Analysis of Enhancer Activity Changes in Late-Stage Prostate Cancer

by

Sıla Akdoğan

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Molecular Biology and Genetics

July 7, 2025

Analysis of Enhancer Activity Changes in Late-Stage Prostate Cancer

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Sıla Akdoğan

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

Assoc. Prof. Nathan Lack

_____

Assist. Prof. Gözde Korkmaz

_____

Assist Prof. Burçak Otlu

_____

Date: 07.07.2025

Eşit, adil ve aydınlık yarınlara...

ABSTRACT

Analysis of Enhancer Activity Changes in Late-Stage Prostate Cancer

Sıla Akdoğan

Master of Science in Molecular Biology and Genetics

July 7, 2025

Prostate cancer is one of the most common cancer types in Turkish men. At nearly all stages of the disease, the androgen receptor (AR) is the main driver of initiation and growth. When activated, the AR binds to enhancer cis-regulatory elements and induces gene expression through chromatin loops that connect enhancers with target promoters. Given its critical function, targeting AR activity is the standard of care to treat metastatic or recurrent prostate cancer. While treatment is initially effective, resistance inevitably occurs. Yet even in these resistant tumors, AR remains the critical driver of proliferation in the vast majority of patients. There is increasing evidence that alterations to enhancer regulatory elements, either through cistrome reprogramming or somatic mutations, can act as acquired drivers of resistance. While both chromatin looping and AR's transcription factor function play an important role in enhancer activity, the extent of their contribution is not fully understood. In this work, we hypothesize that resistance in late-stage prostate cancer is driven by alterations in enhancer activity. Enhancer activity can be influenced by both changes in transcription factor activity and alterations in enhancer-promoter interactions. To investigate the changes in transcription factor activity, we functionally characterized the enhancer activity of a constitutively active AR splice variant using a massively parallel reporter assay. Unexpectedly we found that ARv7, a variant linked with resistance in clinical disease, repressed enhancer activity of AR. Next, we integrated the chromatin looping data from models of distinct stages of prostate cancer and identified significant changes in enhancer-promoter interactions in later stages of the disease. We detected new enhancer-promoter interactions in the late-stage model that are absent in the early-stage counterpart. This suggests that new chromatin interactions can change enhancer activity and may potentially contribute to resistance in late-stage prostate cancer. These results highlight the complex interplay between AR and chromatin organization, shedding light on novel mechanisms driving late-stage prostate cancer.

ÖZETÇE

İleri Evre Prostat Kanserinde Hızlandırıcı Aktivitesi Değişimlerinin Analizi

Sıla Akdoğan

Moleküler Biyoloji ve Genetik, Yüksek Lisans

7 Temmuz 2025

Prostat kanseri, Türkiye'de erkeklerde görülen en yaygın kanser türüdür. Hastalığın her evresinde, androjen reseptörü (AR), hastalık başlangıcı ve ilerlemesinde öncü etkendir. AR, aktif halinde, cis-düzenleyici elemanlara bağlanarak, hızlandırıcı-promotörler arasındaki 3 boyutlu kromatin interaksiyonları aracılığıyla gen ifadesine etkide bulunur. Bu kritik işlevi göz önüne alındığında, metastatik ve tekrarlayan prostat kanseri karşısında AR'ı hedef alan tedaviler yaygın bir şekilde uygulanır. Bu tedaviler başlangıçta etkili olsa da, hastalar genellikle tedaviye karşı direnç geliştirir. Direnç geliştiği durumlarda bile, hastaların çoğunda AR kritik işlevini ve tümör büyümesindeki etkisini korur. Sistrom yeniden programlanması ve/veya somatik mutasyonlar sonucu hızlandırıcı düzenleyici elementlerdeki değişimler, bu direncin oluşumuna katkıda bulunur. Kromatin interaksiyonları ve AR'ın transkripsiyon faktörü işlevleri, hızlandırıcı aktivitesinde büyük rol oynar. Ancak, bu faktörlerin katkısı henüz tam olarak anlaşılabilmiş değildir. Bu tezde hipotezimiz, ileri evre prostat kanserinde oluşan direncin, hızlandırıcı aktivitesindeki değişimlere bağlı olduğudur. Hızlandırıcı aktivitesi, transkripsiyon faktörü aktivitesi ve hızlandırıcı-promotör interaksiyonları tarafından etkilenebilir. Transkripsiyon faktörlerindeki değişimin etkisini anlamak amacıyla, bir AR varyantı olan ARv7'nın sebep olduğu hızlandırıcı aktivitesini analiz ettik. Beklenmedik bir şekilde, daha önce hastalarda tedavi direnciyle işilkili olduğu gösterilmiş ARv7'nın, normal AR aktivitesini baskıladığını gördük. Bir sonraki adımda, prostat kanserinin erken ve ileri evrelerini temsil eden iki farklı hücre hattından elde edilmiş kromatin interaksiyon verisini analiz ettik. Analiz sonucunda erken ve ileri evreler arasında büyük interaksiyon farkları tespit ettik. İleri evrede, erken evre hücre hattında bulunmayan yeni hızlandırıcı-promotör interaksiyonları keşfettik. Ulaştığımız bu sonuçlar, yeni kromatin interaksiyonlarının hızlandırıcı aktivitesine olan etkisini ve bu etkinin tedaviye direnç gelişimine olası katkısını destekler niteliktedir. Hep birlikte ele alındığında sonuçlarımız, AR ve kromatin organizasyonu arasındaki karmaşık ve hassas dengeyi açığa çıkararak, ileri evre prostat kanserinin arkasındaki olası özgün mekanizmalara ışık tutmaktadır.

ACKNOWLEDGEMENTS

Looking back, I now realize how much I've grown as a person during my master's. This journey has taught me many valuable lessons. Surely, I've grown as a scientist and gained several technical skills. But more importantly, I've seen the importance of love, friendship, support, mentorship and teamwork. Science truly thrives on collaboration!

Working in Lack Laboratory has been the best experience of my professional life. I'd like to thank my advisor Dr. Nathan Lack for always believing in me and pushing me to achieve my best. Thanks to Nathan, I now know I should learn to walk before I try to run :). I'd like to thank my mentor Dr. Tunç Morova for guiding me when I was a newbie in the lab. I'd like to thank Ivan Yu for being the best colleague I could ever imagine. Our virtual cup of coffees taught me a lot; it was a privilege to work with you! I'd like to thank Emirhan Tekoğlu for always being there for me when I had a million questions. I thank Umut Berkay Altıntaş for his enthusiasm, our brainstorms helped me a lot! I'd like to thank Lynn Masri for her support throughout these years, we started this journey together and she was there for me all the way. I thank Brandon Fasy for our endless conversations about life; you are a brave warrior for sure! I'd like to thank all my lab mates: Dr. Flora Huang, Hatice Kıbar, Kerim Yavuz, Uğur Meriç Dikbaş, Olka Missaghimamaghani, Dr. Doğancan Özturan, Dr. Bengül Gökbayrak, James Wong, Roselyn Ma and Daniel Robinson. I have learnt something new from each and everyone of you. Our differences and collaboration make us stronger!

A huge thanks goes to my friends for always supporting me and helping me stay sane :D. My high school friends' (Kraliçeler) friendship means the world to me. I couldn't have done it without you! My childhood friends from Gülümkıyı were always there for me. Thank you so much for hang my back, no matter what! My dear friends from ODTÜ, İdil İlayda Duran and

Ekin Su Erdem, it was a pleasure being together with you here at Koç too, thank you for your endless support!

Last but not least, I'd like to thank my family. I am forever grateful to my parents Melda Gençtürk and Birol Gençtürk, and my dearest grandmother Melahat Meşinci. Thank you for always supporting me on my journey to pursue my dreams. I can't find the words to express how proud I am to be your daughter. Finally, my beloved partner Can Bayraktar, thank you for making me feel at home in a new city with your endless love and patience. I know we will continue to build a beautiful life for ourselves. I have a feeling that we'll explore the world together :).

I can't wait to see what the future holds for me and my loved ones. Cheers!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

PCa	Prostate Cancer
AR	Androgen Receptor
ADT	Androgen Deprivation Therapy
CRPC	Castration-Resistant Prostate Cancer
NTD	N-Terminal Domain
DBD	DNA-Binding Domain
LBD	Ligand-Binding Domain
3D	Three-dimensional
DHT	5α-dihydrotestosterone
TF	Transcription Factor
ARBS	AR-Binding Site
ARv7	Androgen Receptor Splice Variant 7
flAR	Full Length Androgen Receptor
ARE	Androgen Response Element
CRE	Cis-Regulatory Element
ARBS	Androgen Receptor Binding Sites
ADLD	Autosomal Dominant Adult-Onset Leukodystrophy
kb	Kilobase
NGS	Next-Generation Sequencing
MPRA	Massively Parallel Reporter Assay
STARR-Seq	Self-Transcribing Active Regulatory Region Sequencing
RNA-seq	RNA Sequencing
H3K4me1	Histone H3 Lys4 Monomethylation
H3K27ac	Histone H3 Lys27 Acetylation
ChIP-seq	Chromatin Immunoprecipitation Sequencing
3C	Chromosome Conformation Capture
qPCR	Quantitative Polymerase Chain Reaction
4C	Chromosome Conformation Capture-on-Chip
5C	Chromosome Conformation Capture Carbon Copy
SV	Structural Variant
SNV	Single-Nucleotide Variant
Indel	Insertion-Deletion
CNV	Copy Number Variation
Mb	Megabase
ecDNA	Extrachromosomal DNA
LN95	LNCaP95 Cell Line
HD	Heterodimer

FC	Fold Change
padj	Adjusted p-value
bp	Base Pair
IGV	Integrative Genomics Viewer
CNR	Copy Number Ratio
FDR	False Discovery Rate
CNV	Copy Number Variation
KLK3	Kallikrein-Related Peptidase 3
PSA	Prostate-Specific Antigen

Chapter 1
INTRODUCTION

1.1 Prostate Cancer Progression

Prostate cancer (PCa) is the most common cancer type in Turkish men and is one of the leading causes of cancer-associated deaths [1]. The age of diagnosis depends on accessibility of healthcare services [2]. A 2009 study done in Türkiye explains that PCa was the fifth most common cancer type in Izmir between the years 1993 and 1997 [3]. In a broader and more recent study from Türkiye, median diagnosis age of PCa is reported as 68 [4].

PCa has distinct stages (Figure 1.1) with a disease burden that can be classified as local or metastatic. In parallel, the hormone sensitivity of PCa changes as it progresses. PCa progression is dependent on androgen signaling at early stages whereas in most cases, hormone insensitivity develops [5]. Hormone sensitivity status of PCa is a key determinant in treatment decisions for patients [1]. It is very important for researchers to understand the mechanisms behind hormone insensitivity i.e., castration resistance in order to develop potential novel therapies.

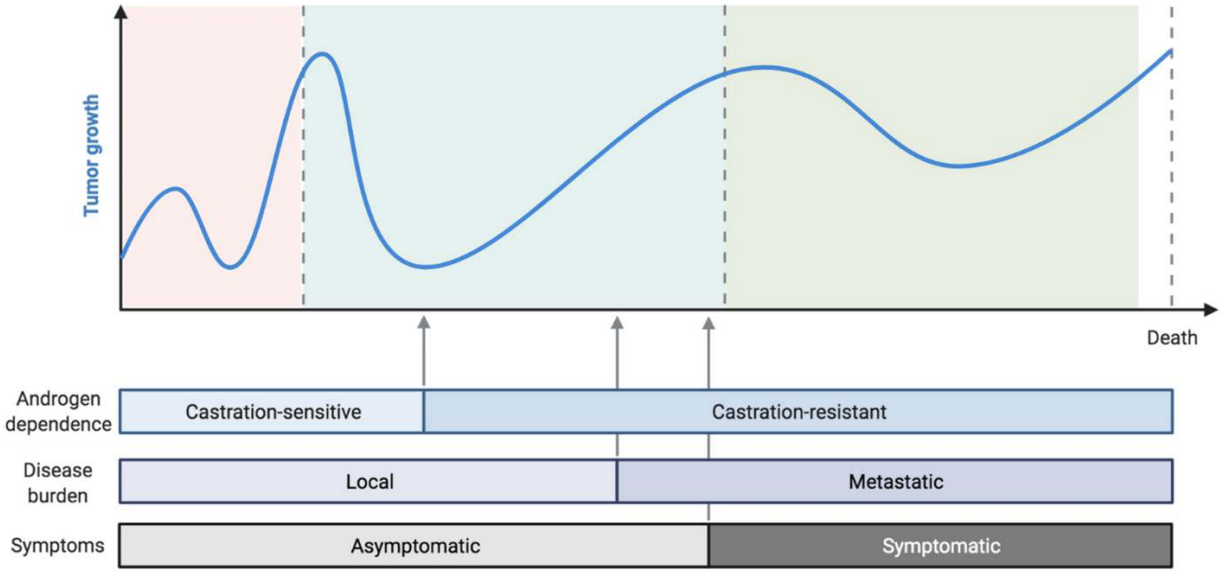

Figure 1.1: General course of prostate cancer progression with androgen dependence, disease burden and symptoms (Source: [6]).

1.1.1 Early-Stage Prostate Cancer and Hormone Sensitivity

In early-stage PCa, the tumor is generally localized within the prostate gland and dependent on hormone signaling via androgen receptor (AR). If the tumor is not metastasized, local therapies such as surgery or radiotherapy can be curative [7]. However, if the disease is metastasized, other therapies are applied. Since tumor growth and PCa progression are androgen dependent, androgen deprivation therapy (ADT) is preferred for patients with recurrent and metastatic PCa [8]. While tumor size and growth rate decrease with ADT, this response is temporary, and the disease will eventually advance into castration-resistant prostate cancer (CRPC) where androgen deprivation is no longer able to regress PCa progression [5].

1.1.2 Late-Stage Prostate Cancer and Castration Resistance

Even following the progression to CRPC, tumor growth and disease progression is dependent on AR signaling for most patients. However, these tumors are resistant to androgen deprivation therapies and continue to grow even with castrate levels of androgens. Unfortunately, there are no fully curative therapies for CRPC yet [5]. Several mechanisms related to AR signaling in CRPC have been proposed including AR overexpression, elevated AR co-regulator expression and intratumoural hormone production [5]. One major factor

contributing to progression into CRPC is the presence of androgen receptor splice variants [9].

1.2 Role of Androgen Receptor in Prostate Cancer

Androgen receptor is a ligand-activated nuclear receptor that is activated following ligand binding [10]. The *AR* gene is composed of 8 exons and is located on chromosome X. The AR protein has three main functional domains including the N-terminal domain (NTD), DNA-binding domain (DBD), and ligand-binding domain (LBD) [10]. The DBD is responsible for direct DNA binding of AR [10]. LBD is where ligand binding occurs by the androgens, such as testosterone and 5α-dihydrotestosterone (DHT) [10, 11].

1.2.1 Androgen Receptor's Transcription Factor Role

AR is a transcription factor (TF) and is involved in transcriptional regulation of target genes [10]. In healthy prostate tissue, activation of AR allows development of reproductive and secondary sexual characteristics, and differentiation in males. However, AR-mediated transcription dramatically changes in PCa and causes tumor growth and proliferation Hence, AR's TF activity is heavily studied in the context of PCa progression [12].

Androgen receptor in its inactive form resides in cell cytoplasm. When testosterone uptake occurs in the cell, it is reduced to DHT and binds to the LBD of AR to activate it. Upon activation, AR translocates into the nucleus, dimerizes and binds to AR-binding sites (ARBS) on DNA. Following DNA binding, AR recruits other TFs, RNA polymerase II, co-activators, and chromatin modifier proteins to the ARBS [13, 14]. This current model can also be seen in Figure 1.2. AR-binding sites correspond to non-coding regulatory enhancer regions on DNA [14]. Enhancers interact with AR target gene promoters via chromatin looping [14]. Through this process, AR performs its transcription factor role, leading to transcriptional regulation of its target genes.

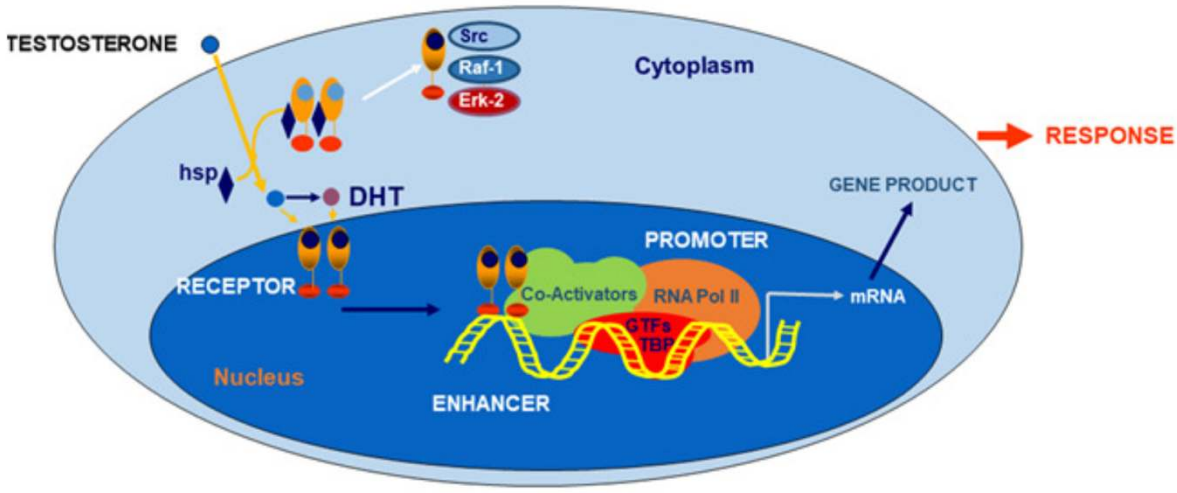

Figure 1.2: Androgen receptor's transcription factor activity, current model (Source: [13]).

1.2.2 Androgen Receptor Splice Variant 7

Alternative splicing events result in expression of androgen receptor splice variants that lack an LBD. These truncated AR variants are ligand-independent and constitutively active even under castration conditions [15]. Presence of AR splice variants strongly correlates with progression to CRPC. Among these, androgen receptor splice variant 7 (ARv7) is the most common AR splice variant found in CRPC cases [5].

ARv7 protein has the NTD and DBD in common with the full-length AR (flAR). But it lacks the hinge region (H) and the ligand-binding domain [15]. Instead, ARv7 utilizes a cryptic exon (CE3) right after the DBD that includes a nuclear localization signal (Figure 1.3). ARv7 is still able to bind to DNA since the DBD is conserved. Consequently, ARv7 is proposed to induce transcription independent of androgen activation. Therefore, target gene expression can potentially be carried out even in the presence of ADT by ARv7 (Figure 1.4) [5]. Research has been and is being done on the details about ARv7 and whether its function is independent of flAR or whether its binding sites differ vastly from known ARBS [15]. Clinical studies show that there is an association between the presence of ARv7 in circulating tumor cells and resistance to AR signaling pathway inhibitor therapies which are employed in patients with metastatic PCa [16]. Also, ARv7 expression is higher in CRPC cases, compared to primary PCa [17]. Lastly, presence of ARv7 is correlated with poor prognosis and lower survival rates in CRPC patients [16].

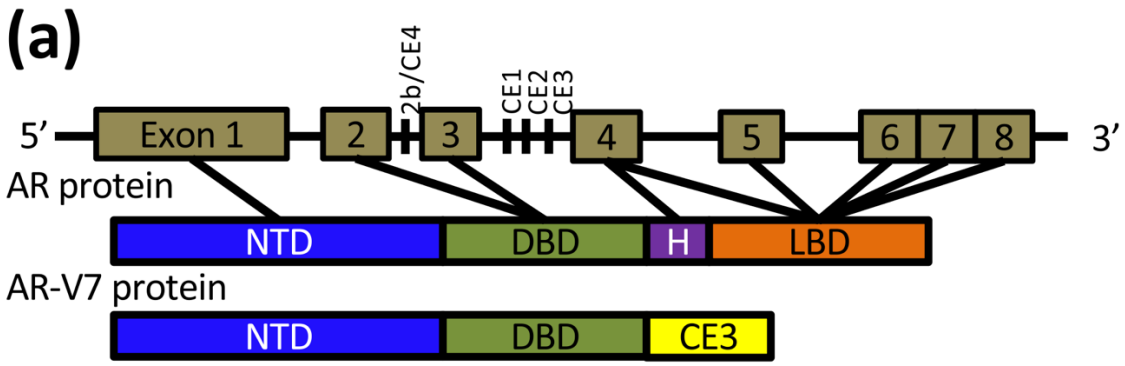

Figure 1.3: AR exon composition with protein domains of full-length androgen receptor and androgen receptor variant 7 (Source: [15]).

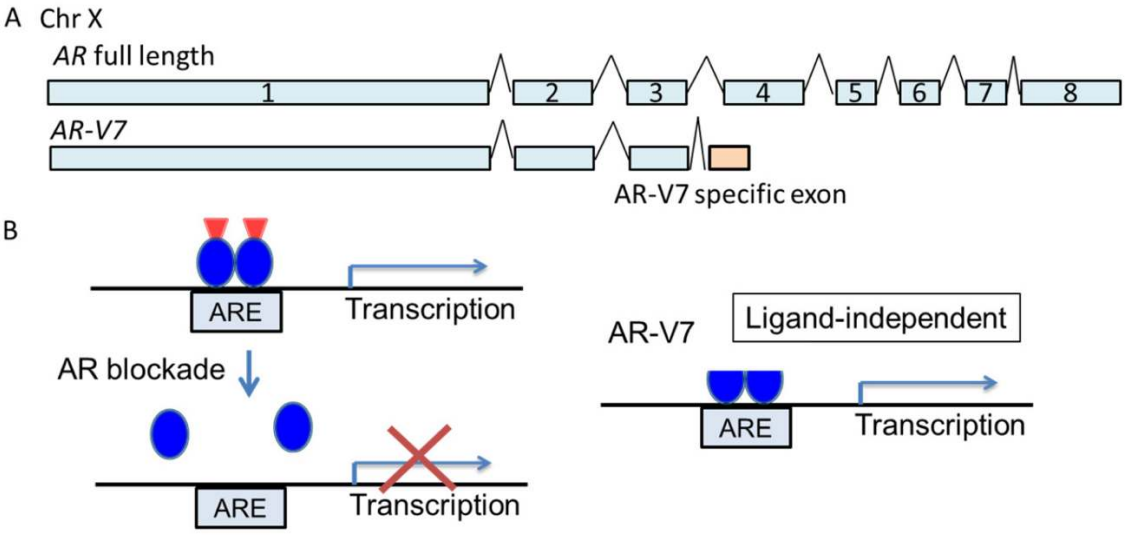

Figure 1.4: Full length AR and AR splice variant 7: (A) Comparison of exon composition (B) Representation of ligand-dependent and independent transcription factor activity of the full-length protein and the variant, respectively. ARE: Androgen response element (Source: [5]).

1.3 Cis-Regulatory Elements

Transcriptional regulation is central in genetics and epigenetics cancer research. Research on regulation of gene expression is crucial since it directly affects protein levels that are involved in disease related pathways. Cis-regulatory regions in the genome play a huge role in transcriptional regulation and there are changes observed in their level of activity comparing healthy tissue to cancerous tissue as well as potentially early-stage disease to late-stage counterparts [18,19].

Cis-regulatory regions are composed of non-coding DNA sequences that have gene expression regulatory roles called cis-regulatory elements (CREs). Transcription factors and other regulatory proteins bind to those elements and affect the levels of transcription of their target genes [20]. CREs also interact with each other via changes in 3D chromatin conformation [21]. Most widely studied CREs are enhancers and promoters [20].

1.3.1 Promoters

Promoters are a class of cis-regulatory regions that define where gene transcription starts. Eukaryotic genes have promoter sequences that are needed for transcription to start. Most genes have a single canonical promoter at a specific site close to their transcription start site [22,23]. It is also known for some genes that multiple promoter sequences are involved in initiation and regulation of transcription. Accuracy and efficiency of transcription initiation and regulation are dependent on promoter sequences [23]. Promoters also interact with other CREs such as enhancers. In fact, promoters alone can only cause basal levels of gene transcription [22]. Enhancers are needed in contact with them to fully regulate and increase the transcriptional activity to needed levels [22].

1.3.2 Enhancers

Enhancers are cis-regulatory elements that increase transcription efficiency of their target genes [24,25]. An enhancer is defined as a DNA sequence that activates and/or increases transcription of its target gene [25,26]. Importantly, an enhancer's activity is not dependent on its orientation or distance to its target promoter [24,25,26]. Transcription factors bind to enhancers and trigger their activity. TF bound enhancers recruit other coregulator proteins

that are involved in chromatin remodeling and transcription initiation [27]. Gene regulation by enhancers is a complicated process as different TFs can bind to a single enhancer and multiple enhancers contribute to the regulation of a single gene. This combinatorial TF binding forms an intricate mechanism that introduces another level of complexity when it comes to characterization of disease-related enhancers [28]. Enhancers contribute to transcriptional regulation by interacting with promoters. Since multiple enhancers can interact, they are able to possess combinatorial effects on their target promoters [28].

1.3.3 Enhancer-Promoter Interactions

Chromatin is organized into 3D topologically associated domains (TAD) in the nucleus (Figure 1.5) [29]. These are important for transcription regulation since chromatin interaction frequencies are higher within TADs. Nevertheless, it is possible to observe chromatin loops outside TAD boundaries and they allow longer-range interactions between different regions of chromatin [30]. This enables long-range enhancer-promoter interactions to occur (Figure 1.6). Changes in the looping landscape allow CREs to come in close physical contact with each other.

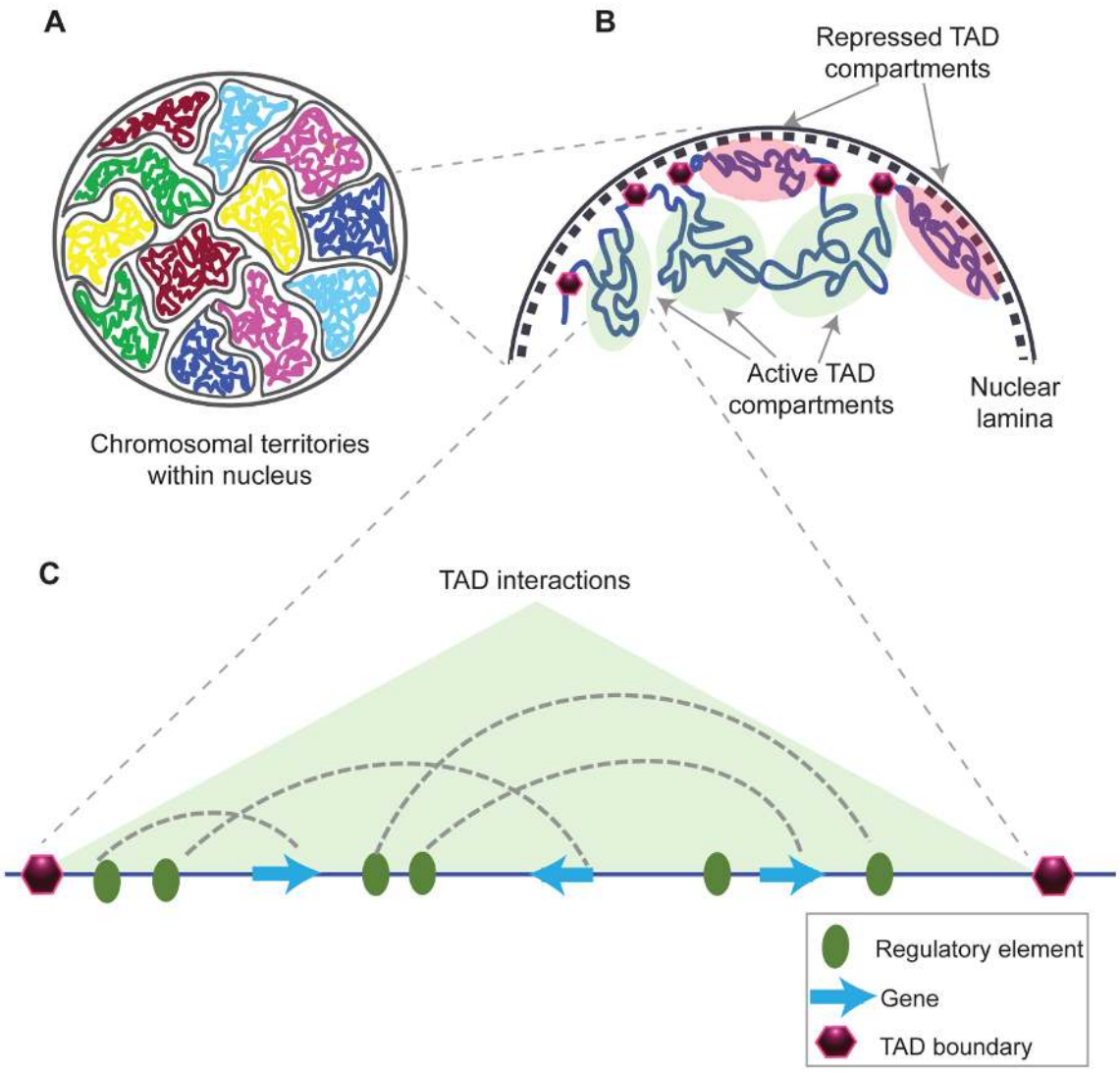

Figure 1.5: Structural organization of chromatin. (A) Representation of chromosomal territories within nucleus (B) Active and inactive topologically associated domains (TAD) (C) Interactions between regulatory elements and genes within the TAD boundary (Source: [29]).

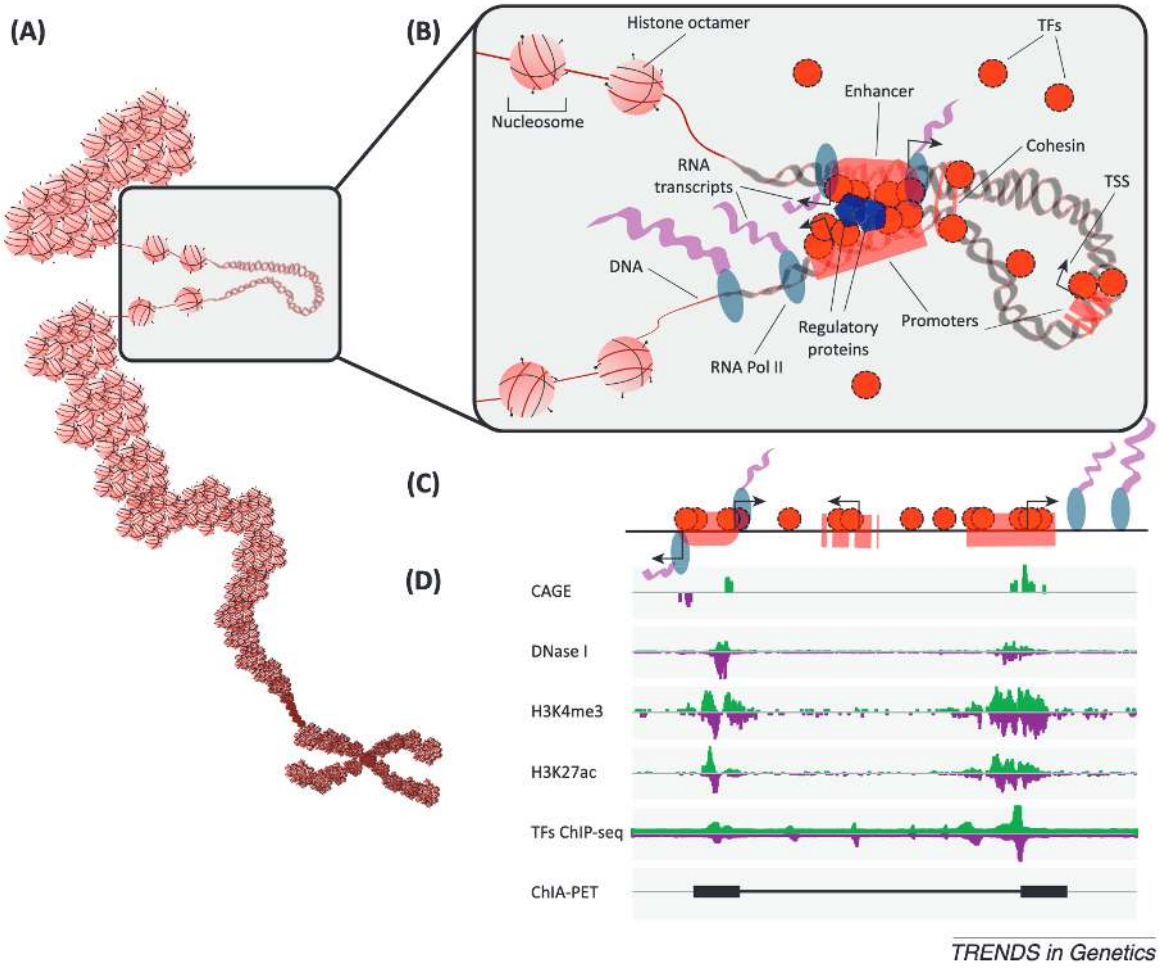

Figure 1.6: Summary of cis-regulatory element function: (A) Overview of 3D chromatin organization with histone octamers (B) Enhancer-promoter interactions via chromatin looping, TF and regulatory protein recruitment (C) Two-dimensional representation of the model in (B) (D) Experimental methods giving insights about transcriptional regulation (H3K4me3, H3K27ac, TF ChIP-seq), gene expression and enhancer activity (CAGE), chromatin accessibility (DNase I hypersensitivity), and chromatin conformation (RNA Pol II ChIA-PET). Green bars correspond to reads mapped to the positive strand and purple bars correspond to reads mapped to the negative strand (Source: [20]).

1.4 Functional Genomics

Functional genomics is a molecular biology discipline focusing on understanding the role of the genome in biological systems. Disease phenotypes can be investigated from a functional genomics perspective. Thanks to improvements in next-generation sequencing (NGS) technologies and bioinformatic analysis, functional genomics data can be used in uncovering underlying genomic mechanisms in human cancers [1]. These powerful tools can improve our understanding of cancer diagnosis, prognosis and treatment.

For the analyses performed in this thesis, data from multiple experimental functional genomics approaches are used. Namely, these are approaches to characterize enhancer activity, transcription factor binding and enhancer-promoter interactions.

1.4.1 Enhancer Activity

Enhancer activity can be measured by massively parallel reporter assays (MPRAs). MPRAs focus on an enhancer's capacity to activate a reporter gene's expression (Figure 1.7) [26]. MPRAs can be used to assess enhancer activities across the whole genome or a library of DNA fragments (capture libraries) containing regions of interest which are possible enhancer candidates [26]. One MPRA method to assess enhancer activity is self-transcribing active regulatory region sequencing (STARR-Seq) [26,31]. The design of STARR-Seq plasmids allows enhancers to induce their own transcription (Figure 1.7). This enables measuring enhancer activity directly from enhancer RNA sequencing (RNA-seq), in a quantitative manner [26,31]. Sequencing data coming from STARR-Seq is then analyzed using appropriate bioinformatics tools to map and quantify enhancer activities of the regions assessed.

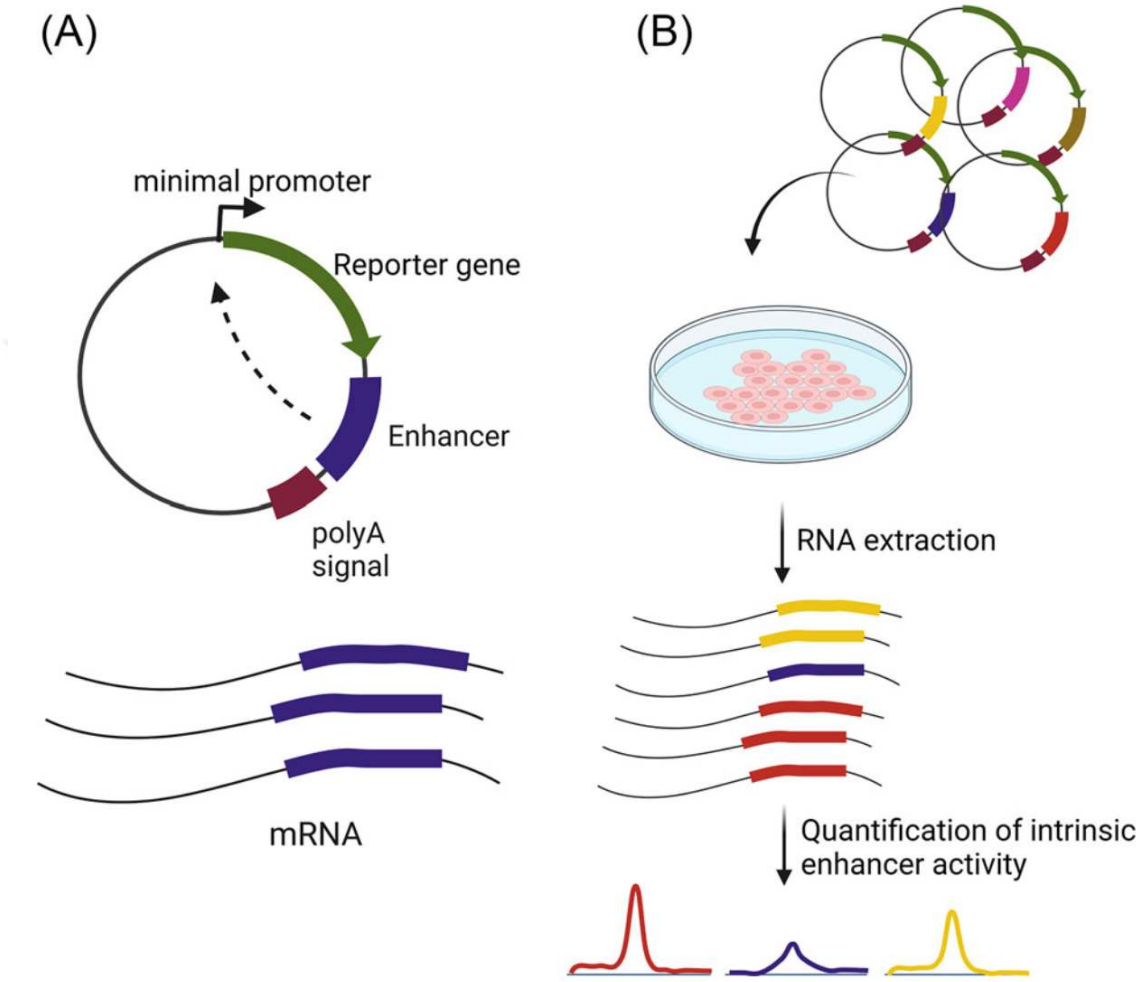

Figure 1.7: STARR-Seq experimental overview (A) Design of STARR-Seq plasmids allows enhancers to induce their own transcription resulting in mRNA coming from enhancers. This way, enhancer activity can be quantified via RNA-seq (B) Plasmids are transferred to a population of cells followed by RNA extraction and NGS. NGS data is then analyzed, and intrinsic enhancer activities of library regions are quantified (Source: [26]).

1.4.2 Transcription Factor Binding

In addition to MPRAs that directly aim to quantifying enhancer activity, there are other chromatin features that support the presence of enhancer activity. Chromatin accessibility, specific histone modifications such as histone H3 Lys4 monomethylation (H3K4me1) and H3K27 acetylation (H3K27ac), and the presence of TF binding are good indicators of enhancer activity [32]. TF binding and histone modifications are commonly identified with a method called chromatin immunoprecipitation sequencing (ChIP-seq) [32,33,34]. ChIP-seq combines chromatin immunoprecipitation and massively parallel sequencing to detect protein binding sites on DNA [34].

First step in ChIP-seq is the crosslinking of DNA and proteins (Figure 1.8). Next, samples containing crosslinked DNA-TF pairs are fragmented and sonication is performed, or exonucleases are added. By sonication or exonucleases, the unbound DNA pieces are cleaved from the DNA-TF complexes. Next, immunoprecipitation of DNA-protein complexes is performed by antibodies specific to TFs used in the procedure. Lastly, DNA extraction is performed and extracted DNA is sequenced. This sequencing data is then analyzed using bioinformatics tools to reveal TF binding sites [32,33,34].

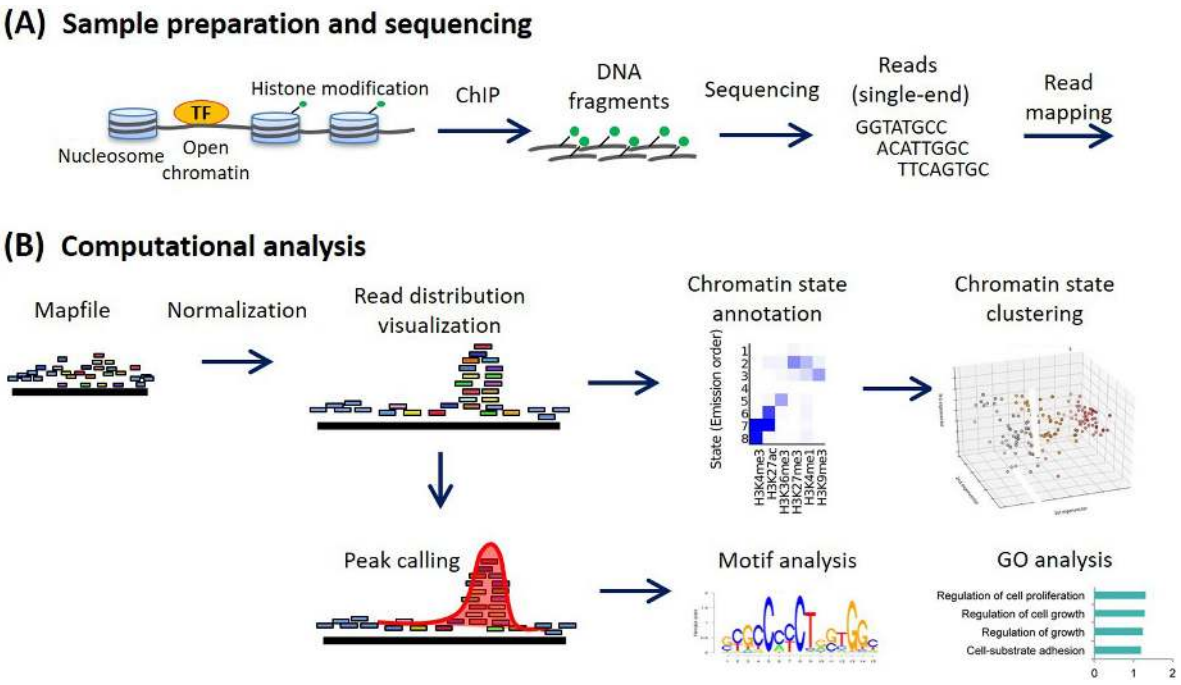

Figure 1.8: Overview of ChIP-seq steps: (A) Summary of experimental steps in ChIP-seq. Chromatin immunoprecipitation followed by sequencing (B) Typical computational analysis steps of ChIP-seq data: mapping, normalization, peak calling, motif analysis, chromatin state annotation, chromatin state clustering and gene ontology (GO) analysis (Source: [35]).

1.4.3 Enhancer-Promoter Interactions

Several methods have been developed to detect 3D chromatin interactions. First method is called chromosome conformation capture (3C) which is published in 2002 [36]. Basic principle of 3C starts with crosslinking interacting chromatin loci. This is followed by cell lysis and restriction digestion resulting in fragments of cross-linked chromatin loci. Following the dilution of samples, ligation is performed in a way that is preferential for cross-linked DNA to ligate, over noninteracting loci. Lastly, crosslinking is reversed, and quantitative polymerase chain reaction (qPCR) is performed to understand interaction frequencies between anchor loci [36]. Since qPCR requires primers specific to known loci, 3C is a suitable method to detect interactions between regions of interest. So, it is a validation method rather than being fully exploratory. This is why it is usually classified as a "one-vs-one" method focusing on singular interacting anchors. Inspired by 3C, several other methods are published such as chromosome conformation capture-on-chip (4C) [37] and chromosome conformation capture carbon copy (5C) [38] in 2006.

In 2009, an "all-vs-all" method is published that allows a genome-wide interaction screening. This method is composed of chromosome conformation capture followed by high-throughput sequencing (Hi-C) [39]. Hi-C combines 3C with NGS and is the first method that can perform an unbiased chromatin interaction search across the whole genome (Figure 1.9). First steps of Hi-C are similar with 3C which are chromatin crosslinking and digestion. Different from 3C, resulting 5′-overhangs are filled with biotinylated nucleotides and then blunt ends are ligated after diluting samples. Ligation again favors cross-linked DNA fragments over others. At this point, samples contain ligated DNA fragments that were close in 3D space in the nucleus, and the junction points are marked with biotin tags [39]. Next, these fragments are pulled down leading to purification and shearing to prepare a Hi-C library. Finally, the resulting library is sequenced [39].

Most recently, a 2016 study introduces a new technique to discover chromatin interactions [40]. This method, HiChIP, adopts a protein-centered approach to discover DNA contacts (Figure 1.10). It combines approaches from previous chromatin contact methods like Hi-C with chromatin immunoprecipitation (ChIP) [40]. Crosslinking, biotinylation, and ligation are performed inside the nucleus before lysing the cells. This minimizes false positives and increases the efficiency of this method compared to previous approaches. DNA contact

capture efficiency of HiChIP is also high because of the ChIP step -a protein of interest guides the process. After reverse crosslinking, DNA extraction and biotin pull down are performed, captured DNA interactions are also associated with the protein of interest. Lastly, paired-end sequencing is performed. Finally, the sequencing data is analyzed by appropriate bioinformatics tools and packages [40].

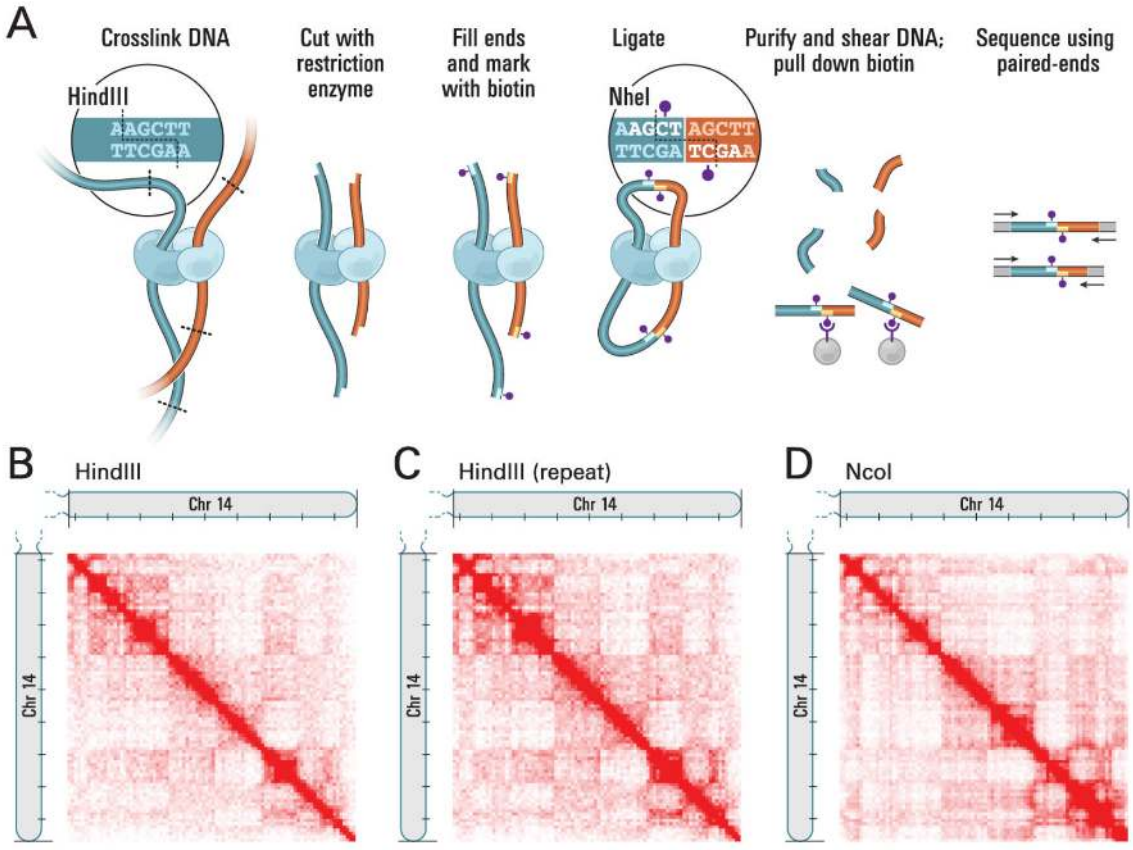

Figure 1.9: Hi-C Overview: (A) Steps of Hi-C: Crosslinking, restriction digestion, biotinylation, ligation, library preparation and sequencing (B) A contact matrix generated using Hi-C data. Intrachromosomal interactions from chromosome 14 are shown (C) Contact matrix coming from Hi-C data, biological repeat of (B) (D) Contact matrix showing interactions within chromosome 14, a different restriction enzyme is used (NcoI) (Source: [39]).

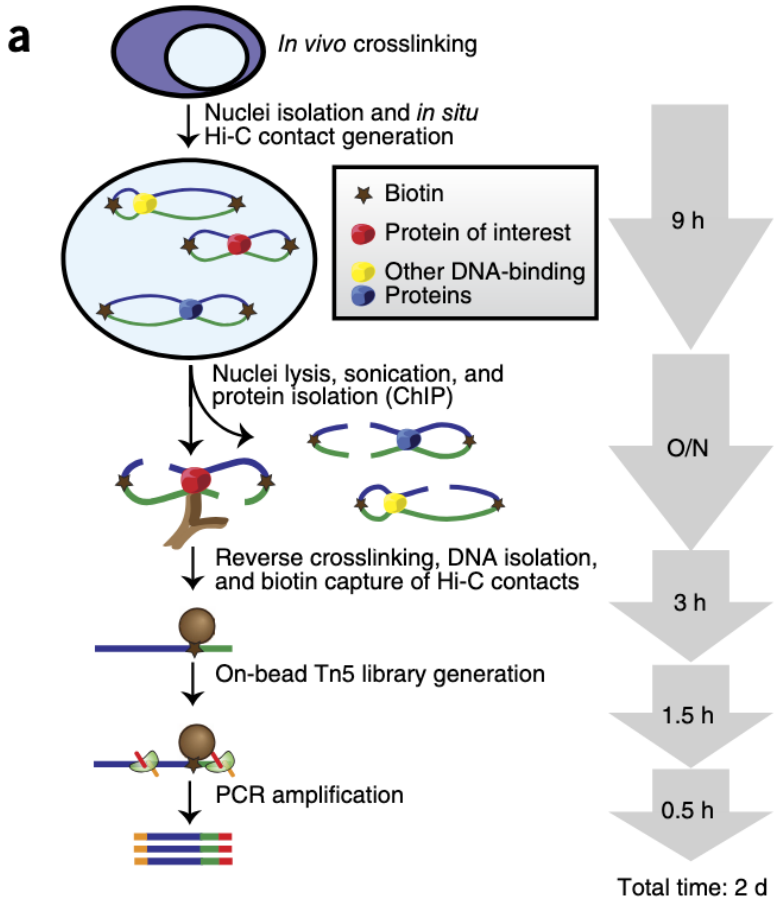

Figure 1.10: Experimental overview of HiChIP. Steps here explain procedures performed up until sequencing. Crosslinking, biotinylation and ligation performed inside the nucleus. Cell lysis, sonication and protein isolation (ChIP) is then carried out. Reverse crosslinking and library preparation is followed by sequencing. Time needed for each step is also provided. O/N: Overnight (Source: [40]).

1.5 Dysregulation of Enhancers as A Driver of Castration-Resistant Prostate Cancer

There are several mechanisms behind the changes in cis-regulatory element activity. Mutations in CRE sequences, changes in chromatin accessibility and in 3D chromatin conformation are the most pronounced ones [22]. 3D conformational changes allow new enhancer-promoter pairs to come in contact or cause loss of preexisting ones, resulting in a change in transcriptional regulation.

Latest studies highlight the importance of CREs in evolutionary, developmental and disease contexts [22]. They have a massive effect in progression of many diseases by contributing to transcriptional regulation. A 2025 study analyzes the CRE landscape of known oncogenes across many human cancers including PCa [41]. They report that targeting specific CREs leads to a change up to 60% in oncogene expression and an effect up to 50% in cell growth [41].

In the context of prostate cancer, AR binds to enhancer CREs and mediate transcriptional regulation [42]. It is shown, in one of our laboratory's previous works, that AR binding sites (ARBS) have an essential role in enhancing interactions with other ARBS and gene promoters [42]. Enhancers play a clear role in numerous diseases including prostate cancer [43]. In CRPC, presence of AR variants, including ARv7, allows a constitutive enhancer activity even in the absence of androgens. In summary, enhancers are a crucial part of AR focused PCa research. For these reasons, it is important to characterize CRE activity in PCa and elucidate their interactions in 3D space to better understand the disease progression.

Also, dysregulation of chromatin loops can dramatically alter gene transcription. Changes in TAD boundaries or loss of TADs as a whole are reported in human diseases [44]. An example comes from an autosomal dominant adult-onset leukodystrophy (ADLD) patient [44]. A 660 kilobase (kb) deletion results in the loss of a TAD which then allows an enhancer to regulate the lamin B1 gene. Adoption of the enhancer which normally does not regulate this gene, causes changes in its expression which finally leads to ADLD [44].

TAD disruptions are also observed in cancers. Altered TADs act as oncogene drivers by introducing new enhancer-promoter interactions leading to oncogene upregulation or tumor suppressor downregulation [45]. TAD alterations can occur via different mechanisms. In addition to changes in TAD boundaries, another prominent mechanism is structural

variations within individual TADs. Structural genomic rearrangements allow new enhancer-promoter interactions within TADs via breakpoints, breakage, and fusions [45].

A 2016 comparative study analyzes differences in 3D chromatin organization between healthy and prostate cancer cells [46,47]. They report the overall chromatin organization and TAD positions are similar in two groups, but TADs appear smaller in PCa cells [47]. Also, they detect new interactions that are PCa-specific and are absent in healthy prostate cells [47]. In a more recent study, 3D chromatin landscapes are analyzed in multiple PCa cell lines that represent different disease stages including CRPC [46,48]. They report changes in TAD organization and transcriptional activity of some genes as the disease progresses into later stages [48]. Overall, such examples are increasing in number in the literature. Relevance of enhancer-promoter interactions to cancer phenotypes appears noteworthy.

As mentioned above, changes in enhancer activity are associated with disease phenotypes including cancer [43]. In castration-resistant prostate cancer, AR chromatin binding and AR-mediated enhancer activity are altered [49]. It is shown that AR has different binding preferences in CRPC compared to early-stage PCa [50,51]. This suggests the differences in AR-bound enhancers are possible drivers or CRPC.

To summarize, dysregulation of enhancers can occur via multiple mechanisms including changes in TF binding, chromatin accessibility and enhancer-promoter interactions in 3D space. Enhancer activity can also be altered by the presence of specific genomic structural variants. Since enhancers are regulatory regions that have a direct effect on transcription, formation of new regulatory networks has a non-negligible importance in prostate cancer research.

1.5.1 Extrachromosomal DNA and Enhancer Hijacking

Somatic structural variants (SVs) are common in cancer cells. Studies show that they act as drivers in many cancer types [52]. Traditionally, somatic point mutations are investigated such as single-nucleotide variants (SNVs) and insertions and deletions (indels) in cancer [52]. However, recent pan-cancer studies show that SVs probably contribute more to cancer initiation and progression, compared to somatic point mutations [53]. SVs appear as deletions, insertions, duplications, copy number variations (CNV), inversions and translocations that affect a region usually larger than 1 megabases (Mb) on a chromosome [54]. In that sense, their scale is larger than somatic point mutations and typically smaller than chromosomal abnormalities [54].

One interesting translocation event causes extrachromosomal DNA (ecDNA) formation in cancer cells. ecDNAs are circular, double-stranded DNA fragments that are found in cancer cell nuclei [55]. They are shown to facilitate rapid evolution of cancer cells and resistance to therapies [55]. When first discovered, the importance of ecDNAs was not well understood and they were thought to be rare occurrences. Latest advancements allowed researchers to show that ecDNAs are more common than previously thought in human cancers. They serve different functions for cancer cells which correlate with bad patient prognosis (Figure 1.11) [55]. Out of these mechanisms, enhancer hijacking is particularly striking in the scope of this thesis. Enhancer hijacking allows usage of enhancers by oncogene promoters that are not capable of using them normally. This leads oncogenes to “hijack” enhancers to upregulate their expression. ecDNA formation mediates enhancer hijacking since it gives a physical proximity to enhancer-promoter pairs that do not have access to each other on chromosomal DNA [56]. This leads to upregulation of some oncogenes in cancer [55,56].

Overall, presence of ecDNAs leads to a worsened prognosis. ecDNA formation and enhancer hijacking events are newly emerging hallmarks of advanced stage disease including prostate cancer.

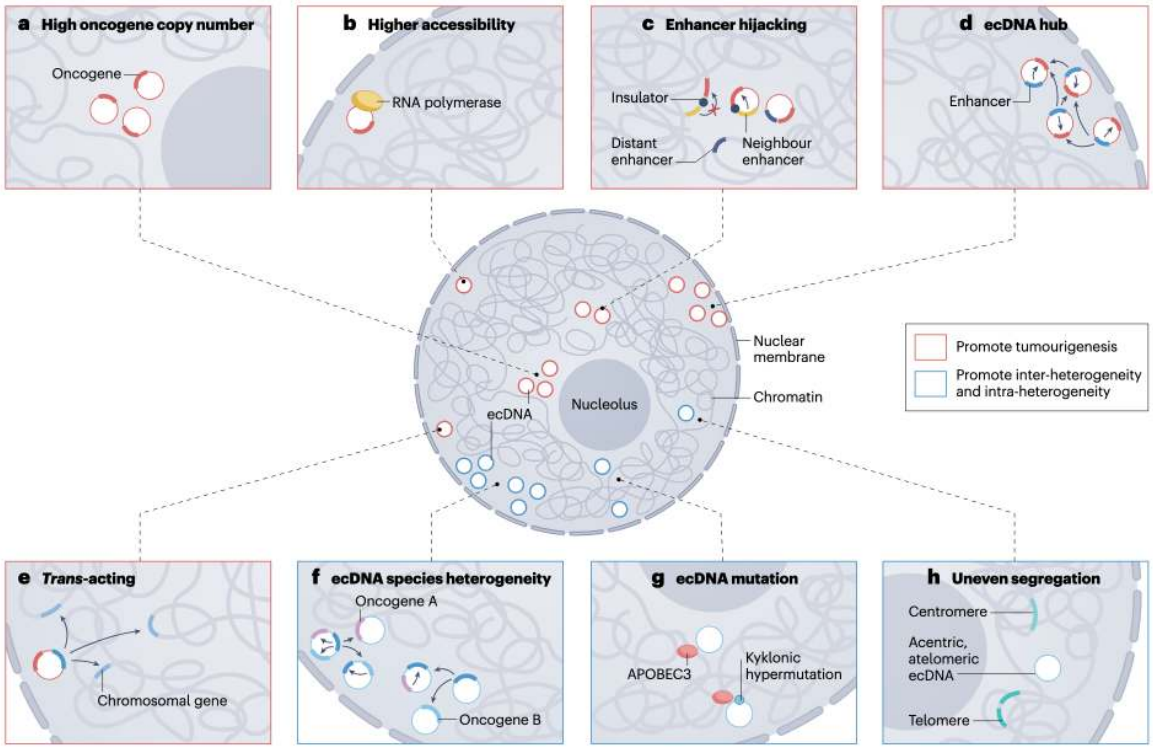

Figure 1.11: ecDNA in cancer: ecDNAs promote tumorigenesis and tumor heterogeneity by distinct mechanisms including: (A) Increase in oncogene copy number (B) Higher oncogene accessibility (C) Enhancer hijacking by bringing distant enhancers and gene promoters together (D) Forming ecDNA hubs composed of multiple interacting ecDNAs (E) Trans-acting ecDNA, interacting with chromosomal genes (F) Increased cellular and tumoral heterogeneity caused by ecDNAs (G) Mutations on ecDNAs (H) Uneven chromosome segregation during cell division (Source: [55])

1.6 Aim of This Work

The cis-regulatory landscape has a significant contribution to transcriptional regulation in all stages of PCa. This thesis focuses on changes in enhancer activity and enhancer-promoter interactions, in early and late-stage PCa. In this work, we hypothesize that resistance in late-stage prostate cancer is partly driven by alterations in enhancer activity. We aim to investigate the changes in enhancer activity in late-stage prostate cancer compared to its early-stage counterpart. The analysis is performed by focusing on two factors affecting enhancer activity: transcription factor binding and enhancer-promoter interactions. To analyze the effect of different transcription factors, we examine enhancer activity caused by flAR and ARv7, supported by TF binding data. To compare changes in enhancer-promoter interactions, we analyze 3D genome features for two cell lines that are representative of early and late-stage disease, separately. Lastly, we support our findings by detecting structural variants from those cell lines.

Chapter 2
METHODS

2.1 Enhancer Activity

To observe the enhancer activity caused by AR and ARv7, STARR-Seq is performed using a prostate cancer cell model, LNCaP95 (LN95), which expresses both flAR and ARv7.

2.1.1 LN95 STARR-Seq Experimental Overview and Preliminary Data Analysis

LN95 STARR-Seq experiments are performed by Lack Laboratory member Dr. Chia-Chi Flora Huang. An overview of different constructs, proteins expressed, and treatments can be found in Figure 2.1. Initial STARR-Seq data analysis and peak calling are performed by Lack Laboratory alumni Dr. Tunc Morova. We collaborated with Lack Laboratory member Ivan Yu on how to proceed with rest of the data analysis. STARR-Seq is performed on a capture library composed of 7371 regions which contain 4138 ARBS, positive and negative control regions [42].

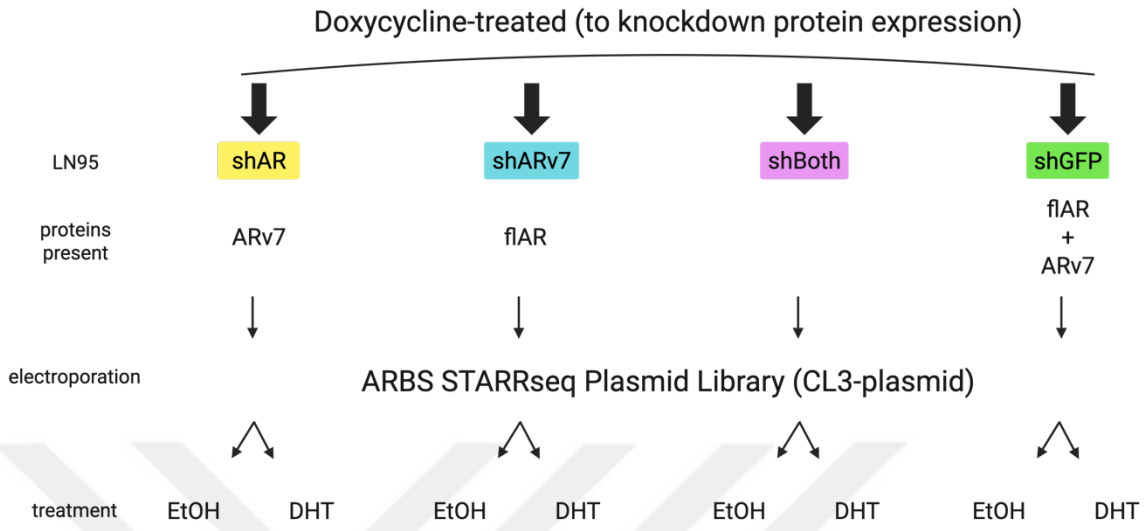

Figure 2.1: Experimental overview of LN95 STARR-Seq. In the shAR sample, flAR is knocked down, only ARv7 protein is present. In the shARv7 sample, only flAR is present. In the shBoth sample, both AR proteins are knocked down. In the shGFP sample, both proteins are present. Experiments are performed by Dr. Chia-Chi Flora Huang. Figure created in https://BioRender.com

2.1.2 LN95 ChIP-seq Data

Published LN95 ChIP-seq data was provided by Dr. Myles Brown's group (GEO: GSE106559). Data analysis and data analysis was performed by Dr. Tunc Morova.

2.1.3 Classifying Androgen Receptor Binding Sites based on STARR-Seq Signals

Androgen binding sites from the STARR-Seq capture library are classified into three categories: flAR regulated regions, ARv7 regulated regions, and heterodimer (HD) regulated regions. This classification is performed based on fold changes of STARR-Seq signal scores between different treatments and p-values. For all classes, the log2 fold change (FC) threshold is 1 for upregulated regions and -1 for downregulated regions. Called STARR-Seq are also filtered to have an adjusted p-value (padj) threshold of 0.05. At the end, all classified regions have a padj value smaller than 0.05.

To identify flAR regulated regions, STARR-Seq signal scores from shARv7 DHT and EtOH treatments were compared. The log2FC value for each ARBS was calculated using the following formula:

$$log2FC = \log_2 \frac{shARv7{\sim}DHT}{shARv7{\sim}EtOH} \quad (2.1.1)$$

ARBS with a log2FC value smaller than -1 are defined as flAR-downregulated regions whereas greater than 1 are defined as flAR-upregulated regions.

To define ARv7 regulated regions, signal scores from shAR and shBoth -both EtOH tretaed- are used. Similarly, regions with a log2FC value smaller than -1 are defined as ARv7-downregulated regions whereas greater than 1 are defined as ARv7-upregulated regions. For this class, the log2FC values are calculated using the formula below:

$$log2FC = \log_2 \frac{shAR{\sim}EtOH}{shBoth{\sim}EtOH} \quad (2.1.2)$$

Lastly, shGFP and shBoth -both EtOH treated- are compared to identify HD regulated regions. Regions with a log2FC value smaller than -1 are defined as HD-downregulated regions whereas greater than 1 are defined as HD-upregulated regions. Below formula is used for each region to obtain log2FC values for this class:

$$log2FC = \log_2 \frac{shGFP{\sim}EtOH}{shBoth{\sim}EtOH} \quad (2.1.3)$$

Regions of each class are shown on separate volcano plots.

Filtering based on padj and log2FC calculations are performed using the R base functions and sqldf R package functions [57, 58]. Volcano plots of regions are generated using the package ggplot2 [59].

2.1.4 Androgen Receptor Variant 7 Specific Upregulated Regions

The overlaps between ARv7 upregulated regions, flAR upregulated regions and previously known 286 enhancers from the LNCaP cell line [42] were detected and plotted as Venn diagrams using functions from the R package eulerr [60].

2.1.5 Visualizing STARR-Seq Activity and Androgen Receptor Binding

Different classes of regions identified in chapter 2.1.3 are separately visualized to see enhancer activity and flAR and ARv7 protein binding.

flAR upregulated regions (n=138) were retrieved from the bed file containing all ARBS, using commands from bedtools [9] and basic Linux commands like “awk” and are saved in a new bed file. Matrices for the STARR-Seq and ChIP-seq heatmaps are calculated over respective bigWig files using the “computeMatrix” application from deepTools2 [62]. Regions are centered and a +/- 1500 base pair (bp) window around the center is included for calculations. Heatmaps are plotted by from the output matrix files using the “plotHeatmap” application from deepTools2 [62]. Regions are sorted in descending order based on the mean STARR-seq signal across the 3000 bp window, in the first heatmap. The following two heatmaps are plotted using the same region order, with the mean flAR and ARv7 ChIP-seq signal score across regions this time.

For ARv7 regulated regions, ARv7 upregulated and downregulated regions are retrieved from the ARBS bed file using bedtools commands and basic Linux commands [61] and are saved in new separate files. Matrices for the STARR-Seq heatmaps with all genotypes and treatments are calculated using respective bigWig files using the “computeMatrix” application from deepTools2 [62]. Again, these regions are centered and and a +/- 1500 base pair (bp) window around the center is used for calculations are visualizations. These two sets of regions are sorted separately in descending mean STARR-Seq signal across the 3000 bp window.

Lastly for HD regulated regions, upregulated and downregulated region lists are extracted from the ARBS bed file using the same commands as above [61]. These two sets of regions are first used to calculate matrix values for the STARR-Seq heatmaps with all genotypes and treatments as explained above [62]. Similarly, the regions are centered a 3000 bp window is

visualized. The two region sets are sorted separately in descending mean STARR-Seq signal across the windows on the final heatmaps.

2.1.6 Correlation Analyses

To understand the relationship between enhancer activity and flAR binding, a correlation analysis is performed. Also, a similar analysis is done to see the correlation between flAR and ARv7 binding.

To understand the correlation between enhancer activity and flAR in the 138 flAR upregulated regions compared to rest of the ARBS, the bed file containing all 4138 ARBS is divided into two separate bed files. 138 flAR upregulated regions are retrieved from the bed file containing all ARBS, using commands from bedtools [61] and basic Linux commands like “awk” and are saved in a new bed file. Remaining 4000 ARBS are saved in a different bed file.

The shGFP-DHT STARR-Seq bigWig file is used to compute mean scores across these sets of regions separately. Mean scores over bed files are calculated by using the bigWigAverageOverBed program [61]. The output bed file with the mean score column appended is used for the downstream correlation analysis. The flAR ChIP-seq bigWig is retrieved from the corresponding narrowPeak file, using bedtools and bedGraphToBigWig [61, 63]. Similarly, the output bigWig file is used to calculate the mean scores across regions by the bigWigAverageOverBed program [63]. Again, the output bed file is used for downstream analysis. Correlation coefficients and p-values for the correlation between STARR-Seq and flAR ChIP-seq scores are calculated using the Spearman correlation method that is available via ggpubr in R [64]. Results are plotted separately for the two sets of regions using the ggplot2 package [59].

Similar to the flAR ChIP-seq, ARv7 ChIP-seq narrowPeak file is converted to bigWig, using bedtools and bedGraphToBigWig [61, 63]. Then the output bigWig file is used to calculate the mean scores across regions by the bigWigAverageOverBed [63]. This time, the correlation between flAR and ARv7 ChIP-seq scores is analyzed. Correlation coefficients and p-values are calculated in the same way above [64] and the results are plotted via ggplot2 [59].

2.2 Enhancer-Promoter Interactions and Structural Variants

LNCaP ChIP-seq whole genome input library is prepared, LNCaP and LuCaP H3K27ac HiChIP experiments are performed by Dr. Ji-Heui Seo from Dr. Matthew L. Freedman's group at Dana-Farber Cancer Institute. LuCaP ChIP-seq whole genome input library is provided by Emma Minnee from Dr. Wilbert Zwart's group at Netherlands Cancer Institute. Additional higher coverage LNCaP ChIP-seq whole genome input library for benchmarking is taken from Seim et al (2017), with SRA accession 'SRX2541290' [65]. We collaborated with Umut Berkay Altıntaş from Lack Laboratory on how to perform the data analysis. Figures showing copy number ratios (CNR) are generated by Lack Laboratory member Lynn Masri. An overview of the analysis pipeline and main steps are summarized in Figure 2.2.

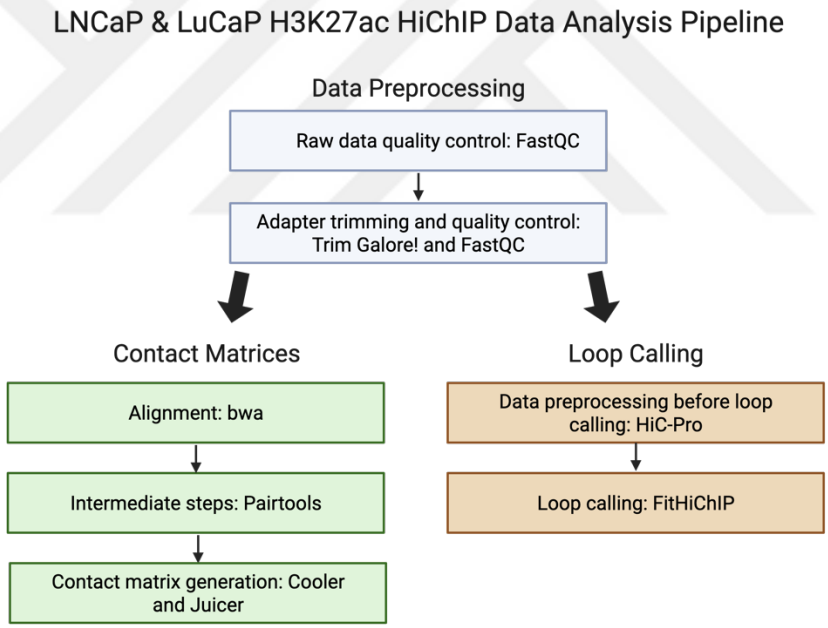

Figure 2.2: HiChIP data analysis pipeline. Quality control and adapter trimming are followed by two separate lines of pipelines: contact matrix generation and loop calling. Figure created in https://BioRender.com

2.2.1 LNCaP and LuCaP H3K27Ac HiChIP Data Preprocessing

Read quality control is performed for LNCaP and LuCaP paired end H3K27Ac HiChIP reads, using the FastQC software [66]. Adapter trimming for both datasets is done by using the Trim Galore tool default settings [67]. After trimming, a second round of quality control is done by running FastQC within Trim Galore.

2.2.2 Alignment and Intermediate Steps

Mapping to reference genome is performed using the Burrows-Wheeler Alignment tool [68]. The *Homo sapiens* genome assembly GRCh38 (hg38) is used as the reference genome for all alignments in this part.

Next, the following intermediate steps are performed separately for data from both cell lines. Output sam files are parsed using the parse tool from the command-line framework Pairtools [69]. Resulting parsed pairsam files are processed with the sorting tool from Pairtools [69]. PCR duplicates are removed using the Pairtools dedup tool [69]. Mapped pairs and unsorted reads are split by the Pairsam split tool [69]. Lastly, the final bam files are generated and indexed using SAMtools commands [70].

2.2.3 Contact Matrix Generation

For both cell lines, mapped.pairs files from the Pairtools split step are used as input and they are preprocessed using bgzip and pairix tools before matrix generation [70, 71]. Low resolution contact matrices showing all chromosomes together are generated using the Cooler tools [72]. Higher resolution, region-specific contact matrices are generated by the Juicer tool and visualized using the Juicebox desktop software [73, 74].

2.2.4 Valid Pair Generation

For valid pair generation to be used in loop calling, the HiC-Pro software is used [75]. Raw H3K27ac HiChIP sequencing data is provided as input from both cell lines separately. Input data file format is FASTQC, compressed by SAMtools bgzip [70]. HiC-Pro configuration file is edited to include the reference genome (hg38), genome size, Bowtie2 index and genome files. Bowtie2 is run for alignment within HiC-Pro [76].

2.2.5 Loop Calling

Loop calling is performed using the software FitHiChIP, separately for both cell line data [77]. HiC-Pro output valid pairs file is provided as input. In total, 8 different loop callings are performed, 4 for each cell line with the following varying parameters of bin size and false discovery rate (FDR) thresholds:

Bin size = 16 kb and FDR = 0.05

Bin size = 2 kb and FDR = 0.05

Bin size = 16 kb and FDR = 0.01

Bin size = 2 kb and FDR = 0.01

FitHiChIP configuration file is edited to include these parameters along with hg38 chromosome sizes and cell line ChIP-seq narrowPeak file paths. Peak-to-all loop calling is performed. Lower and upper loop distance thresholds are set to 5 kb and 7 Mb, respectively. Coverage bias regression is applied.

2.2.6 Amplicon Detection

For amplicon detection, the AmpliconSuite-pipeline software is used [78]. LNCaP and LuCaP ChIP-seq input libraries are used as input. Reads are aligned using BWA-MEM to the human reference genome hg38, run within the pipeline [79]. CNVkit is used to detect candidate amplicon regions with copy number variations (CNV) [80]. Amplicons are called from regions of interest by AmpliconArchitect [81]. Called amplicons are classified into different groups by AmpliconClassifier [78]. Main analysis steps before amplicon classification are shown in Figure 2.3.

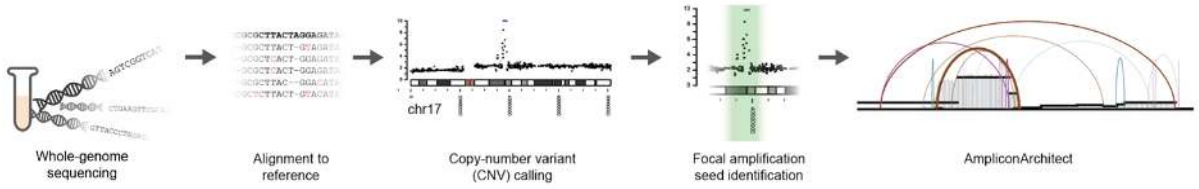

Figure 2.3: AmpliconSuite-pipeline. Whole genome sequencing data is aligned to reference genome. CNV calling is performed by CNVkit and amplicons are detected by AmpliconArchitect. Figure retrieved from AmpliconSuite-pipeline documentation https://github.com/AmpliconSuite/AmpliconSuite-pipeline/

Same pipeline with same steps is also run for the additional higher coverage LNCaP whole genome ChIP-seq input library. Four runs are concatenated to obtain a even higher coverage, and the concatenated data is provided as input to the pipeline. Coverages of each input library, i.e., low coverage LNCaP, low coverage LuCaP and high coverage LNCaP, are calculated using the following formula:

$$Input\ library\ coverage = \frac{(2 * (number\ of\ reads) * (length\ of\ reads))}{(length\ of\ the\ reference\ genome\ (hg38))} \quad (2.2.1)$$

Lastly, CycleViz is used to visualize the called ecDNA with a circular representation [82]. Cycles and graph output files from the AmpliconSuite-pipeline are used as input for CycleViz. Cycle ID and reference genome (hg38) are also provided as input. Other than these, no arguments are changed and default plotting program is run (CycleViz.py) [82].

Chapter 3
RESULTS

3.1 Changes in Enhancer Activity in Late-Stage Prostate Cancer

In this chapter, the effect of flAR and ARv7 on enhancer activity will be analyzed individually and together. Detecting enhancer activity caused by AR proteins separately has an importance since it reflects the enhancer activity changes between primary PCa and CRPC. Activity caused by flAR represents primary PCa, whereas ARv7 regulated enhancer activity reflects CRPC.

It is seen that our protein knockdown system works as expected for all genotypes in the Western blot provided in Figure 3.1.A, which are later used in STARR-Seq experiments. In the shAR sample, flAR is knocked down when protein overexpression is induced by Doxycycline treatment, and ARv7 is clearly expressed in both conditions. In the shV7 sample, ARv7 is knocked down and flAR is expressed. In the shBoth sample, under Doxycycline treatment, it is seen that both AR proteins are knocked down. Lastly in the shGFP sample, both proteins are present. In addition, samples group together according to their genotypes and DHT treatments, based on STARR-Seq signal scores of library regions which represent enhancer activity (Figure 3.1.B). In this section, results from identification of different classes of ARBS, investigation of enhancer activity regulation by flAR and ARv7, analysis of flAR upregulated enhancer activity and its relationship with ARv7, correlation between flAR and ARv7 binding are shared. In the last part, enhancer activity changes in flAR upregulated regions are investigated along with flAR and ARv7 binding.

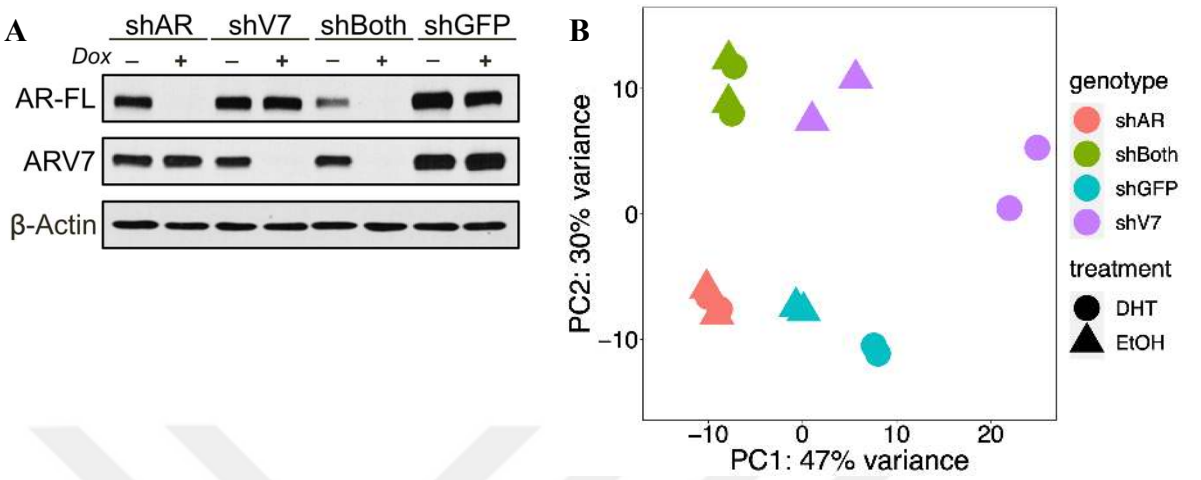

Figure 3.1: Western blot and principal component analysis (PCA) of samples (A) Western blot bands of flAR and ARv7 under each silencing RNA treatment. Western blot was performed by Dr. Chia-Chi Flora Huang (B) PCA plot of samples with different genotypes and treatments. Each genotype and treatment pair has 2 replicates. Genotype-treatment groups cluster together. PCA is done based on signal scores from each sample. (AR-FL: flAR, ARV7: ARv7).

3.1.1 Different Classes of Regions based on Enhancer Activity

To understand enhancer activity of our library regions caused by different AR proteins, we identified different classes of regions. Since we have STARR-Seq data from different samples in which flAR and/or ARv7 are knocked down, this allowed us to see the effect of these proteins individually and together on enhancer activity. Classifications are done based on fold changes in STARR-Seq signals of regions in samples with different genotypes and treatments.

Three different classes of regions are identified based on their enhancer activities, by comparing fold changes of STARR-Seq signals as explained in the Methods (Chapter 2.1.3). By comparing the shARv7-DHT sample to its EtOH treated counterpart, flAR regulated regions are identified. In the shARv7 treatment, only flAR is present in cells (Figure 2.1). In DHT treated samples, the observed enhancer activity is expected to be caused by flAR, in comparison to the EtOH treated sample. Out of 4138 ARBS, 138 regions show upregulated enhancer activity by flAR. Similar to published work, there are no flAR downregulated regions [42]. Rest of the regions (4000) remain neither downregulated nor upregulated (Figure 3.2.A). Next, to understand the regulation mediated by ARv7, shflAR-EtOH and

shBoth-EtOH samples are compared. EtOH treated samples are preferred since ARv7 is not activated by DHT binding due to its lack of LBD. In the shflAR sample there is only ARv7 and in shBoth there is no AR protein (Figure 2.1). There are 11 regions that show upregulated enhancer activity caused by ARv7 and 68 regions are ARv7 downregulated while the remaining 4059 regions fall into neither of these categories (Figure 3.2.B). Lastly, to identify HD regulated regions, shGFP-EtOH and shBoth-EtOH samples are compared, where the first one contains both flAR and ARv7, and the latter one has no AR protein (Figure 2.1). Consequently, 12 HD upregulated, and 98 HD downregulated regions are detected where rest of the regions (4028) appear unaffected (Figure 3.2.C).

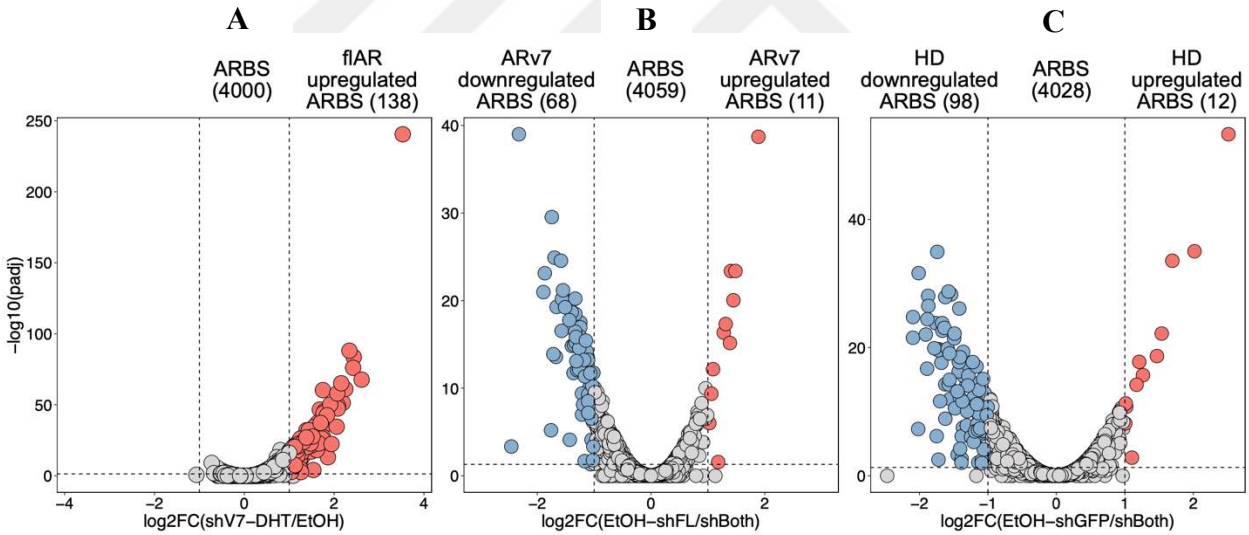

Figure 3.2: Identification of different classes of ARBS based on enhancer activity (A) flAR regulated regions: Identified from the comparison between shARv7 DHT and shARv7 EtOH samples (B) ARv7 regulated regions: Identified from the comparison between shflAR EtOH and shBoth EtOH samples (C) HD regulated regions: Identified from the comparison between shGFP EtOH and shBoth EtOH samples.

3.1.2 Enhancer Activity Caused by Androgen Receptor Variant 7

In the classification above, 11 ARv7 upregulated and 68 ARv7 downregulated ARBS are identified. To understand the enhancer activity of these regions separately under different treatments, we checked the mean STARR-Seq signal across them (Figure 3.3). The 11 ARv7 upregulated regions show a slight increase in activity in the shGFP-DHT sample compared to its EtOH treated counterpart, where both flAR and ARv7 are present (Figure 3.3.A). On the other hand, the 68 downregulated regions show a slightly decreased activity in the DHT treated sample. In the shARv7 sample, ARv7 downregulated regions show a higher enhancer activity both under EtOH and DHT treatments, as expected. Upregulated regions show an increased activity in the DHT treated sample, compared to the shARv7-EtOH sample (Figure 3.3.B). Both ARv7 upregulated and downregulated regions show a similar levels of enhancer activity in both EtOH and DHT treated shflAR samples, where upregulated regions have a higher activity (Figure 3.3.C). In the shBoth samples ARv7 downregulated regions show a much higher activity (Figure 3.3.D). Identification of ARv7 regulated regions is done comparing the shflAR and shBoth samples. Firstly, these results show us there is no difference in enhancer activity of these regions between EtOH and DHT treatments for these samples. This supports our approach to use EtOH samples for identification since ARv7 lacks the LBD. Next, supporting our definition of these regions, it is seen that ARv7 upregulated regions have a much higher enhancer activity in the shflAR sample where only ARv7 is present, compared to the shBoth sample where both AR proteins are absent. However, it is also seen that they have comparable levels of activity in shARv7 and shGFP samples, which brings us to investigate the specificity of the observed activity's dependence on ARv7.

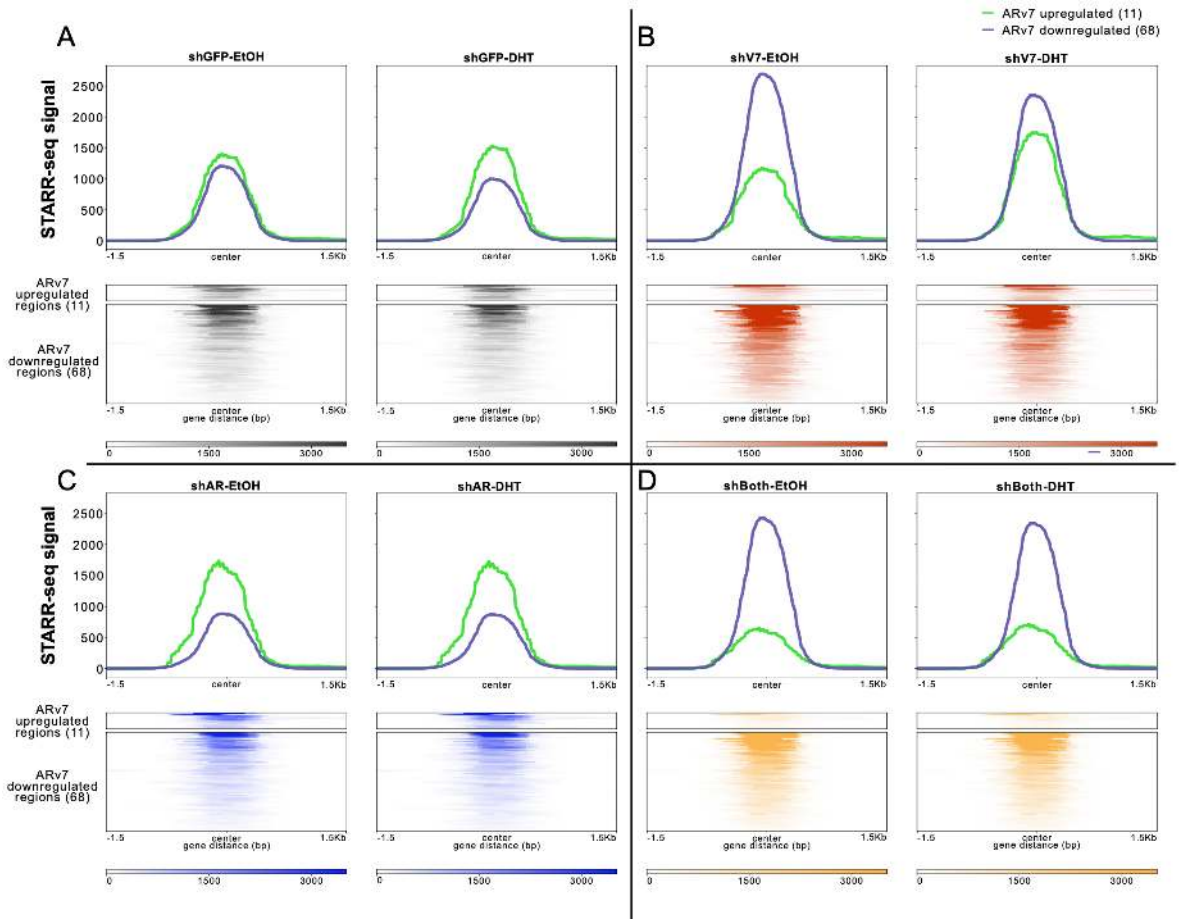

Figure 3.3: Enhancer activity of ARv7 regulated regions. (A) Enhancer activity of ARv7 upregulated and downregulated regions in shGFP EtOH and DHT samples (B) Enhancer activity of ARv7 upregulated and downregulated regions in shARv7 EtOH and DHT samples (C) Enhancer activity of ARv7 upregulated and downregulated regions in shflAR EtOH and DHT samples (D) Enhancer activity of ARv7 upregulated and downregulated regions in shBoth EtOH and DHT samples.

To understand if the upregulation observed in the 11 ARv7 upregulated ARBS is specifically due to ARv7, we checked the overlap between those regions, 286 previously known LNCaP enhancers [42] and 138 flAR upregulated regions. This analysis revealed that 5 of those ARBS overlap with LNCaP enhancers, 3 regions overlap with flAR upregulated ARBS and 3 of them overlap with both groups (Figure 3.4). Therefore, even though we observe an upregulation in enhancer activity of these 11 regions in the presence of only ARv7, these overlaps suggest that this upregulation is not specifically caused by ARv7. So, there are no ARv7-specific upregulated ARBS detected in this analysis. Also given the significantly greater number of ARv7-downregulated CREs we identified, our findings suggest that ARv7, unlike flAR, is not a strong transcriptional activator.

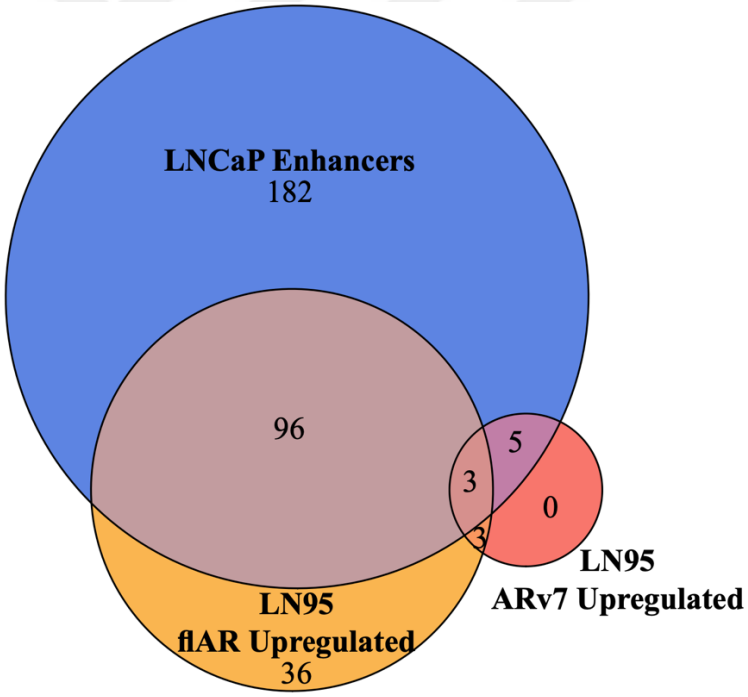

Figure 3.4: ARv7-specific upregulation. The 11 ARv7 upregulated ARBS either overlap with 286 LNCaP enhancers or 138 flAR upregulated ARBS. There are no ARv7-specific upregulated ARBS.

3.1.3 Heterodimer Regulated Enhancer Activity

In the classification of ARBS, 12 HD upregulated and 98 HD downregulated regions are identified. HD regulation expresses the regulation of enhancer activity by both flAR and ARv7 simultaneously. To see the enhancer activity of these regions separately under different treatments, we checked the mean STARR-Seq signal across them (Figure 3.5). The 12 HD upregulated regions and 98 downregulated regions all show a similar level of enhancer activity in the shGFP sample under both EtOH and DHT treatments (Figure 3.5.A). In the shGFP sample, both flAR and ARv7 are present. In shARv7 and shflAR samples, we see an increased activity of HD downregulated regions (Figure 3.5.B and Figure 3.5.C), compared to upregulated regions. This can possibly be explained by the loss of ARv7 and flAR in those samples, respectively. Loss of them causes a decrease in enhancer activity of HD upregulated regions since these regions are upregulated by flAR and ARv7 together. In the shARv7 DHT sample, we see a slight increase in HD upregulated region activity compared to its EtOH treated counterpart. This again demonstrates the dependance of flAR on DHT. Since in the shARv7 sample only flAR is expressed, it is able to recover some amount of the enhancer activity in presence of DHT (Figure 3.5.B). We do not observe this increase in the shflAR sample. Both upregulated and downregulated samples show a comparable activity under EtOH and DHT treatments (Figure 3.5.C). Lastly, in the shBoth sample, HD downregulated regions have much higher enhancer activity compared to upregulated regions, regardless of DHT treatment.

These regions are defined comparing EtOH treated shGFP and shBoth samples. The mean signal profiles in Figure 3.5 supports our definition. HD upregulated regions have a higher activity in the shGFP sample compared to the shBoth sample. In contrast, HD downregulated regions have a higher activity in the shBoth sample. However, it is also seen that in the shGFP sample, where both flAR and ARv7 are expressed, both HD upregulated and downregulated regions show similar levels of enhancer activity. This suggests that our HD upregulated and downregulated definitions are valid between shGFP and shBoth samples but not within the shGFP sample.

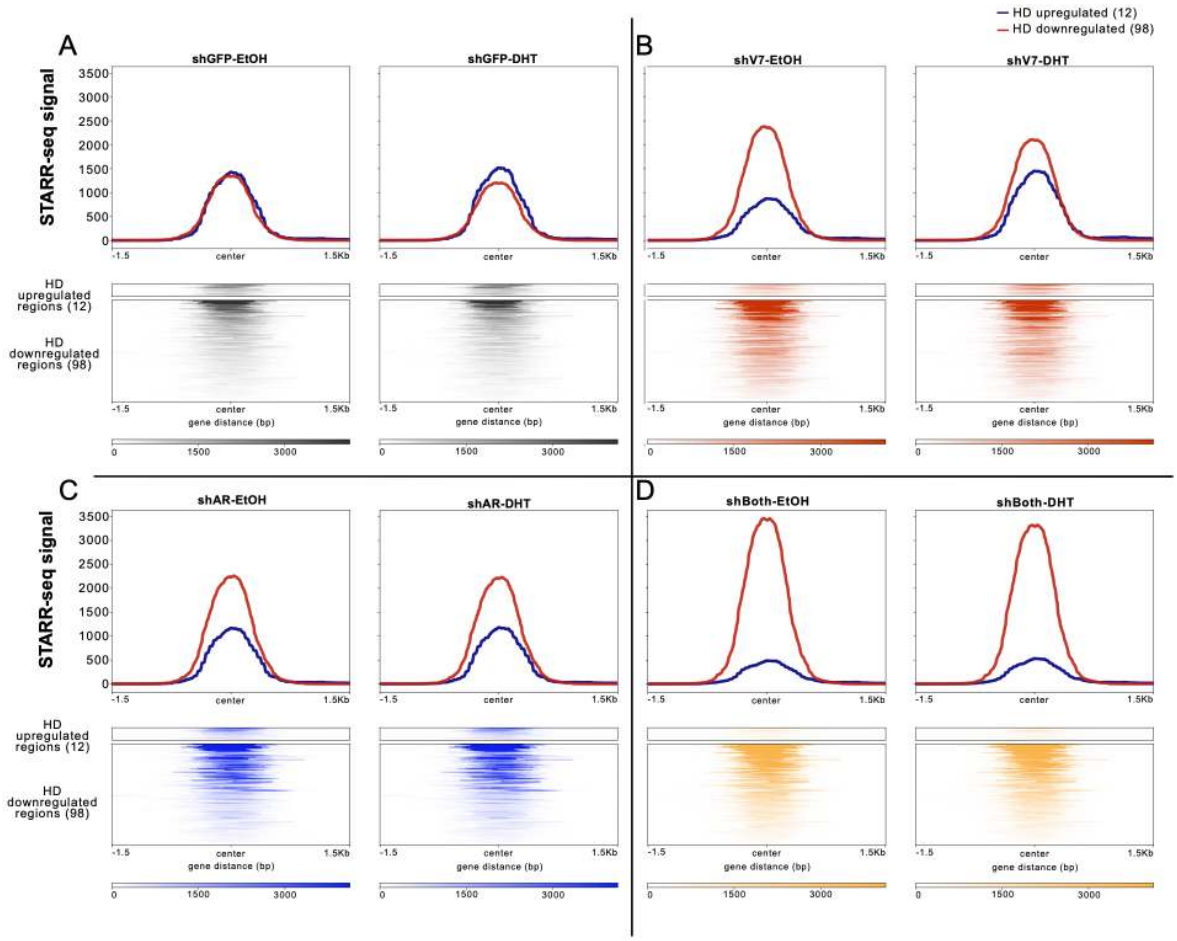

Figure 3.5: Enhancer activity of HD regulated regions. (A) Enhancer activity of HD upregulated and downregulated regions in shGFP EtOH and DHT samples (B) Enhancer activity of HD upregulated and downregulated regions in shARv7 EtOH and DHT samples (C) Enhancer activity of HD upregulated and downregulated regions in shflAR EtOH and DHT samples (D) Enhancer activity of HD upregulated and downregulated regions in shBoth EtOH and DHT samples.

3.1.4 Full Length Androgen Receptor Upregulated Enhancer Activity

In the ARBS classification in Chapter 3.1.1, 138 flAR upregulated regions are identified. No flAR downregulated regions are detected similar to previous work [42].

To see he enhancer activity of these regions under different treatments, we checked the mean STARR-Seq signal across them (Figure 3.6.A). And since we did not detect any ARv7-specific upregulated regions, we decided to proceed with investigating ARv7's effect on flAR upregulated regions. To do this, enhancer activity and AR binding (STARR-Seq and ChIP-seq signals) of 138 flAR upregulated regions are visualized. Visualization is performed for shGFP, shARv7 and shflAR samples for both EtOH and DHT treatments (Figure 3.6). Enhancer activity of flAR upregulated regions is slightly higher in the DHT treated shGFP sample, compared to the EtOH treated one (Figure 3.6.A). Higher enhancer activity is observed in shARv7-DHT sample, compared to its EtOH treated counterpart (Figure 3.6.A). As expected, in the shflAR samples, there are low levels of activity for these regions and no difference is observed between EtOH and DHT treatments (Figure 3.6.A). When flAR binding is investigated, it is observed to be higher in the DHT treated shGFP sample compared to EtOH treatment (Figure 3.6.B). There is a similar difference in flAR binding in the shARv7 samples. Lastly, ARv7 binding is higher in the shGFP DHT sample than in the EtOH treated shGFP sample (Figure 3.6.C). ARv7 binding levels are comparable in shflAR DHT and EtOH samples.

These regions are defined by comparing the shARv7 DHT sample to its EtOH counterpart. The fold change in enhancer activity between these samples are apparent in Figure 3.6.A. In addition, when the shGFP-DHT sample is analyzed, although enhancer activity is slightly higher than the EtOH treated counterpart, it is much lower than the shARv7-DHT sample. Also, interestingly, enhancer activity of these regions in the shGFP-DHT sample is close to shflAR-DHT. This suggests that ARv7 may have a repressive effect on flAR since these regions show higher enhancer activity in the absence of ARv7. This is further supported by the binding signals (Figure 3.6.B and Figure 3.6.C). In the shGFP-DHT sample, both flAR and ARv7 are present and binding signal of flAR is higher compared to flAR binding in the shARv7-DHT sample. Also, high ARv7 binding is observed in this sample.

To investigate these results more in detail, we checked the correlation between enhancer activity and flAR binding, and flAR and ARv7 binding.

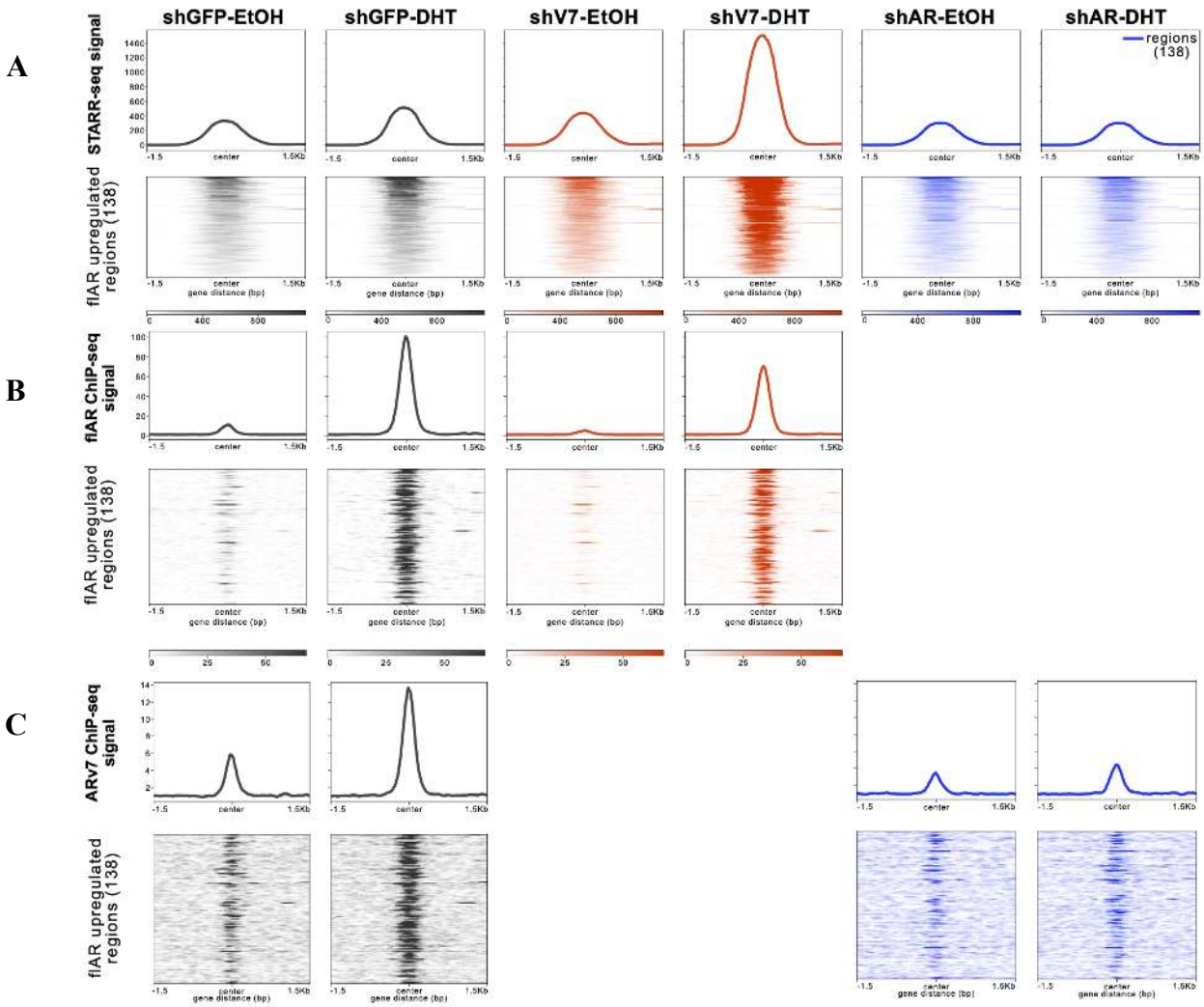

Figure 3.6: STARR-Seq and ChIP-seq signals of flAR upregulated regions in different samples (n=138). (A) Enhancer activity in shGFP, shARv7 and shflAR samples (B) flAR binding in shGFP and shARv7 samples (C) ARv7 binding in shGFP and shflAR samples.

3.1.5 Enhancer Activity and Androgen Receptor Binding

To understand the correlation between enhancer activity of 138 flAR upregulated regions and flAR binding, we divided the full set of ARBS into two sets: 138 flAR upregulated regions and 4000 remaining ARBS. Then, the correlation between enhancer activity of these regions and flAR binding is checked separately. By separating these regions, we aim to create a background set composed of remaining ARBS, as a comparison to flAR upregulated regions. These regions are analyzed separately to see the difference between abovementioned correlation in flAR upregulated regions compared to remaining ARBS.

It is seen that in the DHT treated shGFP sample, there is a statistically significant (p-value<0.05) correlation between enhancer activity and flAR binding for both sets of regions (Figure 3.7). However, for the 138 flAR upregulated regions, we observe a stronger correlation, compared to remaining 4000 ARBS. Although the correlation coefficient is 0.24 for the flAR upregulated regions (Figure 3.7.A), it still appears stronger compared to remaining ARBS which has a correlation coefficient of 0.036 (Figure 3.7.B). This is an expected result since the set of remaining ARBS contains regions that are ARv7, HD regulated or regions that are not found to be regulated by any AR proteins in this analysis.

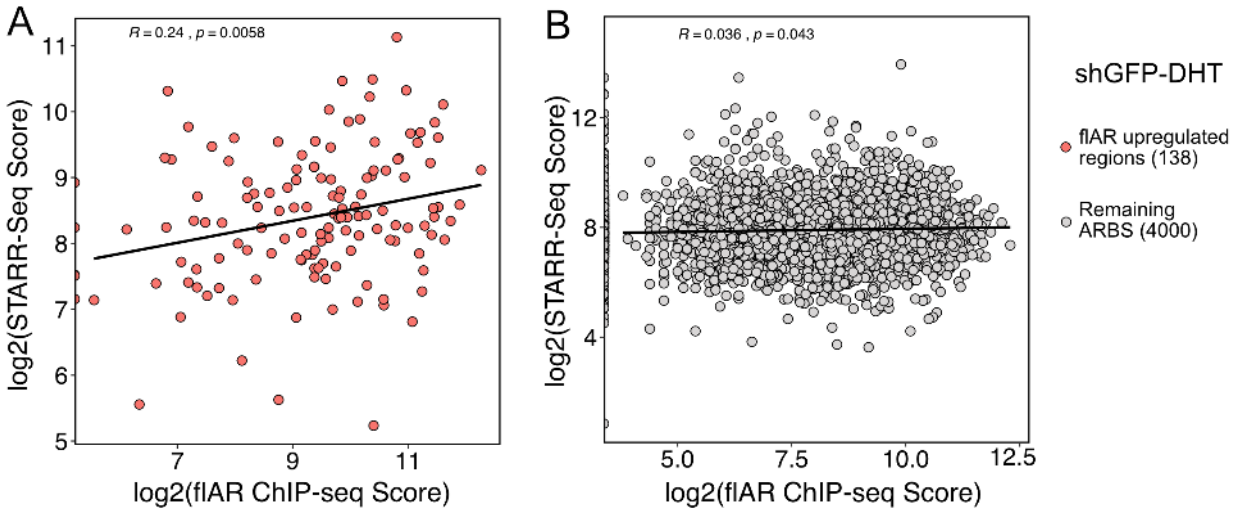

Figure 3.7: Correlation between enhancer activity and flAR binding. (A) Correlation in 138 flAR upregulated regions in shGFP DHT treatment (B) Correlation in remaining 4000 ARBS in shGFP DHT.

To further understand whether flAR and ARv7 act on these regions together, we checked the correlation between binding scores of these proteins. Again, the 138 flAR upregulated regions and 4000 remaining ARBS are analyzed separately, and it is seen that there is a statistically significant correlation between flAR and ARv7 binding for both sets of regions (Figure 3.8). This correlation is stronger for flAR upregulated regions with a correlation coefficient of 0.76 (Figure 3.8.A), compared to remaining ARBS which has a correlation coefficient of 0.45 (Figure 3.8.B).

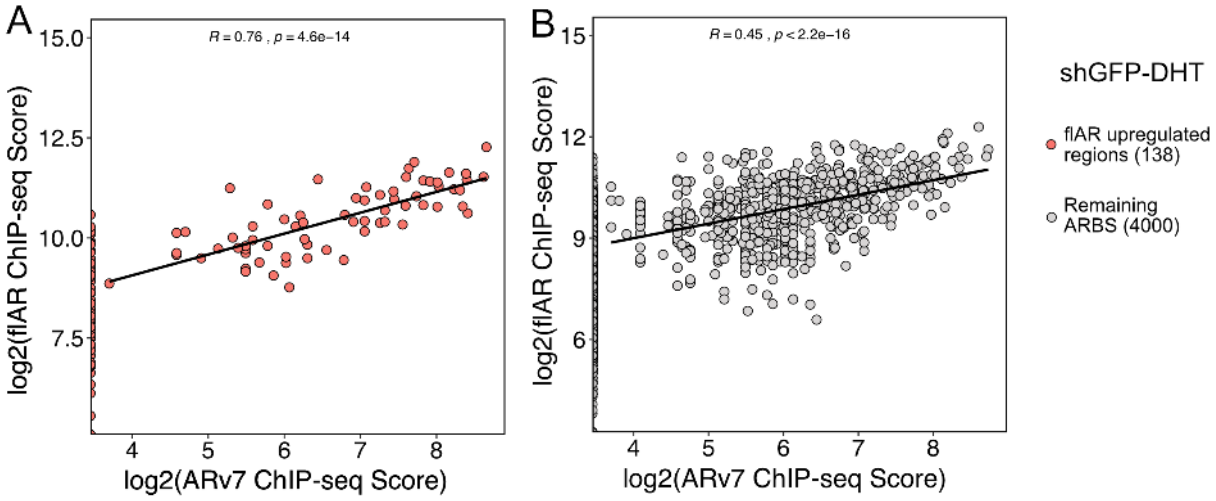

Figure 3.8: Correlation between flAR and ARv7 binding. (A) Correlation in 138 flAR upregulated regions in shGFP DHT treatment (B) Correlation in remaining 4000 ARBS in shGFP DHT.

These results together demonstrate that for flAR upregulated regions, enhancer activity and flAR binding correlate. More importantly, the strong correlation between flAR and ARv7 binding in these regions further supports that ARv7 may have a repressive effect on flAR's upregulatory role on these regions, combined with results from the previous section (Figure 3.6). Since these analyses are done in the shGFP-DHT sample, both AR proteins are expressed, and these regions have a higher enhancer activity in the shARv7-DHT sample where only flAR is expressed (Figure 3.6).

3.2 Changes in Enhancer-Promoter Interactions and Structural Variants in Late-Stage Prostate Cancer

In this chapter, 3D chromatin interactions in LNCaP and LuCaP are analyzed, and major differences are investigated. Interaction differences between these cell lines have importance since they demonstrate different stages of PCa. LNCaP represents primary PCa, whereas the LuCaP cell line reflects CRPC. Understanding differences in interaction profiles has a potential to shed light on possible mechanisms behind the transition from primary PCa to CRPC.

3.2.1 Overview: Genome-Wide Contact Matrices

As a result of H3K27ac HiChIP data analysis, genome-wide differences are observed in LNCaP and LuCaP cell lines (Figure 3.9). In the genome-wide view, big interchromosomal interactions are visible. In LuCaP, chromosome 8 has both intra- and interchromosomal interactions that are absent in LNCaP (Figure 3.9.B). Also, chromosome X shows novel interchromosomal interactions in LuCaP, compared to LNCaP (Figure 3.9.B). A more informative and detailed view of these chromosomes and some regions of interest are shared in the next subchapter.

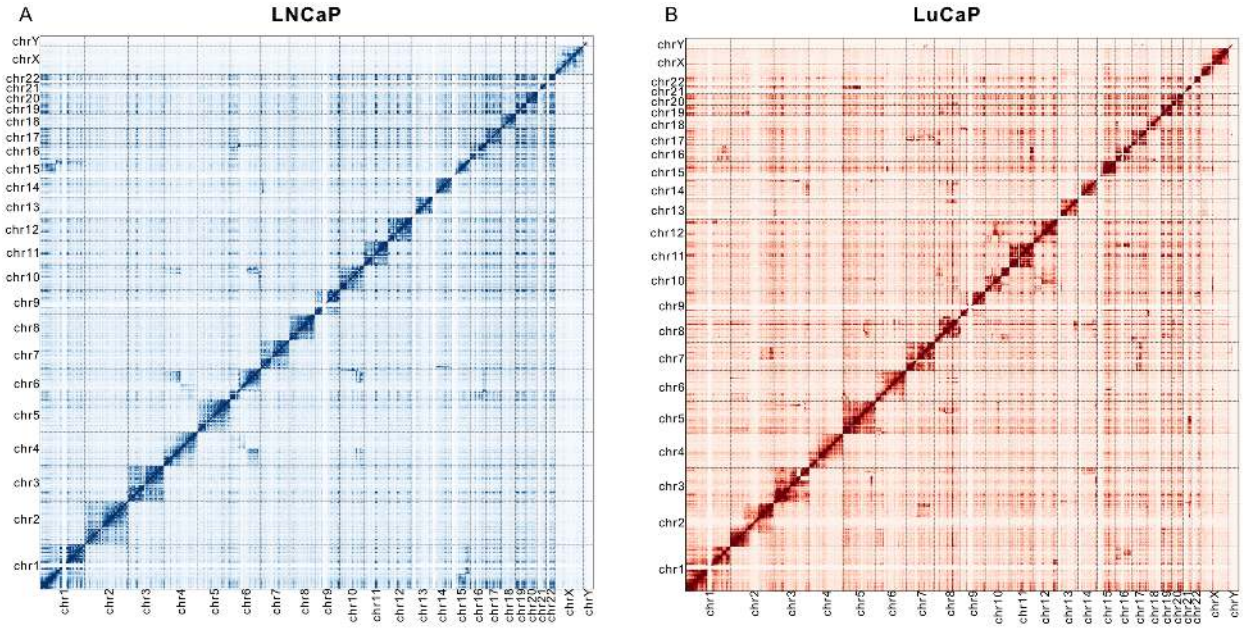

Figure 3.9: Overview of genome-wide 3D interactions (A) Interactions across all chromosomes in LNCaP (B) Interactions across all chromosomes in LuCaP. Both contact matrix resolutions are 640 kb.

3.2.2 Closer Look: Interactions of Regions of Interest

In this part, closer analyses of some regions are shared to understand major interaction differences between LNCaP and LuCaP cell lines. From the general view, chromosome X and the area around *AR* seems to bear differential interactions between the two cell lines. Another differentially interacting locus is the proto-oncogene *MYC* on chromosome 8, which is a candidate to harbor big SVs. Lastly, a closer look of the area around Kallikrein-related peptidase 3 (*KLK3*) on chromosome 19 is provided as a neutral example that has comparable interaction profiles for the two cell lines.

First, interactions on chromosome X, around *AR* are investigated closely. When the two cell lines are compared, it is seen that *AR* has enriched interactions in LuCaP that are not observed in LNCaP (Figure 3.10). Enriched interaction intensity also increases in areas closer to *AR* compared to its surroundings in LuCaP. These results suggest that *AR* has interactions with other chromosome X loci in LuCaP that are absent in LNCaP. CNR plots across the investigated window are also aligned with contact matrices and an amplification is seen in the area that has enriched interactions in LuCaP (Figure 3.10). In contrast, the whole investigated window shows mild deletion in the LNCaP cell line.

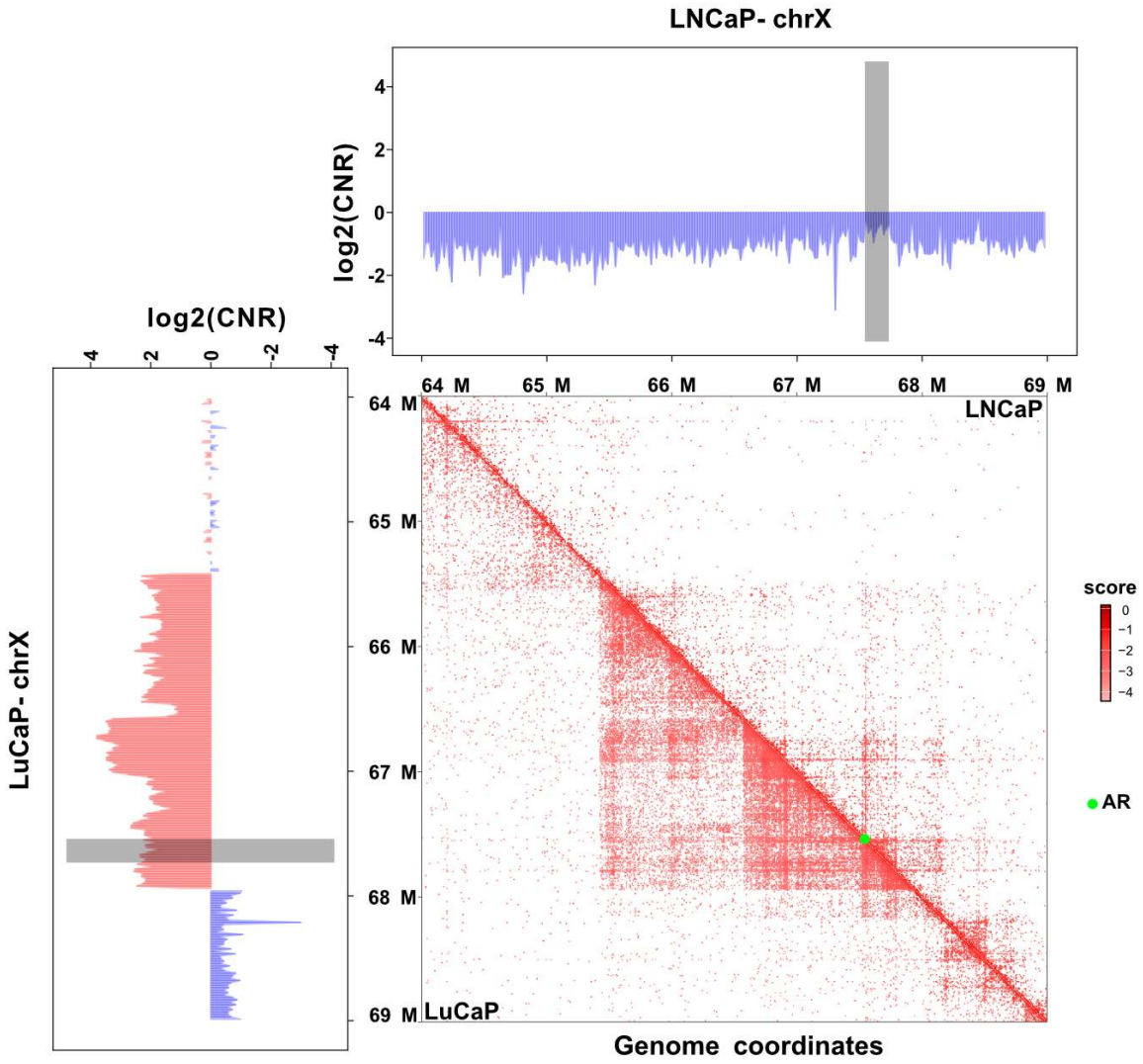

Figure 3.10: LNCaP and LuCaP interactions and aligned CNR around *AR* on chromosome X. Upper right triangle of the matrix is from LNCaP and lower left triangle is from LuCaP. Contact matrix resolution is 8 Kb. Visualized window is chromosome X: 64 Mb-69 Mb. *AR* is highligted with a green dot. Reference genome hg38 is used. *AR*'s hg38 coordinates: chrX:67,544,021-67,730,619.

Next, the area around a known proto-oncogene *MYC* on chromsome 8 is investigated closely. It is seen that in LuCaP, there are big deleted regions around *MYC*, that are not present in LuCaP (Figure 3.11). These deletions in LuCaP are also clearly seen in CNR graphs, consistent with the contact matrix. In addition to deletions, *MYC* shows enriched interactions with the area between 125 – 125.5 Mb in LuCaP, which is absent in LNCaP (Figure 3.11). Also, the close area around *MYC*, between two big deletions, seems to be amplified.

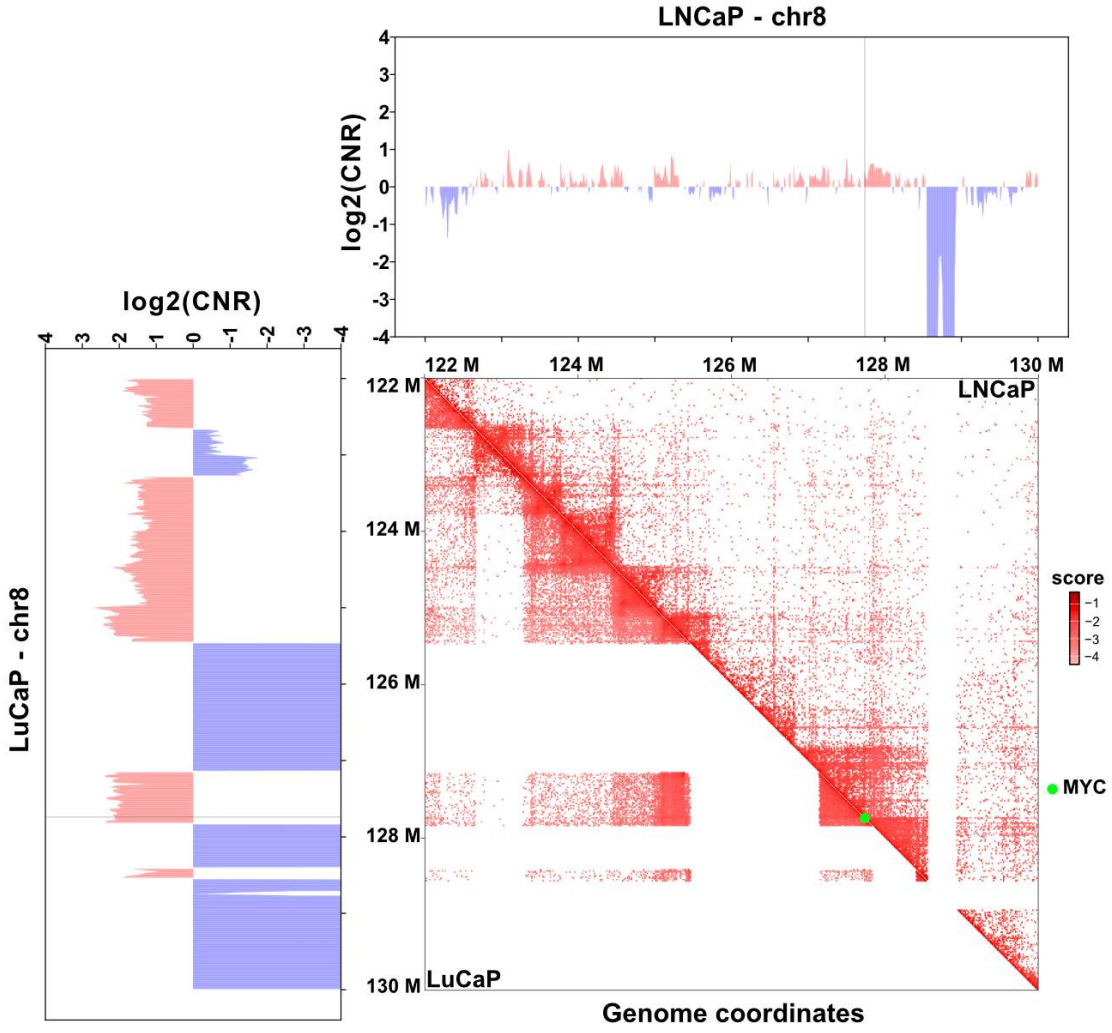

Figure 3.11: LNCaP and LuCaP interactions and aligned CNR around *MYC* on chromosome 8. Upper right triangle of the matrix is from LNCaP and lower left triangle is from LuCaP. Contact matrix resolution is 16 Kb. Visualized window is chromosome 8: 122 Mb-130 Mb. *MYC* is highligted with a green dot. Reference genome hg38 is used. *MYC*'s hg38 coordinates: chr8:127,735,434-127,742,951.

Lastly, the interactions of *KLK3* on chromosome 19 are investigated. *KLK3* is also known as prostate-specific antigen (PSA) [83]. Compared to *AR* and *MYC*, the area around *KLK3* shows a comparable interaction profile in both cell lines (Figure 3.12). There are 2 deleted common regions in both cell lines around 48 Mb and 50 Mb, which are also supported by CNR graphs (Figure 3.12).

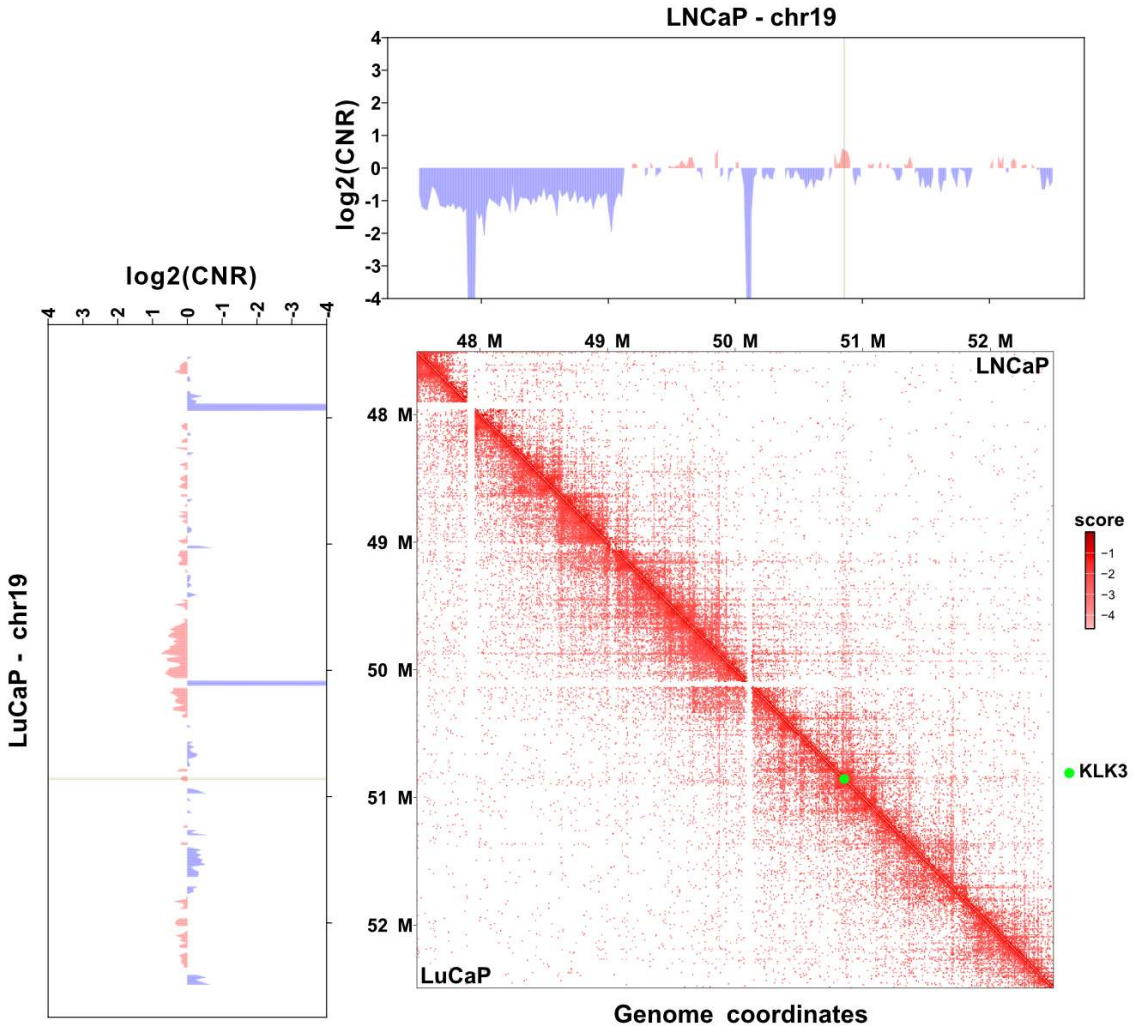

Figure 3.12: LNCaP and LuCaP interactions and aligned CNR around *KLK3* on chromosome 19. Upper right triangle of the matrix is from LNCaP and lower left triangle is from LuCaP. Contact matrix resolution is 8 Kb. Visualized window is chromosome 19: 47.5 Mb-52.5 Mb. *KLK3* is highligted with a green dot. Reference genome hg38 is used. *KLK3*'s hg38 coordinates: chr19:50,854,915-50,860,764.

3.2.3 Called Loops

Differential enriched interactions observed in contact matrices in LuCaP are also supported by called loops from the same HiChIP data.

Interaction differences between the two cell lines are especially striking in chromosome 8 around the MYC loci, with big deletions and intense enriched interactions (Figure 3.11). Therefore, interactions between *MYC* and the area between 125-125.5 Mb on chromosome 8 are closely investigated also in the loop calling. Consequently, and consistently with the findings from contact matrices, in LuCaP, there is enriched looping between *MYC* and the above-mentioned area (Figure 3.13). These loops are absent in LNCaP.

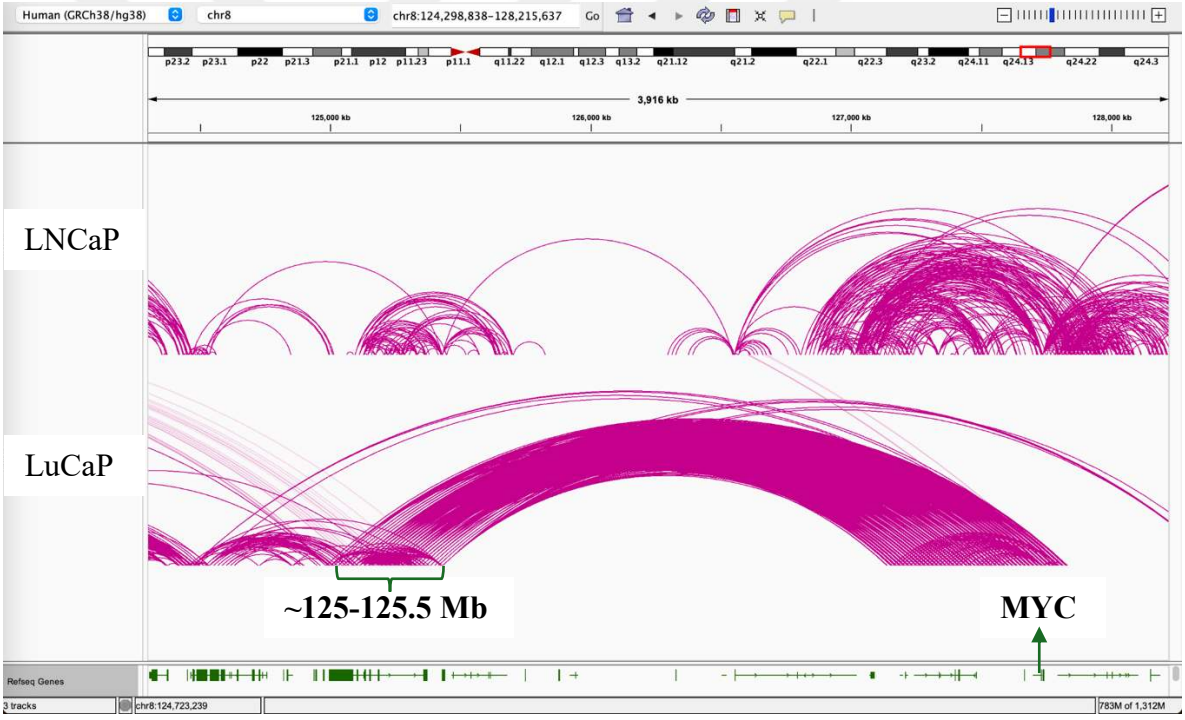

Figure 3.13: Differential looping of *MYC* in LuCaP, compared to LNCaP. Loop calling is performed with FitHiChIP. Bin size is 16 Kb. Lower loop distance threshold is 5 kb. Upper loop distance threshold is 7 Mb. FDR threshold is 0.01. Differential loops between *MYC* and chromosome 8: 125-125.5 Mb are observed in LuCaP, compared to LNCaP, consistent with contact matrices.

3.2.4 Benchmarking between Input Libraries with Different Coverages

To understand the effect of input library coverages on amplicon detection by the AmpliconSuite-pipeline, a benchmarking is performed comparing low and high coverage LNCaP ChIP-seq input libraries. Four runs are concatenated from the LNCaP data received from Seim et al (2017) [65]. LNCaP input library coverage is calculated to be ~76x for the concatenated data, using the formula (2.2.1) given in the Methods part. Coverage for the low coverage LNCaP input library is calculated as ~2.9x, using the same formula. Lastly, the LuCaP ChIP-seq input library coverage is calculated to be ~3.1x by the same formula. Amplicon detection results are compared between high (76x) and low (2.9x) LNCaP input libraries and it is seen that neither of them has any amplicons detected.

3.2.5 Detected Amplicons and Extrachromosomal DNA

Lastly, since there are many differential interactions in LuCaP that potentially harbor SVs and other genomic rearrangements, we proceeded with analyzing the amplicon content in the two cell lines. The aim is to systematically detect amplicons including linear amplicons and ecDNAs in both cell lines. Especially ecDNAs have an importance since they mediate new enhancer-promoter interactions and allow enhancer hijacking [55].

The AmpliconSuite-pipeline is run with input ChIP-seq library data from both cell lines. The pipeline detected 16 amplicons in LuCaP (Table 3.1), and no amplicons are detected in LNCaP. Out of detected LuCaP amplicons, 4 are later classified as invalid by the pipeline. Out of 12 remaining amplicons, 11 are classified as linear. One of the amplicons, “Amplicon3” is classified as harboring a cyclic amplicon and that is found to be an ecDNA (Table 3.1).

Table 3.1: LuCaP amplicon classification profiles. Full list of amplicons detected in LuCaP ChIP-seq input library. Initially, 16 amplicons are detected. 4 of them are later classified as invalid. Out of 12 remaining, 11 of them are linear amplicons. Amplicon3 harbors a cyclic amplicon and is marked positive for ecDNA presence.

LuCaP_ld_amplicon_classification_profiles

sample_name	amplicon_number	amplicon_decomposition_class	ecDNA+	BFB+	ecDNA_amplicons
LuCaP_ld	amplicon10	No amp/Invalid	None detected	None detected	0
LuCaP_ld	amplicon11	Linear	None detected	None detected	0
LuCaP_ld	amplicon12	Linear	None detected	None detected	0
LuCaP_ld	amplicon13	Linear	None detected	None detected	0
LuCaP_ld	amplicon14	Linear	None detected	None detected	0
LuCaP_ld	amplicon15	Linear	None detected	None detected	0
LuCaP_ld	amplicon16	Linear	None detected	None detected	0
LuCaP_ld	amplicon1	No amp/Invalid	None detected	None detected	0
LuCaP_ld	amplicon2	No amp/Invalid	None detected	None detected	0
LuCaP_ld	amplicon3	Cyclic	Positive	None detected	1
LuCaP_ld	amplicon4	Linear	None detected	None detected	0
LuCaP_ld	amplicon5	Linear	None detected	None detected	0
LuCaP_ld	amplicon6	Linear	None detected	None detected	0
LuCaP_ld	amplicon7	Linear	None detected	None detected	0
LuCaP_ld	amplicon8	No amp/Invalid	None detected	None detected	0
LuCaP_ld	amplicon9	Linear	None detected	None detected	0

Amplicon3's composition can be seen in Figure 3.14. It has 4 amplified oncogenes: PTK2B, TRIB1, PVT1, MYC (Figure 3.14). This is consistent with the results from HiChIP data analysis, where we observed differential 3D interactions around MYC on chromosome 8 in LuCaP. The gained interaction in LuCaP between MYC and 125-125.5 Mb is also visible in Amplicon3 as an interaction between MYC and TRIB1. Also consistent with previous results, a deletion is detected in the area between MYC and TRIB1 by the AmpliconSuite-pipeline (Figure 3.14).

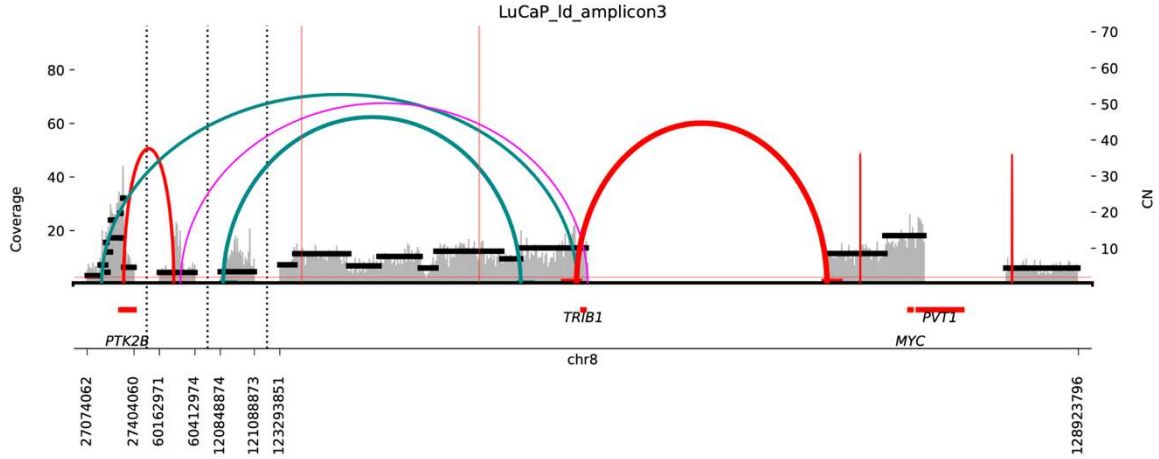

Figure 3.14: LuCaP Amplicon3, chromosome 8, AmpliconSuite-pipeline output. Amplified oncogenes: PTK2B, TRIB1, PVT1, MYC. Average amplified copy count = 7.92. Gray vertical histogram bars represent coverage. Black horizontal lines represent binning of the intervals based on coverage and copy numbers. Y-axis position of lines represent the copy numbers. Colored arcs represent read pair clusters, colors change based on orientation of the read. Thickness of the arcs qualitatively represent the amount of paired-end read support. Vertical colored lines correspond to connections to source vertex. Amplified oncogenes are annotated.

The circular structure of the detected Amplicon3 ecDNA can be found in Figure 3.15. It hosts the genes CASC11, MYC and PVT1, which are all from chromosome 8.

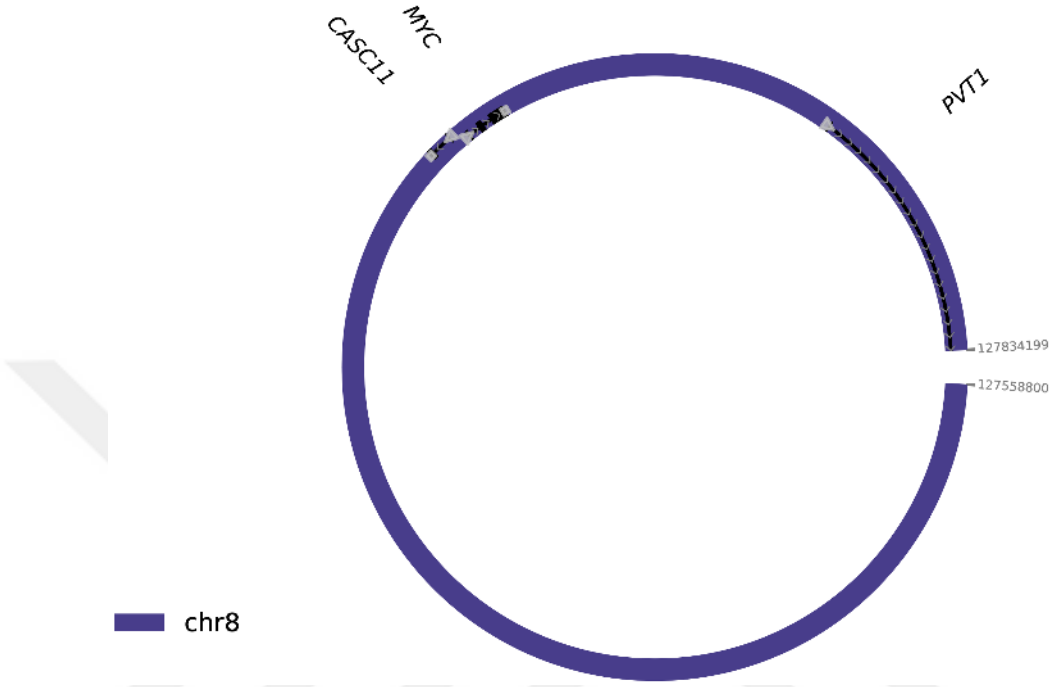

Figure 3.15: Detected ecDNA in Amplicon 3. It hosts the genes CASC11, MYC and PVT1, all from chromosome 8. Copy count of this detected ecDNA is reported as 7.89.

Results from this part collectively support that there are SVs in the LuCaP cell line which are not present in LNCaP. This demonstrates that there are genomic rearrangements in CRPC that are absent in primary PCa. Presence of an ecDNA which harbors oncogenes in LuCaP is noteworthy. These are important findings since these rearrangements -especially the detected ecDNA- may play a role in underlying mechanisms employed in progression from primary PCa to CRPC, like enhancer hijacking. These spots can be used as therapy targets after careful downstream analyses.

Chapter 4
DISCUSSION

It is shown in the literature that CREs play an important role in cancer initiation and progression [22,43,49,50,51]. Enhancer activity is very important in gene expression regulation, and it differs in cancers compared to healthy counterparts. It is previously shown by many studies that presence of AR variants is associated with resistance to AR signaling pathway inhibitor therapies in PCa. Expression of ARv7 correlates with this resistance in CRPC. Also, TAD composition and enhancer-promoter interactions are altered in cancers, including PCa. Presence of other SVs such as ecDNAs that allow enhancer hijacking events are shown to increase oncogenic activity of involved genes. These events are associated with late-stage disease and poorer prognosis. So, in this work we hypothesize that resistance in CRPC is partly driven by alterations in enhancer activity. Enhancer activity changes in CRPC are investigated via analyzing enhancer activity caused by flAR and ARv7, and differential 3D interactions in cell lines that represent primary PCa and CRPC.

In chapter 3.1, we analyzed STARR-Seq data from the LN95 cell line with a conditional isoform knockdown of flAR and Arv7. We detected 138 flAR upregulated, 11 ARv7 upregulated and 12 HD upregulated ARBS. Also, 68 ARv7 downregulated, and 98 HD downregulated ARBS are found. For the 138 flAR upregulated regions, ARv7 seems to exert a repressive effect accompanied by both flAR and ARv7 binding. The 11 ARv7 upregulated regions overlap with either flAR upregulated regions or previously known LNCaP enhancers. So, we did not detect any ARv7 specific upregulation. It is important to note that LN95 STARR-Seq is performed on a capture library composed of composed of 7371 regions which

contain previously known 4138 ARBS, positive and negative control regions [42]. These ARBS are defined in one of our laboratory's previous works in the LNCaP cell line [42]. We identified different classes of regions that are mentioned above from these ARBS. Since LNCaP does not express ARv7, it is expected that these regions are more prone to regulation by flAR, rather than ARv7. So, using this library might be introducing a bias and be the main reason why there is no ARv7 specific upregulation in LN95. Another effect of this bias is the small number of ARv7 and HD upregulated regions. It is hard to interpret or do further statistical analyses due to small sample sizes with these regions.

In summary, these results suggest that ARv7 has a repressive effect on enhancer activity of flAR's target regions (138 ARBS). This suggests that ARv7 could potentially suppresses enhancer activity in primary PCa and activates other enhancer regions in the genome in CRPC. But since we do not have those potential regions in our library, we were not able to detect them. This is a potential mechanism behind the transition from primary PCa to CRPC.

In chapter 3.2, results from H3K27ac HiChIP data analysis are shared from two cell lines: LNCaP and LuCaP, which represent primary PCa and CRPC, respectively. From the overview, we observed several interactions that are present in LuCaP which are absent in LNCaP. In closer analysis of these interactions, we verified the presence of gained interactions in LuCaP. On chromosome X, AR and the area around it have gained interactions in LuCaP that are absent in LNCaP. Importantly, on chromosome 8, MYC has very strong gained interactions in LuCaP. These are also supported by copy number analyses. There is a big deletion around MYC in LuCaP that is not present LNCaP, which introduces the possibility of SVs in the LuCaP cell line.

For SV detection, an amplicon detection pipeline is run with input ChIP-seq libraries from both cell lines. To understand the effect of library coverages on amplicon detection, a benchmarking is performed with a very high coverage LNCaP whole genome sequencing data and the results are compared with our LNCaP library's results. In both high coverage and low coverage libraries, no amplicons are detected in the LNCaP cell line. While this reassures us to some extent, having negative results does not fully eliminate the possibility of low coverage libraries' effect on amplicon detection. We detected 12 amplicons in the LuCaP cell line. One of them harbors an ecDNA. This is an important finding since the presence of ecDNAs is associated with worse prognosis and is an indicator of advanced stage

disease. The detected ecDNA is composed of CASC11, MYC and PVT1. CASC11 (Cancer Susceptibility 11), is associated with cancers such as hepatocellular carcinoma [84]. It functions in chromatin looping and enhancer-promoter loop anchoring. MYC is a well known proto-oncogene [85]. It is employed in cell cycle progression, apoptosis and cellular transformation. Lastly, PVT1 is a candidate oncogene [86]. Its amplification and overexpression are associated with many cancer types including breast and ovarian cancers, acute myeloid leukemia and Hodgkin lymphoma [86]. PVT1's transcription is regulated by the tumor suppressor p53. Also, it is reported before that PVT1 is involved in MYC's regulation to boost tumorigenesis. Presence of these genes on an ecDNA together may be leading to an amplified oncogenic activity, with possible enhancer hijacking events. So, this ecDNA may be employed in transition from primary PCa to CRPC.

These results together suggest that there is differential enhancer-promoter looping and SVs in CRPC that are not observed in primary PCa. These can be possible drivers of disease progression from primary PCa to CRPC. After careful downstream analyses, these events may appear as possible therapy targets for CRPC patients.

Results from this thesis collectively show the importance of enhancer activity in PCa progression. In this work, possible mechanisms behind enhancer activity changes in late-stage PCa compared to primary PCa were investigated. With future in-depth and other lines of research, these results can contribute to identification of new therapy targets for CRPC.

Chapter 5
FUTURE DIRECTIONS

In the first part of our results (Chapter 3.1), we observed that ARv7 has a repressive effect on flAR, rather than having an independent upregulating activity in the LN95 cell line. As explained in the discussion part (Chapter 4), this could be the case since our STARR-Seq library is mostly composed of previously known ARBS. Performing a whole genome STARR-Seq in the LN95 cell line may allow us to achieve more unbiased results in that sense. Results from a whole genome STARR-Seq in the same cell line can be interpreted in an integrative manner with our current results. This would potentially lead to detection of ARv7 regulated specific enhancer activity in LN95 cells.

In the second part (Chapter 3.2), we detected differential 3D interactions between LuCaP and LNCaP cell lines. Also, in LuCaP, we detected 12 amplicons, 1 of which harbors an ecDNA composed of oncogenes. A method that is able to detect differential looping systematically can be developed. In our pipeline, CNVkit is run within the AmpliconSuite-pipeline to detect candidate regions for amplicon detection. However, AmpliconSuite-pipeline has another mode where it accepts an input bed file of regions of interest. These regions of interest can be detected by the differential interaction script that is mentioned above. This input of regions can potentially change the AmpliconSuite-pipeline results. Another point is the coverage issue of the input ChIP-seq libraries used for amplicon detection, for reasons explained in the discussion part (Chapter 4). Although the AmpliconSuite-pipeline documentation states that a coverage as low as 0.5x is sufficient for amplicon detection [78], using higher coverage libraries than the ones we used (2.9x for LNCaP and 3.1x for LuCaP) would yield more reliable results.

Separately, these HiChIP data offer much more. A script to classify highly interactive regions within sample along with interchromosomal looping can be very useful for downstream analyses. The relationship between genomic distance between interacting anchors, and interaction intensities can be analyzed. Also, the correlation between a locus' number of interactions and interaction intensities can be investigated. Some loci may have many 3D interactions, but the supporting read numbers may be low. These results would give us a better understanding of the relative importance of genomic distance, interaction frequency and interaction intensity. At the end, highly interactive interaction hubs can be detected. With the integration of clinical and gene expression data, essentiality scores can be generated for different loci. These scripts written to analyze interaction frequencies, interchromosomal interactions and differential looping can be developed into separate tools and published.

Lastly, STARR-Seq and RNA-seq data from LNCaP and LuCaP can be analyzed to see if increased enhancer activity indicates increased gene expression. These results can be interpreted in parallel with HiChIP analysis results. After all necessary downstream analyses and experimental validation, as an end goal, if oncogenes on detected ecDNAs show increased enhancer activity and upregulated gene expression, they can be therapy targets in CRPC.

BIBLIOGRAPHY

[1] Chen, J., Zhang, D., Yan, W., Yang, D., & Shen, B. (2013). Translational bioinformatics for diagnostic and prognostic prediction of prostate cancer in the next-generation sequencing era. *BioMed research international*, *2013*, 901578. https://doi.org/10.1155/2013/901578

[2] Sekhoacha, M., Riet, K., Motloung, P., Gumenku, L., Adegoke, A., & Mashele, S. (2022). Prostate Cancer Review: Genetics, Diagnosis, Treatment Options, and Alternative Approaches. *Molecules (Basel, Switzerland)*, *27*(17), 5730. https://doi.org/10.3390/molecules27175730

[3] Eser, S., Zorlu, F., Divtik, R. T., Cal, C., Ozkan, M., & Kirkali, Z. (2009). Incidence and epidemiological features of cancers of the genitourinary tract in Izmir between 1993-2002. *Asian Pacific journal of cancer prevention: APJCP*, *10*(3), 491–496. Available on https://journal.waocp.org/?sid=Entrez:PubMed&id=pmid:19640197&key=2009.10.3.491

[4] Zorlu, F., Zorlu, R., Divrik, R. T., Eser, S., & Yorukoglu, K. (2014). Prostate cancer incidence in Turkey: an epidemiological study. *Asian Pacific journal of cancer prevention : APJCP*, *15*(21), 9125–9130. https://doi.org/10.7314/apjcp.2014.15.21.9125

[5] Takayama K. I. (2019). Splicing Factors Have an Essential Role in Prostate Cancer Progression and Androgen Receptor Signaling. *Biomolecules*, *9*(4), 131. https://doi.org/10.3390/biom9040131

[6] Mollica, V., Marchetti, A., Rosellini, M., Nuvola, G., Rizzo, A., Santoni, M., Cimadamore, A., Montironi, R., & Massari, F. (2021). An Insight on Novel Molecular Pathways in Metastatic Prostate Cancer: A Focus on DDR, MSI and AKT. *International journal of molecular sciences*, *22*(24), 13519. https://doi.org/10.3390/ijms222413519

[7] Cornford, P., van den Bergh, R. C. N., Briers, E., Van den Broeck, T., Brunckhorst, O., Darraugh, J., Eberli, D., De Meerleer, G., De Santis, M., Farolfi, A., Gandaglia, G.,

Gillessen, S., Grivas, N., Henry, A. M., Lardas, M., van Leenders, G. J. L. H., Liew, M., Linares Espinos, E., Oldenburg, J., van Oort, I. M., … Tilki, D. (2024). EAU-EANM-ESTRO-ESUR-ISUP-SIOG Guidelines on Prostate Cancer-2024 Update. Part I: Screening, Diagnosis, and Local Treatment with Curative Intent. *European urology*, *86*(2), 148–163. https://doi.org/10.1016/j.eururo.2024.03.027

[8] Gómez Rivas, J., Fernandez, L., Abad-Lopez, P., & Moreno-Sierra, J. (2023). Androgen deprivation therapy in localized prostate cancer. Current status and future trends. *Actas urologicas espanolas*, *47*(7), 398–407. https://doi.org/10.1016/j.acuroe.2022.08.009

[9] Daniels, V. A., Luo, J., Paller, C. J., & Kanayama, M. (2024). Therapeutic Approaches to Targeting Androgen Receptor Splice Variants. *Cells*, *13*(1), 104. https://doi.org/10.3390/cells13010104

[10] Tan, M., Li, J., Xu, H. *et al.* Androgen receptor: structure, role in prostate cancer and drug discovery. *Acta Pharmacol Sin* **36**, 3–23 (2015). https://doi.org/10.1038/aps.2014.18

[11] Chen, Y., & Lan, T. (2024). N-terminal domain of androgen receptor is a major therapeutic barrier and potential pharmacological target for treating castration resistant prostate cancer: a comprehensive review. *Frontiers in pharmacology*, *15*, 1451957. https://doi.org/10.3389/fphar.2024.1451957

[12] Özturan, D., Morova, T., & Lack, N. A. (2022). Androgen Receptor-Mediated Transcription in Prostate Cancer. *Cells*, *11*(5), 898. https://doi.org/10.3390/cells11050898

[13] McEwan IJ, Brinkmann AO. Androgen Physiology: Receptor and Metabolic Disorders. [Updated 2021 Jul 2]. In: Feingold KR, Anawalt B, Blackman MR, et al., editors. Endotext [Internet]. South Dartmouth (MA): MDText.com, Inc.; 2000-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK279028/

[14] Yuan, X., Cai, C., Chen, S. *et al.* Androgen receptor functions in castration-resistant prostate cancer and mechanisms of resistance to new agents targeting the androgen axis. *Oncogene* **33**, 2815–2825 (2014). https://doi.org/10.1038/onc.2013.235 **use it also for 1.2.1

[15] Basil, P., Robertson, M.J., Bingman, W.E. *et al.* Cistrome and transcriptome analysis identifies unique androgen receptor (AR) and AR-V7 splice variant chromatin binding and transcriptional activities. *Sci Rep* **12**, 5351 (2022). https://doi.org/10.1038/s41598-022-09371-x

[16] Antonarakis, E. S., Lu, C., Wang, H., Luber, B., Nakazawa, M., Roeser, J. C., Chen, Y., Mohammad, T. A., Chen, Y., Fedor, H. L., Lotan, T. L., Zheng, Q., De Marzo, A. M., Isaacs, J. T., Isaacs, W. B., Nadal, R., Paller, C. J., Denmeade, S. R., Carducci, M. A., Eisenberger, M. A., … Luo, J. (2014). AR-V7 and resistance to enzalutamide and abiraterone in prostate cancer. *The New England journal of medicine*, *371*(11), 1028–1038. https://doi.org/10.1056/NEJMoa1315815

[17] Sharp, A., Coleman, I., Yuan, W., Sprenger, C., Dolling, D., Rodrigues, D. N., Russo, J. W., Figueiredo, I., Bertan, C., Seed, G., Riisnaes, R., Uo, T., Neeb, A., Welti, J., Morrissey, C., Carreira, S., Luo, J., Nelson, P. S., Balk, S. P., True, L. D., … Plymate, S. R. (2019). Androgen receptor splice variant-7 expression emerges with castration resistance in prostate cancer. *The Journal of clinical investigation*, *129*(1), 192–208. https://doi.org/10.1172/JCI122819

[18] Visel, A., Rubin, E. & Pennacchio, L. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009). https://doi.org/10.1038/nature08451

[19] Ward, L., Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**, 1095–1106 (2012). https://doi.org/10.1038/nbt.2422

[20] Mathelier, A., Shi, W., & Wasserman, W. W. (2015). Identification of altered cis-regulatory elements in human disease. *Trends in genetics : TIG*, *31*(2), 67–76. https://doi.org/10.1016/j.tig.2014.12.003

[21] Li, Y., Hu, M., & Shen, Y. (2018). Gene regulation in the 3D genome. *Human molecular genetics*, *27*(R2), R228–R233. https://doi.org/10.1093/hmg/ddy164

[22] Wittkopp, P., Kalay, G. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13**, 59–69 (2012). https://doi.org/10.1038/nrg3095

[23] Ayoubi, T. A., & Van De Ven, W. J. (1996). Regulation of gene expression by alternative promoters. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, *10*(4), 453–460.

[24] Khoury, G., & Gruss, P. (1983). Enhancer elements. *Cell*, *33*(2), 313–314. https://doi.org/10.1016/0092-8674(83)90410-5

[25] Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, *27*(2 Pt 1), 299–308. https://doi.org/10.1016/0092-8674(81)90413-x

[26] Thomas, H. F., & Buecker, C. (2023). What is an enhancer?. *BioEssays : news and reviews in molecular, cellular and developmental biology*, *45*(10), e2300044. https://doi.org/10.1002/bies.202300044

[27] Stallcup, M. R., & Poulard, C. (2020). Gene-Specific Actions of Transcriptional Coregulators Facilitate Physiological Plasticity: Evidence for a Physiological Coregulator Code. *Trends in biochemical sciences*, *45*(6), 497–510. https://doi.org/10.1016/j.tibs.2020.02.006

[28] Spitz, F., Furlong, E. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012). https://doi.org/10.1038/nrg3207

[29] Matharu, N., & Ahituv, N. (2015). Minor Loops in Major Folds: Enhancer-Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease. *PLoS genetics*, *11*(12), e1005640. https://doi.org/10.1371/journal.pgen.1005640

[30] Kadauke, S., & Blobel, G. A. (2009). Chromatin loops in gene regulation. *Biochimica et biophysica acta*, *1789*(1), 17–25. https://doi.org/10.1016/j.bbagrm.2008.07.002

[31] Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (New York, N.Y.)*, *339*(6123), 1074–1077. https://doi.org/10.1126/science.1232542

[32] Catarino, R. R., & Stark, A. (2018). Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes & development*, *32*(3-4), 202–223. https://doi.org/10.1101/gad.310367.117

[33] Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)*, *316*(5830), 1497–1502. https://doi.org/10.1126/science.1141319

[34] Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., & Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using

chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, *4*(8), 651–657. https://doi.org/10.1038/nmeth1068

[35] Nakato, R., & Sakata, T. (2021). Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods (San Diego, Calif.)*, *187*, 44–53. https://doi.org/10.1016/j.ymeth.2020.03.005

[36] Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science (New York, N.Y.)*, *295*(5558), 1306–1311. https://doi.org/10.1126/science.1067799

[37] Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., & de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics*, *38*(11), 1348–1354. https://doi.org/10.1038/ng1896

[38] Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., & Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, *16*(10), 1299–1309. https://doi.org/10.1101/gr.5571506

[39] Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, *326*(5950), 289–293. https://doi.org/10.1126/science.1181369

[40] Mumbach, M., Rubin, A., Flynn, R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**, 919–922 (2016). https://doi.org/10.1038/nmeth.3999

[41] Caragine, C.M., Le, V.T., Mustafa, M. *et al.* Comprehensive dissection of *cis*-regulatory elements in a 2.8 Mb topologically associated domain in six human cancers. *Nat Commun* **16**, 1611 (2025). https://doi.org/10.1038/s41467-025-56568-5

[42] Huang, C. F., Lingadahalli, S., Morova, T., Ozturan, D., Hu, E., Yu, I. P. L., Linder, S., Hoogstraat, M., Stelloo, S., Sar, F., van der Poel, H., Altintas, U. B., Saffarzadeh, M.,

Le Bihan, S., McConeghy, B., Gokbayrak, B., Feng, F. Y., Gleave, M. E., Bergman, A. M., Collins, C., … Lack, N. A. (2021). Functional mapping of androgen receptor enhancer activity. *Genome biology*, *22*(1), 149. https://doi.org/10.1186/s13059-021-02339-6

[43] Sur, I., Taipale, J. The role of enhancers in cancer. *Nat Rev Cancer* **16**, 483–493 (2016). https://doi.org/10.1038/nrc.2016.62

[44] Giorgio, E., Robyr, D., Spielmann, M., Ferrero, E., Di Gregorio, E., Imperiale, D., Vaula, G., Stamoulis, G., Santoni, F., Atzori, C., Gasparini, L., Ferrera, D., Canale, C., Guipponi, M., Pennacchio, L. A., Antonarakis, S. E., Brussino, A., & Brusco, A. (2015). A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). *Human molecular genetics*, *24*(11), 3143–3154. https://doi.org/10.1093/hmg/ddv065

[45] Valton, A. L., & Dekker, J. (2016). TAD disruption as oncogenic driver. *Current opinion in genetics & development*, *36*, 34–40. https://doi.org/10.1016/j.gde.2016.03.008

[46] Zhou, T., & Feng, Q. (2022). Androgen receptor signaling and spatial chromatin organization in castration-resistant prostate cancer. *Frontiers in medicine*, *9*, 924087. https://doi.org/10.3389/fmed.2022.924087

[47] Taberlay, P. C., Achinger-Kawecka, J., Lun, A. T., Buske, F. A., Sabir, K., Gould, C. M., Zotenko, E., Bert, S. A., Giles, K. A., Bauer, D. C., Smyth, G. K., Stirzaker, C., O'Donoghue, S. I., & Clark, S. J. (2016). Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome research*, *26*(6), 719–731. https://doi.org/10.1101/gr.201517.115

[48] San Martin, R., Das, P., Dos Reis Marques, R., Xu, Y., Roberts, J. M., Sanders, J. T., Golloshi, R., & McCord, R. P. (2022). Chromosome compartmentalization alterations in prostate cancer cell lines model disease progression. *The Journal of cell biology*, *221*(2), e202104108. https://doi.org/10.1083/jcb.202104108

[49] Stelloo, S., Bergman, A. M., & Zwart, W. (2019). Androgen receptor enhancer usage and the chromatin regulatory landscape in human prostate cancers. *Endocrine-related cancer*, *26*(5), R267–R285. https://doi.org/10.1530/ERC-19-0032

[50] 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., & Neal, D. E. (2013). The

androgen receptor induces a distinct transcriptional program in castration-resistant prostate cancer in man. *Cancer cell*, *23*(1), 35–47. https://doi.org/10.1016/j.ccr.2012.11.010

[51] Stelloo, S., Nevedomskaya, E., van der Poel, H. G., de Jong, J., van Leenders, G. J., Jenster, G., Wessels, L. F., Bergman, A. M., & Zwart, W. (2015). Androgen receptor profiling predicts prostate cancer outcome. *EMBO molecular medicine*, *7*(11), 1450–1464. https://doi.org/10.15252/emmm.201505424

[52] Cosenza, M. R., Rodriguez-Martin, B., & Korbel, J. O. (2022). Structural Variation in Cancer: Role, Prevalence, and Mechanisms. *Annual review of genomics and human genetics*, *23*, 123–152. https://doi.org/10.1146/annurev-genom-120121-101149

[53] ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–93. https://doi.org/10.1038/s41586-020-1969-6

[54] Feuk, L., Carson, A. & Scherer, S. Structural variation in the human genome. *Nat Rev Genet* **7**, 85–97 (2006). https://doi.org/10.1038/nrg1767

[55] Yan, X., Mischel, P. & Chang, H. Extrachromosomal DNA in cancer. *Nat Rev Cancer* **24**, 261–273 (2024). https://doi.org/10.1038/s41568-024-00669-8

[56] Wang, X., & Yue, F. (2024). Hijacked enhancer-promoter and silencer-promoter loops in cancer. *Current opinion in genetics & development*, *86*, 102199. https://doi.org/10.1016/j.gde.2024.102199

[57] R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[58] Grothendieck G (2017). sqldf: Manipulate R Data Frames Using SQL. R package version 0.4-11, https://CRAN.R-project.org/package=sqldf

[59] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

[60] Larsson J (2024). eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses. R package version 7.0.2, https://CRAN.R-project.org/package=eulerr

[61] Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

[62] Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research*, *44*(W1), W160–W165. https://doi.org/10.1093/nar/gkw257

[63] Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., & Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics (Oxford, England)*, *26*(17), 2204–2207. https://doi.org/10.1093/bioinformatics/btq351

[64] Kassambara A (2023). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.6.0, https://CRAN.R-project.org/package=ggpubr

[65] Seim, I., Jeffery, P. L., Thomas, P. B., Nelson, C. C., & Chopin, L. K. (2017). Whole-Genome Sequence of the Metastatic PC3 and LNCaP Human Prostate Cancer Cell Lines. *G3 (Bethesda, Md.)*, *7*(6), 1731–1741. https://doi.org/10.1534/g3.117.039909

[66] Andrews S. (2023). FastQC: a quality control tool for high throughput sequence data. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[67] Krueger F. (2023). Trim Galore!. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

[68] Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

[69] Open2C, Abdennur, N., Fudenberg, G., Flyamer, I. M., Galitsyna, A. A., Goloborodko, A., Imakaev, M., & Venev, S. V. (2023). Pairtools: from sequencing data to chromosome contacts. *bioRxiv : the preprint server for biology*, 2023.02.13.528389. https://doi.org/10.1101/2023.02.13.528389

[70] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

[71] Lee, S., Bakker, C. R., Vitzthum, C., Alver, B. H., & Park, P. J. (2022). Pairs and Pairix: a file format and a tool for efficient storage and retrieval for Hi-C read pairs. *Bioinformatics (Oxford, England)*, *38*(6), 1729–1731. https://doi.org/10.1093/bioinformatics/btab870

[72] Abdennur, N., & Mirny, L. A. (2020). Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics (Oxford, England)*, *36*(1), 311–316. https://doi.org/10.1093/bioinformatics/btz540

[73] Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell systems*, *3*(1), 95–98. https://doi.org/10.1016/j.cels.2016.07.002

[74] Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell systems*, *3*(1), 99–101. https://doi.org/10.1016/j.cels.2015.07.012

[75] Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., Heard, E., Dekker, J., & Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology*, *16*, 259. https://doi.org/10.1186/s13059-015-0831-x

[76] Langmead, B., Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012). https://doi.org/10.1038/nmeth.1923

[77] Bhattacharyya, S., Chandra, V., Vijayanand, P., & Ay, F. (2019). Identification of significant chromatin contacts from HiChIP data by FitHiChIP. *Nature communications*, *10*(1), 4221. https://doi.org/10.1038/s41467-019-11950-y

[78] Luebeck, J., Huang, E., Kim, F., Liefeld, T., Dameracharla, B., Ahuja, R., Schreyer, D., Prasad, G., Adamaszek, M., Kenkre, R., Agashe, T., Torvi, D., Tabor, T., Giurgiu, M., Kim, S., Kim, H., Bailey, P., Verhaak, R. G. W., Deshpande, V., Reich, M., Mischel, P. S., Mesirov, J. & Bafna, V. (2024). AmpliconSuite: an end-to-end workflow for analyzing focal amplifications in cancer genomes. *bioRxiv : the preprint server for biology*, 2024.05.06.592768. https://doi.org/10.1101/2024.05.06.592768

[79] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Preprint available on arXiv*. https://doi.org/10.48550/arXiv.1303.3997

[80] Talevich, E., Shain, A. H., Botton, T., & Bastian, B. C. (2016). CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS computational biology*, *12*(4), e1004873. https://doi.org/10.1371/journal.pcbi.1004873

[81] Deshpande, V., Luebeck, J., Nguyen, N. D., Bakhtiari, M., Turner, K. M., Schwab, R., Carter, H., Mischel, P. S., & Bafna, V. (2019). Exploring the landscape of focal

amplifications in cancer using AmpliconArchitect. *Nature communications*, *10*(1), 392. https://doi.org/10.1038/s41467-018-08200-y

[82] Luebeck, J., Coruh, C., Dehkordi, S. R., Lange, J. T., Turner, K. M., Deshpande, V., Pai, D. A., Zhang, C., Rajkumar, U., Law, J. A., Mischel, P. S., & Bafna, V. (2020). AmpliconReconstructor integrates NGS and optical mapping to resolve the complex structures of focal amplifications. *Nature communications*, *11*(1), 4374. https://doi.org/10.1038/s41467-020-18099-z

[83] Sayers, E. W., Beck, J., Bolton, E. E., Brister, J. R., Chan, J., Connor, R., Feldgarden, M., Fine, A. M., Funk, K., Hoffman, J., Kannan, S., Kelly, C., Klimke, W., Kim, S., Lathrop, S., Marchler-Bauer, A., Murphy, T. D., O'Sullivan, C., Schmieder, E., Skripchenko, Y., … Pruitt, K. D. (2025). Database resources of the National Center for Biotechnology Information in 2025. *Nucleic acids research*, *53*(D1), D20–D29. https://doi.org/10.1093/nar/gkae979

[84] Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary, D, Warshawsky D, Guan - Golan Y, Kohn A, Rappaport N, Safran M, and Lancet D. *The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses* Current Protocols in Bioinformatics(2016), 54:1.30.1 - 1.30.33.doi: 10.1002 / cpbi.5, CASC11: https://www.genecards.org/cgi-bin/carddisp.pl?gene=CASC11

[85] Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary, D, Warshawsky D, Guan - Golan Y, Kohn A, Rappaport N, Safran M, and Lancet D. *The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses* Current Protocols in Bioinformatics(2016), 54:1.30.1 - 1.30.33.doi: 10.1002 / cpbi.5, MYC: https://www.genecards.org/cgi-bin/carddisp.pl?gene=MYC

[86] Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary, D, Warshawsky D, Guan - Golan Y, Kohn A, Rappaport N, Safran M, and Lancet D. *The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses* Current Protocols in Bioinformatics(2016), 54:1.30.1 - 1.30.33.doi: 10.1002 / cpbi.5, PVT1: https://www.genecards.org/cgi-bin/carddisp.pl?gene=PVT1