).
28. Costello, L. C., Liu, Y., Zou, J. & Franklin, R. B. Mitochondrial aconitase gene expression is regulated by testosterone and prolactin in prostate epithelial cells. *Prostate* **42**, 196–202 (2000).
29. Liu, Q. et al. Malate dehydrogenase 2 confers docetaxel resistance via regulations of JNK signaling and oxidative metabolism. *Prostate* **73**, 1028–1037 (2013).
30. Liu, Y. et al. Visnagin protects against doxorubicin-induced cardiomyopathy through modulation of mitochondrial malate dehydrogenase. *Sci. Transl. Med.* **6**, 266ra170 (2014).
31. Thomas, P. D. et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).
32. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
33. Dweep, H. & Gretz, N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nat. Methods* **12**, 697 (2015).
34. Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.* **11**, R90 (2010).
35. Anders, G. et al. doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.* **40**, D180–D186 (2012).
36. Agarwal, V., Bell, G. W., Nam, J. W., Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).
37. Lin, J. et al. POMO–plotting omics analysis results for multiple organisms. *BMC Genom.* **14**, 918 (2013).
38. Leinonen, K. A. et al. Loss of PTEN is associated with aggressive behavior in ERG-positive prostate cancer. *Cancer Epidemiol. Biomark. Prev.* **22**, 2333–2344 (2013).

Acknowledgements

We thank Paula Kosonen, Riina Kylätie, Artturi Lassila, Katja Liljeström, Saara Lähdekorpi, Päivi Martikainen, and Marika Vähä- Jaakkola for their technical assistance and M.D. Teemu Tolonen for his professional assistance. This work was supported by grants from the Academy of Finland (Project Nos. 269474 MN, 127187 TV), Sigrid Juselius Foundation (M.N., T.V.), Cancer Society of Finland (M.N., T.V.), Competitive State Research Financing of the Expert Responsibility area of Tampere University Hospital (TV, MN, TT), TEKES (Project No. 66/31/2012 U.A., R.B., A.J., J.N., H.U.), and Elsemay Björn Fund (U.A., A.J., J.N., H.U.). We would also like to acknowledge CSC—IT Center for Science Ltd. (https://www.csc.fi/csc) for providing the computational resources.

Author contributions

U.A., R.B., M.N., H.U., and T.V. conceived and supervised the study. All authors designed and discussed experiments. A.J. carried out the mass spectrometry and SWATH analysis. A.J. and J.N. performed proteomics data analysis. E.A. performed integrative bioinformatics analyses. M.A. and K.W. performed mutation analyses. L.L. and J.N. performed pathway analysis. L.L. carried out western, immunohistochemical, and cellular analyses, and prepared the manuscript. All authors contributed to writing of the manuscript, as well as reviewed and accepted the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-03573-6.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

PUBLICATION
III

Chromatin accessibility analysis uncovers regulatory element landscape in prostate cancer progression

Uusi-Mäkelä J.*, Afyounian E. *, Tabaro F.*, Häkkinen T.*, Lussana A., Shcherban A., Annala M., Nurminen R., Kivinummi K., Tammela T.L., Urbanucci A., Latonen L., Kesseli J., Granberg K. J., Visakorpi T., Nykter M.

bioRxiv (2020)
DOI: 10.1101/2020.09.08.287268

Chromatin accessibility analysis uncovers regulatory element landscape in prostate cancer progression

Joonas Uusi-Mäkelä[1,2]*, Ebrahim Afyounian[1,2]*, Francesco Tabaro[1,2]*, Tomi Häkkinen[1,2]*, Alessandro Lussana[1,2], Anastasia Shcherban[1,2], Matti Annala[1,2], Riikka Nurminen[1,2], Kati Kivinummi[1,2], Teuvo L.J. Tammela[1,2,3], Alfonso Urbanucci[4], Leena Latonen[5], Juha Kesseli[1,2], Kirsi J. Granberg[1,2], Tapio Visakorpi[1,2,6], Matti Nykter[1,2]✦

[1] Prostate Cancer Research Center, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
[2] Tays Cancer Center, Tampere University Hospital, Tampere, Finland
[3] Department of Urology, Tampere University Hospital, Tampere, Finland
[4] Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway
[5] Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland
[6] Fimlab Laboratories Ltd, Tampere, Finland

*These authors contributed equally.

✦Corresponding author

Abstract

Aberrant oncogene functions and structural variation alter the chromatin structure in cancer cells. While gene regulation by chromatin states has been studied extensively, chromatin accessibility and its relevance in aberrant gene expression during prostate cancer progression is not well understood. Here, we report a genome-wide chromatin accessibility analysis of clinical tissue samples of benign prostatic hyperplasia (BPH), untreated primary prostate cancer (PC) and castration-resistant prostate cancer (CRPC) and integrative analysis with transcriptome, methylome, and proteome profiles of the same samples to uncover disease-relevant regulatory elements and their association to altered gene expression during prostate cancer progression. While promoter accessibility is consistent during disease initiation and progression, at distal sites chromatin accessibility is variable enabling transcription factors (TFs) binding patterns that are differently activated in different patients and disease stages. We identify consistent progression-related chromatin alterations during the progression to CRPC. By studying the TF binding patterns, we demonstrate the activation and suppression of androgen receptor-driven regulatory programs during PC progression and identify complementary TF regulatory modules characterized by e.g. MYC and glucocorticoid receptor. By correlation analysis we assign at least one putative regulatory region for 62% of genes and 85% of proteins differentially expressed during prostate cancer progression. Taken together, our analysis of the chromatin landscape in PC identifies putative regulatory elements for the majority of cancer-associated genes and characterizes their impact on the cancer phenotype.

Introduction

Prostate cancer (PC) is a common malignancy with heterogeneous phenotypes in men. In 18% of patients, disease progresses to lethal castration-resistant prostate cancer (CRPC) (Siegel et al., 2018). Recurrent genomic alterations in primary and metastatic PC have been identified and their role in disease progression has been studied extensively (Armenia et al., 2018; Espiritu et al., 2018; Grasso et al., 2012; Gundem et al., 2015; Peng et al., 2015; Quigley et al., 2018; Robinson et al., 2015). In addition to implicating cancer genes, genome sequencing studies have revealed structural variation in non-coding regions, including enhancer elements driving oncogene expression (Takeda et al., 2018; Viswanathan et al., 2018). Epigenetic characterization studies have further extended understanding of the non-coding genome by revealing the role of DNA methylation patterns (Bedford and van Helden, 1987; Börno et al., 2012; Friedlander et al., 2012; Jimenez et al., 2000; Lee et al., 1997; Mahapatra et al., 2012; Varambally et al., 2002; Xu et al., 2012; Zhao et al., 2020), specific transcription factor (TF) binding sites and histone modifications, including the characterization of the active enhancer landscape in PC tissues (Kron et al., 2017; Pomerantz et al., 2015, 2020; Stelloo et al., 2018; Urbanucci et al., 2012, 2017; Yu et al., 2010). Still, how the chromatin landscape evolves during PC progression and drives aberrant transcriptome (Cancer Genome Atlas Research Network, 2015) and proteome (Latonen et al., 2018; Sinha et al., 2019), is unclear.

Genomic aberrations and epigenetic regulation alter chromatin structure in cancer cells (Flavahan et al., 2017; Losada, 2014). Different chromatin accessibility analysis methods have been used to identify the chromatin landscape across cell lines (Thurman et al., 2012), tissues (Roadmap Epigenomics Consortium et al., 2015), and, most recently, tumor tissues (Corces et al., 2018). In PC, the study by Corces et al. uncovered chromatin accessibility changes at single-nucleotide polymorphism that are associated with increased PC susceptibility and illustrated androgen receptor (AR) binding site enrichment in regulatory regions specific to primary PC (Corces et al., 2018). A recent epigenetic study further demonstrated an association between prostate lineage-specific regulatory elements and PC risk loci and somatic mutation density in different stages of PC (Pomerantz et al., 2020). Binding of AR prominently occurs at distal regulatory elements (Massie et al., 2011; Yu et al., 2010), and AR-driven regulatory programs are context-dependent (Sharma et al., 2013; Wang et al., 2009)(Pomerantz et al., 2020)(Sharma et al., 2013; Wang et al., 2009). In PC cells, AR (Urbanucci et al., 2012; Yu et al., 2010), FOXA1 (Adams et al., 2019; Parolia et al., 2019; Sahu et al., 2011), HOXB13 (Chen et al., 2018; Pomerantz et al., 2015), ERG, and CHD1 (Augello et al., 2019) have emerged as epigenetic drivers of disease (Stelloo et al., 2018). More specificly, ERG fusion-positive tumors have a cis-regulatory landscape that is distinct from other tumors (Kron et al., 2017), and aberrant ERG expression has been shown to alter chromatin conformation and regulation in prostate cells (Rickman et al., 2012; Sandoval et al., 2018; Yu et al., 2010).

To gain insight into the role of the chromatin dynamics in determining phenotypes in PC progression, we analyzed chromatin accessibility in a cohort of clinical patient samples of human PC from benign prostatic hyperplasia (BPH), untreated primary prostate cancer (PC), and locally recurrent castration-resistant prostate cancer (CRPC). By integrating DNA, RNA, protein, and DNA methylation data (Annala et al., 2015; Latonen et al., 2018; Ylipää et al.,

2015) from the same samples, we provide a comprehensive catalogue of chromatin-related alterations in PC development and progression. Our results highlight high heterogeneity of regulatory elements utilization, complementarity of chromatin accessibility with DNA methylation, and extensive chromatin-driven reprogramming of the AR activity. In this study, we uncover putative regulatory elements for 65% and 85% of progression-related genes and proteins, respectively.

Results

ATAC-seq data from human prostate tissues

To study the chromatin landscape's role in PC development and progression, we first optimized the assay for transposase-accessible chromatin using a sequencing (ATAC-seq) protocol (Buenrostro et al., 2013) for frozen tissue samples. We characterize chromatin accessibility in 11 BPH, 16 PC, and 11 CRPC prostate tissue samples (see **Methods, Supplementary Table 1**). In earlier studies, we have analyzed these same samples using DNA, RNA, and DNA methylation sequencing and SWATH proteomics (**Supplementary Table 1**) (Latonen et al., 2018; Ylipää et al., 2015). Here, these data types were integrated with the ATAC-seq data (**Figure 1A**). ATAC-seq data depth varied from 69 to 204 million reads per sample. Quality control illustrated that there was no significant association between the sequencing depth and key quality parameters such as transcription start site (TSS) enrichment or number of detected peaks (**Supplementary Figure 1A-C**). On the contrary, we observed a good correlation between high quality autosomal alignments (HQAA) and TSS enrichment, indicating a good signal to noise ratio.

Chromatin accessibility at distal sites is heterogeneous in prostate cancer

To identify accessible and progression-related chromatin features, we used two complementary approaches. In the first approach, accessible chromatin regions in each sample were identified by peak calling using MACS2 peak calling algorithm (Zhang et al., 2008). We identified 23,840 to 138,942 raw peaks per sample (**Supplementary Table 1**). The number of detected peaks was not characteristic to a specific sample group, but samples with high and low peak count were observed throughout BPH, PC, and CRPC groups (**Figure 1B**). To obtain a robust set of reproducible peaks across samples, we used a previously proposed approach to unify raw peak calls (see **Methods**)(Corces et al., 2018). This approach resulted in the compilation of 178,206 peaks across the sample set (**Supplementary Table 1**). This is consistent with previous estimates for the number of cancer type-specific peaks in chromatin accessibility data (Corces et al., 2018). In the second approach, we performed genome-wide analysis to identify differentially accessible regions (DARs) by comparing samples in BPH to PC and PC to CRPC groups (see **Methods**). As a result, we identified 1,727 and 3,498 differentially accessible regions (DARs) for BPH to PC and PC to CRPC, respectively, with false discovery rate (FDR) below 10% (**Supplementary Table 2**). For peaks and DARs, a clear chromatin accessibility signal is detected (**Figure 1C, Supplementary Figure 2A**) and DNA methylation is depleted (**Figure 1C, Supplementary Figure 2A**) consistent with previous studies reporting decreased DNA methylation at accessible chromatin loci (Corces et al., 2018; Urbanucci et al., 2017).

Of the 180,442 identified chromatin features, 72% overlapped with regulatory regions found in normal tissues (Corces et al., 2018; Roadmap Epigenomics Consortium et al., 2015) or

TCGA data (Corces et al., 2018; Roadmap Epigenomics Consortium et al., 2015) (**Figure 1D**). The overlap was consistent for both peaks and DARs (**Supplementary Figure 2B**). TCGA data included 20 primary PC samples and within these samples 65.8% of their peaks overlap with our peak set. Taken together, our data showed consistency with earlier chromatin accessibility studies and we were able to expand the known regulatory landscape by discovering 38,157 new prostate cancer related chromatin features.

Of all identified chromatin features 7.4% were in promoters, 6.6% were in exons and untranslated regions, 39.3% were in introns (51.8% overlapping previously marked enhancers), and 46.7% were intergenic (32.5% overlapping previously marked enhancers) (Fishilevich et al., 2017) (**Figure 1D**). Peaks and DARs were distributed similarly, except for the promoter region in which 7.4% of the peaks but only 1.9% to 2.4% of the DARs were located (**Supplementary Figure 2B**). Furthermore, the peaks located at promoters had higher signal intensity than peaks in other genomic annotation groups (**Supplementary Figure 2C-D**). In addition, 60% of the peaks common to all the samples are located on promoters (**Figure 1E, Supplementary Figure 2E**). When assigning the peaks to a sample group or groups based on if they are present in a specific sample (**Figure 1F**), we observed that most peaks are not group-specific. For the peaks assigned to each sample group, the annotation distribution is similar (**Supplementary Figure 2F**). Importantly, we did not observe any peaks that would be group-specific and present in all the samples of that group (**Supplementary Figure 2G**). These data show that while promoters are robustly open across samples, accessibility at other genomic regions is highly variable between samples and sample groups. This indicates that, while accessibility remains robust during PC progression, most of the chromatin alterations occur at intronic and intergenic regions.

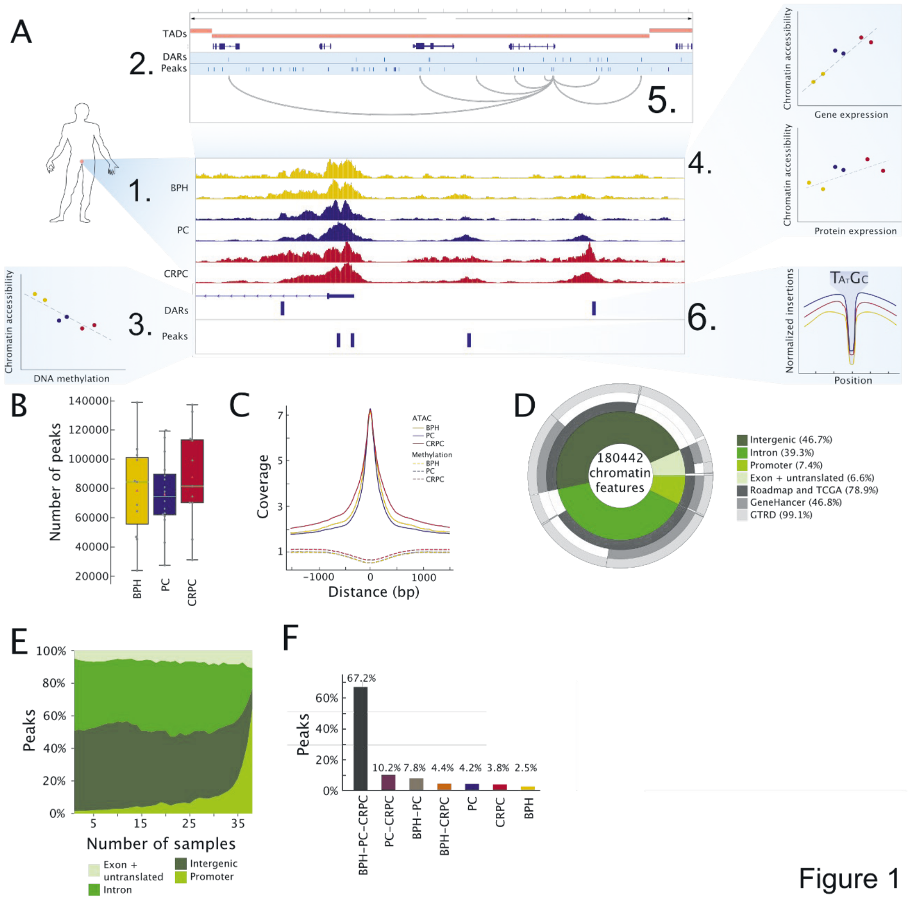


Figure 1

Figure 1: Chromatin accessibility in promoters is robust during prostate cancer progression

A. Cartoon illustration of ATAC-seq data analysis. After (1) generating ATAC-seq data from human prostate tissues, we (2) identified peaks and differentially accessible regions (DARs) between BPH, PC and CRPC groups. We (3) compared chromatin accessibility to DNA methylation and (4) gene and protein expression. Next, we associated (5) accessible chromatin regions with correlating target genes within the same topologically associating domains (TADs). Finally, (6) transcription factor binding at accessible chromatin was analyzed using TF footprinting, integration with ChIP-seq data, and using deep learning models to uncover binding context. **B.** Boxplots of the number of raw peaks in each sample (grey dots) in BPH, PC, and CRPC groups are shown. Peak counts in each group are comparable. **C.** Background-corrected coverages from ATAC-seq data at peak locations show a strong signal. Background-corrected DNA methylation data in the same locations is slightly depleted. Distances are relative to peak center. Median signals from BPH, PC, and CRPC samples are shown. **D.** Chromatin features are ordered in the donut plot based on their annotation to genomic location categories: intergenic, intron, promoter, and exon and untranslated regions (5’-UTR, 3’-UTR, transcription termination sites, non-coding RNA).

Majority of the features are located in intergenic and intronic regions. For each category, the proportion of previously identified areas of accessible chromatin (Roadmap and TCGA), known enhancer regions (GeneHancer), and detected TF binding sites from ChIP-seq data (GTRD) are shown. Fraction of chromatin features belonging to each region is shown in the donut plot with percentages given in labels. **E.** The proportion of peaks located in genomic location categories is shown for peaks present in the different number of samples. Most consistently observed ATAC peaks are located at promoters, and peaks in the distal regions are more heterogeneous across the samples. **F.** Percentage of peaks in different sample group combinations. Although the majority of peaks are present in samples from all three sample groups, a subset of peaks are sample group-specific.

Progression-related chromatin alterations are consistent

Having characterized chromatin features, we looked into its alterations over disease progression. Comparing DARs in BPH to PC and PC to CRPC we found little overlap (**Figure 2A**) suggesting that differential accessibility-related chromatin changes are specific to PC initiation and to progression of CRPC (**Figure 2B, Supplementary Figure 3A**). DARs in PC to CRPC comparison show a clear increase in untranslated and exon regions (**Supplementary Figure 2B, Supplementary Figure 3B**). These loci are not usually reported to harbor gene regulatory elements, but this combined with the finding that CRPC samples show more opening DARs than the other group (**Figure 2A, Supplementary Figure 2B**) may reflect overall chromatin relaxation (Urbanucci et al., 2017) or events related to chromatin reorganization.

Using methylated DNA immunoprecipitation sequencing (MeDIP-seq) data on the same clinical samples, we also called progression-related differentially methylated regions (DMRs) (see **Methods**). Comparing BPH to PC and PC to CRPC, we found 2,061 and 2,723 DMRs (**Supplementary Table 2, Figure 2C**). Comparing DARs and DMRs, we detected only 13 (0.6%) and 23 (0.8%) overlapping features in each comparison(**Figure 2C, Supplementary Figure 2B, Supplementary Figure 3C, Supplementary Table 2**). Little overlap between DARs and DMRs suggests that regulation of chromatin accessibility and DNA methylation might work as distinct epigenetic regulatory mechanisms in PC, affecting different transcriptional outputs.

Heterogeneity in chromatin accessibility is associated with disease-relevant regulators

As most of the observed chromatin alterations occur at intronic and intergenic regions, to understand how the heterogeneity of chromatin relates to disease progression, we first focused on the cancer-specific peaks (**Figure 1F**) with highest variance in signal across the samples. This includes mostly peaks distal from TSS, whilst promoter peaks are depleted in this set (**Supplementary Figure 3D)**. Unsupervised analysis of these peaks (see **Methods**) separated the samples in three clusters containing 273 to 1655 peaks, but failed to separate PC and CRPC samples in a data-driven manner (**Figure 2D, Supplementary Figure 3E)**. The three clusters did not correlate with tumor class/state, Gleason score, or ERG fusion status. However, the peaks separated in seven clusters based on consensus clustering (**Supplementary Figure 3F**). Enrichment analysis using TF binding site predictions in each of the seven peak clusters was used to evaluate whether these contained regulatory regions for specific TFs (**Supplementary Figure 3E, Supplementary Table 1**). Interestingly, each peak cluster is associated with DNA binding of different PC-related TFs. ERG-enriched and AR- and FOXA1-enriched clusters showed a similar activity pattern across samples. Likewise epithelial to mesenchymal transition (EMT) associated Wnt/β-Catenin signaling and TEAD1 and SNAI1 clusters behave similarly (Odero-Marah et al., 2018; Zhou et al., 2016). AR pioneering factors GATA2 and HOXB13 (Hankey et al., 2020; Pomerantz et al., 2015) were enriched into the same cluster, which showed the highest accessibility in the CRPC-rich sample group. Other clusters represent sample specific signals, for example, immune response related TFs were highly accessible only in one CRPC sample, possibly due to the patient's immune response.

To further study the effect of chromatin accessibility variation on TF activity, we performed TF footprint analysis in each sample for expressed TFs with available binding motif (see **Methods**, **Figure 2E, Supplementary Table 3**). Quantification of TF footprint by "flanking accessibility" (FA) and "footprint depth" (FD) allows the study of TF activities in a genome-wide manner (Baek et al., 2017). For the majority of TFs, FA and FD correlate. Notably, we do not detect any TF e.g. with low FA and high FD. Several disease relevant TFs, including AR and FOXA1, are among the ones with largest change in FA and FD during progression (**Figure 2E**). AR, and related co-factors FOXA1, and HOXB13 have similar Tn5 insertion patterns with the highest accessibility in PC (**Figure 2F**) while ERG accessibility is similar in all sample groups. During progression to CRPC, CTCF displays a large change in FA which might reflect relaxation or other alterations of chromatin structure.

Taken together, these results highlight a highly heterogeneous chromatin landscape across samples, and demonstrate that the observed regulatory patterns are associated with known disease-relevant processes and regulators. Furthermore, changes in disease relevant TF activities are consistent over progression.

Similar TF binding syntaxes are conserved across tumor samples

To understand if the heterogeneity in chromatin accessibility leads to variability in TF binding syntax, we utilized the recently developed BPNET model (Avsec et al.)(see **Methods**). BPNET builds predictive models of chromatin accessibility, and recursively decomposes the output to assign base-pair contribution scores to every input sequence that can be combined to obtain binding syntax motifs. We tested the model with cell line data and were able to discover highly detailed binding patterns, e.g. different forms of known AR binding configuration, demonstrating the feasibility of the approach with ATAC-seq data (**Supplementary Figure 4**). With application to data from patient samples we observed that model performance is dependent on both the signal-to-noise ratio and the number of available training peaks (**Figure 2G**). When we trained the model on individual samples using the whole reproducible peak set, we were able to recover motifs that match with known TFs, including AR, FOXA1, CTCF, GRHL2 and SP family (**Figure 2H**, **Supplementary Table 3**). Despite high heterogeneity in peaks across samples, detected binding syntaxes are consistent. When performing model training using only peaks with a known DNA binding site (see **Methods**) for key TFs AR, FOXA1, or HOXB13, we observed a consensus binding motif for all tested TFs across high quality samples (**Supplementary Table 3, Supplementary Figure 5-8**). This observation further supports the idea that TFs binding properties do not change despite heterogeneity in chromatin accessibility. In addition, we were able to identify disease state-related factors co-occurring with selected driver TFs (**Supplementary Table 3, Supplementary Figure 5-8**).

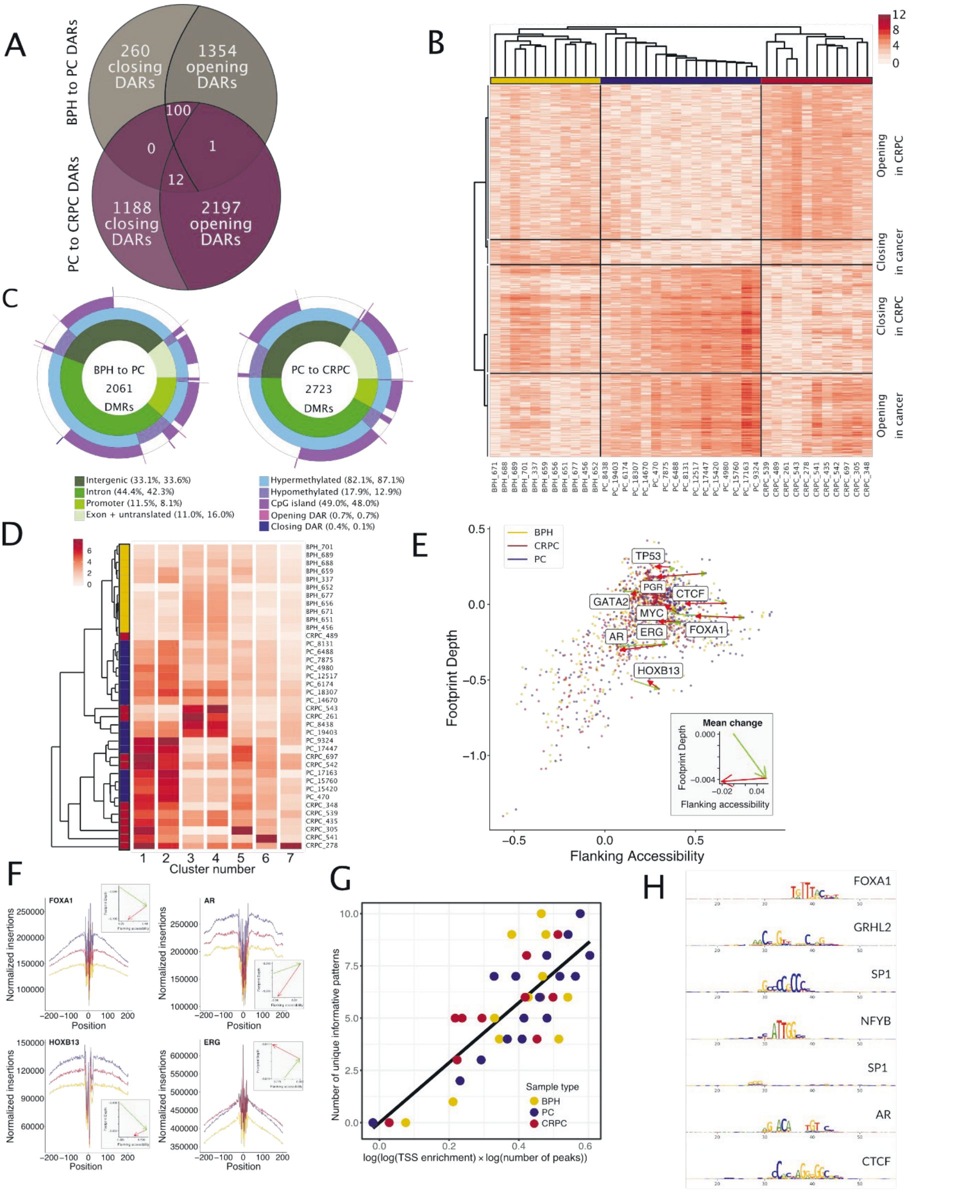


Figure 2

Figure 2: Differential accessibility is concentrated on regulatory regions

A. Venn diagram showing the numbers BPH to PC and PC to CRPC DARs and their overlap. Only a small portion of DARs are shared between comparisons. **B.** Clustering of samples using ATAC-seq signal of DARs separates them into BPH, PC, and CRPC groups, and identifies progression-related chromatin accessibility patterns. Scale bar shows log2 of normalized ATAC-seq signal. Pearson correlation was used as the distance metric, and linkage was calculated using the Weighted Pair Group Method with Arithmetic Mean (WPGMA) algorithm. **C.** Donut plots show genomic location categories for DMRs from BPH to PC and PC to CRPC comparison groups. Rings show whether DMR is hypermethylated or hypomethylated and whether it is located in a known CpG-island (within +/2kb). The outermost wedges show overlap with opening and closing DARs. In labels, percentages are given for BPH to PC and PC to CRPC DMRs, respectively. Differential accessibility and DNA methylation during progression occur at distinct loci. **D.** Unsupervised clustering of cancer-specific peaks shows clear clusters but fails to separate PC samples from CRPC samples. **E.** TF footprinting based on Tn5 transposase insertion sites was done for all expressed TFs with HOCOMOCO motif to quantify flanking accessibility and footprint depth. Averages from BPH, PC, and CRPC samples are shown and transitions in footprinting space (BPH to PC in green, PC to CRPC in red) are illustrated for PC-related TFs and those with the largest change between groups. Mean change is shown in the inset. **F.** Detailed TF footprints for key TFs AR, FOXA1, HOXB13, and ERG to illustrate the change in chromatin accessibility during progression. Quantification of footprint depth and flanking accessibility are shown in the insets. **G.** Motif discovery with BPNET correlates with the signal to noise (TSS enrichment) and the number of peaks used in training. **H.** Example of discovered motifs with BPNET on high quality sample PC_9324.

Distal regulatory elements accessibility correlate with expression of disease relevant genes

To gain insight into the functional role of accessible chromatin, we integrated ATAC-seq data with RNA and protein expression data from the same samples (see **Methods**). While promoter accessibility was consistent across samples, correlation between gene expression and transcription start sites (TSS) accessibility is very moderate (Spearman correlation $\rho = 0.11$ and $\rho = 0.04$ for RNA and protein data, respectively; **Figure 3A**, **Supplementary Table 4**). Analysis of gene groups with different expression levels (high, moderate, low, and housekeeping genes) suggests that this is due to promoters of expressed genes being mostly open in basal state (**Figure 3B**). However, differential chromatin accessibility at TSS and differential expression between groups are still co-occurring. For differentially expressed (D.E.) genes in the BPH to PC comparison, we observed an enrichment of genes with association between accessibility and expression (Fisher's exact test $p < 10^{-16}$, **Figure 3C**). These included several PC-related oncogenes such as *AR*, *MYC*, and *BCL11A* (**Figure 3D**). In the PC to CRPC comparison, there was an enrichment of genes in which TSS closing was associated with decreased expression (Fisher's exact test $p = 9.19 * 10^{-16}$, **Figure 3C**). Overall, from the promoter-proximal regions (-1kbp/+100bp), we detected 418 peaks, one BPH to PC DAR, and 9 PC to CRPC DARs with strong correlation (|correlation coefficient| > 0.5) to expression for the adjacent gene (**Supplementary Table 4**). For the PC to CRPC comparison, eight out of nine DARs showed increased accessibility. The remainder DAR shows reduced accessibility in CRPC and is located in the promoter of the *MIR30A* gene, which codes for a tumor suppressor miRNA (Jiang et al., 2018) downregulated in CRPC (log2 fold change -1.2389, Spearman $\rho = 0.7$ $p = 6.18*10^{-5}$). Thus, while global correlation is moderate, the expression of several disease-relevant genes is strongly correlated with promoter accessibility, suggesting reconfiguration of the promoter state during disease progression.

Next, we focused on understanding how distal accessible chromatin sites, that vary the most across the sample set, associate with gene expression changes. Co-regulated genes are found within topologically associating domains (TAD, (Pombo and Dillon, 2015)). Therefore, to limit the target gene associations to a biologically meaningful context, we used previously published annotations of TADs (Pombo and Dillon, 2015) from PC cells (see **Methods**). Within these TADs boundaries, we identified all peak-gene and DAR-gene pairs with a strong correlation (see **Methods**, **Supplementary Figure 9**). All together 9.6% (17,066) of all peaks and 25.4% (1300) of DARs were assigned to putative target genes based on correlation (**Supplementary Table 4**), including 8977 unique genes from 1871 TADs. We found that 29.6% of PC to CRPC DARs correlate with gene expression while only 16.4% of BPH to PC DARs correlate. Ingenuity Pathway Analysis (IPA) performed separately for genes associated with either DARs or peaks showed several PC-related and cancer-related pathways enriched (**Supplementary Table 4**), demonstrating that chromatin-related changes reflect disease-relevant target gene alterations.

When looking into associations with specific PC genes, we found 5 PC to CRPC DARs and 48 peaks with strong correlation to *AR* expression (**Figure 3E**, **Supplementary Table 4**). DARs correlated with *AR* expression are located within 2 Mbp region around the *AR* locus, indicating regulatory potential throughout the TAD area. These DARs harbor binding sites for key TFs including *AR*, *FOXA1*, *HOXB13*, and *ERG* (**Figure 3E**). Peaks correlating with *AR* expression are mainly upstream of TSSs (41/48) (**Figure 3E**). Identified peaks include an enhancer known to be amplified in advanced PC ((Takeda et al., 2018; Viswanathan et al., 2018)). The expression of 42 known oncogenes (e.g. *EGFR, ERBB2, JUN, FGFR1* and *FGFR2*), 27 tumor suppressor genes (e.g. *NOTCH1, BRCA1, BRCA2, IL2*) and 22 genes related to chromatin regulation (e.g. *HDAC1, HDAC2, HDAC5, HDAC6, HDAC9, HDAC10,* and *SMARCD1*) correlated with the chromatin accessibility of at least one peak (**Supplementary Table 4**). In addition, the expression of 4 oncogenes (*JUN, PIM1, CARD11,* and *TFG*), 5 tumor suppressor genes (*PTEN, NOTCH1, CDK6, FH,* and *WT1*) and 2 factors involved in chromatin regulation (*HDAC7* and *CHRAC1*) were strongly associated with DARs irrespective of the comparison group (**Supplementary Table 4**).

Analyses of distal and promoter areas identified 418 associations from TSS signal to gene expression as well as 27,353 peak–gene and 3,513 DAR-gene pairs. Expression-associated areas of chromatin accessibility are mostly located close to TSSs (median distance 4.7 kbp upstream of TSS) (**Figure 3F**). Also, 45.8% (4,124) of genes with expression correlating with chromatin accessibility are linked to exactly one regulatory element, while 97 genes (1.07%) can be associated to 30 or more regulatory elements (mean= 3.4, **Figure 3F, middle**). Likewise, 72.4% (13,359) of peaks or DARs correlating with gene expression are associated with a single gene and 35 are linked to 30 or more unique genes, indicating that those might be regulatory hubs (mean=1.7, **Figure 3F, right**). Taken together, we could associate at least one peak or DAR to 45.5% of genes and 30.8% of proteins (**Figure 3G**, **Supplementary Table 4**). When focusing on the genes with differential expression patterns, 62.4% and 84.7% of genes and proteins were associated, respectively. As reported earlier, correlations at the transcript and protein levels are not consistent (Latonen et al., 2018; Sinha et al., 2019), but both data levels support the conclusion that the majority of differential expression in progression-related genes can be correlated with chromatin accessibility.

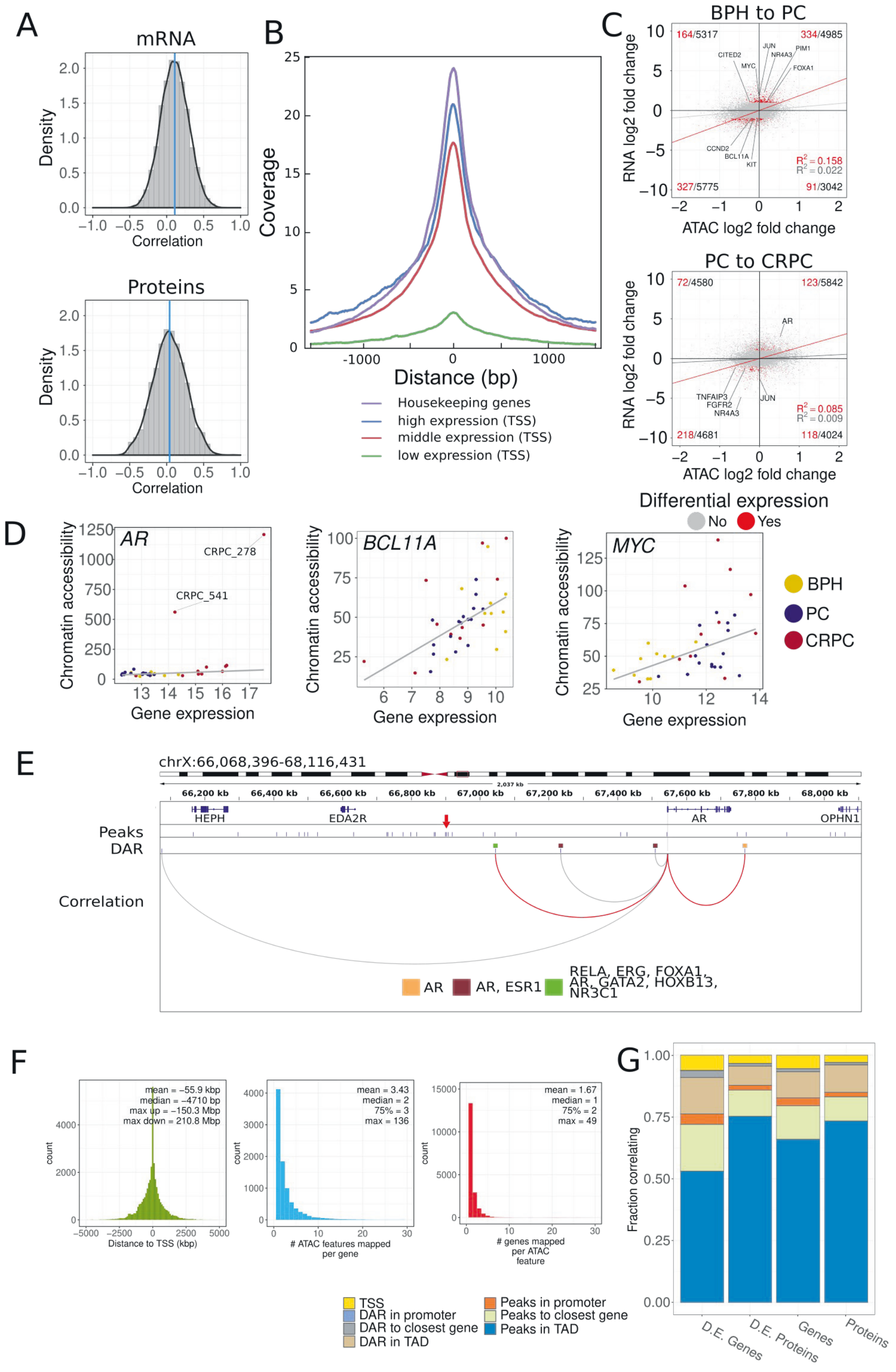

A
mRNA
Density
2.0
1.5
1.0
0.5
0.0
−1.0 −0.5 0.0 0.5 1.0
Correlation
Proteins
Density
Correlation
B
Coverage
25
20
15
10
5
0
-1000
0
1000
Distance (bp)
Housekeeping genes
high expression (TSS)
middle expression (TSS)
low expression (TSS)
C
BPH to PC
RNA log2 fold change
164/5317
334/4985
327/5775
91/3042
CITED2
MYC
JUN
NR4A3
PIM1
FOXA1
CCND2
BCL11A
KIT
$R^2 = 0.158$
$R^2 = 0.022$
ATAC log2 fold change
PC to CRPC
72/4580
123/5842
218/4681
118/4024
AR
TNFAIP3
FGFR2
NR4A3
JUN
$R^2 = 0.085$
$R^2 = 0.009$
ATAC log2 fold change
Differential expression
No
Yes
D
Chromatin accessibility
AR
CRPC_278
CRPC_541
1250
1000
750
500
250
0
13 14 15 16 17
Gene expression
BCL11A
100
75
50
25
6 7 8 9 10
Gene expression
MYC
125
100
75
50
25
10 12 14
Gene expression
BPH
PC
CRPC
E
chrX:66,068,396-68,116,431
66,200 kb
66,400 kb
66,600 kb
66,800 kb
67,000 kb
67,200 kb
67,400 kb
67,600 kb
67,800 kb
68,000 kb
2,037 kb
HEPH
EDA2R
AR
OPHN1
Peaks
DAR
Correlation
AR
AR, ESR1
RELA, ERG, FOXA1, AR, GATA2, HOXB13, NR3C1
F
mean = −55.9 kbp
median = −4710 bp
max up = −150.3 Mbp
max down = 210.8 Mbp
count
4000
2000
0
−5000 −2500 0 2500 5000
Distance to TSS (kbp)
mean = 3.43
median = 2
75% = 3
max = 136
count
4000
3000
2000
1000
0
0 10 20 30
# ATAC features mapped per gene
mean = 1.67
median = 1
75% = 2
max = 49
count
15000
10000
5000
0
0 10 20 30
# genes mapped per ATAC feature
G
Fraction correlating
1.00
0.75
0.50
0.25
0.00
D.E. Genes
D.E. Proteins
Genes
Proteins
TSS
DAR in promoter
DAR to closest gene
DAR in TAD
Peaks in promoter
Peaks to closest gene
Peaks in TAD

Figure 3

Figure 3: Distal features detected by ATAC-seq are correlated with gene expression in prostate cancer samples

A. Correlation between TSS chromatin accessibility and gene expression is moderate at the genome-wide scale. Density plot of Spearman correlation coefficient between gene (top, median=0.11) or protein (bottom, median=0.04) expressions and normalized ATAC-seq signal at the TSS. **B.** ATAC-seq background-corrected coverage on TSS. TSSs are grouped based on expression of the gene (high, middle, and low (see **Methods**)) or annotation to housekeeping genes. Chromatin of expressed genes is accessible at TSS. Low expression genes show minimal chromatin accessibility. **C.** Differential expression and chromatin accessibility have positive association. Scatter plots visualize the association between differential RNA expression and TSS accessibility in BPH to PC (top panel) and PC to CRPC (bottom panel) comparisons. Differentially expressed genes are shown in red. Gray and red lines show regression lines fitted to their corresponding data points to demonstrate the association between data types. Selected oncogenes are labeled. Numbers in the corners of each quadrant of the scatter plot report counts of differentially expressed and total genes. Differentially expressed genes are enriched for opening chromatin and increased expression, and closing and decreased expression in BPH to PC comparison. In PC to CRPC comparison, enrichment is seen only in closing and decreased expression quadrant. **D.** Correlation between chromatin accessibility and gene expression for the selected oncogenes demonstrate increasing (*AR, MYC*) and decreasing (*BCL11A*) accessibility during progression. For *AR*, outlier chromatin accessibility is observed for samples with high-level amplification identified from DNA-seq data (CRPC_278, CRPC_541). **E.** Correlation analysis between chromatin accessibility and gene expression identifies putative regulatory elements. In total 48 peaks and 5 DARs are detected in a 2 Mbp TAD region around the *AR* locus. Known associations from GeneHancer database are shown in red. Binding sites for selected TFs from GTRD database within associated DARs are shown. Red arrow indicates a peak detected at recently reported *AR* enhancer locus. **F.** Characterization of correlations shows that associations between regulatory elements and genes are specific. Left panel shows the distance of correlating chromatin features from TTS. Middle panel indicates the number of chromatin features mapped to each gene. Finally, the last panel gives the number of genes mapped to each chromatin feature. Summary statistics are given in the insets. Mean, median, and maximum upstream (max up) and downstream (max down) distances are reported for the distance distribution. For the middle and right panels, mean, median, upper quartile and maximum number of associations are reported. **G.** Summary of all correlation analyses. Fraction of genes and proteins correlating with ATAC-seq features across all analyses is reported. Data for all and differentially expressed gene subsets are shown.

Chromatin accessibility alterations during disease progression are associated with different transcription factors regulatory modules

To gain understanding on how the chromatin accessible sites direct transcriptional programs during PC progression, we generated TF–gene expression regulatory network. TFs were connected to their target genes through known binding sites in accessible chromatin regions (see **Methods**). We focused this analysis specifically on DARs that correlate with gene expression (**Figure 4A**). From the TF-gene network that we generated, we identified regulatory modules, defined as a set of TFs that share a set of target genes (see **Methods**). Two clear modules with 1082 and 799 target genes emerged from the analysis. The module with the largest number of target genes represents the well-characterized AR regulatory program, including AR, FOXA1, and ERG (**Figure 4B, Supplementary Figure 10A**). The second module contains a number of TFs with known function in driving aggressive prostate cancer e.g. glucocorticoid receptor (NR3C1) as well as TF coding genes MYC, HOXB13, GATA2, NKX3-1, and PGR (Chen et al., 2018; Grindstad et al., 2018; Isikbay et al., 2014; Koh et al., 2010; Rodriguez-Bravo et al., 2017). Surprisingly, genes targeted by this second module are a subset of AR module target genes (**Figure 4C**). We validated this by repeating the analysis using peaks instead of DARs (**Figure 4D**, **Supplementary Figure 10B-D**). IPA analysis of target genes confirmed AR as an upstream regulator for both modules (**Supplementary Table 4**), but in the second module, AR activity is predicted to be inhibited. This suggests that this second TF module could compensate for reduced AR activity e.g. due to androgen deprivation treatment. This was clearly shown for glucocorticoid receptor which is upregulated in CRPC especially resistant to enzalutamide treatment (Arora et al., 2013).

To elucidate the interplay of TFs in more detail, we performed a comparative analysis of TF binding sites, identified from prostate cancer cells, in opening and closing DARs (**Figure 4A**). In DARs from BPH to PC comparison AR, FOXA1 and HOXB13 binding sites are the most abundant and are co-occurring within 53.1% and 2.1% of opening and closing AR sites, respectively (**Figure 4E**, **Supplementary Figure 10E**). In PC to CRPC DARs, we observed the opposite pattern with 1.6% and 36.1% of opening and closing AR sites, respectively (**Figure 4F**, **Supplementary Figure 10E**). Again, we observed consistent correlations when repeating the analysis using peaks instead of DARs (**Supplementary Figure 5F).** These results suggest that chromatin opening in PC remains mostly accessible also in CRPC and harbour AR binding sites. Moreover, in CRPC new chromatin opening events enable additional TFs to bind the regulatory regions (**Supplementary Figure 10G-H**). Concomitantly, in CRPC several AR binding sites are closing, consistent with reduced AR activity in CRPC samples (p=0.02, **Supplementary Figure 10I**).

To test whether the chromatin in CRPC is selectively closed in AR binding sites related to canonical AR regulation, we used publicly available cell line data (Massie et al. 2011). To study the interplay between AR chromatin binding, androgen stimulation and chromatin accessibility we evaluated the overlap between androgen-induced AR binding sites in cell lines and DARs (**Figure 4G**, **Supplementary Figure 11A**). The majority of DARs are open in BPH to PC and closed in PC to CRPC comparison, which confirms our hypothesis that the canonical AR regulation is suppressed during progression to CRPC. In agreement with this observation, more PC-specific ATAC-seq peaks overlap these AR binding sites than CRPC- or BPH-specific peaks (**Supplementary Figure 11B**). We also note that the AR binding site locations from the cell line have most accessible chromatin in PC samples (**Figure 4H, Supplementary Figure 11C-F**).

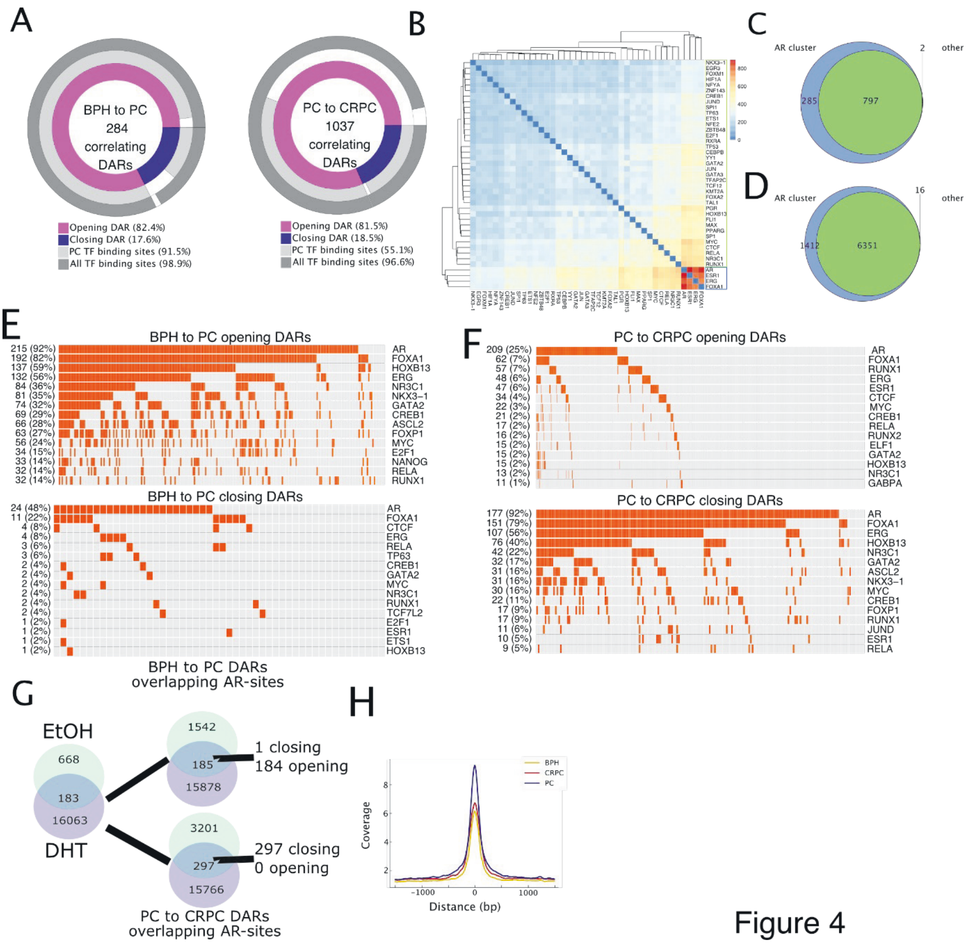


Figure 4

Figure 4: Disease progression alters prostate cancer-specific transcription factor binding site accessibility and regulatory programs

A. Donut plots showing numbers of gene expression correlating DARs in BPH to PC (left) and PC to CRPC (right) comparisons. Shown are also percentages of opening and closing sites and whether they harbour TF binding sites as characterized in the GTRD database. **B.** Hierarchical clustering of TF gene expression network uncovers two groups of TFs: a core cluster composed of AR, ERG, FOXA1 and ESR1, and a second cluster sharing a high number of target genes with the AR core cluster. Complete linkage and euclidean distance were used in clustering. Scale bar encodes the number of shared genes. **C**. Venn diagram shows that the two TF clusters indicated in B share a substantial amount of target genes. **D.** Repeating the intersection analysis with genes linked to peaks, a similar pattern as in C is observed. **E.** Oncoprints illustrate 15 TFs with the highest number of binding sites (taken from GTRD prostate cancer subset) overlapping with gene expression correlating DARs. Panels represent sites from BPH to PC opening (top) and closing (bottom) DARs. AR binding sites are present in almost all (92%) opening sites in this comparison. **F.** Similar oncoprints as in E but for PC to CRPC opening (top) and closing (bottom) DARs. In this comparison, most of the closing sites (92%) include AR binding sites **G.** Androgen-induced

AR binding sites taken from Massie et al. 2011 within DARs are present in opening regions in BPH to PC comparison and in closing regions in PC to CRPC comparison. **H.** Background-corrected ATAC-seq coverage of AR binding sites from androgen-treated (DHT) cells (Massie et al. 2011) is stronger in PC samples than other sample groups.

Discussion

In this study we integrated for the first time data on chromatin accessibility, DNA methylation, transcriptome, and proteome in clinical BPH, PC, and CRPC tissue samples. We used ATAC-seq to define a catalogue of accessible genomic regions and to characterize changes in chromatin accessibility during PC progression. The identified open chromatin regions are consistent with previous chromatin accessibility studies (Roadmap Epigenomics Consortium et al., 2015); (Corces et al., 2018). Furthermore, the number of detected peaks is consistent with earlier predictions of cancer type-specific peaks (Corces et al., 2018). Our analysis extended the known chromatin landscape by 38,157 reproducible previously uncovered accessible chromatin sites specific for PC. The majority of these sites have previously reported TF binding activity.

The chromatin accessibility of PC shows inter-sample heterogeneity. While we observed consistent accessibility at promoter regions during disease progression, accessibility does not correlate well with gene expression at the genome-wide level. As gene expression is regulated by the repressive or activating functions of the TFs binding to the promoters and distal regulatory elements, it is clear that promoter accessibility signal alone cannot be highly predictive of expression, as reported also by several earlier ATAC-seq studies across different systems (Rajbhandari et al., 2018; Scharer et al., 2018; Toenhake et al., 2018; Wu et al., 2018). This also highlights the important role of enhancers and their regulation in driving tumor development and progression. We did observe strong correlation with promoter or putative enhancer accessibility to gene expression for a subset of PC-related genes. At least one putative accessible regulatory element was found for 62.4% of protein coding genes and 84.7% of proteins with a differential expression. The majority of these regulatory elements are from the peaks and DARs that correlate with genes within the same TAD, providing a rich resource of candidate genes and regulatory elements for future investigation.

Still, a large fraction of putative regulatory regions could not be associated with genes. This might be explained by our utilization of stringent criteria for detecting target genes because of the limited cohort size. In addition, we used predefined TAD structure in the analysis and thus, our analysis could not detect associations resulting from altered TAD boundaries (Taberlay et al., 2016). Furthermore, many of the identified regions might contribute to functions other than direct regulation of gene expression. For example, it is known that higher order chromatin structure alterations may occur in PC tumorigenesis (Gerhauser et al., 2018), such as chromatin compartment formation and looping (Gerhauser et al., 2018; Rowley et al., 2018; Weischenfeldt et al., 2017). We did observe a large number of CTCF binding sites in peaks and DARs that may partially reflect these phenomena. Moreover, the majority of DARs in BPH to PC and PC to CRPC were opening (84.3% and 63.2%, respectively), supporting the idea that chromatin in PC initiation and progression undergoes a process of continued relaxation (Urbanucci et al., 2017)(Braadland and Urbanucci, 2019).

Chromatin accessibility and DNA methylation had the expected inverse relationship at the genome-wide level. The increase in the number of DMRs in the PC to CRPC comparison was not as significant as the two-fold increase in the number of DARs. This indicates that methylation-independent changes in chromatin accessibility are more prevalent during the progression to CRPC. Furthermore, DMRs and DARs overlapped in only a few regions, suggesting that these two epigenetic mechanisms are driving different transcriptional regulatory programs. Earlier work has shown the interplay between chromatin modifications and DNA methylation through interaction of EZH2 with DNA methyltransferases (DNMTs) (Viré et al., 2006). Further studies are needed to better understand how differential regulation of DNA methylation and chromatin accessibility are targeted.

Integration of TF binding data and predictions with accessible chromatin areas allowed us to analyze the regulatory programs that are associated with the identified peaks and DARs. Analysis of TF binding patterns demonstrated that despite high variability in chromatin accessibility, the observed motifs and TF enrichments are consistent during PC evolution. This suggests that there are a number of different chromatin configurations that can lead to similar phenotypes. For instance, AR was identified among the top candidate regulators but at the same time, the AR gene was one of the most targeted genes by chromatin remodelling during PC progression. A number of sites with accessibility were present in the genomic neighborhood of *AR,* including a previously reported AR-enhancer site, which was shown to be activated by structural rearrangement (Takeda et al., 2018). The analysis revealed that the interplay between AR, FOXA1, and HOXB13 TFs (Pomerantz et al., 2015) was the most prominent PC initiation-associated transcriptional regulatory module. FOXA1 is known to pioneer TFs binding to chromatin, including AR (Lupien et al., 2008) (Jozwik and Carroll, 2012). HOXB13 is a prostate lineage-specific TF and germline alterations have been shown to increase PC risk (Ewing et al., 2012). Previous studies with PC cell-lines identified alternative AR programs in CRPC (Sharma et al., 2013; Wang et al., 2009). Here we were able to show that this AR, FOXA1, HOXB13 program is initially activated in PC then depleted during progression to CRPC, when it is substituted by the activation of alternative regulatory modules composed of several TFs previously reported to be important in progression to CRPC. These TFs include glucocorticoid receptor, known to have a role in developing resistance to antiandrogens (Arora et al., 2013), and progesterone receptor that has been associated with disease progression (Grindstad et al., 2015, 2018). Overall, these analyses demonstrate that epigenetic chromatin reprogramming during CRPC progression enables binding sites for disease driving TFs, in addition to AR.

In summary, we demonstrated how transcriptional regulatory programs are altered in PC progression by characterizing the chromatin accessibility landscape and its alterations in human PC tissue. We reveal regulatory elements that are activated in PC and identify putative regulators for known oncogenic and tumor suppressive genes.

Methods

Sample collection

Fresh frozen tissue specimens were acquired from Tampere University Hospital (Tampere, Finland). 11 BPH, 16 untreated PC, and 11 CRPC samples were used for ATAC-seq library generation. BPH samples included were collected either by transurethral resection of the prostate (TURP; n=4) or radical prostatectomy (RP; n=7) (**Supplementary Table 1**). PC samples were obtained by radical prostatectomy. Locally recurrent CRPC samples were obtained by transurethral resection of the prostate. Samples were snap-frozen and stored in liquid nitrogen. Histological evaluation and Gleason grading was performed by a pathologist based on hematoxylin/eosin-stained slides. All samples contained a minimum of 70% cancerous or hyperplastic cells. The use of clinical material was approved by the ethical committee of the Tampere University Hospital. Written informed consent was obtained from the donors.

Tissue sample processing

Samples were cut from the frozen blocks as 2x50 µm sections. Nuclei were isolated from these sections. All the steps were performed on ice. First 6 ml of ice cold lysis buffer (10 mM Tris·Cl, pH 7.4, 10 mM NaCl, 3 mMMgCl2, 0.1% (v/v) Igepal CA-630, 1× protease inhibitors (Roche, cOmplete)) was added to pre-cooled petri dish and sections were moved from tube to petri dish with 1 ml of lysis buffer. Sections were cut into smaller pieces with a scalpel. Buffer and sections pieces were moved to a 15 ml Falcon tube. Each sample was pulled through a 16 G needle 15 times. Larger pieces were let to sink to the bottom. Supernatant was moved into a new tube and centrifuged at 700 g for 10 min at 4 °C. Supernatant was removed and the pellet was dissolved in a PBS buffer. Nuclei were counted and 50,000 nuclei were transferred to a new tube. Nuclei were pelleted by centrifugation at 700 g for 10 min at 4 °C. Supernatant was removed.

Processing of cell lines

VCaP cells were cultured in culbecco's modified eagle's medium with 10% fetal bovine serum and 1% L-glutamine. Cells were harvested using trypsin and counted. We took 50,000 cells and centrifuged them at 500 x g for 5 min, 4°C. Cells were washed once with 50 µl cold 1xPBS buffer and centrifuged again with the same settings. Supernatant was removed and cells resuspended to 50 µl of cold lysis buffer followed by centrifugation with the same settings. Supernatant was removed.

ATAC-seq library generation and sequencing

ATAC-seq libraries were generated as presented earlier (Buenrostro et al., 2013). Briefly, transposition mix (25 µl 2× TD buffer, 2.5 µl transposase (Tn5, 100 nM final), 22.5 µl water) was added to the nuclear pellet. Reaction was incubated at 37 °C for 45 minutes and amplified using PCR. Samples were purified using Qiagen MinElute PCR Purification Kit and again using Agencourt AMPure XP magnetic beads. For primer sequences, see **Supplementary Table 1**.

Samples were sequenced using Illumina NextSeq high output 2x75 bp settings. Seven samples were sequenced per run. Number of obtained sequencing reads is provided in **Supplementary Table 1**.

ATAC-seq data quality control, alignment, and peak detection

Raw sequencing reads were inspected using fastqc version 0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and subsequently trimmed with Trim Galore version 0.5.0 (https://github.com/FelixKrueger/TrimGalore) using parameters --fastqc --paired --length 20 -q 20. Sequence alignment was performed using Bowtie2 version

2.3.4.1 (Langmead and Salzberg, 2012) against GRCh38 reference genome. During alignment, parameters --sensitive-local and -X 2000 were used. Additional filtering (-q 20), sorting and indexing was done with Samtools version 1.8 (Li et al., 2009). Finally duplicates were marked using Picard Markduplicates tool version 2.9.2 (http://broadinstitute.github.io/picard/), with parameters VALIDATION_STRINGENCY=LENIENT and REMOVE_DUPLICATES=FALSE. For filtered alignments, peak calling was done with MACS2 v2.1.0 (Zhang et al., 2008) using parameters -g hs --llocal 160000 --slocal 147 -q 0.05 -f BAMPE --nomodel --broad --bgd --call-summits. Final quality control was performed for aligned samples after peak calling using ataqv toolkit (version 1.0.0, https://github.com/ParkerLab/ataqv).

Identification of artefact regions

As significant number of ATAC-seq reads originate from mitochondria, this can bias analysis at loci which have homology to autosomal or sex chromosome sequences. To exclude these regions from the analysis, we generated 100 copies of all the 30-mer sequences from mitochondrial DNA and aligned them to CRGh38 genome reference from which mitochondrial DNA had been excluded. Bowtie2 with --very-sensitive parameter was used. Alignments were converted to bed ranges using bedtools version 2.27.1 genomecov and merge tools (Quinlan and Hall, 2010).

ATAC-seq signal quantification

We binned the genome into overlapping windows of size 500 bp and steps of 250 bp. To obtain read counts in each window, we used *bedtools coverage -counts*. For robust quantification of the signal in loci of interest, background correction, normalization and bias correction steps were performed. To obtain background corrected read count *c* for a given window at position *x*, we used the following formula:

$$c(x)=max(0,(R(x)-max(Q_1(P_{10kbp}(x)),Q_1(P_{100kbp}(x)),Q_1(P_{chr}(x)))))$$

where *R* is the read count for the window at position *x, $P_{10kbp}()$*, *$P_{100kbp}()$*, and *$P_{chr}()$* are lists of read counts for all windows within the range of +/-5kbp, +/-50kbp, and chromosome arm, respectively, from position *x*, excluding the window at position *x. Q_1* is the value corresponding to the first quartile. This correction compensates for the variation in local background between samples and also enables detection of DARs from copy number aberrated genome areas (**Supplementary Figure 12A**). After background correction, we applied the median of ratio normalization (Anders and Huber, 2010), where sites with geometric mean below 1 were excluded from the calculation of the ratios, to obtain normalized read counts.

To compensate for potential bias due to sample collection procedure (RP and TURP), we divided the samples in the two groups. In the TURP group, we randomly assigned 4 BPH and 4 CRPC samples and in the RP group 4 BPH and 4 PC samples to keep the group sizes fixed. For each window, we applied the two-sided Wilcoxon rank-sum test. Random assignment of samples and significance testing was repeated 100 times. If 5th percentile of p-value distribution for a given window was less than p=0.01, we calculated the difference between medians of all the TURP and RP samples normalized read counts and subtracted this difference from all the TURP samples normalized read counts (717930 sites were corrected i.e. 6.5% of all sites). Application of this correction to normalized read counts resulted in quantified ATAC-seq signal.

Identification of the differentially accessible regions (DARs)

To identify DARs, we compared the samples from two different groups (BPH to PC or PC to CRPC). We calculated the log2-ratio of the median value of each group (eg. log2(median(PC) / median(BPH))), absolute median difference between two groups (e.g | median(PC) - median(BPH)|), and used the two-sided Wilcoxon rank-sum test of two groups. For each window, we checked whether all the following 3 criteria were satisfied: |log2-ratio| > 2; p-value < 0.01; absolute-median-difference > 14. These thresholds were derived based on false discovery rate (FDR) analysis and correspond to FDR 9.7% and 9.14% in BPH to PC and PC to CRPC comparisons, respectively. If the log2-ratio of a DAR was positive, we called it an opening DAR and if the log2-ratio of a DAR was negative, we called it a closing DAR.

Copy number aberration analysis

Raw sequencing reads from the whole genome sequencing experiment (DNA-seq) were aligned to the GRCh38 reference genome using Burrows-Wheeler Aligner (BWA) version 0.7.17 (Li and Durbin, 2009). Duplicate reads were marked using SAMBLASTER version 0.1.22 (Faust and Hall, 2014). Alignments were converted to BAM format and sorted using Samtools. We used Segmentum (Afyounian et al., 2017) to perform copy number analysis for the samples for which we had whole genome sequencing data (i.e. 4 BPH, 15 PC, 7 CRPC samples). Copy numbers were called using pooled BPH samples as reference with the following parameters: read depth were extracted for windows of width 500 bp, *window_size=15*, *clogr_threshold=0.8*, *min_read=35*, *logr_merge=0.2*. We used the reported log2-ratios for each genomic segment from Segmentum's result to infer the copy number of that segment. This data was used to confirm that quantified ATAC-signal was not confounded by copy number alterations (**Supplementary Figure 12A**).

Identification of the differentially methylated regions (DMRs)

Methylated DNA immunoprecipitation (meDIP) sequencing data was aligned to GRCh38 using Bowtie2 (settings: --score-min L,0,-0.15.), alignments were converted to BAM format and sorted using Samtools. Duplicated reads were marked with Picard Markduplicates. Samtools was used to filter out the duplicate reads. Differentially methylated regions were identified as described above for DARs using meDIP samples for which we had ATAC-seq data available. In the median of ratio normalization step, sites with geometric mean below 2 were excluded from calculating the ratios. DMRs were called with criteria |log2-ratio| > 2; p-value < 0.01; absolute-median-difference > 10, corresponding to FDR 4.61% and 7.90% for BPH to PC and PC to CRPC comparisons, respectively. If the log2-ratio of a DMR was positive, we called it a hypermethylated DMR and if the log2-ratio of a DMR was negative, we called it a hypomethylated DMR.

Compilation and quantification of the peak set

In order to compile a consensus set of peaks across all samples, we adapted the approach from (Corces et al., 2018). For each individual sample, we used the summits position of peaks called by MACS2 (Zhang et al., 2008) and extended them by +/- 250 bp to acquire the raw peak set for that sample. Preliminary signal for each raw peak was obtained using the above presented ATAC-seq signal quantification. If a raw peak was overlapping several adjacent windows, the weighted average based on the amount of overlap between the peak and overlapping windows, was used. For each sample, if there were overlapping raw peaks, the raw peak with the highest preliminary signal was selected. To standardize the peak signals across samples, these were further scaled in each sample by the sum of the signals of all the peaks divided by 10^6 (i.e. $\left(\sum_{i=1}^{n} \square i_{th}\, peak\ signal\right)/10^6$where *n* is the number of raw peaks in a given sample). Next, we pooled the peaks across all samples and removed their overlaps with the above approach using scaled signal values. Further, we removed raw peaks from the set if they were only present in one sample. This resulted in a peak set without overlaps.

To quantify the peaks signal, we used the approach above at the peak coordinates. A peak was removed from the peaks set, if all samples had standardized signals below a data-driven threshold (*t=5*) for that peak (**Supplementary Figure 12B**). Using this filtering criterion, we removed 4,935 loci. Finally, 127 peaks overlapping the artefact regions were removed. This resulted in a final 178,206 peak set for analysis.

Quantification of chromatin accessibility at Transcription Start Sites (TSSs)

We extracted TSS coordinates for 18,537 protein coding genes and 1,471 miRNA (see quantification of gene and smallRNA expression below) from Ensembl version 90. For each gene and miRNA, we quantified chromatin accessibility within +/-500bp window from TSS using the same signal quantification approach as with the above peak set. The larger window size was used to account for the shape of the ATAC signal at the TSS sites (Supplementary Figure 1A).

Visualization of the coverage at peaks, DARs and DMRs

All boxplots show the quantified ATAC-seq signal at peak or DAR locations. In all boxplots, the median is shown with a green line and mean with a red triangle. Lower and upper whiskers have been set to first quartile (Q1) - 1.5*IQR (interquartile range) and third quartile (Q3) + 1.5*IQR, respectively.

In coverage plots, we extended midpoints of loci of interest by +/-1.5 kbp. For the resulting regions in each sample, we extracted the read counts in bins of size 10bp using *bedtools coverage*. Next, we subtracted an estimated global background from the read count of each bin to acquire the background corrected read counts. To estimate the global background, we randomly selected 50,000 loci of size 500 bp excluding those that overlap with the loci of interest using *bedtools random*. We extended, binned and quantified each of these loci as above. The global background was calculated by the arithmetic mean across all the binned read counts from random loci. In case of meDIP data, if a bin had a background corrected read count above 50 across all samples, it was considered as an artefact region and the read counts for that locus were set to zero. To generate sample-specific profiles, we calculated the arithmetic mean of background corrected values across the corresponding bins for all loci of interest.

Finally, we calculated the median across the corresponding bins of sample-specific profiles for each group (i.e. either BPH, PC, or CRPC).

Annotation of peaks and DARs and DMRs

We annotated the loci of interest using *annotatePeaks.pl* routine from Hypergeometric Optimization of Motif EnRichment tool (HOMER; (Heinz et al., 2010)). We grouped regions annotated as 3' UTR, TTS, non-coding, 5' UTR, and exon under the term “Exon + untranslated”. We further annotated the loci of interest using *bedtools intersect* or *closest* with the following data sets. GeneHancer version 4.7 (Fishilevich et al., 2017) was used to annotate known regulatory elements (enhancers and promoters) and predicted regulatory region target gene associations. Pan-cancer peak set and PRAD peak calls from ATAC-seq data generated from TCGA samples (Corces et al., 2018), and Roadmap Epigenomics project DNase-seq data (Roadmap Epigenomics Consortium et al., 2015) were used to annotate previously identified accessible chromatin areas. Roadmap Epigenomics data was downloaded from reg2map (https://personal.broadinstitute.org/meuleman/reg2map/HoneyBadger_release/) and data from 3 distinct sets of regions (i.e. promoters, enhancers and dyadic regions) were combined. Duplicates were removed and LiftOver (Hinrichs et al., 2006) was used to convert the GRCh37 coordinates to GRCh38 (only 0.03% of the sites were lost due to LiftOver). To annotate ATAC features with experimentally validated transcription factor binding sites (TFBS), we downloaded the data from the Gene Transcription Regulation Database version 18.06 (Yevshin et al., 2019) which collects 5,068 ChIP-seq experiments and data from 846 unique TF. From the entire database, we subset GTRD ChIP-seq data for binding sites detected in prostate cancer cell lines and use it as a prostate-specific set, including 40 unique TFs from 1818 experiments. We assigned TFBS to ATAC features using R *findOverlaps* function.

We checked the overlap between DMRs and the CpG islands using the information obtained from UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/cpgIslandExt.txt.gz) by using *bedtools closest*. If the distance of a locus and its closest CpG island was below 2 kbp, we marked that locus as a CpG island.

We retrieved a list of 3,804 human housekeeping genes from (Eisenberg and Levanon, 2013). We mapped gene names to Ensembl gene id version 90 using the R *merge* function. The resulting list included 3,662 genes.

Quantification of gene expression

Previously published transcriptome sequencing data, including 12 BPH, 30 PC and 13 CRPC samples (Ylipää et al., 2015) was aligned to GRCh38 and quantified by STAR version 2.5.3a using Ensembl version 90 annotations. We obtained quantification of 58243 genes. Samples with high quality data were matched with ATAC-seq samples (**Supplementary Table 1**). Lower quartile value of expression distribution across all the samples was used as a threshold to remove low expressed genes, resulting in 18537 protein coding genes (mitochondrial excluded) available for the analysis. The DESeq2 version 1.20 Bioconductor package was used to model the data and extract differentially expressed genes. We fit the model taking into account both RNA isolation methods (Qiagen™ Trizol™ and Qiagen™ All

Prep™) and stages of prostate cancer progression (BPH, PC, CRPC). To address the bias introduced by different extraction protocol, we used coefficients estimated from the model (Ylipää et al., 2015): we extracted coefficients for the RNA isolation method covariate using DESeq2 *coef* function and subtracted these values from library size corrected read counts of Trizol-treated samples in log2 scale. We detected differential expression in two comparisons, BPH to PC and PC to CRPC. A gene was considered differentially expressed (D.E.) if the absolute median difference of normalized read counts between the groups was greater than 180, the log2 fold change was greater than 1 and the FDR corrected p-value lower than 0.05. In the BPH to PC comparison and PC to CRPC comparisons 933 and 533 D.E. genes were detected. If the log2-ratio of a D.E. was positive, we called it an overexpressed D.E. gene. If the log2-ratio of a D.E. was negative, we called it an underexpressed D.E. gene.

Quantification of small RNA expression

Previously published small RNA sequencing data (Ylipää et al., 2015) was re-analysed by mapping sequence tags to human sequences from mirBase version 22. We mapped sequencing tags allowing for single base deletion at the 3' or insertion at either 3' or 5'. Modified sequences mapping to the same mirBase identifier were collapsed and their abundance summed. This process yielded data for 1471 annotated miRNA sequences. The resulting data matrix was normalized using median of ratios normalization, genes with geometric mean lower than 15 were discarded. Differentially expressed miRNA were detected in BPH to PC and PC to CRPC comparisons. A miRNA was considered differentially expressed if showing a log2 fold change greater than 1 and a FDR adjusted t-test p-value lower than 0.05. This analysis yielded 26 and 51 differentially expressed miRNA for BPH to PC and PC to CRPC comparisons, respectively.

Protein expression data

We used our previously published sequential window acquisition of all theoretical mass spectra (SWATH-MS) data and defined differentially expressed proteins as described earlier (Latonen et al., 2018).

Quantification of AR activity score

AR activity score was determined using a publicly available gene expression signature composed of 27 genes (Hieronymus et al., 2006). Of these, 21 genes are upregulated in the first 24 hours after androgenic treatment: PSA, TMPRSS2, NKX3-1, KLK2, GNMT, TMEPAI, MPHOS9, ZBTB10, EAF2, BM039, SARG, ACSL3, PTGER4, ABCC4, NNMT, ADAM7, FKBP5, ELL2, MED28, HERC3, MAF.

Normalized gene expression values in log2 scale were converted to z-scores:

$$z_i = \frac{g_i - \mu_i}{\sigma_i}$$

where i represents a gene from the list above, g the gene expression, μ the arithmetic mean of expression values and $\sigma$the standard deviation of the gene. Both mean and standard deviation were computed using all samples. For each sample, AR activity score was computed by summing genes scores:

$$s_j = \sum_{i}^{\square} \square z_{ij}$$

Scores were split according to sample groups: BPH, PC and CRPC. Score distributions were visualized as violin plots. Both upper and lower quartiles and the mean activity score value were overlaid on the figures. P-values were computed with two tails Mann-Whitney U-test to assess the statistical significance between BPH and PC or PC and CRPC scores under the null hypothesis of no difference between groups.

Association of chromatin accessibility with target genes

To link chromatin features (peaks and DARs) to putative target genes, we performed correlation analysis across all samples with RNA-seq, smallRNA-seq, or SWATH-MS protein data. We calculated Pearson and Spearman correlations between ATAC-seq signal and gene or protein expression in four different contexts: 1) at transcription start site (TSS), defined as +/-500bp from TSS annotation, we computed correlation between TSS ATAC-seq signal and corresponding gene expression; 2) we defined a region of 1 kbp upstream and 100 bp downstream of TSS as a promoter, we searched for ATAC features overlapping this region and computed correlation between their signal and corresponding gene or protein expression, if available; 3) for each ATAC feature we searched for the closest gene using annotations from the HOMER tool and computed correlation between their signal and gene or protein expression, if available; 4) we used all ATAC features and genes falling within same TAD to compute correlation between all pairs. To define TAD boundaries, we used annotations from ENCODE consortium based on data from LNCaP cell line (ENCODE Project Consortium, 2012), GEO accession: GSE105557, downloaded from http://promoter.bx.psu.edu/hi-c/ (Wang et al., 2018)). We extended this list with genomic intervals included between each pair of TAD using *bedtools complement* and merged the resulting list with the initial one.

To derive a threshold for significant associations, in each context, we computed null distributions by randomizing sample order prior to correlation coefficient computation (**Supplementary Figure 9A**). To enable false positive rate estimation, randomization was repeated 10 times for each pair of comparisons in each context. Based on evaluation of the distributions, we chose to set thresholds to |correlation coefficient| > 0.5 for genes and |correlation coefficient| > 0.6 for proteins resulting in false positive rate from 5.4 * 10^{-3} to 1.3 * 10^{-3}, respectively (**Supplementary Figure 9A**). Associations with either Pearson or Spearman correlation above threshold were kept for the downstream analysis. The above analysis was implemented by custom script using standard Unix tools, Python 3.6.8, R version 3.5.2 and packages from the Bioconductor framework managed via the BiocManager package version 3.8, HOMER tool and *bedtools*.

TF gene expression network

Each gene was assigned to one or more ATAC-seq features from previous correlation analysis. Transcription factor binding sites in ATAC-seq features were detected during the annotation step. A gene was defined as the target of a transcription factor if its expression showed correlation with accessibility of an ATAC-seq feature carrying a binding site for the TF. For each pair of TFs, the number of co-regulated genes was calculated resulting in a contingency matrix of 845x845 TFs. This matrix was filtered to retain TFs sharing at least 100 genes, leaving a 192x192 contingency matrix. Hierarchical clustering was applied and two clusters were detected. Manhattan distance was used as distance metric and UPGMA as clustering algorithm (**Supplementary Figure 10A**). The smallest cluster, containing the majority of detected connections, was extracted. Another filter was implemented similarly to the one described above: TFs sharing at least 300 genes with at least one other factor were retained. This filtering procedure resulted in a 41x41 contingency matrix. For visualization hierarchical clustering was calculated using Euclidean distance as distance metric and complete linkage as clustering algorithm. The analysis was implemented in R 3.5.2 using the *pheatmap* package for visualization and basic clustering functions.

TF binding site enrichment analysis

We determined the number of clusters for k-means clustering using consensus clustering with elbow method. For clustering, we used top 20% peaks with highest variance. For relative TF enrichment analysis, each cluster was compared against all the others. Enrichment analysis was performed using HOMER *findMotifs.pl* version 4.10. We used the full HOCOMOCO version 11 human TF (p 0.001) (Kulakovskiy et al., 2018) database as a known input TF list. Plotting was done using the R 3.5.2 and ggplot package.

TF footprinting and accessibility

For TF footprint depth and flanking accessibility analysis Tn5 cut sites were counted using custom R scripts. Pooled samples for BPH, PC and CRPC groups were generated using *Picard MergeSamFiles* and used for group level analysis. In other analysis, individual BAM files were used directly. Possible TF binding locations were predicted using FIMO version 5.0.2 (Grant et al., 2011) with HOCOMOCO v11 database and *--thresh 0.001* parameter. Predicted sites were intersected with peaks and DARs from BPH to PC and PC to CRPC comparison groups. We filtered the TF list by gene expression across samples. TF belonging to the lower quartile of this distribution were discarded. We quantified footprint base as mean count of insertions at the motif positions, while for flanking height, we considered 25 bases around each detected motif. To quantify each motif background, we used a set of 25 bp windows 200 bp upstream and downstream of the motif center. We computed flanking accessibility as log2(flanking height/background) and footprint depth as log2(footprint base/flanking height). For expression association, Pearson correlation between these footprint parameters and TF expression was calculated. In footprint visualization, the number of cutting sites were scaled according to read numbers in respective phenotypes.

Motifs discovery from accessible chromatin sites

We used the BPNET Python package version 0.0.21 (Avsec et al.) to train and interpret sequence-to-profile convolutional neural networks from sample-specific ATAC-seq data. In BPNET recurring patterns with high contribution scores are clustered based on sequence identity to build contribution weight matrices (CWMs). We first tested the applicability of the BPNET model with data from VCaP cell lines. We compared the CWMs obtained from models trained with publicly available AR Chip-seq data (Massie et al., 2011) and with ATAC-seq data generated in-house. These data were aligned and peaks detected as presented above. To consider the TF-specific binding context in ATAC-seq data, we extended the ATAC-seq peaks summits by 50 bp in both directions, and intersected the 100 bp regions with the direct AR-DNA interaction map defined in the UniBind database (Gheorghe et al., 2019). We kept the regions having an intersection of at least 1 bp, and selected these peaks as model training sequences. We tested the similarity between the resulting CWMs and all the known Position Weight Matrices (PWMs) collected in the HOCOMOCO v11 database (Kulakovskiy et al., 2018) using the Tomtom motif comparison tool (Gupta et al., 2007). Tomtom results were used to identify the TF or TF family to be associated with each CWM. We observed that the motifs discovered by the model trained with Chip-seq data were also discovered by the model trained with ATAC-seq data (**Supplementary Figure 4**). We trained BPNET models for each of the 38 clinical samples using the above presented peaks set summits to define the training sequences and the ATAC-seq data. We then applied the procedure we tested on cell lines to build and interpret TF-specific models for 4 TFs - namely AR, FOXA1, HOXB13, and ERG - on the highest quality samples (6 BPH, 4 CRPC, 8 PC) having TSS enrichment > 3.5. ATAC-seq BPNET models were trained on canonical chromosomes with default hyper-parameters, and chromosomes 2, 3, and 4 as validation chromosomes. Chip-seq BPNET models were trained using the same configuration, except an increased kernel size of 50 for the transposed convolution layer. The models trained in cell lines and the models trained in clinical samples using the above presented peaks set summits were trained with a patience of 5 epochs. The TF-specific models trained in clinical samples were trained with a patience of 20 epochs. We represented the information content of the discovered motifs as sequence logos using the built-in BPNET function; when more than one meta cluster was reported by BPNET, we omitted the meta clusters with no matching TFs if at least one pattern in another meta cluster had a TF or a TF family assigned to it (**Supplementary Figures S5-S8**).

Data and code availability

Sequencing data has been deposited in European Genome-phenome Archive under accession number EGAS00001000526. Code used for the analysis is available at https://github.com/nykterlab/Tampere_PC/

Authors contributions

Initiated the study: MN
Supervised the work: MN, TV, KG, JK
Designed the analysis: MN, KG, JK, LL, AU
Produced the data: JU
Contributed data / samples / analysis tools: MA, KK, TT
Performed analysis: JU, EA, FT, TH, AL, AS, KG, RN
Drafted the manuscript: JU
Wrote the paper: JU, EA, FT, TH, KG, MN
All authors have read and approved the final version of the paper.

Declaration of interests

All authors declare that they have no conflicts of interest.

Acknowledgements

We thank Paula Kosonen, Hanna Selin, Riina Kylätie, Marja Pirinen, and Päivi Martikainen for their technical assistance. Financial support is acknowledged from: Academy of Finland project no. 312043 (MN), project no. 310829 (MN), project no. 333545 (KG), project no. 317871 (LL), Cancer Foundation Finland (MN, KG, LL), Competitive State Research Financing of the Expert Responsibility area of Tampere University Hospital (MN, KG), Norwegian Cancer Society project no. 198016-2018 (AU), Pirkanmaa Regional Fund (JU), Tampere University doctoral school (EA, FT), Emil Aaltonen Foundation (KG), Sigrid Juselius Foundation (MN, KG, LL), Foundation of the Finnish Cancer Institute (LL). CSC—IT Center for Science Ltd. (https://www.csc.fi/csc) provided computational resources for the work. The results published here are in part based upon data generated by The Cancer Genome Atlas project (dbGaP Study Accession: phs000178.v9.p8) established by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov. We acknowledge ENCODE Consortium and the ENCODE production laboratories for generating the Hi-C data.

References

Adams, E.J., Karthaus, W.R., Hoover, E., Liu, D., Gruet, A., Zhang, Z., Cho, H., DiLoreto, R., Chhangawala, S., Liu, Y., et al. (2019). FOXA1 mutations alter pioneering activity, differentiation and prostate cancer phenotypes. Nature *571*, 408–412.

Afyounian, E., Annala, M., and Nykter, M. (2017). Segmentum: a tool for copy number analysis of cancer genomes. BMC Bioinformatics *18*, 215.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Annala, M., Kivinummi, K., Tuominen, J., Karakurt, S., Granberg, K., Latonen, L., Ylipää, A., Sjöblom, L., Ruusuvuori, P., Saramäki, O., et al. (2015). Recurrent SKIL-activating rearrangements in ETS-negative prostate cancer. Oncotarget *6*, 6235–6250.

Armenia, J., Wankowicz, S.A.M., Liu, D., Gao, J., Kundra, R., Reznik, E., Chatila, W.K., Chakravarty, D., Han, G.C., Coleman, I., et al. (2018). The long tail of oncogenic drivers in prostate cancer. Nat. Genet. *50*, 645–651.

Arora, V.K., Schenkein, E., Murali, R., Subudhi, S.K., Wongvipat, J., Balbas, M.D., Shah, N., Cai, L., Efstathiou, E., Logothetis, C., et al. (2013). Glucocorticoid receptor confers resistance to antiandrogens by bypassing androgen receptor blockade. Cell *155*, 1309–1322.

Augello, M.A., Liu, D., Deonarine, L.D., Robinson, B.D., Huang, D., Stelloo, S., Blattner, M., Doane, A.S., Wong, E.W.P., Chen, Y., et al. (2019). CHD1 Loss Alters AR Binding at Lineage-Specific Enhancers and Modulates Distinct Transcriptional Programs to Drive Prostate Tumorigenesis. Cancer Cell *35*, 817–819.

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., et al. Base-resolution models of transcription factor binding reveal soft motif syntax.

Bedford, M.T., and van Helden, P.D. (1987). Hypomethylation of DNA in pathological conditions of the human prostate. Cancer Res. *47*, 5274–5276.

Börno, S.T., Fischer, A., Kerick, M., Fälth, M., Laible, M., Brase, J.C., Kuner, R., Dahl, A., Grimm, C., Sayanjali, B., et al. (2012). Genome-wide DNA methylation events in TMPRSS2-ERG fusion-negative prostate cancers implicate an EZH2-dependent mechanism with miR-26a hypermethylation. Cancer Discov. *2*, 1024–1035.

Braadland, P.R., and Urbanucci, A. (2019). Chromatin reprogramming as an adaptation mechanism in advanced prostate cancer. Endocr. Relat. Cancer *26*, R211–R235.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Cancer Genome Atlas Research Network (2015). The Molecular Taxonomy of Primary Prostate Cancer. Cell *163*, 1011–1025.

Chen, Z., Wu, D., Thomas-Ahner, J.M., Lu, C., Zhao, P., Zhang, Q., Geraghty, C., Yan, P.S., Hankey, W., Sunkel, B., et al. (2018). Diverse AR-V7 cistromes in castration-resistant prostate cancer are governed by HoxB13. Proc. Natl. Acad. Sci. U. S. A. *115*, 6810–6815.

Corces, M.R., Ryan Corces, M., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., et al. (2018). The chromatin accessibility landscape of primary human cancers. Science *362*, eaav1898.

Eisenberg, E., and Levanon, E.Y. (2013). Human housekeeping genes, revisited. Trends Genet. *29*, 569–574.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Espiritu, S.M.G., Liu, L.Y., Rubanova, Y., Bhandari, V., Holgersen, E.M., Szyca, L.M., Fox, N.S., Chua, M.L.K., Yamaguchi, T.N., Heisler, L.E., et al. (2018). The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression. Cell *173*, 1003–1013.e15.

Ewing, C.M., Ray, A.M., Lange, E.M., Zuhlke, K.A., Robbins, C.M., Tembe, W.D., Wiley, K.E., Isaacs, S.D., Johng, D., Wang, Y., et al. (2012). Germline mutations in HOXB13 and prostate-cancer risk. N. Engl. J. Med. *366*, 141–149.

Faust, G.G., and Hall, I.M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics *30*, 2503–2505.

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., et al. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database *2017*.

Flavahan, W.A., Gaskell, E., and Bernstein, B.E. (2017). Epigenetic plasticity and the hallmarks of cancer. Science *357*.

Friedlander, T.W., Roy, R., Tomlins, S.A., Ngo, V.T., Kobayashi, Y., Azameera, A., Rubin, M.A., Pienta, K.J., Chinnaiyan, A., Ittmann, M.M., et al. (2012). Common structural and epigenetic changes in the genome of castration-resistant prostate cancer. Cancer Res. *72*, 616–625.

Gerhauser, C., Favero, F., Risch, T., Simon, R., Feuerbach, L., Assenov, Y., Heckmann, D., Sidiropoulos, N., Waszak, S.M., Hübschmann, D., et al. (2018). Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. Cancer Cell *34*, 996–1011.e8.

Gheorghe, M., Sandve, G.K., Khan, A., Chèneby, J., Ballester, B., and Mathelier, A. (2019). A map of direct TF-DNA interactions in the human genome. Nucleic Acids Res. *47*, 7715.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

Grasso, C.S., Wu, Y.-M., Robinson, D.R., Cao, X., Dhanasekaran, S.M., Khan, A.P., Quist, M.J., Jing, X., Lonigro, R.J., Brenner, J.C., et al. (2012). The mutational landscape of lethal castration-resistant prostate cancer. Nature *487*, 239–243.

Grindstad, T., Andersen, S., Al-Saad, S., Donnem, T., Kiselev, Y., Nordahl Melbø-Jørgensen, C., Skjefstad, K., Busund, L.-T., Bremnes, R.M., and Richardsen, E. (2015). High progesterone receptor expression in prostate cancer is associated with clinical failure. PLoS One *10*, e0116691.

Grindstad, T., Richardsen, E., Andersen, S., Skjefstad, K., Rakaee Khanehkenari, M., Donnem, T., Ness, N., Nordby, Y., Bremnes, R.M., Al-Saad, S., et al. (2018). Progesterone Receptors in Prostate Cancer: Progesterone receptor B is the isoform associated with disease progression. Sci. Rep. *8*, 11358.

Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L.B., Tubio, J.M.C., Papaemmanuil, E., Brewer, D.S., Kallio, H.M.L., Högnäs, G., Annala, M., et al. (2015). The evolutionary history of lethal metastatic prostate cancer. Nature *520*, 353–357.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W. (2007). Quantifying similarity between motifs. Genome Biology *8*, R24.

Hankey, W., Chen, Z., and Wang, Q. (2020). Shaping Chromatin States in Prostate Cancer by Pioneer Transcription Factors. Cancer Res. *80*, 2427–2436.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Hieronymus, H., Lamb, J., Ross, K.N., Peng, X.P., Clement, C., Rodina, A., Nieto, M., Du, J., Stegmaier, K., Raj, S.M., et al. (2006). Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. Cancer Cell *10*, 321–330.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590–D598.

Isikbay, M., Otto, K., Kregel, S., Kach, J., Cai, Y., Vander Griend, D.J., Conzen, S.D., and Szmulewitz, R.Z. (2014). Glucocorticoid receptor activity contributes to resistance to androgen-targeted therapy in prostate cancer. Horm. Cancer *5*, 72–89.

Jiang, L.-H., Zhang, H., and Tang, J.-H. (2018). MiR-30a: A Novel Biomarker and Potential Therapeutic Target for Cancer. J. Oncol. *2018*, 5167829.

Jimenez, R.E., Fischer, A.H., Petros, J.A., and Amin, M.B. (2000). Glutathione S-transferase pi gene methylation: the search for a molecular marker of prostatic adenocarcinoma. Adv. Anat. Pathol. *7*, 382–389.

Jozwik, K.M., and Carroll, J.S. (2012). Pioneer factors in hormone-dependent cancers. Nat. Rev. Cancer *12*, 381–385.

Koh, C.M., Bieberich, C.J., Dang, C.V., Nelson, W.G., Yegnasubramanian, S., and De Marzo, A.M. (2010). MYC and Prostate Cancer. Genes Cancer *1*, 617–628.

Kron, K.J., Murison, A., Zhou, S., Huang, V., Yamaguchi, T.N., Shiah, Y.-J., Fraser, M., van der Kwast, T., Boutros, P.C., Bristow, R.G., et al. (2017). TMPRSS2–ERG fusion co-opts master transcription factors and activates NOTCH signaling in primary prostate cancer. Nat.

Genet. *49*, 1336.

Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. *46*, D252–D259.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Latonen, L., Afyounian, E., Jylhä, A., Nättinen, J., Aapola, U., Annala, M., Kivinummi, K.K., Tammela, T.T.L., Beuerman, R.W., Uusitalo, H., et al. (2018). Integrative proteomics in prostate cancer uncovers robustness against genomic and transcriptomic aberrations during disease progression. Nat. Commun. *9*, 1176.

Lee, W.H., Isaacs, W.B., Bova, G.S., and Nelson, W.G. (1997). CG island methylation changes near the GSTP1 gene in prostatic carcinoma cells detected using the polymerase chain reaction: a new prostate cancer biomarker. Cancer Epidemiol. Biomarkers Prev. *6*, 443–450.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Losada, A. (2014). Cohesin in cancer: chromosome segregation and beyond. Nat. Rev. Cancer *14*, 389–393.

Lupien, M., Eeckhoute, J., Meyer, C.A., Wang, Q., Zhang, Y., Li, W., Carroll, J.S., Liu, X.S., and Brown, M. (2008). FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell *132*, 958–970.

Mahapatra, S., Klee, E.W., Young, C.Y.F., Sun, Z., Jimenez, R.E., Klee, G.G., Tindall, D.J., and Donkena, K.V. (2012). Global methylation profiling for risk prediction of prostate cancer. Clin. Cancer Res. *18*, 2882–2895.

Massie, C.E., Lynch, A., Ramos-Montoya, A., Boren, J., Stark, R., Fazli, L., Warren, A., Scott, H., Madhu, B., Sharma, N., et al. (2011). The androgen receptor fuels prostate cancer by regulating central metabolism and biosynthesis. EMBO J. *30*, 2719–2733.

Odero-Marah, V., Hawsawi, O., Henderson, V., and Sweeney, J. (2018). Epithelial-Mesenchymal Transition (EMT) and Prostate Cancer. In Cell & Molecular Biology of Prostate Cancer: Updates, Insights and New Frontiers, H. Schatten, ed. (Cham: Springer International Publishing), pp. 101–110.

Parolia, A., Cieslik, M., Chu, S.-C., Xiao, L., Ouchi, T., Zhang, Y., Wang, X., Vats, P., Cao, X., Pitchiaya, S., et al. (2019). Distinct structural classes of activating FOXA1 alterations in advanced prostate cancer. Nature *571*, 413–418.

Peng, L., Bian, X.W., Li, D.K., Xu, C., Wang, G.M., Xia, Q.Y., and Xiong, Q. (2015). Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types. Sci. Rep. *5*, 13413.

Pombo, A., and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. Nat. Rev. Mol. Cell Biol. *16*, 245–257.

Pomerantz, M.M., Li, F., Takeda, D.Y., Lenci, R., Chonkar, A., Chabot, M., Cejas, P., Vazquez, F., Cook, J., Shivdasani, R.A., et al. (2015). The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. Nat. Genet. *47*, 1346–1351.

Pomerantz, M.M., Qiu, X., Zhu, Y., Takeda, D.Y., Pan, W., Baca, S.C., Gusev, A., Korthauer, K.D., Severson, T.M., Ha, G., et al. (2020). Prostate cancer reactivates developmental epigenomic programs during metastatic progression. Nat. Genet.

Quigley, D.A., Dang, H.X., Zhao, S.G., Lloyd, P., Aggarwal, R., Alumkal, J.J., Foye, A., Kothari, V., Perry, M.D., Bailey, A.M., et al. (2018). Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. Cell *175*, 889.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rajbhandari, P., Thomas, B.J., Feng, A.-C., Hong, C., Wang, J., Vergnes, L., Sallam, T., Wang, B., Sandhu, J., Seldin, M.M., et al. (2018). IL-10 Signaling Remodels Adipose Chromatin Architecture to Limit Thermogenesis and Energy Expenditure. Cell *172*, 218–233.e17.

Rickman, D.S., Soong, T.D., Moss, B., Mosquera, J.M., Dlabal, J., Terry, S., MacDonald, T.Y., Tripodi, J., Bunting, K., Najfeld, V., et al. (2012). Oncogene-mediated alterations in chromatin conformation. Proc. Natl. Acad. Sci. U. S. A. *109*, 9083–9088.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Robinson, D., Van Allen, E.M., Wu, Y.-M., Schultz, N., Lonigro, R.J., Mosquera, J.-M., Montgomery, B., Taplin, M.-E., Pritchard, C.C., Attard, G., et al. (2015). Integrative clinical genomics of advanced prostate cancer. Cell *161*, 1215–1228.

Rodriguez-Bravo, V., Carceles-Cordon, M., Hoshida, Y., Cordon-Cardo, C., Galsky, M.D., and Domingo-Domenech, J. (2017). The role of GATA2 in lethal prostate cancer aggressiveness. Nat. Rev. Urol. *14*, 38–48.

Rowley, M.J., Jordan Rowley, M., and Corces, V.G. (2018). Organizational principles of 3D genome architecture. Nature Reviews Genetics *19*, 789–800.

Sahu, B., Laakso, M., Ovaska, K., Mirtti, T., Lundin, J., Rannikko, A., Sankila, A., Turunen, J.-P., Lundin, M., Konsti, J., et al. (2011). Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. EMBO J. *30*, 3962–3976.

Sandoval, G.J., Pulice, J.L., Pakula, H., Schenone, M., Takeda, D.Y., Pop, M., Boulay, G., Williamson, K.E., McBride, M.J., Pan, J., et al. (2018). Binding of TMPRSS2-ERG to BAF Chromatin Remodeling Complexes Mediates Prostate Oncogenesis. Mol. Cell *71*, 554–566.e7.

Scharer, C.D., Barwick, B.G., Guo, M., Bally, A.P.R., and Boss, J.M. (2018). Plasma cell differentiation is controlled by multiple cell division-coupled epigenetic programs. Nat. Commun. *9*, 1698.

Sharma, N.L., Massie, C.E., Ramos-Montoya, A., Zecchini, V., Scott, H.E., Lamb, A.D., MacArthur, S., Stark, R., Warren, A.Y., Mills, I.G., et al. (2013). The androgen receptor induces a distinct transcriptional program in castration-resistant prostate cancer in man. Cancer Cell *23*, 35–47.

Siegel, R.L., Miller, K.D., and Jemal, A. (2018). Cancer statistics, 2018. CA: A Cancer Journal for Clinicians *68*, 7–30.

Sinha, A., Huang, V., Livingstone, J., Wang, J., Fox, N.S., Kurganovs, N., Ignatchenko, V., Fritsch, K., Donmez, N., Heisler, L.E., et al. (2019). The Proteogenomic Landscape of Curable Prostate Cancer. Cancer Cell *35*, 414–427.e6.

Stelloo, S., Nevedomskaya, E., Kim, Y., Schuurman, K., Valle-Encinas, E., Lobo, J., Krijgsman, O., Peeper, D.S., Chang, S.L., Feng, F.Y.-C., et al. (2018). Integrative epigenetic taxonomy of primary prostate cancer. Nat. Commun. *9*, 4900.

Taberlay, P.C., Achinger-Kawecka, J., Lun, A.T.L., Buske, F.A., Sabir, K., Gould, C.M., Zotenko, E., Bert, S.A., Giles, K.A., Bauer, D.C., et al. (2016). Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. Genome Res. *26*, 719–731.

Takeda, D.Y., Spisák, S., Seo, J.-H., Bell, C., O'Connor, E., Korthauer, K., Ribli, D., Csabai, I., Solymosi, N., Szállási, Z., et al. (2018). A Somatically Acquired Enhancer of the Androgen Receptor Is a Noncoding Driver in Advanced Prostate Cancer. Cell *174*, 422–432.e13.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Toenhake, C.G., Fraschka, S.A.-K., Vijayabaskar, M.S., Westhead, D.R., van Heeringen, S.J., and Bártfai, R. (2018). Chromatin Accessibility-Based Characterization of the Gene Regulatory Network Underlying Plasmodium falciparum Blood-Stage Development. Cell Host Microbe *23*, 557–569.e9.

Urbanucci, A., Sahu, B., Seppälä, J., Larjo, A., Latonen, L.M., Waltering, K.K., Tammela, T.L.J., Vessella, R.L., Lähdesmäki, H., Jänne, O.A., et al. (2012). Overexpression of androgen receptor enhances the binding of the receptor to the chromatin in prostate cancer. Oncogene *31*, 2153–2163.

Urbanucci, A., Barfeld, S.J., Kytölä, V., Itkonen, H.M., Coleman, I.M., Vodák, D., Sjöblom, L., Sheng, X., Tolonen, T., Minner, S., et al. (2017). Androgen Receptor Deregulation Drives Bromodomain-Mediated Chromatin Alterations in Prostate Cancer. Cell Rep. *19*, 2045–2059.

Varambally, S., Dhanasekaran, S.M., Zhou, M., Barrette, T.R., Kumar-Sinha, C., Sanda, M.G., Ghosh, D., Pienta, K.J., Sewalt, R.G.A.B., Otte, A.P., et al. (2002). The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature *419*, 624–629.

Viré, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.-M., et al. (2006). The Polycomb group protein EZH2 directly controls DNA methylation. Nature *439*, 871–874.

Viswanathan, S.R., Ha, G., Hoff, A.M., Wala, J.A., Carrot-Zhang, J., Whelan, C.W., Haradhvala, N.J., Freeman, S.S., Reed, S.C., Rhoades, J., et al. (2018). Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. Cell *174*, 433–447.e19.

Wang, Q., Li, W., Zhang, Y., Yuan, X., Xu, K., Yu, J., Chen, Z., Beroukhim, R., Wang, H., Lupien, M., et al. (2009). Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. Cell *138*, 245–256.

Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M.N.K., Li, Y., Hu, M., et al. (2018). The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biology *19*.

Weischenfeldt, J., Dubash, T., Drainas, A.P., Mardin, B.R., Chen, Y., Stütz, A.M., Waszak, S.M., Bosco, G., Halvorsen, A.R., Raeder, B., et al. (2017). Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nat. Genet. *49*, 65–74.

Wu, J., Xu, J., Liu, B., Yao, G., Wang, P., Lin, Z., Huang, B., Wang, X., Li, T., Shi, S., et al. (2018). Chromatin analysis in human early development reveals epigenetic transition during ZGA. Nature *557*, 256–260.

Xu, K., Wu, Z.J., Groner, A.C., He, H.H., Cai, C., Lis, R.T., Wu, X., Stack, E.C., Loda, M., Liu, T., et al. (2012). EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent. Science *338*, 1465–1469.

Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y., and Kolpakov, F. (2019). GTRD: a database on gene transcription regulation—2019 update. Nucleic Acids Res. *47*, D100–D105.

Ylipää, A., Kivinummi, K., Kohvakka, A., Annala, M., Latonen, L., Scaravilli, M., Kartasalo, K., Leppänen, S.-P., Karakurt, S., Seppälä, J., et al. (2015). Transcriptome Sequencing Reveals PCAT5 as a Novel ERG-Regulated Long Noncoding RNA in Prostate Cancer. Cancer Res. *75*, 4026–4031.

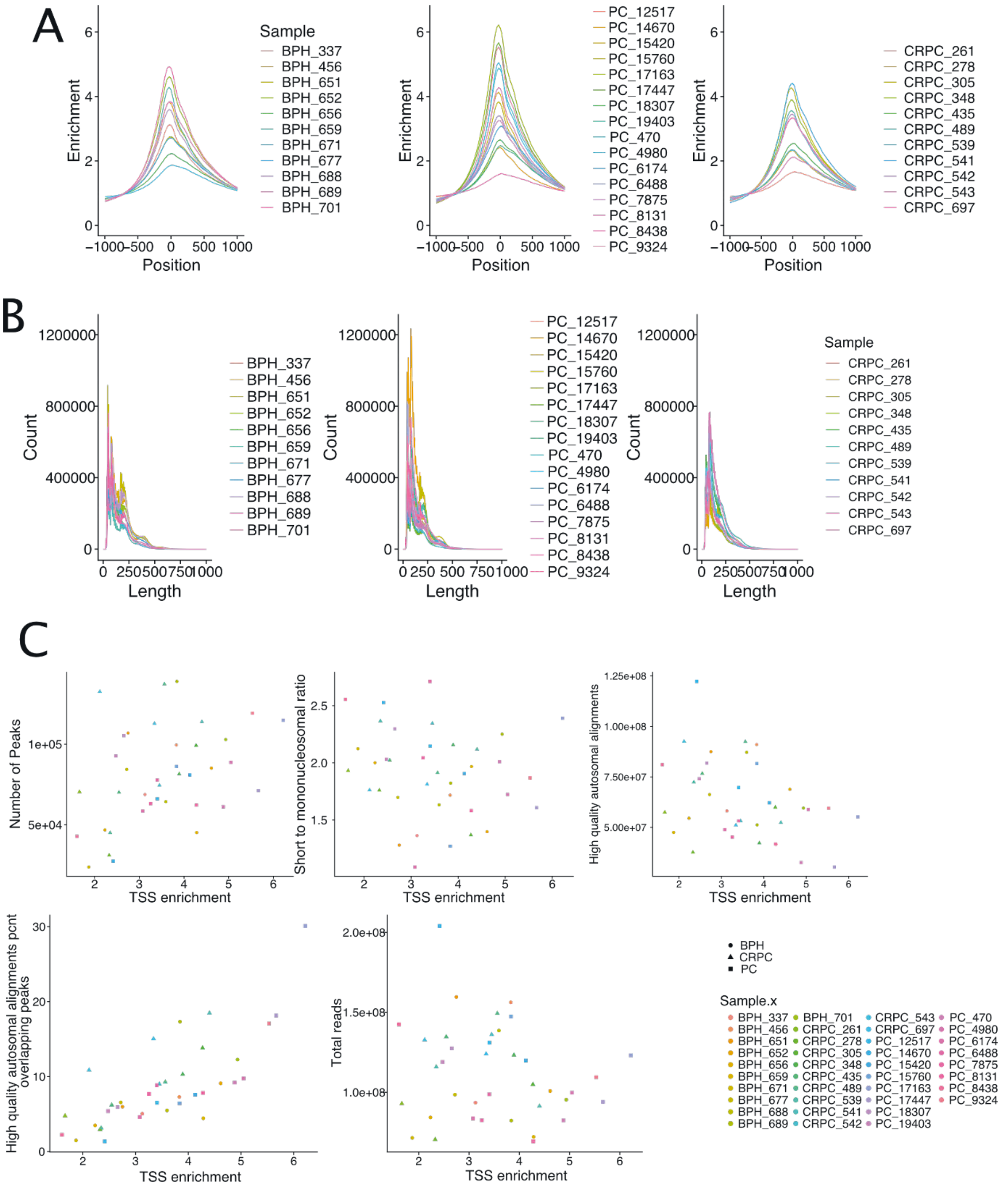
Yu, J., Yu, J., Mani, R.-S., Cao, Q., Brenner, C.J., Cao, X., Wang, X., Wu, L., Li, J., Hu, M., et al. (2010). An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. Cancer Cell *17*, 443–454.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhao, S.G., Chen, W.S., Li, H., Foye, A., Zhang, M., Sjöström, M., Aggarwal, R., Playdle, D., Liao, A., Alumkal, J.J., et al. (2020). The DNA methylation landscape of advanced prostate cancer. Nat. Genet. *52*, 778–789.

Zhou, Y., Huang, T., Cheng, A.S.L., Yu, J., Kang, W., and To, K.F. (2016). The TEAD Family and Its Oncogenic Role in Promoting Tumorigenesis. Int. J. Mol. Sci. *17*, 138.

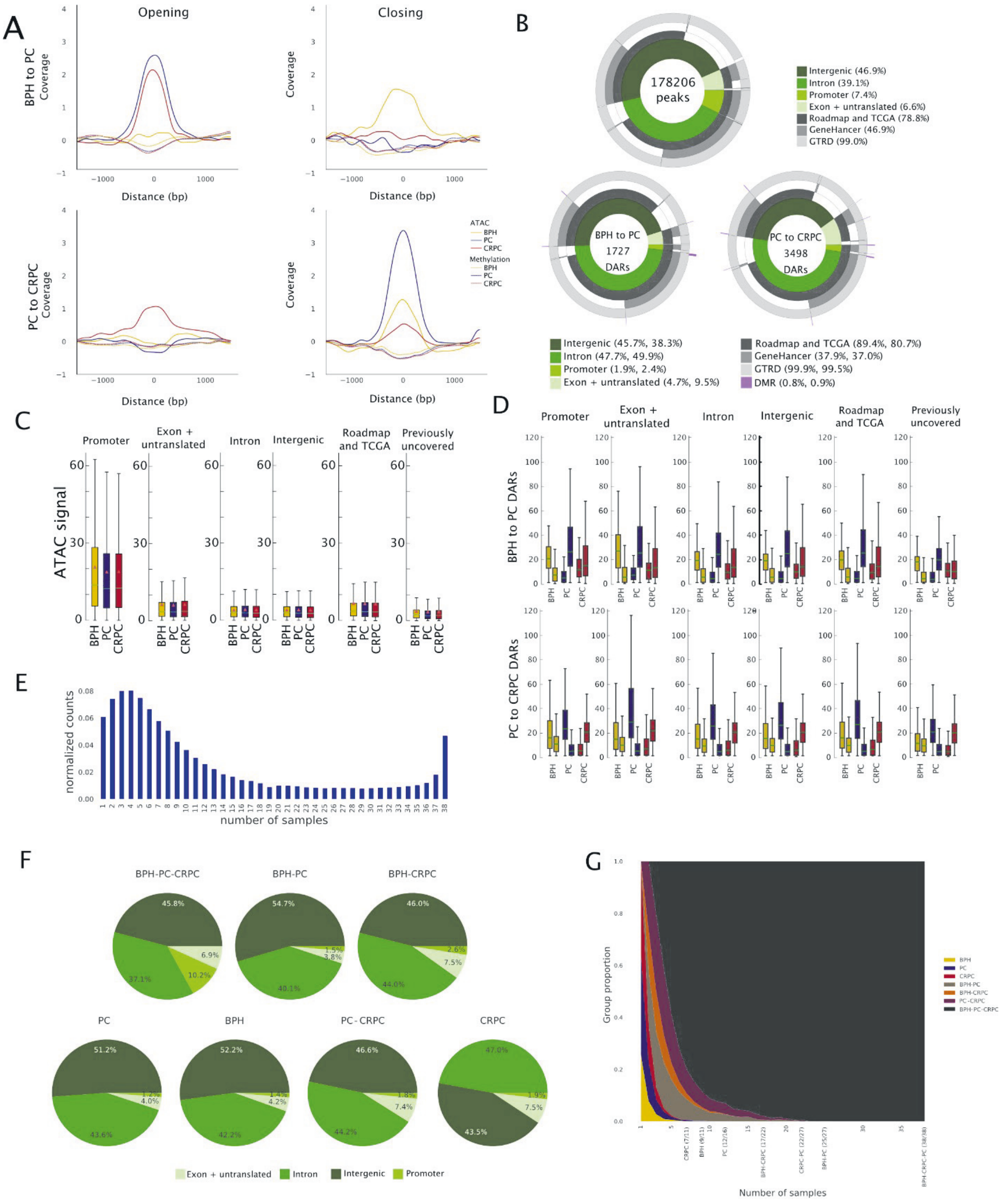
Supplementary Figures



Supplementary Figure 1

Figure S1

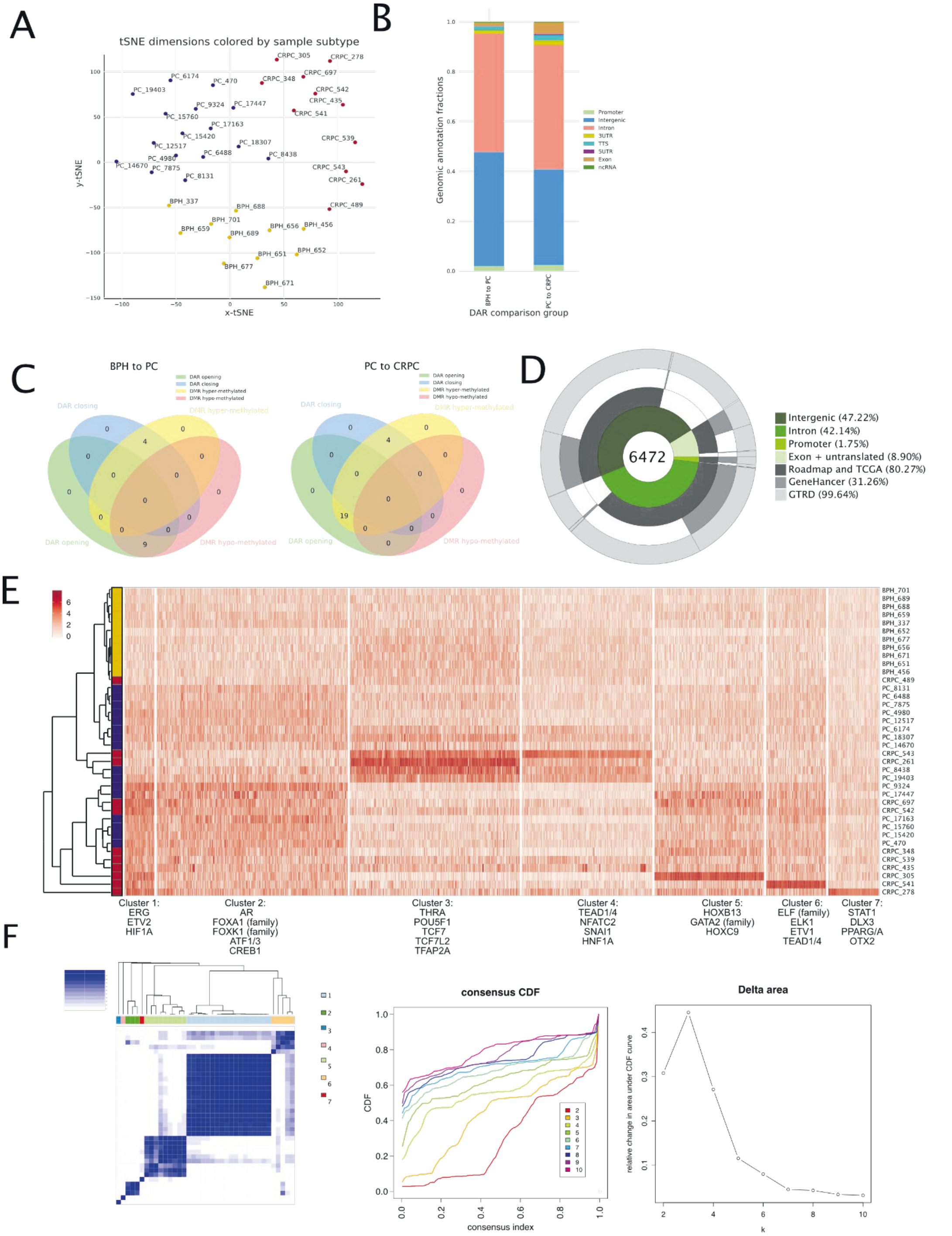
A. ATAC-seq enrichments at TSS for all samples. Sample groups have comparable quality and variability. **B.** Fragment length patterns in all samples show the first peak around 160 bp, matching single nucleosome size. A second peak is observed for the second nucleosome. **C.** Correlations between TSS enrichment and several key quality control values. The “high-quality autosomal alignments percentage overlapping peaks” value correlates with TSS enrichments, as expected. These show that no bias was introduced by the sequencing step.



Supplementary Figure 2

Figure S2

A. Background-corrected ATAC-seq and MeDIP coverage for loci corresponding to opening (left column) and closing (right column) DARs for BPH to PC (top row) and PC to CRPC (bottom row) comparisons. Each curve's baseline has been shifted to zero for presentation clarity. The curves have been smoothed with a Gaussian filter with a standard deviation set to 7. **B.** Donut plots for locations of peaks and DARs from both comparisons. Majority of peaks and DARs are located in intergenic and intronic regions but there is a clear difference in promoter regions where ~7.5% of peaks are located compared to ~2% in both DAR comparisons. Overlaps with DMR regions are shown in the DAR donut plots. **C.** Peaks ATAC-signal in different annotation categories. Strongest ATAC-seq signal is detected at the promoters. **D.** Comparable ATAC-seq signal across different genome annotation areas. For each sample group, the left boxplot shows the signal in closing and right boxplot in opening DARs from respective comparisons. Data from the group of samples that were not part of the comparison (CRPC in top panels, BPH in lower panels) are shown from the same loci for reference. **E.** Normalized peak counts present in different numbers of samples. **F.** Genomic locations of peaks belonging to each sample group or combination of sample groups. Peaks belonging to the set with all sample groups have over 10% of peaks annotated to promoters, whereas in other groups the promoter fraction is 1.2-2.6%. **G.** Number of samples reporting peaks by group or combination of groups membership. The number of samples in which a peak is present is shown on the X-axis. Labels for sample groups and sample group combinations are reported. In addition, points where the number of peaks in groups or group combinations go to zero are also shown. The proportion of peaks in each group is shown on the Y-axis.



Supplementary Figure 3

Figure S3

A. Two-dimensional t-SNE dimension reduction using normalized ATAC-seq signal from all DARs separates samples to their respective groups. Python's Scikit-Learn package t-SNE algorithm implementation was used with default parameter values except *perplexity=15, metric = Pearson correlation,* and *method = exact*. **B.** Annotation of different genomic locations for DARs detected in BPH to PC and PC to CRPC comparisons. A higher fraction of DARs are located in the gene body and near the gene body in PC to CRPC compared to BPH to PC. **C.** Number of overlaps between DARs and DMRs in BPH to PC and PC to CRPC comparisons show only minimal overlap. **D.** Donut plot showing genomic annotations of cancer-specific peaks and overlap with previously reported features. **E.** K-means consensus clustering of the 20% topmost peaks with the highest variance identifies 7 clusters. Scale bar indicates quantification value. Examples of disease-relevant TFs from TF binding site enrichment analysis are shown for each cluster. **F.** Selection criteria for K=7 clusters in the cluster analysis. Consensus matrix (K=7, left), cumulative distribution function (CDF) plots for K=2-10 (middle), and relative change in CDF (right) are shown. K=7 illustrates stable cluster structure with relative CDF change at elbow point of the curve.

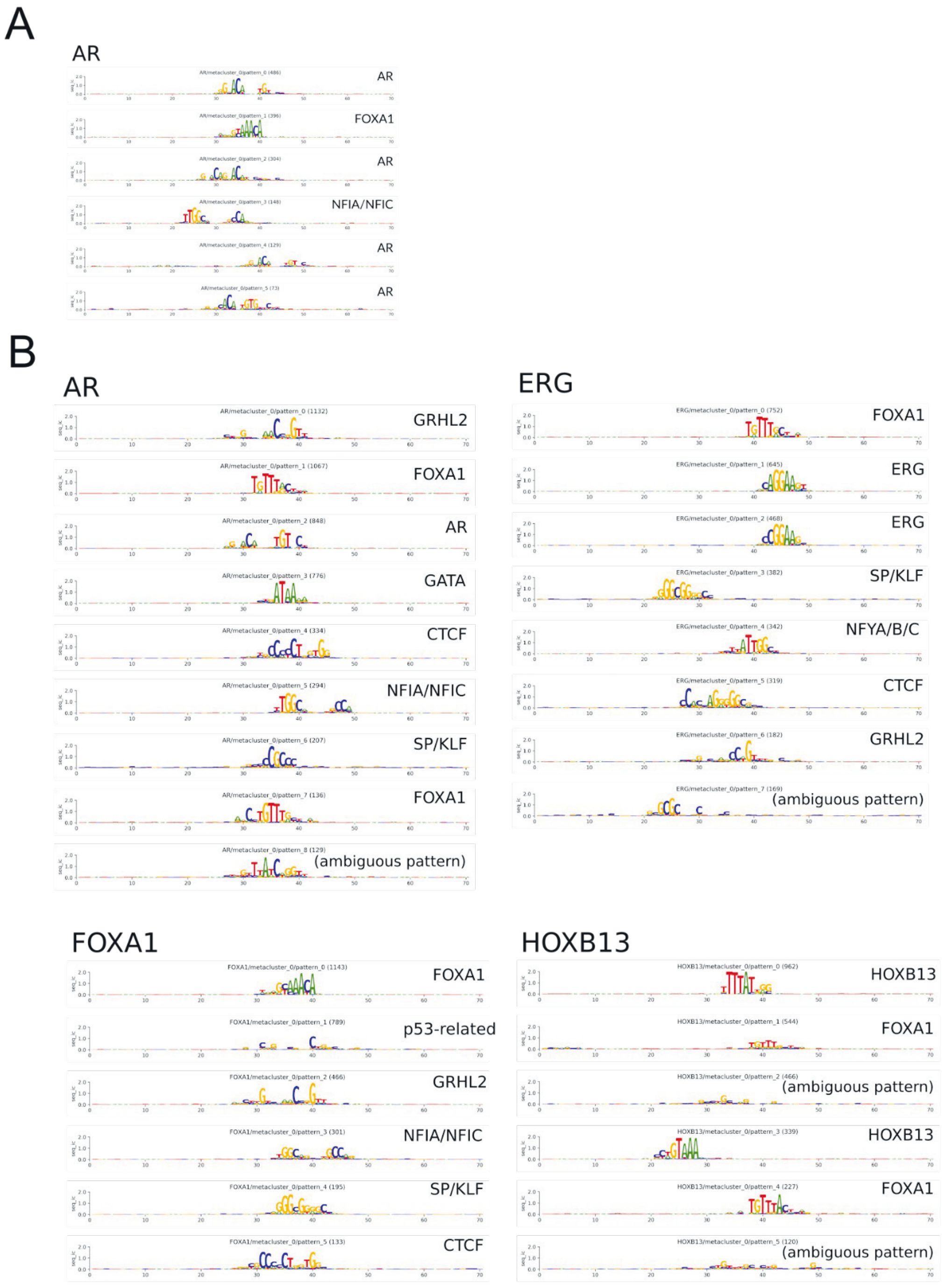

A
AR
AR/metacluster_0/pattern_0 (486) AR
AR/metacluster_0/pattern_1 (396) FOXA1
AR/metacluster_0/pattern_2 (304) AR
AR/metacluster_0/pattern_3 (148) NFIA/NFIC
AR/metacluster_0/pattern_4 (129) AR
AR/metacluster_0/pattern_5 (73) AR
seq_ic
B
AR
AR/metacluster_0/pattern_0 (1132) GRHL2
AR/metacluster_0/pattern_1 (1067) FOXA1
AR/metacluster_0/pattern_2 (848) AR
AR/metacluster_0/pattern_3 (776) GATA
AR/metacluster_0/pattern_4 (334) CTCF
AR/metacluster_0/pattern_5 (294) NFIA/NFIC
AR/metacluster_0/pattern_6 (207) SP/KLF
AR/metacluster_0/pattern_7 (136) FOXA1
AR/metacluster_0/pattern_8 (129) (ambiguous pattern)
ERG
ERG/metacluster_0/pattern_0 (752) FOXA1
ERG/metacluster_0/pattern_1 (645) ERG
ERG/metacluster_0/pattern_2 (468) ERG
ERG/metacluster_0/pattern_3 (382) SP/KLF
ERG/metacluster_0/pattern_4 (342) NFYA/B/C
ERG/metacluster_0/pattern_5 (319) CTCF
ERG/metacluster_0/pattern_6 (182) GRHL2
ERG/metacluster_0/pattern_7 (169) (ambiguous pattern)
FOXA1
FOXA1/metacluster_0/pattern_0 (1143) FOXA1
FOXA1/metacluster_0/pattern_1 (789) p53-related
FOXA1/metacluster_0/pattern_2 (466) GRHL2
FOXA1/metacluster_0/pattern_3 (301) NFIA/NFIC
FOXA1/metacluster_0/pattern_4 (195) SP/KLF
FOXA1/metacluster_0/pattern_5 (133) CTCF
HOXB13
HOXB13/metacluster_0/pattern_0 (962) HOXB13
HOXB13/metacluster_0/pattern_1 (544) FOXA1
HOXB13/metacluster_0/pattern_2 (466) (ambiguous pattern)
HOXB13/metacluster_0/pattern_3 (339) HOXB13
HOXB13/metacluster_0/pattern_4 (227) FOXA1
HOXB13/metacluster_0/pattern_5 (120) (ambiguous pattern)

Figure S4

A. Binding motifs discovered with BPNET from AR ChIP-seq data generated from VCaP cell line. **B.** Discovered binding sites using ATAC-seq data from VCaP cells with binding sites for AR, FOXA1, HOXB13, and ERG.

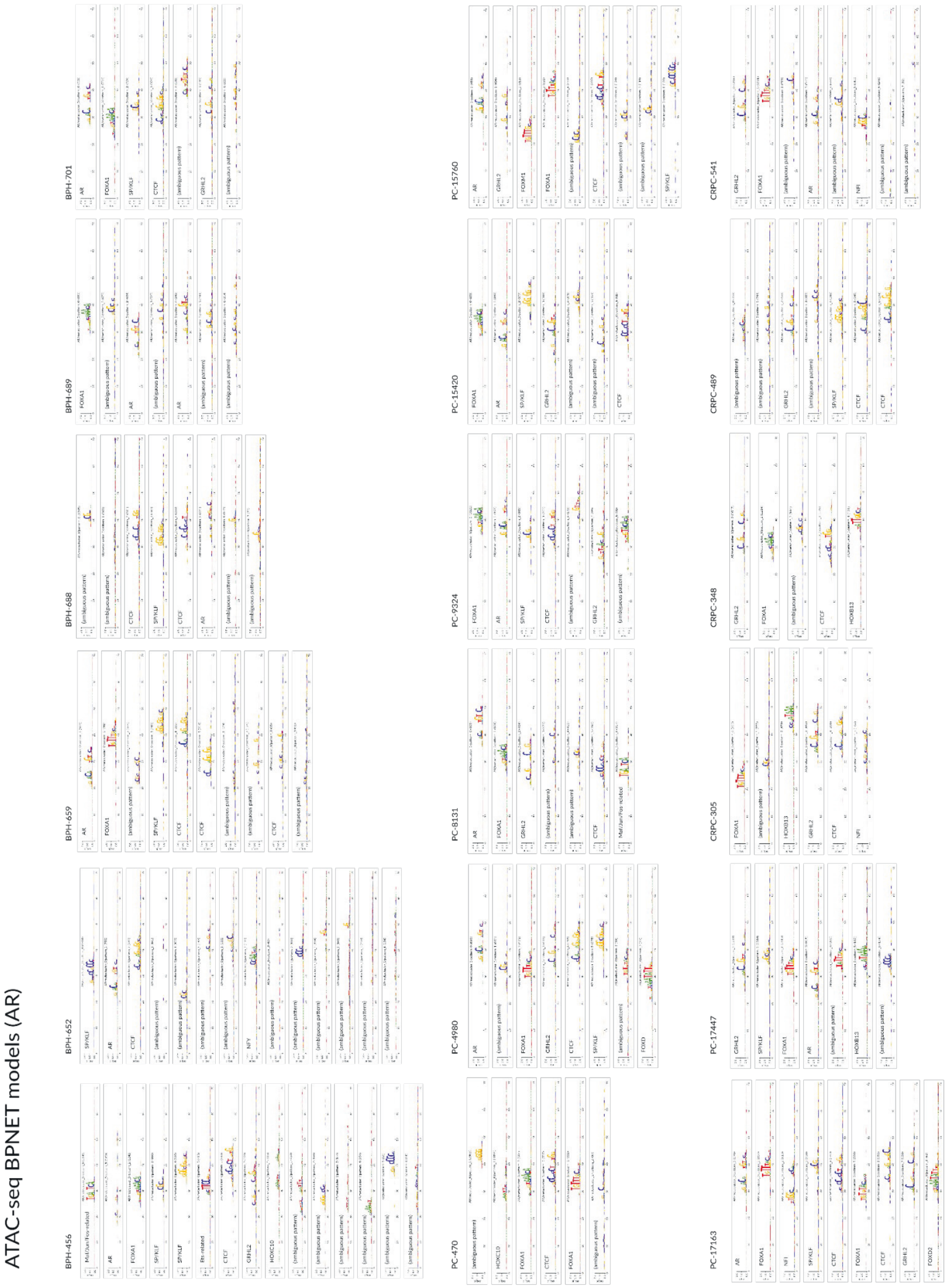

ATAC-seq BPNET models (AR)
BPH-701
AR
FOXA1
SP/KLF
CTCF
(ambiguous pattern)
GRHL2
(ambiguous pattern)
BPH-689
FOXA1
(ambiguous pattern)
AR
(ambiguous pattern)
AR
(ambiguous pattern)
(ambiguous pattern)
BPH-688
(ambiguous pattern)
(ambiguous pattern)
CTCF
SP/KLF
CTCF
AR
(ambiguous pattern)
(ambiguous pattern)
BPH-659
AR
FOXA1
(ambiguous pattern)
SP/KLF
CTCF
CTCF
(ambiguous pattern)
(ambiguous pattern)
CTCF
(ambiguous pattern)
BPH-652
SP/KLF
AR
CTCF
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
NFY
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
BPH-456
Maf/Jun/Fos-related
AR
FOXA1
SP/KLF
SP/KLF
Ets-related
CTCF
GRHL2
HOXC10
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
PC-15760
AR
GRHL2
FOXM1
FOXA1
(ambiguous pattern)
CTCF
(ambiguous pattern)
(ambiguous pattern)
SP/KLF
PC-15420
FOXA1
AR
SP/KLF
GRHL2
(ambiguous pattern)
(ambiguous pattern)
CTCF
PC-9324
FOXA1
AR
SP/KLF
CTCF
(ambiguous pattern)
GRHL2
(ambiguous pattern)
PC-8131
AR
FOXA1
GRHL2
(ambiguous pattern)
(ambiguous pattern)
CTCF
Maf/Jun/Fos related
PC-4980
AR
(ambiguous pattern)
FOXA1
GRHL2
CTCF
SP/KLF
(ambiguous pattern)
FOXD
PC-470
(ambiguous pattern)
HOXC10
FOXA1
CTCF
FOXA1
(ambiguous pattern)
CRPC-541
GRHL2
FOXA1
(ambiguous pattern)
AR
(ambiguous pattern)
NFI
(ambiguous pattern)
(ambiguous pattern)
CRPC-489
(ambiguous pattern)
(ambiguous pattern)
GRHL2
(ambiguous pattern)
SP/KLF
CTCF
CTCF
CRPC-348
GRHL2
FOXA1
(ambiguous pattern)
CTCF
HOXB13
CRPC-305
FOXA1
(ambiguous pattern)
HOXB13
GRHL2
CTCF
NFI
PC-17447
GRHL2
SP/KLF
FOXA1
AR
(ambiguous pattern)
HOXB13
(ambiguous pattern)
PC-17163
AR
FOXA1
NFI
SP/KLF
CTCF
FOXA1
CTCF
GRHL2
FOXD2

Figure S5

Binding motifs discovered with BPNET from clinical ATAC-seq samples using AR binding sites overlapping with peaks from ATAC-seq data. For each pattern, the number of BPNET *seqlets* contributing to that pattern is reported in parenthesis (see **Methods**).

ATAC-seq BPNET models (FOXA1)

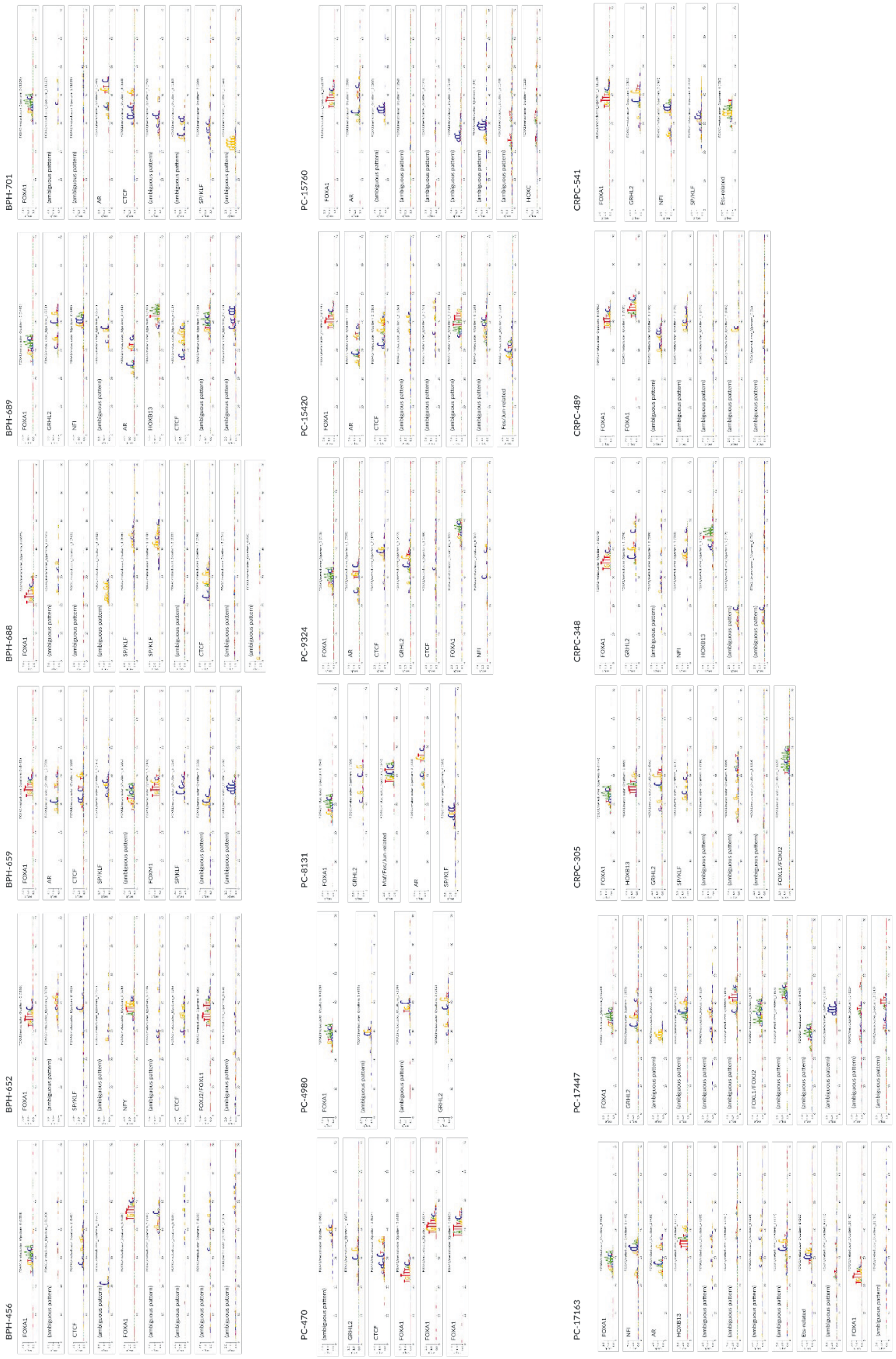


Figure S6

Binding motifs discovered with BPNET from clinical ATAC-seq samples using FOXA1 binding sites overlapping with peaks from ATAC-seq data. For each pattern, the number of BPNET *seqlets* contributing to that pattern is reported in parenthesis (see **Methods**).

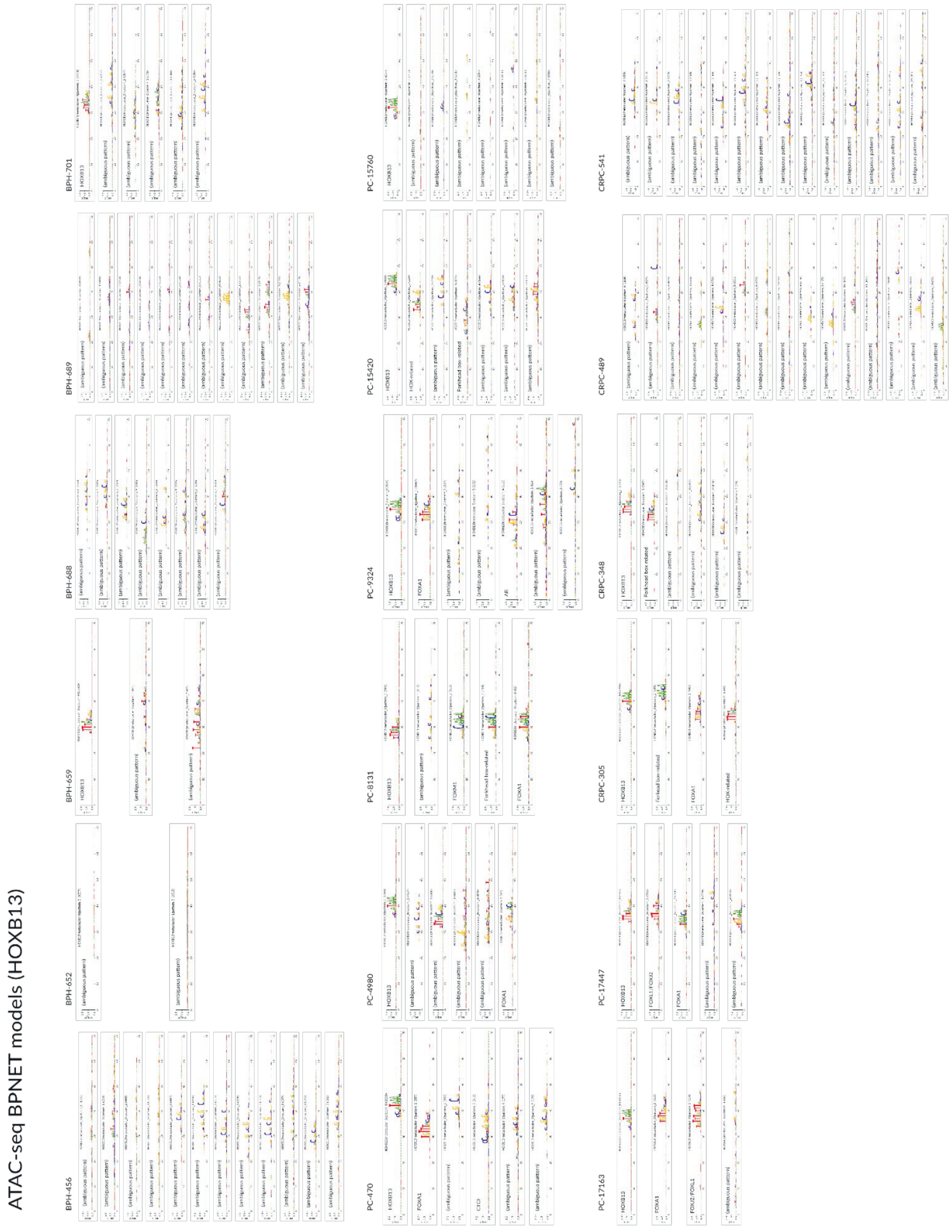

ATAC-seq BPNET models (HOXB13)
BPH-701
BPH-689
BPH-688
BPH-659
BPH-652
BPH-456
PC-15760
PC-15420
PC-9324
PC-8131
PC-4980
PC-470
CRPC-541
CRPC-489
CRPC-348
CRPC-305
PC-17447
PC-17163
HOXB13
FOXA1
FOXM1
FOXL1 FOXJ2
FOXJ2/FOXL1
CTCF
AR
HOX-related
Forkhead box-related
(ambiguous pattern)

Figure S7

Binding motifs discovered with BPNET from clinical ATAC-seq samples using HOXB13 binding sites overlapping with peaks from ATAC-seq data. For each pattern, the number of BPNET *seqlets* contributing to that pattern is reported in parenthesis (see **Methods**).

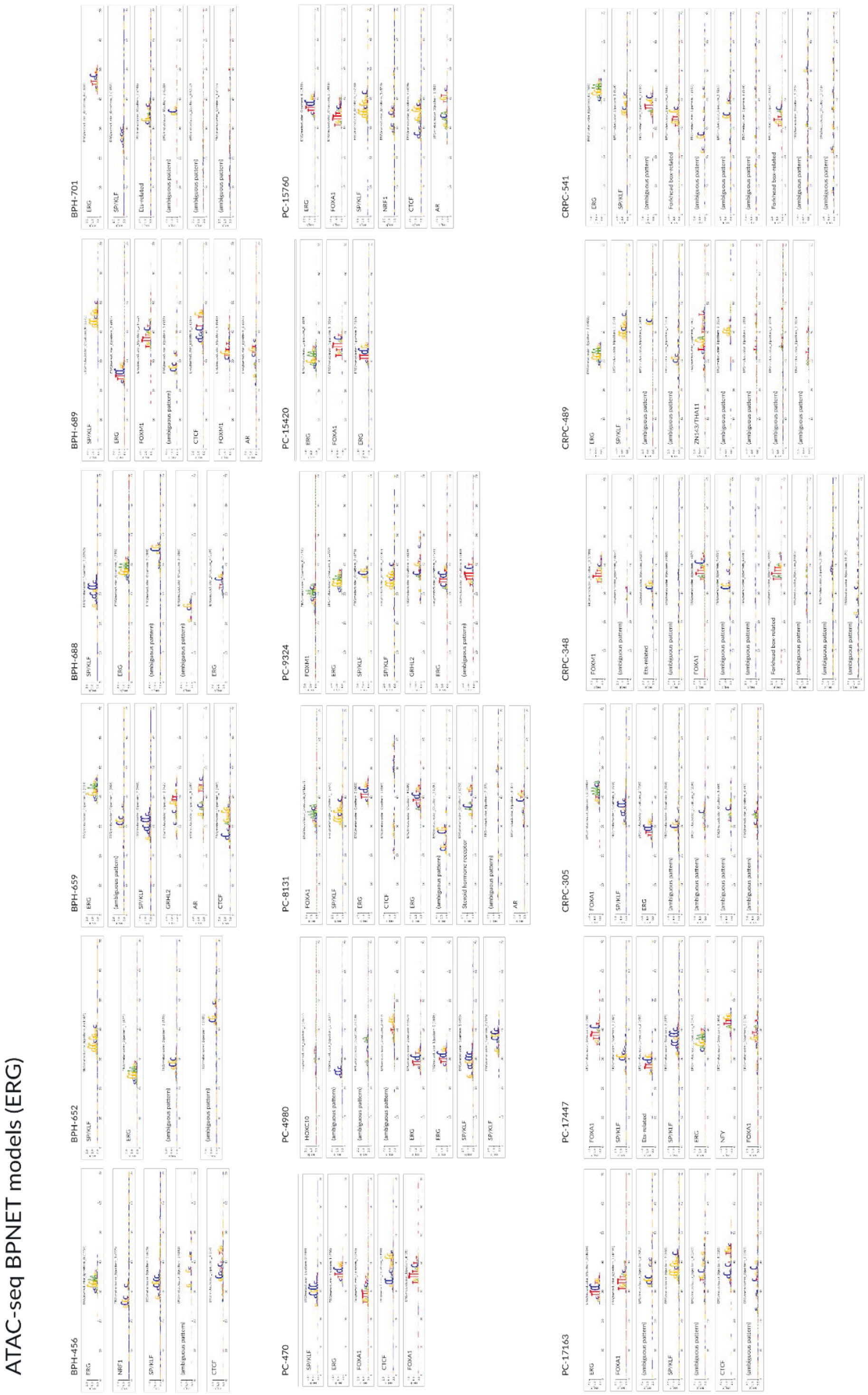
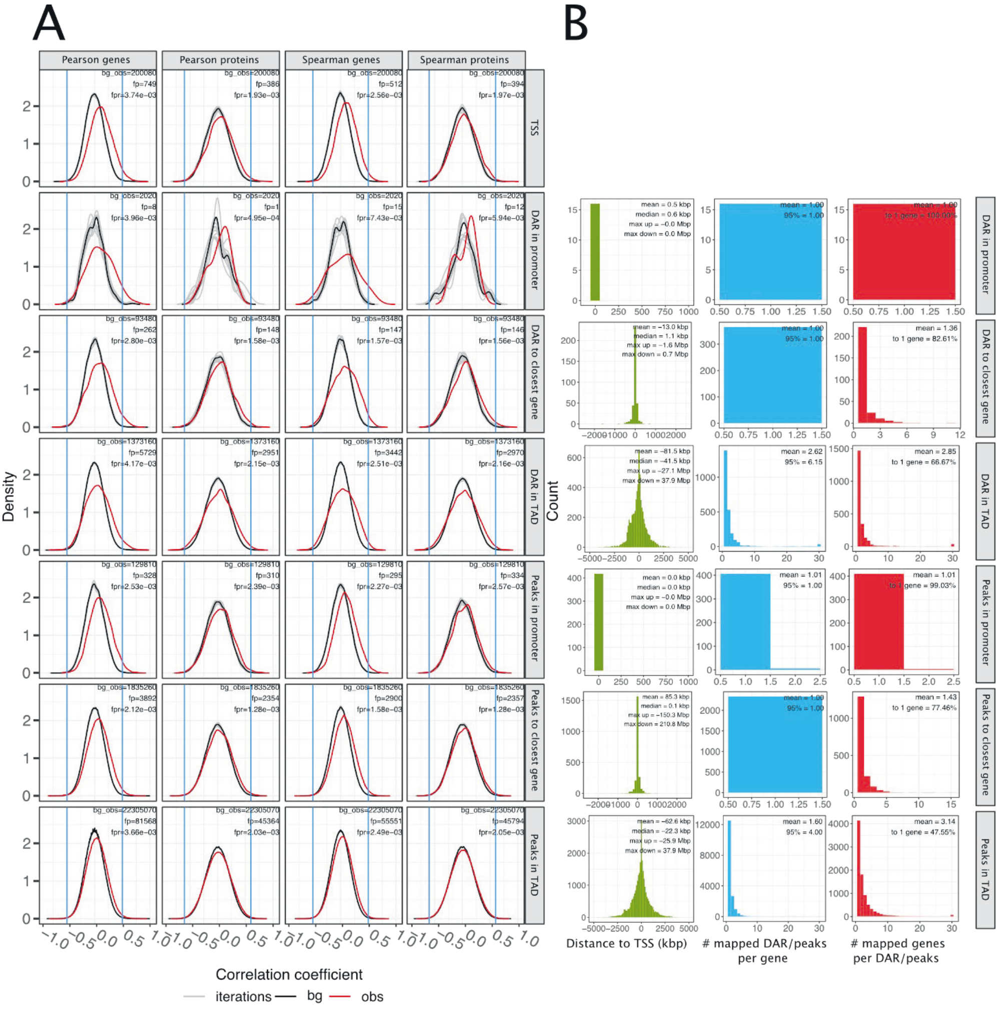

ATAC-seq BPNET models (ERG)
BPH-701
ERG
SP/KLF
Ets-related
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
BPH-689
SP/KLF
ERG
FOXM1
(ambiguous pattern)
CTCF
FOXM1
AR
BPH-688
SP/KLF
ERG
(ambiguous pattern)
(ambiguous pattern)
ERG
BPH-659
ERG
(ambiguous pattern)
SP/KLF
GRHL2
AR
CTCF
BPH-652
SP/KLF
ERG
(ambiguous pattern)
(ambiguous pattern)
BPH-456
ERG
NRF1
SP/KLF
(ambiguous pattern)
CTCF
PC-15760
ERG
FOXA1
SP/KLF
NRF1
CTCF
AR
PC-15420
ERG
FOXA1
ERG
PC-9324
FOXM1
ERG
SP/KLF
SP/KLF
GRHL2
ERG
(ambiguous pattern)
PC-8131
FOXA1
SP/KLF
ERG
CTCF
ERG
(ambiguous pattern)
Steroid hormone receptor
(ambiguous pattern)
AR
PC-4980
HOXC10
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
ERG
ERG
SP/KLF
SP/KLF
PC-470
SP/KLF
ERG
FOXA1
CTCF
FOXA1
CRPC-541
ERG
SP/KLF
(ambiguous pattern)
Forkhead box-related
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
Forkhead box-related
(ambiguous pattern)
(ambiguous pattern)
CRPC-489
ERG
SP/KLF
(ambiguous pattern)
(ambiguous pattern)
ZNF143/THAP11
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
CRPC-348
FOXM1
(ambiguous pattern)
Ets-related
(ambiguous pattern)
FOXA1
(ambiguous pattern)
(ambiguous pattern)
Forkhead box-related
(ambiguous pattern)
(ambiguous pattern)
CRPC-305
FOXA1
SP/KLF
ERG
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
(ambiguous pattern)
PC-17447
FOXA1
SP/KLF
Ets-related
SP/KLF
ERG
NFY
FOXA1
PC-17163
ERG
FOXA1
(ambiguous pattern)
SP/KLF
(ambiguous pattern)
CTCF
(ambiguous pattern)

Figure S8

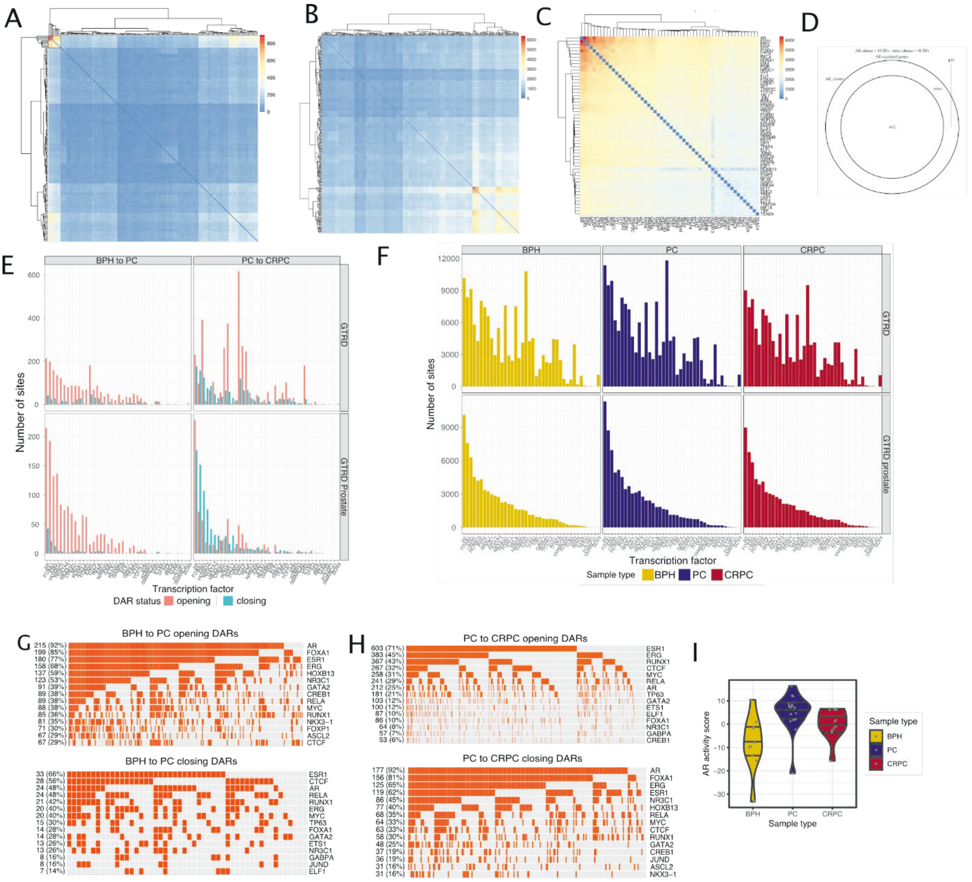
Binding motifs discovered with BPNET from clinical ATAC-seq samples using ERG binding sites overlapping with peaks from ATAC-seq data. For each pattern, the number of BPNET *seqlets* contributing to that pattern is reported in parenthesis (see **Methods**).



Supplementary Figure 9

Figure S9

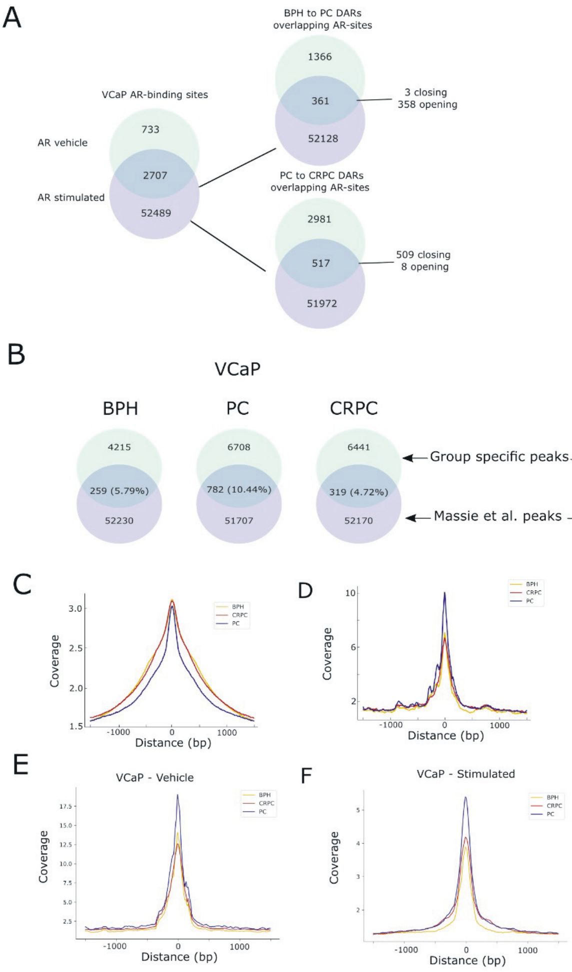
A. Pearson and Spearman correlation coefficients for all associations between chromatin accessibility and gene or protein expression. Data are shown for both gene (RNA-seq and smallRNA-seq) and protein expressions (SWATH-MS). Number of correlation coefficients used for null distribution, false positives and false positive ratio are reported in the inset. **B**. Histograms of distances between ATAC-seq features and TSS of the associated gene are shown across all comparisons. In addition, numbers of peaks/DARs associated to a given gene and also the number of genes associated to a given peak/DAR are shown as histograms (truncated at 30). Mean of these histograms is given in the figure. In addition, 95th percentile and fraction of peaks/DARs linked to a single gene are given in respective comparisons. Mean, median and maximum upstream (max up) and downstream (max down) distances are reported for the distance distribution. The percentage of ATAC-features linked to exactly one gene is also reported for the right panel.



Supplementary Figure 10

Figure S10

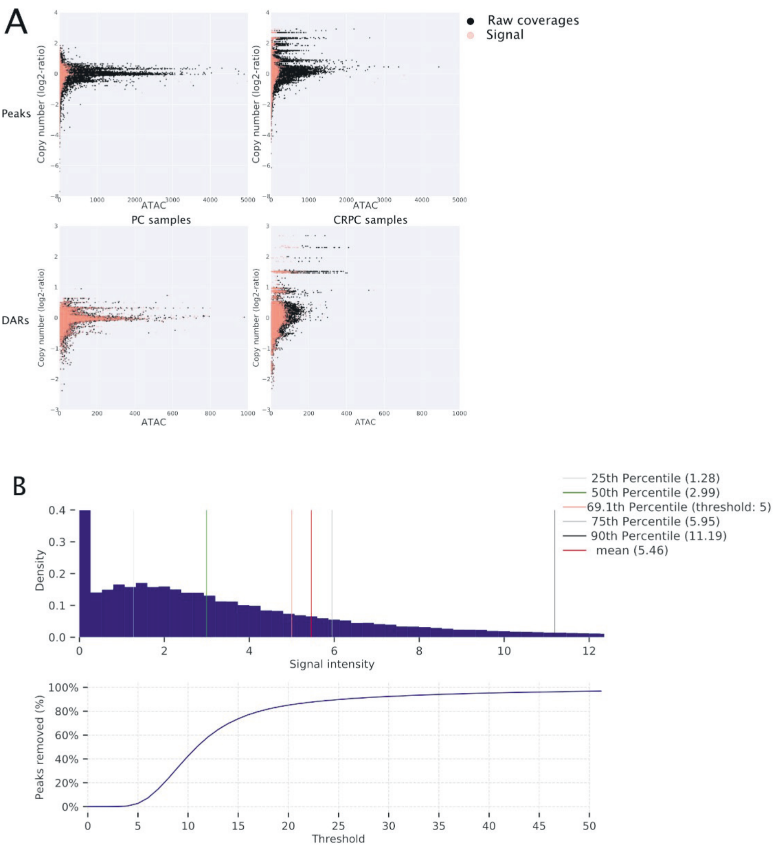
A. The heatmap represents a gene expression regulatory network created by using DARs and genes that have correlation: rows and columns are transcription factors (nodes), each cell in the matrix represents an edge, the weight of the edge is given by the number of shared genes which is encoded in the color. Rows and columns are filtered to have at least one cell with a value greater than 100. **B.** Same as panel A but using peaks. **C.** Subset of TFs with the highest number of genes (from data shown in panel B). The highest number of genes can be seen in the top right corner where there are the same four TFs as in **Figure 4B**. **D.** AR cluster-regulated genes from C are a superset of the genes regulated by the other cluster of TFs. The Venn diagram reports the agreement between the sets of genes regulated by the two clusters. **E.** Number of TFs that have binding site in DARs with associated target genes. Data are shown using all the data from GTRD (top panels) and using only prostate cancer-specific subset (GTRD prostate; bottom panels). Shown are both BPH to PC (left panels) and PC to CRPC (right panels) comparisons. **F.** Number of TFs that have binding site in peaks with associated target genes. Data are shown using all the data from GTRD (top panels) and using only prostate cancer-specific subset (GTRD prostate; bottom panels). In the GTRD prostate, we see that several key TFs like AR, FOXA1, ERG and HOXB13 are among the most common ones. **G.** Oncoprints representing TF binding sites overlapping DARs correlated with gene expression in BPH to PC comparison using the complete GTRD dataset. Number of binding sites for each TF is shown. In parentheses, the percentage of DARs reporting that binding site is also shown. **H.** Same as panel G but for PC to CRPC comparison. **I.** Violin plots of AR activity scores for each sample group. Individual samples are shown as grey dots.



Supplementary Figure 11

Figure S11

A. Number of sites and their overlap from DHT- and vehicle-treated VCaP cells. Sites that are present in DHT-stimulated cells are compared to DARs. In BPH to PC, most of the sites overlap with opening DARs and in PC to CRPC with closing DARs. **B.** Comparison of BPH, PC, and CRPC group-specific peaks to AR-stimulated peaks from LNCaP and VCaP cell lines (Massie et al., 2011) shows the highest overlap with the PC group. **C.** Background-corrected ATAC-seq coverage of AR binding sites from all GTRD AR binding sites. **D.** Background-corrected ATAC-seq coverage of AR binding sites from vehicle-treated cells (Massie et al. 2011). **E.** Background-corrected ATAC-seq coverage at AR binding sites from vehicle-treated VCaP cells. **F.** Same as panel E but with DHT stimulation. Signal is stronger in PC samples than other sample groups.



Supplementary Figure 12

Figure S12

A. Effect of the background correction and normalization (Signal, red dots) in relation to raw ATAC data (black dots) and DNA copy number. Background correction and normalization successfully removes the linear relationship between copy number and ATAC coverage. **B.** Distribution of peak quantifications across samples. Different percentiles and the utilized threshold 5 are shown (top). The number of sites that would be removed with a given threshold (bottom).

Supplementary Table legends

Supplementary Table 1: Quality control metrics and peak detection results

Table contains information about samples, and relevant information from sequencing such as quality control metrics and primers used. In addition, it contains information about peaks and their clustering.

Supplementary Table 2: Differentially accessible and differentially methylated regions

Table contains information about differentially accessible and differentially methylated regions in different comparison groups.

Supplementary Table 3: TF binding analysis

Table contains information about transcription factor footprint analysis, correlation between footprint depth, flanking accessibility and gene expression as well as information about motifs discovered using BPNET.

Supplementary Table 4: Correlations of accessible chromatin regions and gene expression

Table contains information about peaks and DARs correlation coefficients computed against gene and protein expression in different biological contexts. Table also reports gene names of transcription factors with binding site overlapping peaks and DARs and basic annotations of correlated genes.