Androgen receptor binding sites are highly mutated in prostate cancer

by

Tunç Morova

A Dissertation Submitted to the
Graduate School of Science and Engineering in
Partial Fulfillment of the Requirements for
The Degree of
Master of Science
in
Biomedical Science and Engineering

KOÇ
UNIVERSITY

DATE: 05.07.2018

Androgen receptor binding sites are highly mutated in 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

Tunç Morova

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 A. LACK

_____

Prof. Dr. Attila GÜRSOY

_____

Asst. Prof. Öznur TAŞTAN

_____

Date: _____

I dedicate this thesis to my biggest supporters, Mom and Melis. Thank you for your help and support through this journey...

Abstract

Cancer arises through the sequential accumulation of mutations that induce neoplastic transformation and uncontrolled proliferation. These somatic mutations however do not occur in a normal distribution across the genome and are affected by several variables including GC content, replication time, distance to telomere and chromatin compaction. Recent study demonstrated that TF binding to DNA increases the rate of mutations in melanoma due the impairment of nuclear excision repair mechanism. However, it is unclear if this phenomenon is specific to only high-mutation rate cancers or certain chromosomal regions.

Prostate cancer (PCa) is the most frequently diagnosed cancer in European men and the second leading cause of cancer-related death. Androgen receptor (AR) mediated transcription is critical at all stages of PCa progression. Following activation, AR binds to specific DNA response elements where it recruits numerous co-activators that induce gene transcription. However, despite the importance of AR signaling in prostate cancer it plays no role in almost all other cancer types. Given this specificity, the AR is an ideal model to study TF mediated DNA damage characteristics and frequency.

Therefore, the aim of this work was to investigate how TF binding affects somatic mutations in prostate cancer (PCa). To test this, we analyzed whole genome sequencing data from prostate primary PCa (n=196) to investigate both the type and frequency of mutations at AR binding sites (ARBS). We demonstrated that PCa has the highest frequency of mutations at ARBS of any cancer type. Further, among all TF tested, ARBS had by far the highest mutation frequency. Interestingly, we also found a novel mutation signature at ARBS that is different from the remainder of the genome. Specifically, there was a markedly higher rate of TpG->ApG mutations potentially mediated by spontaneous depurination. This mutational signature is independent of the nucleotide composition of ARBS. While speculative, we believe the higher rate of mutations occurs due to an inability of the Base Excision Repair machinery to access the spontaneous mutations at ARBS and not DSB. Herein we reveal distinct mutation phenomena in prostate cancer patients that require validation and further investigation.

Özetçe

Zamanla biriken mutasyonlar kanserin ilerlemesi ve proliferasyonuna büyük ölçüde katkıda bulunmaktadır. Ancak söz konusu mutasyonların oluşmaşı bir çok faktörden etkilenip, insan genomu üzerinde eşit olarak dağılmamaktır. Genom üzerindeki bu faktörler sırasıyla G-C oranı, replikasyon zamanlaması, telomere uzaklık ve kromatin paketlenmesi. Son zamanlarda yapılan bir çalışma ise, transkripsyon faktörlerinin (TF) DNA'ya baglanması sonucunda baz eksizyon onarım mekanizmasında aksaklıklar olduğunu göstermiştir. Bu durum ise melenomadaki mutasyon oranının artışı ile ilişkilendirilmiştir. Ancak bu durumun sadece yüksek mutasyon oranlı kanserlere ya da belirli kromozom bölgelerine spesifik olup olmadığı bilinmemektedir.

Prostat kanseri Avrupalı erkeklerde en sık oranda tanılanmış olmakla birlikte kanser ilişkili ölümlerde ikinci sırada yer almaktadır. Prostat kanserinin ilerlemesinde Androjen reseptörü (AR) aracılığıyla indüklenen transkripsiyon kritik rol oynamaktadır. Androgen aktivasyonunu takiben AR genom üzerinde sorumlu bölgelere baglanır. Bu bağlanma sonucunda gerekli koaktivatörlerin o bölgeye gelmesini sağlayarak, gen transkripsiyonunu başlatır. AR yolakları prostat kanserinde çok önemli olmasına rağmen aynı önem başka kanser türlerinde gözlemlenmemektedir. Bu spesifik etki göz önünde bulundurulduğunda AR, TF-aracılı DNA hasarı karakterizasyonu ve sıklığı çalışmaları için en ideal modeldir.

Bu çalışmanın amacı TF bağlanmasının prostat kanserindeki somatik mutasyonlar üzerindeki etkisini araştırmaktır. Androjen bağlanma bölgelerindeki (ARBB) mutasyonların tür ve sıklığının incelenmesi için, 196 primer prostat kanseri tüm genom dizileme sonuçları analiz edilmiştir. Bulduğumuz sonuçlar neticesinde prostat kanserinde, diğer kanser türlerine oranla ARBB'de mutasyonların daha sık olduğunu gösterdik. Buna ek olarak test edilen tüm TF arasında en yüksek frekanstaki mutasyonlar ARBB'de görülmüştür. İlaveten genomun geri kalanından farklı olarak, ARBB'de özgün bir mutasyon karakteri bulunmuştur. Yüksek oranda saptadığımız TpG->ApG mutasyonları, bu bölgelerdeki kendiliğinden depürinasyon kaynaklıdır. Bu mutasyonlar ARBB'nin nükleotid komposizyonundan bağımsızdır. Kesin olmamakla beraber bu yüksek mutasyon oranın, baz eksizyon onarım mekanizmasının ARBB bölgelerindeki mutasyonlara erişememesi sebebiyle olduğu düşünülmektedir. Bu projede, prostat kanseri hastalarında farklı mutasyon türleri açığa çıkardık. Ancak, bu bulguların deneysel çalışmalarla desteklenmesi gerektiğine inanıyoruz.

Acknowledements

Past three years have been the most grinding period of my life that I wouldn't have been able to successfully finalize it without the support of the people around me. These people deserve credit as much as I deserve.

First of all, I would like to thank Dr. Nathan A. Lack for giving me opportunity to work on this project. From the first and last moment of the project, he has never given up on me and supported me in every aspect. Without his lead, I wouldn't have had high standards on science and any sort of analysis. Besides Dr. Lack, I would like to thank Dr. Tugba Bagci-Önder and Dr. Tamer Önder for including me their projects, which allowed me to experience various cutting edge scientific projects. Besides their scientific contributions, they have always had time for my non-scientific conversations and tolerated my unsual sense of humor. My graduate project contains lots of statistical and computational analysis and whenever I had problems with these parts, Dr. Mehmet Gönen guided me to find a better solution. I would like to mention the support of Dr. Ozlem Keskin and Dr. Attilla Gursoy. They have supported my stay in Koc University and been my co-advisor that guided me in the computational problems in Koc University's High Performance Cluster. Also I would like to thank Dr. Öznur Taştan for accepting to be a committee member in my thesis. I call the aforementioned persons as my scientific mentors who led me to be a better scientist.

In addition to my mentors, my lab colleagues showed outstanding understanding to my ignorant questions about biology. As I graduated from engineering focused department, I was not as developed in biology like my colleagues. However, they helped me relentlessly whenever I needed their support. They also have been very tolerant to uncommon behaviors outside of the lab where they did not laugh at me but they laughed with me. Therefore, I thank Ceren Seref, Fatma Ozgun, Hilal Sarac, Dogancan Ozturan, Derya Cavga, Betul Ersoy, Bengul Cetin, Gizem Hazal Senturk and our alumni Zeynep Kaya, Firat Uyulur, Can Aztekin for being my closest in scientific and social life.

Outside of the lab, I have witnessed my best friends; Gulben Gurhan and Kenan Sevinc, tying up their lives together. This was the best thing after my Dota2 team won the weekend tournament in 2018. Probably I have spent time with them more than any of my relatives and lived wildly interesting moments with these three years. They were also pretty tolerant in my socially awkward behaviors. Through out this time, I can clearly say that I earned two life long friends, which is one of the biggest benefits of my graduate education.

My family made great commitments to support me in this journey. I would like to thank Dad for understanding my busy schedule and supporting me all the way from Ankara. On the other hand, I think Mom deserves special thanks. I am simply here because of mom's extreme support. She has never given up on me before and throughout the graduate school. I learned how to be a good human being and enthusiastic in life and my job. I am aware that I cannot pay anything back to her. However I can say that Mom, I love you and thank you for everything you have given me up now.

Lastly, I have to acknowledge the support of another person. Melis Saygin, you have supported me in every aspect of life for the last 4 years. In fact, you are the reason why I chose Koç University and I am grateful that I followed your word. You tolerated all of my long working hours without any comment and simply supported me even though this affected our beautiful dates.

I also would like to thank 1190 for his absolute support during my thesis. Our annual Wednesday night outs prevented severe psychological damage on my personality and without his support I wont be able to get over from losing the most important person of my life.

Table of Contents

Chapter 1: Introduction

1.1 Prostate Cancer

Prostate cancer (PCa) is the most frequently diagnosed cancer in men and one of the leading causes of cancer-related death. In 2018 there were approximately 165,000 cases diagnosed and 29,430 deaths due to PCa[1]. Almost all prostate cancers are hormone driven. Activation of the androgen receptor (AR) initiates a transcription program, which drives the proliferation of the cancer. Therefore, PCa growth is critically affected by hormone induced transcription factor (TF) activation[2]. Given this essential role, reducing the levels of circulating hormone or directly inhibiting the androgen receptor is the standard of care for recurrent prostate cancer.

1.2 Androgens

Testesterone is the primary circulating androgen in men. Approximately 95% of this hormone is produced in testis (Leyding cells) with the remaining 5% made in the adrenal glands. Testesterone is synthesized via gamma-5 metabolic pathway in both of these tissues[3]. After that it enters cells with diffusion. Testesterone is converted to the more potent form of androgen, 5-alpha-dihydrotestosterone (DHT) by the 5-α-reductase in prostate tissue. DHT has two times more activity in cellular pathways and gets degraded five times slower than testesterone[4,5] In addition to their role in cancer, androgens are required for healthy prostate development and maintenance.

1.3 Androgen receptor

The AR is a member of the nuclear receptor family of ligand-activated transcription factors[6]. Like all nuclear receptors, it contains three domains (**Figure 1**). This includes:

- N-terminal domain (NTD)
- DNA-binding domain (DBD)
- Ligand binding domain (LBD)

The first exon of the AR gene codes the transcriptional regulatory region of the protein, which is the N- terminal domain (NTD). The NTD is the least conserved region of hormone receptors and it is responsible for the transactivation activity of the androgen receptor. This domain contains the activation function 1 (AF-1) region that is critical for gene transcriptional. Exon 2 and 3 code for the central DBD. This region is highly conserved and it is located at the center of AR. The DBD contains two zinc-finger motifs that recognizes specific DNA sequences and facilitates binding to chromatin. Finally, exon 4 - 8 code the C-terminal ligand binding domain (LBD) that mediates the interaction with heat shock chaperone proteins and DHT[4,8]**(Figure 1).** This domain contains the activator function 2 (AF-2) that facilitates interaction with histone acetylases[7] (Discussed later). The hinge region of the LBD is crucial to the nuclear translocation, as it contains the nuclear localization signal

that is exposed upon hormone binding[89].

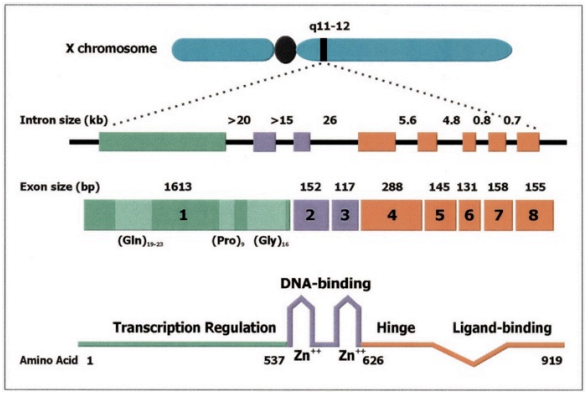

Figure 1: Genomic organization of the AR gene. Diagram of the protein structure demonstrates how the exon organization translates into discrete functional regions of the receptor. Figure taken from[10].

1.4 AR mediated transcription

When inactive, the AR resides in the cytoplasm bound to heat shock proteins. This complex protects AR from proteolysis as well as preparing it for hormone activation by maintaining an apo-confirmation. The activation of the AR follows a well-characterized pathway (**Figure 2**). Upon hormone stimulation, AR binds to DHT and disassociates from heat shock proteins. This causes an allosteric movement of alpha-helix 12 in the LBD, which exposes the NLS in the hinge domain. The AR then forms homodimers and translocate in to the nucleus. Once in the nucleus the AR binds to specific DNA response elements where pioneer factors have "primed" the site for transcription factor binding. These pioneer factors, including FOXA1, GATA2 and HOXB13, bind to DNA prior to AR and alter chromatin structure to allow transcription. Once bound, the AR recruits' co-activator complexes that initiate transcription via RNA polymerase II[2]

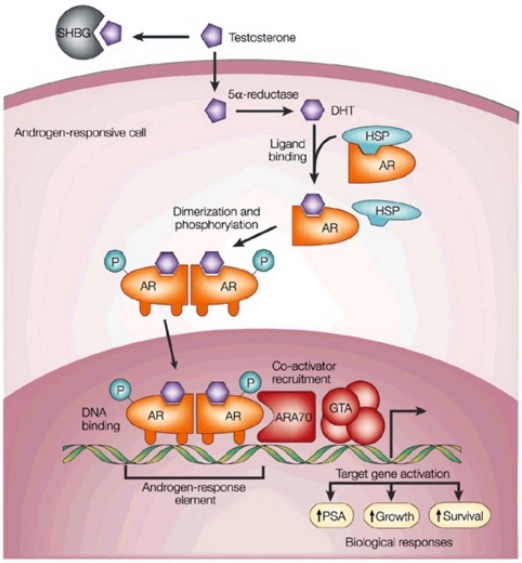

Figure 2: Androgen receptor's transcription program activation. AR stays in complex with heat-shock proteins when there is no androgen stimulation is present. When testosterone goes in to the nucleous it is degraded to dihydtrotestesterone by 5-alpha-reductase. Then DHT binds to AR. This complex dimerizes and translocates in to the nucleus. DHT bound AR binds specific locations on the genome. Thus, transcriptional activation occurs[11].

The AR-DBD binds to DNA at the androgen responsive elements (ARE). Canonical AREs consist of a 15-bp palindromic sequence that contains two hexameric half sites 5'-AGAACA-3'arranged as an inverted repeat with a 3bp spacer. The highest affinity between AR and ARE occurs with this "perfect" motif, however binding is not enough and several factors including local chromosomal environment, chromatin accessibility of specific co-operating factors and basal transcriptional factors all impact transcription[12]. Interestingly, most of the AR binding sites are found far from the transcription start site (TSS) of androgen regulated genes[13]. Wang *et al* indicated that on chromosome 21 and chromosome 22, only 34 of 90 (38%) AR binding sites (ARBS) were located within 500 kb of TSS of AR regulated genes[13]. Later studies with ChIPseq demonstrated that the vast majority of ARBS are located on intronic and intergenic regions on the genome with only 1% of binding sites located either in a promoter or exonic region. Overall this suggests that that ARBS are primarily enhancers that activate transcription by chromosome looping[14-15].

1. 5 Pioneer Factors

In general, pioneer factors bind to closed heterochromatin and make it accessible through histone acetylation. In addition, they also facilitate chromatin looping with AR-bound distal enhancers and AR-regulated gene promoters[16]. *GATA2*, *FOXA1* and *HOXB13* genes are the most commonly studied pioneer factors involved in nuclear receptor mediated transcription (**Figure 3**)[17,16,18]. Initial studies demonstrated that FOXA1 is essential for allowing AR to bind and induce transcriptional program of well-characterized AR regulated gene, *KLK3*[19]. FOXA1 dependence was later demonstrated by various chromatin immunoprecipitation experiments that investigates co-localization of FOXA1 with AR[13]. However, later studies have shown that FOXA1's pioneer activity is more complex throughout the genome. Specifically, it is not essential at canonical ARE regions but rather it is required at low-affinity half-ARE regions. Overexpression of FOXA1 creates excessive open chromatin that allows AR to bind at redundant half-ARE regions, thereby decrease the overall activity ARBS. Therefore down regulation of FOXA1 cause aberrant AR activation that results more AR binding to chromatin[20,21]. Supporting this, silencing FOXA1 did not effect the expression of TMPRSS2 and PSA, which suggests that this pioneer factor is not essential for all AR targets but only some of them[13]. Similar to FOXA1, GATA2 was demonstrated to co-localize with AR at ARE upon androgen stimulation. In contrast, GATA2 itself alters AR expression[22]. Silencing GATA2 decreases the mRNA and protein level of AR in both androgen stimulated and deprived cells. Correspondingly, silencing GATA2 decreased the expression of all known AR regulated genes, suggesting that it more essential than FOXA1[13]. Unlike GATA2, HOXB13 was initially demonstrated to act as an repressor for prostate cancer cell proliferation[23]. However, contrary to this another study demonstrated HOXB13 is

essential for prostate cancer cells proliferation[18]. What is more, three potential mechanisms of action were proposed for HOXB13 regulation of AR-mediated transcription. The first mechanism is a collaborative model (**Figure 4A**), whereby a binding motif of both AR and HOXB13 is present and the AR-HOXB13 complex enhances gene expression. Supporting this model, the expression of NKX3.1 and TMPRSS2 were down-regulated in siHOXB13 treated cells. The second mechanism is a tether model (**Figure 4B**), where HOXB13 binds to the DNA binding domain of AR and activates enhancers that have a HOXB13 binding motif. This is seen with *ORM1*, that has no ARE binding motif but has a HOXB13 binding motif in the gene enhancer. Upon disruption of the AR-DBD domain, ORM1 expression is reduced. The final mechanism is a repressive model (**Figure 4C**). This is based on earlier studies that showed AR activation can actually cause decrease the activity of certain genes that have ARE motifs such as PSA, STEAP4 and FASN. Given that HOXB13 interacts with the DNA binding domain of AR, in this model HOXB13 is proposed to physical block AR interactions with DNA. Despite the reduced binding affinity due to blockage of AR to ARE, some genes will pass the binding affinity threshold and eventually be expressed. This will be determined by AR-ARE motif binding affinity, therefore there will be variability between the AREs[18].

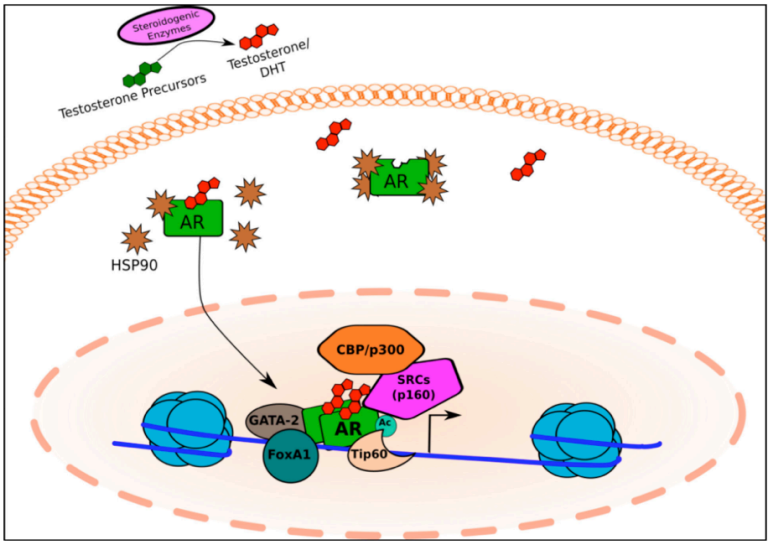

Figure 3: Interaction of AR with co-activators. Figure taken from[20].

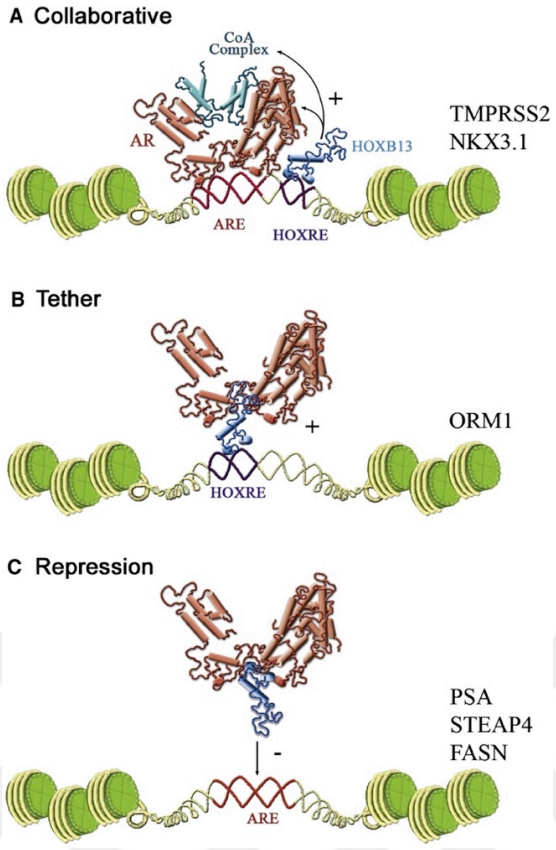

Figure 4: Proposed HOXB13 mediated regulation. Upregulation models (A) and (B) and repression model (C). Figure taken from[18].

1.6 Co-activators

The AR protein does not directly interact with RNA polymerase II, but rather acts as a "hub" that recruits co-activators to initiate transcription of AR regulated genes. Many of these co-activators are epigenetic modifiers that acetylate histone to allow gene activation. Contrastingly, deacetlyation is generally linked with gene repression[24,25]. Histone acetyl transferase (HAT), CREB binding protein (CBP), p160, p300 and pCAF (P300/CBP-associated factor) are found active at promoter of activated genes. They are recruited by DHT-bound AR in the following the order; p160 and p300, CBP and pCAF[26]. For transcription to occur CBP and p160-p300 binding is necessary and it was shown that upon deletion of the interaction domain, transcription does not initiate[26]. In addition, Tat interactive protein (TIP60) has HAT activity that interacts with the AR-LBD. Tip60 interacts with lysine residues of AR hinge region[27]. This interaction increases AR-mediated transcription activity. For transcriptional repression, histone deacetylase (HDAC) proteins SIRT1 (sir2 a class II HDAC) inhibits androgen-stimulated transcription by deacetylating p300 and PCAF acetylated regions[28]. In addition, histone deacetylase 1 (HDAC1) counters aforementioned acetylation of Tip60[29]. Overall, histone 3 and 4 residues are often targeted by HAT and HDAC to modulate AR-mediated transcription (**Table 1**).

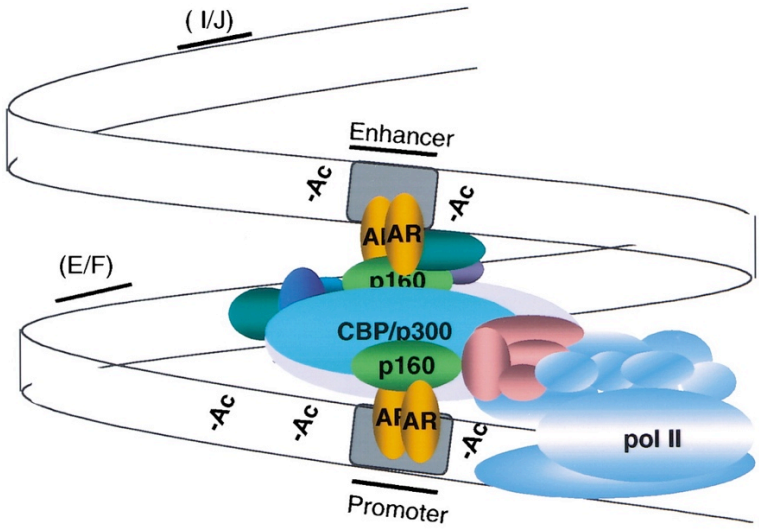

Figure 5: Utilization of p160, CBP/p300 proteins. Figure taken from[26].

Table 1: Aceltylation modifiers and their histone marks.

HISTONE	MODIFIED SITE	MODIFICATION PATTERN	MODIFYING ENZYME
H2A	Lys5 (mammals)	Acetylation	Tip60, p300/CBP
H2B	Lys5	Acetylation	p300
H2B	Lys12 (mammals)	Acetylation	p300/CBP
H2B	Lys15 (mammals)	Acetylation	p300/CBP
H2B	Lys20	Acetylation	p300
H3	Lys14	Acetylation	PCAF, Tip60, p300
H3	Lys18	Acetylation	p300/CBP
H3	Lys23	Acetylation	p300/CBP
H4	Lys5	Acetylation	Hat1, Tip60, p300
H4	Lys8	Acetylation	PCAF, Tip60, p300
H4	Lys12	Acetylation	Hat1, Tip60, p300
H4	Lys16	Acetylation	Tip60
H4	Lys91 (*S. cerevisiae*)	Acetylation	Hat1/Hat2

Even though histone acetylation is generally correlated with transcription activation, histone methylation can also regulate transcription in both a positive and negative manner. Both the methylation position and number of marks (mono-, di-, tri-) determines if this regulation is positive or negative. For instance, H3R2me2 (di methylation of arginine 2 of histone 3)[30] and H3R17me2-3[31]methylations are associated with positive regulation, whereas H3K9me2-3[32] are associated with

repression of genes. Specifically, coactivator-associated arginine methyltranferase-1 (CARM-1) was demonstrated to promote transcription of AR regulated genes in an indirect fashion. Upon androgen stimulation, CARM-1 is recruited to ARE where it interacts with p160 family members and p300/CBP complex to activate gene transcription. It was demonstrated that loss of CARM-1 cause a reduction in AR-mediated gene transcription[33]. Similar to CARM-1, protein arginine methyltransferase (PRMT-5) induces AR-mediated gene expression via p160 family proteins[34]. In contrast, H3K27me3 by EZH2, a member of Polycomb Repressive Complex 2 (PRC2), was related to transcriptional repression[35]. However, EZH2 does not always negatively regulate AR transcription, and a recent study found EZH2 can also act as an AR co-activator, independent of H3K27 methylation[36]. Similar to EZH2, SUZ12 which is another member of PRC2 complex is also necessary for H3K27me3[37]. Demethylation of the histones can correspondingly impact AR-mediated transcription. Lysine specific demethylase 1 (KDM1A) demethylates H3K9me2 and induce AR-dependent transcription. Knockdown of KDM1A decreases androgen stimulated transcription program in prostate cells. Upon androgen stimulation, KDM1A co-localizes with AR and demethylates H3K9me1, H3K9me2 and H3K9me3 at AR target gene promoters. By removing these repressive histone marks there is an increase in of AR-regulated transcription[38]. In addition to histone mark associated co-activators, recent study discovered another co-regulator called Grainyhead-like 2 (GRHL2) transcription factor. GRHL2 is an AR regulated gene that is upregulated in PCa. Surprisingly, GRHL2 maintains AR expression in various PCa cell lines whereby it can bind DNA without AR localization. Suggesting that AR and GRHL2 are linked in a positive feed-back loop that regulate genes towards disease progression[39].

HISTONE	MODIFIED SITE	MODIFICATION PATTERN	MODIFYING ENZYME
H1	Lys26	Methylation	Ezh2
H3	Arg8	Methylation	PRMT5
H3	Arg17	Methylation	CARM1
H3	Lys27	Methylation	Ezh2, G9a
H4	Arg3	Methylation	PRMT5

1.7 Coding and non-coding driver mutations

The growth and proliferation of a cell works in a complex balance that can be affected by genetic variations. Although most potential mutations are corrected by the DNA repair mechanisms, some remain uncorrected and aggregate throughout the time. Depending on the location and type of mutation, these can induce neoplastic transformation, whereby cellular growth becomes uncontrolled and tumorigenesis occurs[40]. Therefore, cancer is thought of a disease of genetic mutations. Comparison of cancer from large-scale sequencing projects demonstrated that few mutations

commonly occur in all cancers[41–43]. Potential reasons for this heterogeneity include 1) the same networks could be affected by different mutations that will give same output, 2) random factors could cause alterations on different regions of the genome. Therefore, mutations can be classified in terms of their damage into two groups; passenger and driver. Passenger mutations are neutral mutations that do not contribute to proliferative advantage. While the vast majority of these mutations are passenger mutations that do not alter cell growth, a few of them cause cancer transformation. These driver mutations give increased genetic fitness to the tumor [44]. These provide an evolutionary advantage by either increasing survival or proliferation of the cancer[45].

The exponential decrease in the cost of DNA sequencing price has allowed a large number of cancers to be sequenced in an effort to identify neoplastic driver mutations. Large-scale sequencing projects such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) conducted whole genome or exome sequencing of >20,000 patients from 22 cancer types. Characterization of the mutations showed that only ~7% percent of the mutations are recurrently enriched among patients. Of the remaining 93% it is impossible to currently know if these are driver mutations in the so-called "long-tail" or simply passenger mutations[46].

Of the validated driver mutations, sequencing studies in prostate cancer have primarily focused on protein coding mutations[47–51]. In these studies, single nucleotide variations (SNVs) on *SPOP*, *FOXA1*, *IDH*, *TP53*, *KDM6A*, *KDMT2D* and gene fusions on ETS family genes are commonly occurring mutations. When, prostate cancer patients are grouped in 7 subtypes based on alteration of those genes, alterations of SPOP and FOXA1 genes were shown to have highest expression of AR-mediated transcripts[50]. Moreover, from the same study, 20% of the patients had inactivated DNA repair genes[50]. For gene fusions, *TMPRSS2*:*ERG* is the most common fusion events in prostate cancer and is seen in 50% of North American patients[52] (**Figure 6**). The resulting fusion produces an oncogene whereby androgen regulated TMPRSS2 causes increased expression of ERG[53]. Even though this fusion is very common among prostate cancer patients, its clinical role is poorly clarified. Some studies found that there is a strong correlation between this fusion with disease outcome[54–56] while others found the opposite[57,58] .

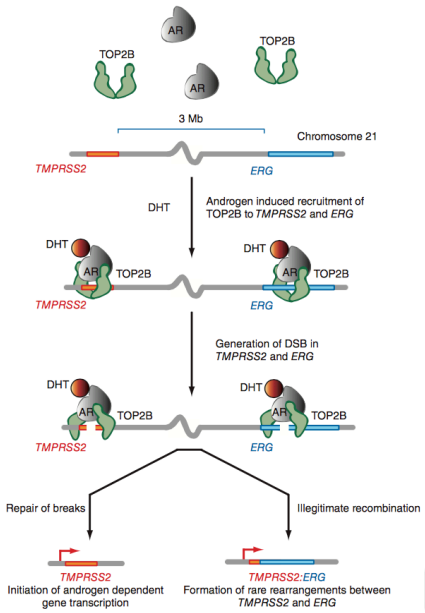

Figure 6: AR-driven TOP2b recruitment causes TMPRSS2-ERG fusion. Genotoxic stress and faulty DNA repair on these regions give rise to illegitimate recombination that give rise to fusion of TMPRSS2 and ERG genes.

Yet, considering that coding regions covers only 2% of the genome, noncoding regions can also be important. Non-coding genome consists of promoters, enhancers and insulators. Although their contribution to tumor progression is more challenging to elucidate, there is increasing evidence that non-coding mutations can also act as drivers. In 2013, a SNV in the promoter region of telomerase reverse transcriptase (TERT) was found to commonly occur in melanoma[59]. This C->T somatic mutation gave rise to a *de novo* ETS transcription motif that increased the expression of *TERT* gene. As a result, the chromosomal telomeres were elongated thereby resulting in a reduced rate of apoptosis[59]. After this initial study, the regulatory regions of the genome have been investigated in many cancer projects. Recurrent non-coding mutations on the enhancer of *TAL1* in T-cell lymphoblastic leukemia was demonstrated that small insertions could cause *de novo* MYB binding sites that acts as super-enhancer initiator[60]. Similar to T-cell lymphoblastic leukemia, chronic lymphotic leukemia patients also have a recurrent mutation on the enhancer of PAX5 transcription factor that is essential for B-cell-development. Finally, same leukemia patients were found to have recurrent mutations on 3' UTR of *NOTCH1* gene. These mutations cause abnormal splicing events that lead to increased *NOTCH1* activity, and correspondingly a more aggressive form of the disease[61]. From 19 prostate cancer patients, 2 recurrent SNVs were found in FAM48A binding site that lies in the promoter of *WDR74* gene[62], which has been previously associated to cell cycle control and apoptosis[63]. Further, in the prostate cancer cohort of ICGC revealed that recurrent mutation on 3'UTR of FOXA1 gene was found in prostate cancer patients[46]. These non-coding driver mutations were obtained based on genome-wide hotspot calculations. However recent work in colorectal cancer, demonstrated that mutation density of CTCF binding regions has higher rate of mutation than average

genome[64]. Therefore the identification of potential non-coding drivers can be influenced by transcription factor binding sites.

1.8 Mutation calling process & limitations on the field

Mutation calling is the first step of any cancer genomic studies. Briefly, the aim of this is to determine the difference between reference and the sample considering multiple parameters such as read quality, flaking read quality, read depth and variant allele frequency (VAF)[65,66]. Ideally, these mutations should be identified regardless of VAF if the read depth is high enough. Further complicating mutation calling, the mappability of the genome is not evenly distributed due to the long repeating segments[67]. Simple mutation calling consists of 3 steps; read processing, mapping and variant calling. Read processing will remove and trim low-quality reads to eliminate noise. The resulting high-quality reads are then mapped to the reference genome with one of the standard tools including BWA or Bowtie[68–70]. Depending on the type of the mutations, splice-aware mappers can also be used in order to study indel and rearrangements. At the mutation-calling step, somatic and germline mutation calling are separated. Briefly, germline variants have either 50% or 100% VAF with a genotype that is AA, AB or BB[71,72]. In contrast, somatic variants VAF are much more heterogeneous and often found at much lower frequencies than germline mutations. This variability is often due to the nature of tumor heterogeneity. Consequently, crude filtrations of germline variants cannot be applied and variables such as; rare tumor subclones, sample impurity and possible sequence errors should be extensively evaluated. While advances in the NGS technology have allowed cheaper sequencing and correspondingly increased read depth, somatic mutation calling still remains challenging. One may to reduce any potential caller bias is to use multiple algorithms and then use the consensus mutation calling. This conservative method increases the confidence of the mutations and reduces false positives. For example, some of the algorithms perform better at low VAF conditions, but may be more vulnerable to normal-tumor sample contamination[73]. Therefore, using a consensus is a simple but effective method to reduce noise and increase confidence[74]. As such it has been extensively used with the ICGC project[46].

1.9 Factors that affect mutation rates

Mutations are not evenly distributed throughout on the genome and can be affected by multiple factors including: sequence context, GC content, distance to telomere, replication timing, chromatin state and transcription factor binding.

1.9.a) GC content and sequence context

The base composition of the genome is not evenly distributed and GC percentage is relatively higher than the AT in humans. Therefore, the base composition is not at equilibrium. To balance the equilibrium, there must be GC→ AT substitutions at GC high regions[75,76]. Therefore, during DNA replication recombinant DNA mismatches have a tendency to be GC-biased. Specifically as there is

greater GC content, there are more GC mutations than AT mutations[77]. In addition to the effect of GC dinucleotide context on mutation rates, the effect of the flanking bases can also impact mutation rate. Specifically, the flanking bases in both directions can alter the mutation rate by stabilizing mismatches[78–80]. Furthermore, the GC content can also enhances methylation-mediated mutation rate as DNA frequently occur on CpG dinucleotides which frequently give rise to C->T substitutions following spontaneous deamination of 5-methylcystosine (**Figure 7**)[81].

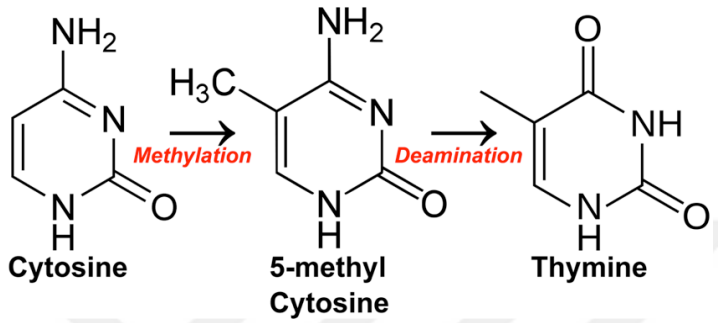

Figure 7: Deamination of methylated cytosine. Upon spontaneous deamination, methylated cytosine turns to thymine base. Figure taken from[82] .

1.9.b) Distance to telomere

Telomeres are six-base pair repeated sequences that protect chromosome ends from chromosomal fusions. It has been previously shown that length of the telomere affects double stranded break sensitivity[83,84]. Multiple modeling studies have shown that distance to telomere is an accurate predictor of variation rate and it is negatively correlated with mutation rates[84–86].

1.9.c) Replication timing

Replication occurs in a segmented process that occurs at different time across the whole genome the timing of replication affects the rate of mutation on the genome. This is due to an accumulation of mutation on single-stranded DNA (ssDNA) regions at replication fork. ssDNA is vulnerable to DNA damage alkylation, oxidation and deamination. Thus, aggregation of mutations is more potent in late replicated regions than early replicated regions[87](**Figure 8**). This has also been demonstrated in colorectal cancer patients whereby regions that late replicating regions have higher rate of mutation than early replicating regions[64].

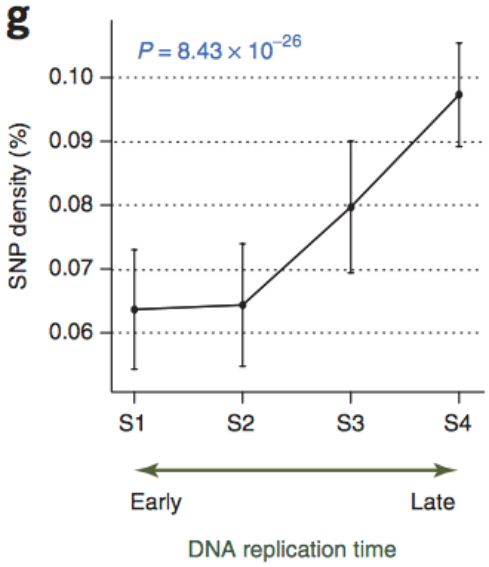

Figure 8: The effect of replication timing on mutation burden. It is clearly seen that mutation rate increases as the part of the genome replicates in the late stages. Figure taken from[87].

1.9.d) Chromatin state

Chromatin structure is very crucial for activation and repression of genes. Euchromatin is lightly packed form of chromatin that allows active transcription[88]. However, despite this accessibility it does not mean there is a higher rate of mutations. In fact, there are generally less mutations at euchromatin as compared to heterochromatin as while mutations may happen more frequently, the damaged DNA is more accessible to DNA repair enzymes. Therefore, euchromatin generally has less mutations[59,64,89,90].

Recent advances in the sequencing technology showed that genome structure is not linear[91], but it is rather 3D. Therefore, two distant locations of the genome could interact with looping under appropriate conditions. These interactions could also give rise to large-scale deletions and rearrangements in cases when high chemotoxic stress is present and DNA repair mechanisms are faulty[92].

1.9.e) Transcription factor

A recent study discovered that there were higher rates of mutations at TF binding sites. Specifically, CTCF binding sites have 3.5 fold enrichment compared to the corresponding relative control regions[93]. Moreover, another study also demonstrated that CTCF binding sites have a significantly higher rate of mutations in colorectal cancer[64]. Furthermore, a landmark paper by Sabarinathan et al. demonstrated that this increase in mutations was due to TF binding and preventing DNA repair. This was demonstrated by combining a published nucleotide excision repair (NER) map with somatic mutation data from melanoma patients[94]. The authors demonstrated that accessibility to UV damaged regions by NER is critical to repair DNA damage and the regions with higher accessibility have

decreased mutation rates. However, when TF is bound to chromatin regions the overall accessibility was limited due to steric conflicts, which thereby prevented NER. This led to an increase in mutations at the TF binding regions[94] (**Figure 9**).

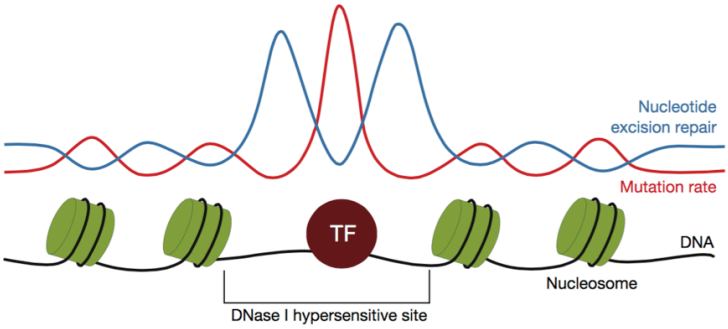

Figure 9: Mutations caused by TF binding blocking NER. In open chromatin, DNA is more accessible by other proteins or mutagens. However, these regions also accessible to DNA repair machinery. Upon TF binding, accessibility of the DNA becomes reduced. DNA repair machinery cannot reach at the mutated regions. Hence mutations are aggregated in proximity of TF protein. Figure taken from[94].

1.10 AR and DNA Damage

In addition to it's role in PCa growth, AR-mediated transcription has also been demonstrated to regulate expression of DNA damage repair (DDR) genes such as *ATM*, *CHK2* and *PARP*[5]. Interestingly, hormone stimulation is not always necessary for AR mediated gene transcription. Upon genotoxic insult, DNA damage occurs and this up regulates *p53* and *E2F1*. These proteins in turn alter AR regulation, causing an increase in AR-mediated transcription via E2F1 and p53 [95,96]. Activated AR the up regulates DDR genes that repair genotoxic stress[97]. Thus instead of hormone activation, AR activity was stimulated by DNA damage.

AR-mediated regulation of repair genes is not specific to one DNA repair mechanism. For example, both Flap endonuclease (*FEN1*), a base excision repair enzyme, and *XRCC6*, a non-homologous end joining component, are both up-regulated by AR[14,98,99]. Expressions of DNAPKC are also important for AR-mediated DNA repair[100]. It was demonstrated that DNAPKC also act as an AR co-activator and when down-regulated the AR transcriptional program is significantly decreased. AR mediated DNA repair genes activation and DNAPKCs are therefore suggested to work in a positive feedback loop that amplifies each other's activity[97] (**Figure 10**).

In addition to the AR's regulation on DNA repair pathways, AR has been shown to directly cause DNA damage. To initiate transcription, AR recruits multiple proteins to the binding site including topoisomerase IIb (TOP2b). This enzyme relieves the torsional stress caused by transcriptional machinery by introducing double stranded breaks (DSB) on DNA[101]. In normal conditions, DSB are not recombinogenic and are easily repaired by DNA repair machinery. However, in neoplastic cells TOP2b induced breaks are not repaired successfully due to the high genotoxic stress and can cause

fusion events [102]. Similar to TOP2b, topoisomerase I (TOP1b) also introduces DNA breaks to relief torsional stress. However, TOP1b only introduces single stranded breaks. *TOP1b* is required at AR/NKX3.1 occupied enhancers to activate androgen-induced genes such as *KLK3* and *KLK2*[101]. Unlike TOP2b, it is not believed that TOP1b induces gene fusions. However, TOP1 could cause deleterious effects if it is not removed after its utilization and act as an obstacle for transcription machinery which could lead to collapse of the replication fork at these regions[103].

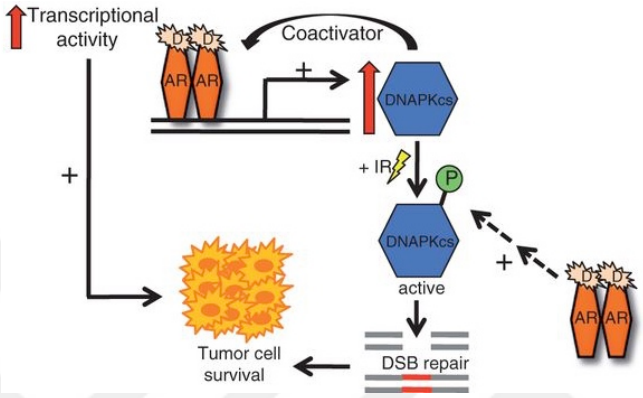

Figure 10: Positive feedback loop of AR and DNA damage response.

Aim of this work

Somatic mutations affect the progression all cancer types. Recent studies have demonstrated that TF binding can increase somatic mutations. However, it is not clear if the increase in mutations is due to the TF or the genetic location. Given the importance of AR in almost all prostate cancers, this nuclear receptor presents a powerful model to investigate TF-mediated mutations. Therefore, the aim of this work is to investigate AR binding site mutations in clinical cancer data and characterize variants in these regions understand their impact on prostate cancer progression.

Chapter 2: Methods

2.1 File Formats

2.1.1 FASTQ

Fastq file format is a common file storage format that keeps biological sequence information with their sequencing quality. This format is human readable therefore can be view with text editors. It is consisted of 4 rows for each read. The first row represents the id of the read, which was generated by the sequencing machine. The second row contains the sequence information. The third row is a placeholder that is built for additional information. The forth row contains quality scores of each complementary base.

2.1.2 BED

BED format is the simplest form of storing genomic location information. Minimum bed file requires at least three-column information that keeps chromosome, start and end information. Depending on the necessity, information such as region id, region score, region strand and desired region colour can also be stored in this format. However, most tools demand only the first 5 column of bed file, therefore additional information is either removed or discarded during analysis.

2.1.3 VCF

Variant call format (VCF) was initialized with 1000 genomes project. The aim was to come up with a standard method of storing variant information. VCF files have 8 fixed fields for each mutation call. These are chromosome, position, unique identifier, reference base, alternated base, quality, filter and info. While most of the fields are self explaining, filter and info fields deeper explanation. Each mutation caller determines variants under specific assumptions whereby a standardisation is required. Filter column tags each variant their eligibility to considered as high confidence mutation as "PASS". Info field contains attributes of that variant such as the overall flanking base quality or variant allele frequency.

2.2 Mutation Information of ICGC Patients

Whole genome sequencing data was obtained from Pan Cancer Analysis of Whole Genomes (PCAWG) release on August 24, 2016[104]. For PCa, only those patients with primary cancer (n=196) were included in the study due to the limited number of patients with metastatic or late-state prostate cancer. SNV and indels were previously called with three different mutation-calling algorithms (Sanger: INDEL=Pindel, SNV=Caveman; DKFZ:INDEL+SNV=Platypus; Broad: INDEL=Snowman, SNV=Mutect). Only those mutations which had been called by two or more callers and not found in dbSNP (v147) were used in this work.

2.3 Transcription factor binding sites

ChIP-seq data was obtained for the following published work: FOXA1(GSM1410788), CHD1(GSM1573653), CTBP1(GSM1410762), CTBP2(GSM1410763), CTCF(GSM1006887), ETV1(GSM1145322), EZH2(GSM969570), GATA2(GSM941194), HOXB13(GSM1716764), MRE11A(GSM1543776), POLR2A(GSM1415124), RUNX1(GSM1527840), SUZ12(GSM969572), TCF7L2(GSM1249449), TET2(GSM1613322), TOP1(GSM1543792), WDR5(GSM1333369), KDM1A(GSM1279769), EP300(GSM686943), MED12(GSM686945), AR(GSE83860), GRLH2(GSM2122802), H3K9ME3(GSM353610), H3K4ME1(GSM1410780), H3K4ME3(ENCODE: ENCFF401MDR), H3K27AC(GSM1249448), H3K36ME3(GSM875814). Clinical ARBS were identified from AR ChIP-seq of 13 tumour and 7 normal human tissue samples (GSE70079). Overlapping peaks were identified by HOMER's (v4.7) *mergePeaks* function (–d

parameter 200)[105]. All binding sites that overlapped with UCSD blacklisted regions were removed. Motif driven peaks were predicted by PWMtools[106] with given positional weight matrixes obtained from JASPAR DB[107].

2.4 Peak Annotation

We used *annotatePeak* function of HOMER[105] to annotate peaks based on their chromosomal location. Hg19 genome build was used during annotation along with HOMER's default libraries. To get proportions of each chromosome location, we added *–annStat* option.

2.5 Investigation of TF binding motifs on genomic regions

We used HOMER's *findMotifsGenome* to search previously published binding motifs on genomic intervals. Similar to peak annotation we used genome build hg19. We did not use any additional options and used *findMotifsGenome* function with default settings[105].

2.6 Determination of intersecting regions

Bedtools[108] (version 2.26.0) and bedops[109] (version 2.4.26) were used to intersect, manipulate and filter specific regions in bed and vcf files. To extend binding regions bedtools *slop* function was used. For intersection and filtration, we used bedtools *intersect* and bedops *bedmap* function.

2.7 Comparing specific region mutation frequency with background

Bedtools[108] *shuffle* function was used to generate randomized regions across the genome. Each bed file was randomized 1000 times to generate a null distribution. All gapped regions (UCSC gapped regions) were removed. To generate random bed files with similar base composition (ATCG) of each random region we extensively randomized the AR binding data and then calculated base composition. We then z-normalized each nucleotide type columns identify those random bed files similar to ARBS 250 bed file (as null value). The peak files which have the base composition that are in the ± 2 standard deviation (sd) range were selected.

2.8 Mutation Signature Analysis

Mutation signature analysis was done using the bioconductor package SomaticSignature (version 2.12.1) with R version 3.4.0[110]. Mutation signature were obtained from *plotMutationSpectrum()* function with default parameters. Those TFs with less than 480 mutations across all patients were not included in our analysis. This value was used as it was demonstrated to have a deciphering accuracy of >0.95 for two mutation signature[111]. As previously published, the *cosine()* function from the 'lsa' package was used to calculate the similarity between signatures obtained from SomaticSignature *motifMatrix()* function[111].

2.9 Mutation aggregation analysis on TF binding regions

For each of the binding regions, overlapping mutations were mapped and mutation distances to the

center of the TF binding region were calculated. For a given TF, each of the binding regions were overlapped based on their center. Mutation densities of 100 bp windows were calculated with smooth kernel density method. Calculation and visualization was conducted with ggplot2 R package.

2.10 Methylation Analysis

CpG positions were identified from a published custom made perl script (https://www.biostars.org/p/68352/#256983). DNA methylation was quantified from whole genome bisulfide sequencing from LNCaP cells (GSE86832). Methylation data points with coverage less than 10 were excluded from our analysis. Those locations with a DNA methylation less than 0.52 (median of LNCaP) were classified as unmethylated. Intersecting CpG of each peak was combined as a vector. Then all of the methylated and unmethylated sites were summed up to obtain single value of overall methylated rate of a TF. For given TFs, the intersection between TF and whole genome CpG was obtained.

2.11 Heatmap

'pheatmap' package (version 1.0.8) was used for drawing heatmap from CRAN package repository *pheatmap* function was used in default settings to produce heat maps based on pairwise cosine similarity values of mutation signatures.

2.12 Statistical Analysis

The distribution of mutation events limits the usage of parametric tests. For preventing biasing, we used R statistical language default *wilcox.test*() function is used for Wilcoxon rank sum test.

2.13 Visualization

Data was visualized with *ggplot2* (version 2.2.1) and Venn diagrams were drawn in RShiny app, https://github.com/jolars/shiny-server.

Chapter 3: Results

Overview

Somatic mutations have diverse affects on cancer progression. While some of them are neutral, even a few of them could drastically increase cancer proliferation. Previous studies have indicated that TF binding can increase somatic mutation rates. While aforementioned factors still affect genomic alteration rates, mechanism behind TF mediated mutations has not been revealed yet. As prostate cancer progression is critically affected by AR's action, this transcription factor could represent TF mediated mutations. Therefore, the aim of this work is to investigate AR binding site mutations in clinical cancer data and characterize variants in these regions understand their impact on prostate cancer progression. To achieve our aim, we three distinct steps in the project as following:

1. Determination of AR binding regions in clinical patients
2. Generation of mutation framework
3. Analysis of Prostate cancer and other Pan-Cancer mutations

3. 1 Determination of AR binding regions

The goal of this work is to investigate mutations at transcription factor binding sites in clinical primary prostate cancer (PCa). Therefore, to identify those clinical binding sites we initially used ChIPseq from patient tumours. We found two projects in the GEO Database that have done AR ChIPseq (GSE28680) with clinical samples and picked the study with higher number of samples (GSE70079). This study from Pomerantz *et al*.[112] contained ChIPseq datasets from 13 tumor and 7 normal prostatic samples (**Figure 11**). Each sample had information of AR binding site (ARBS) in the standard 4-column BED format that contained chromosome, start, end and binding intensity information. As an exploratory analysis, we compared the overlap in all peaks between the tumour/normal tissue and LNCaP cells to find the intersecting binding regions across different sample (**Figure 12**); (See methods). Overall, we found that in commonly occurring peaks that were present in multiple samples generally had an increase in in peak signal strength **(Figure 13.A-B***)*. To limit heterogeneity, which could alter our analysis of TF binding sites, we chose to stratify the peaks into high and low confidence binding regions. High confidence peaks was defined as those binding sites that were found in either all of the tumour or normal samples. In contrast, low confidence peaks were simply the remaining peaks. While strict, this dramatically limited the AR clinical peaks to the most robust 5%.

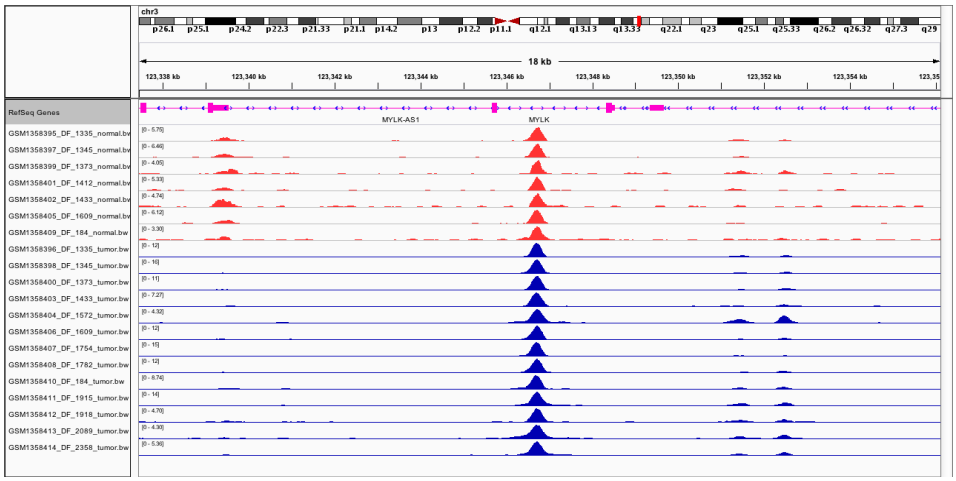

Figure 11: AR ChIPseq in clinical samples. Red colored tracks represent normal samples, where as blue ones are tumor. Binding regions are the ChIPseq peaks.

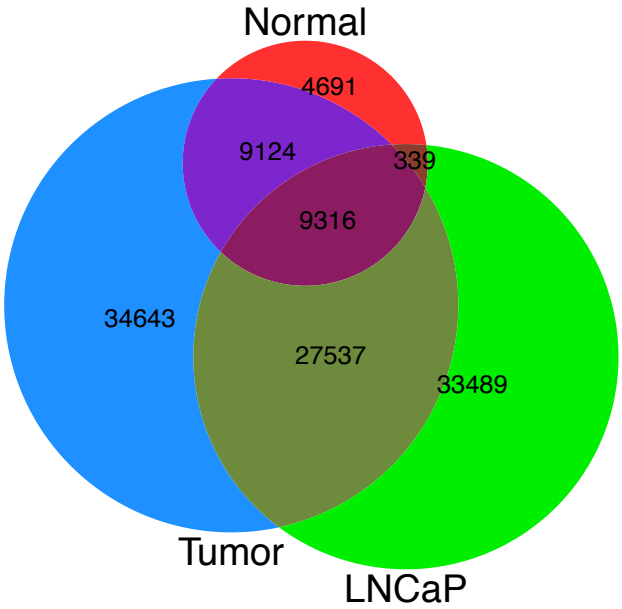

Figure 12: Weighted Venn diagram union of clinical and LNCaP ChIPseq. It is clearly seen that only 50% of the LNCaP regions overlap with tumor or normal samples.

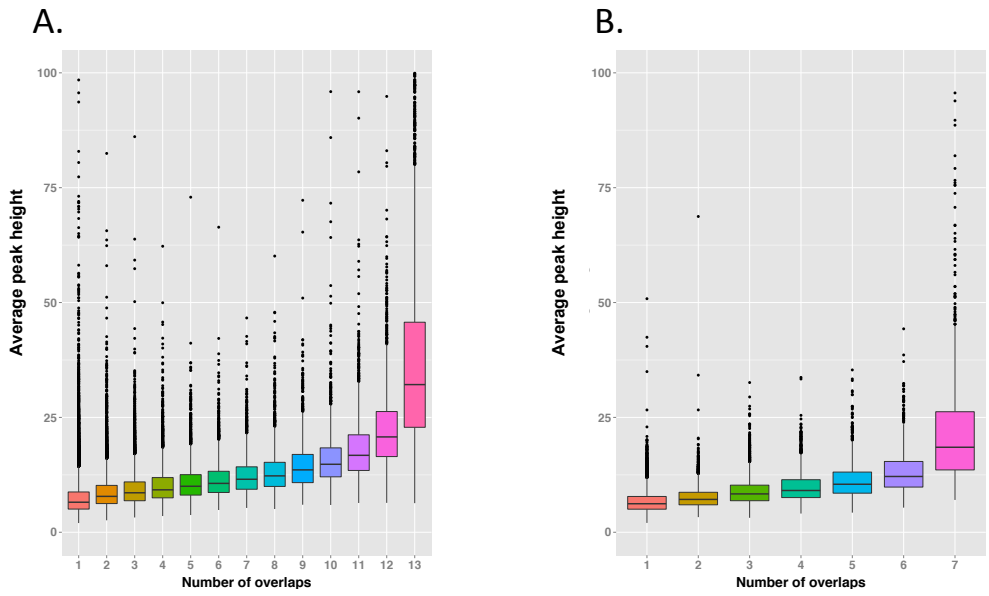

Figure 13: Correlation between peak overlap and height. Each peak are separated based on their overlapping partners and averages of each overlap were calculated. A) 13 tumor samples. B) 7 normal samples.

When we intersected the high and low confidence peaks for tumor and normal we found that there was a similar ratio between normal and tumour peaks suggesting that there was little bias in the peak stratification (**Figure 14**). To reduce signal noise, we therefore continued our project using the high confidence-binding ARBS.

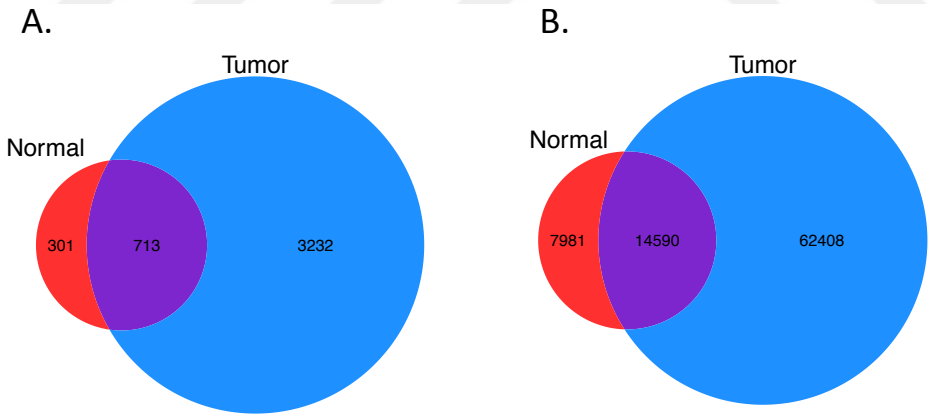

Figure 14: Overlap between tumour and normal peaks. High confidence peak overlaps on A and low confidence peak overlaps on B.

After identifying high-confidence sites we then characterized the properties of these ARBS. First, we looked for DNA motifs at these genomic regions. As expected, the canonical ARE motif was the most common (**Figure 15**). However, this motif was not universal and only 30% of all peaks have an ARE motif, suggesting that AR does not directly binding to DNA at most sites. To investigate the effect of ARE motif to AR binding intensity, we compared peaks height of those samples with an ARE motif (n=1079) or no ARE motif (n=3060). ARE motif peaks appeared to have higher average peak height

than no ARE motif peaks (*Wilcox.test p-value < 2.2e-16*) (**Figure 16**). Next, we annotated the high confidence ARBS based on the location of the binding regions. Similar to previously published work in both clinical samples and cell lines, >95% of the binding regions were located at intronic and intergenic regions (**Figure 17**). This supports the hypothesis that the primary role of AR is to bind enhancers to induce transcription, as few ARBS are located in promoter regions.

Homer Known Motif Enrichment Results (AR_motifoutput)

Homer *de novo* Motif Results
Gene Ontology Enrichment Results
Known Motif Enrichment Results (txt file)
Total Target Sequences = 3416, Total Background Sequences = 45099

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif
1		ARE(NR)/LNCAP-AR-ChIP-Seq(GSE27824)/Homer	1e-844	-1.944e+03	0.0000	1079.0	31.59%
2		GRE(NR),IR3/A549-GR-ChIP-Seq(GSE32465)/Homer	1e-814	-1.876e+03	0.0000	871.0	25.50%
3		GRE(NR),IR3/RAW264.7-GRE-ChIP-Seq(Unpublished)/Homer	1e-803	-1.851e+03	0.0000	1099.0	32.17%
4		PGR(NR)/EndoStromal-PGR-ChIP-Seq(GSE69539)/Homer	1e-652	-1.503e+03	0.0000	1072.0	31.38%
5		FOXM1(Forkhead)/MCF7-FOXM1-ChIP-Seq(GSE72977)/Homer	1e-487	-1.123e+03	0.0000	1929.0	56.47%
6		FOXA1(Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	1e-484	-1.115e+03	0.0000	2131.0	62.38%
7		FOXA1(Forkhead)/MCF7-FOXA1-ChIP-Seq(GSE26831)/Homer	1e-482	-1.110e+03	0.0000	1942.0	56.85%
8		PR(NR)/T47D-PR-ChIP-Seq(GSE31130)/Homer	1e-476	-1.097e+03	0.0000	2308.0	67.56%
9		Foxa2(Forkhead)/Liver-Foxa2-ChIP-Seq(GSE25694)/Homer	1e-410	-9.448e+02	0.0000	1595.0	46.69%
10		Fox:Ebox(Forkhead,bHLH)/Panc1-Foxa2-ChIP-	1e-	-8.655e+02	0.0000	1605.0	46.98%

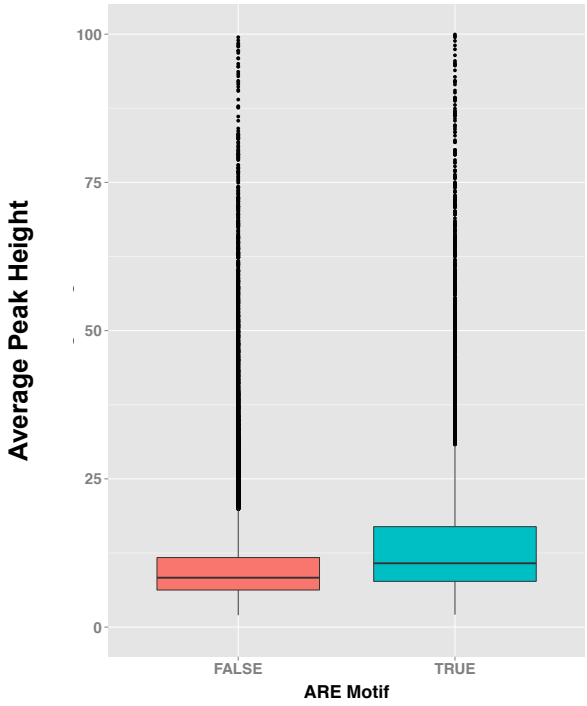

Figure 16: Comparison of height of those peaks with different motifs. Overall those peaks with an ARE motif appeared to have a higher peak height than those peak which have got no ARE motif.

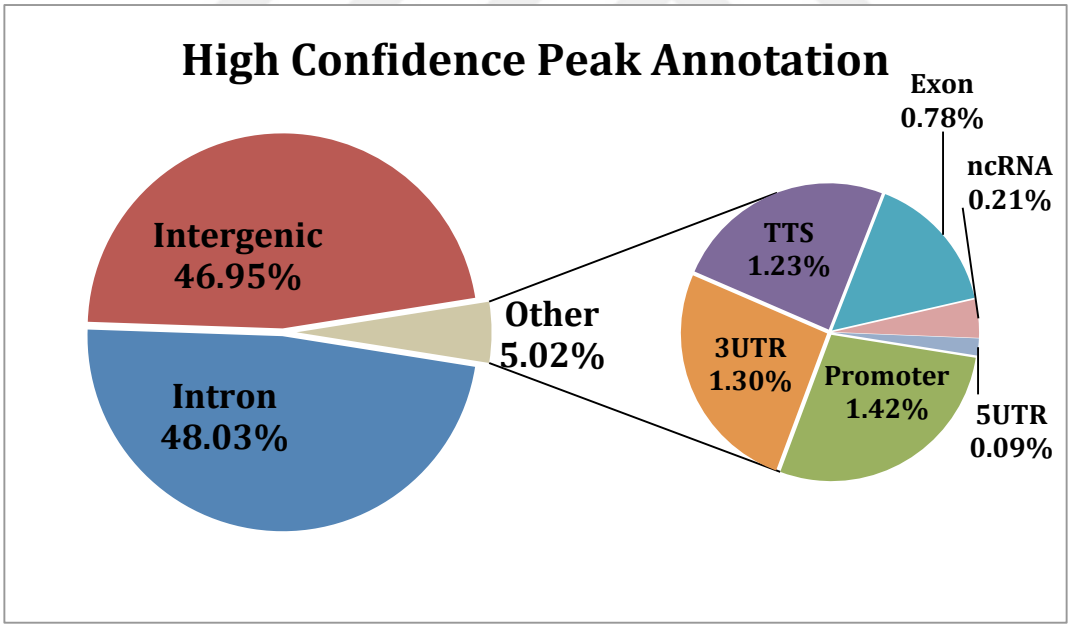

Figure 17: Characterization of AR binding sites. Most of the binding regions are located at intergenic and intronic regions of the genome.

In addition to the high-confidence clinical ARBS we wanted to investigate how other TFs affected mutation rate in prostate cancer. However, due to a lack of clinical ChIPseq we chose to use studies with the common secondary cell line LNCaP. Therefore, we downloaded all

available ChIPseq data for LNCaP cell line *(***Table 3***)*. The majority of these studies were generated from the ENCODE project.

Table 2: List of GEO identifiers for all ChIPseq used in this project

GEO id	TF/Histone Type
GSM353610	H3K9me3
GSM1410780	H3K4me1
ENCFF401MDR	H3K4me3
GSM1249448	H3K27ac
GSM875814	H3K36me3
GSM1410788	FOXA1
GSM1573653	CHD1
GSM1410762	CTBP1
GSM1410763	CTBP2
GSM1006887	CTCF
GSM1145322	ETV1
GSM969570	EZH2
GSM941194	GATA2
GSM1716764	HOXB13
GSM1543776	MRE11A
GSM1415124	POLR2A
GSM1527840	RUNX1
GSM969572	SUZ12
GSM1249449	TCF7L2
GSM1613322	TET2
GSM1543792	TOP1
GSM1333369	WDR5
GSM1279769	KDM1A
GSM686943	EP300
GSM686945	MED12
GSE83860	AR
GSM212280	GRLH2

Summary

We obtained clinical AR ChIP seq data from normal and cancer patients. This was subsetted to focus on high confidence ARBS that were present in all samples from the normal and tumour samples. In these high-confidence ARBS, we showed that the ARE motif was the present in only 30% of the ARBS and peak height increased with motif presence. In addition to the ARBS, we obtained TF binding site information from all available ChIPseq studies in

LNCaP cells. The binding sites of clinical samples in combination with those identified from cell lines will be used for mutation analysis.

3.2) Generation of mutation framework

3.2.1) Mutation calling and optimization

The Pan Cancer Analysis of Whole Genomes (PCAWG) is a recently completed large-scale project to do whole genome sequencing on >2500 cancer patients from 20 different cancers. This large dataset was ideal for our proposed project as it contained 214 prostate cancer samples. However, the initial data releases had only mutation data that been called one of three different algorithms. Given the variability of each algorithm it would not be possible to compare data called with different methods. As we were unsure when the final data would be released we conducted a pilot study to download and call mutations from whole genome sequencing data. Raw sequencing data (FASTQ) was downloaded to the Amazon Cloud Computation framework using the prebuilt mutation calling pipeline, bcbio suite[113]. From these test runs and optimization we calculated that running project on Amazon would be extremely expensive, though time efficient (**Table 4**).

Table 3: Calculated cost of calling 196 patient mutations from whole genome sequencing data.

Process	Rented Entitiy	Unit	Charge	Unit	Cost
Cpu Cost	48	hr	1.7	$/hr	$ 81.60
Storage Cost - EBS	750	Gb	0.1	$/gb	5
Storage Cost - AWS	350	Gb	0.03	$/gb	0.7
					$ 87.30
				+15%'	100.395
		TOTAL			
For 214 Patients	196	**$ 19,677.42**			

Therefore, we attempted to utilize our local High Performance Cluster. In a set of optimization experiments, we first investigated the CPU ~ time relation on chr6 exon data. We found that 4 CPUs per patient was the most optimized value of CPU utilization and increasing the number of cores >4 does not dramatically improve the processing time. With this we calculated the expected time required to process 100 patients. We found that in the most optimized settings it would take 5 days to finish exomic data for all patients. However, WGS is much more complex than only exomic data. When we ran one of the WGS datasets in our HPC with 4 CPUs, it took 2 days. Therefore, it would take ~60 days of 24/7 sequencing and mutation calling for all 200 patients assuming both no delays between processes or mistakes (**Figure 18**). Fortunately, as we began to initiate this process, ICGC released

consensus data of SNV and INDEL mutation calls. Therefore, at this point, mutation calling was not necessary and we could simply build our framework on this published data.

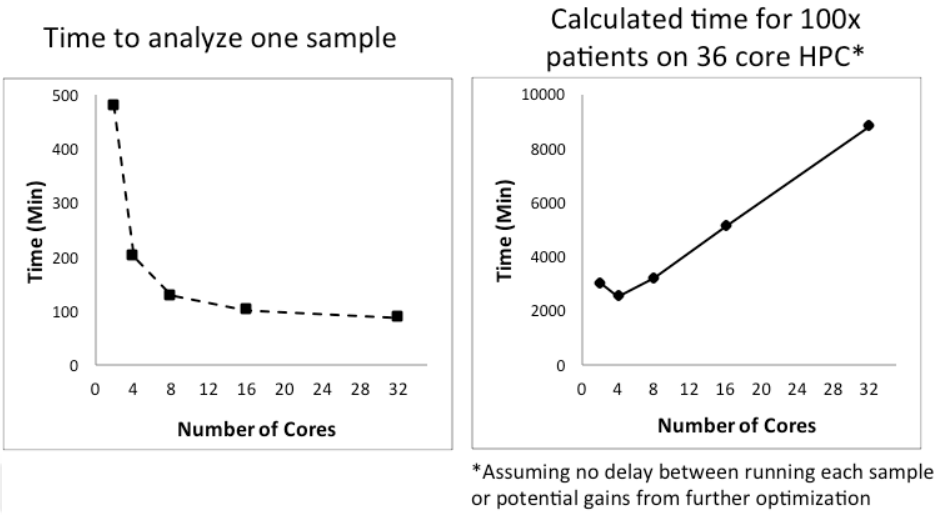

Figure 18: Optimization of somatic mutation calling. (Left) CPU ~ time optimization was investigated. (Right) In sum of 36 cores, time required to finish 100 patient data that x number per patient was shown.

3.2.2) Exploratory analysis of ICGC mutations

With the called ICGC-PCAWG data we combined 49,508,580 SNVs and 3,562,413 indels from 2576 patients with 20 different cancer types in a single database. As expected the overall mutation values were extremely variable between the different cancer types with prostate cancer have a relatively low frequency of both SNVs and INDELs (**Figure 19A)**. Other cancers, such as skin cancer, had several log more mutations than prostate cancer (**Figure 19B**). This low mutation number makes prostate cancer difficult to study as a large patient population is needed to provide sufficient statistical strength.

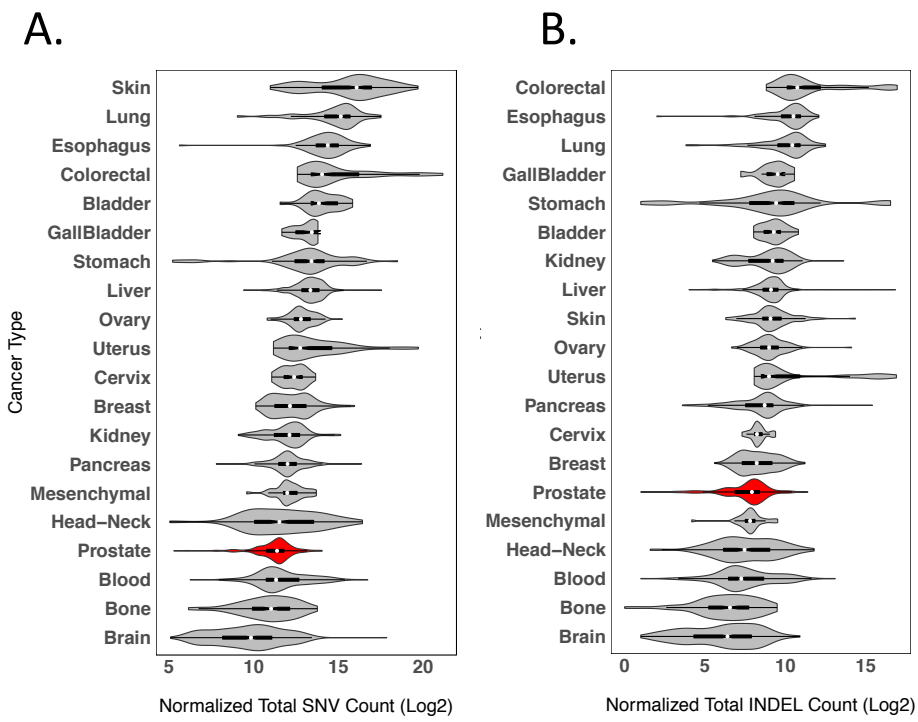

Figure 19: Mutations frequency for different cancer types. On the left side SNVs were compared, where as on the right side INDELs were compared.

Summary

We created a framework to keeps the large number of patient mutations from PCAWG for further analysis.

3.3) Analysis of Prostate cancer and other Pan-Cancer mutations

3.3.1) Investigation of mutation burden at Transcription Factor binding sites

With the previously identified binding sites (**Chapter 2**), we characterized the relative mutation rate at the TF binding sites (**Figure 20**). Initially, these mutation rates were compared to to randomly shuffled genomic regions. Many TF including HOXB13, EZH2 and SUZ12 were found to have more mutations than would be expected randomly. As previously published, DNAse Hypersensitive Sites (DHS) regions had a lower mutation frequency than the null distribution. This has been shown to occur as DNA repair mechanisms can access mutated region more efficiently in open chromatin regions[114]. Contrary to previously published results in melanoma and colorectal cancer, we did not see increased mutation rates at CTCF binding sites (**Figure 21**)[64]. However, ARBS from both LNCaP cell lines and clinical samples were found to have the the highest rate of mutations among all TF binding sites

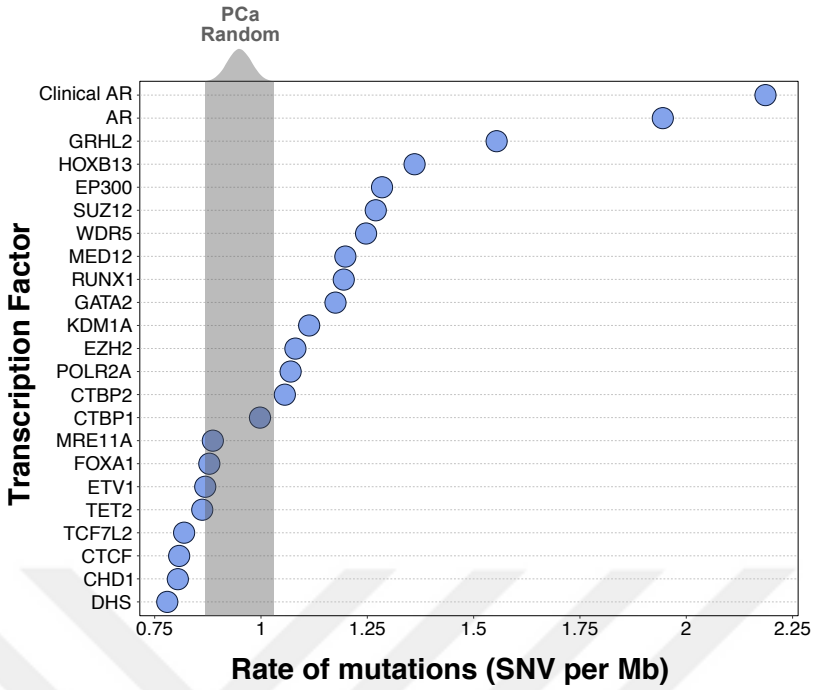

Figure 20: SNV at TF binding sites. The gray bar represents background distribution of prostate cancer patients with the distribution plot seen above the graph.

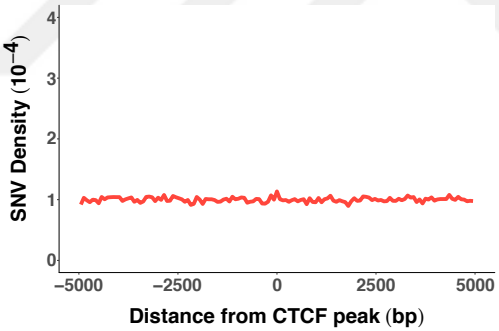

Figure 21: CTCF binding sites mutation density in PCa.

3.3.2) Comparison of AR proximity mutation rates in prostate cancer and other pan-cancers.

AR is essential for the growth of PCa. However, AR plays no role in almost all other cancers. Therefore, if the increased mutations observed are driven by AR binding rather than the chromosomal location, we would expect to see no increase in mutations at ARBS in other cancer types. To test this hypothesis, we compared the mutation frequency in 20 cancer types from 2576 patients. As predicted, prostate cancer ranked highest among Pan-Cancer cohort in ARBS mutations in individual and grouped comparisons (Wilcox t-test; $p<2x10^{-16}$) (**Figure 22A+B**). In fact, no cancer, except prostate, had higher rate of mutation in ARBS than randomly shuffled regions suggesting that ARBS mutations in prostate cancer are due to AR occupancy (**Figure 22C**).

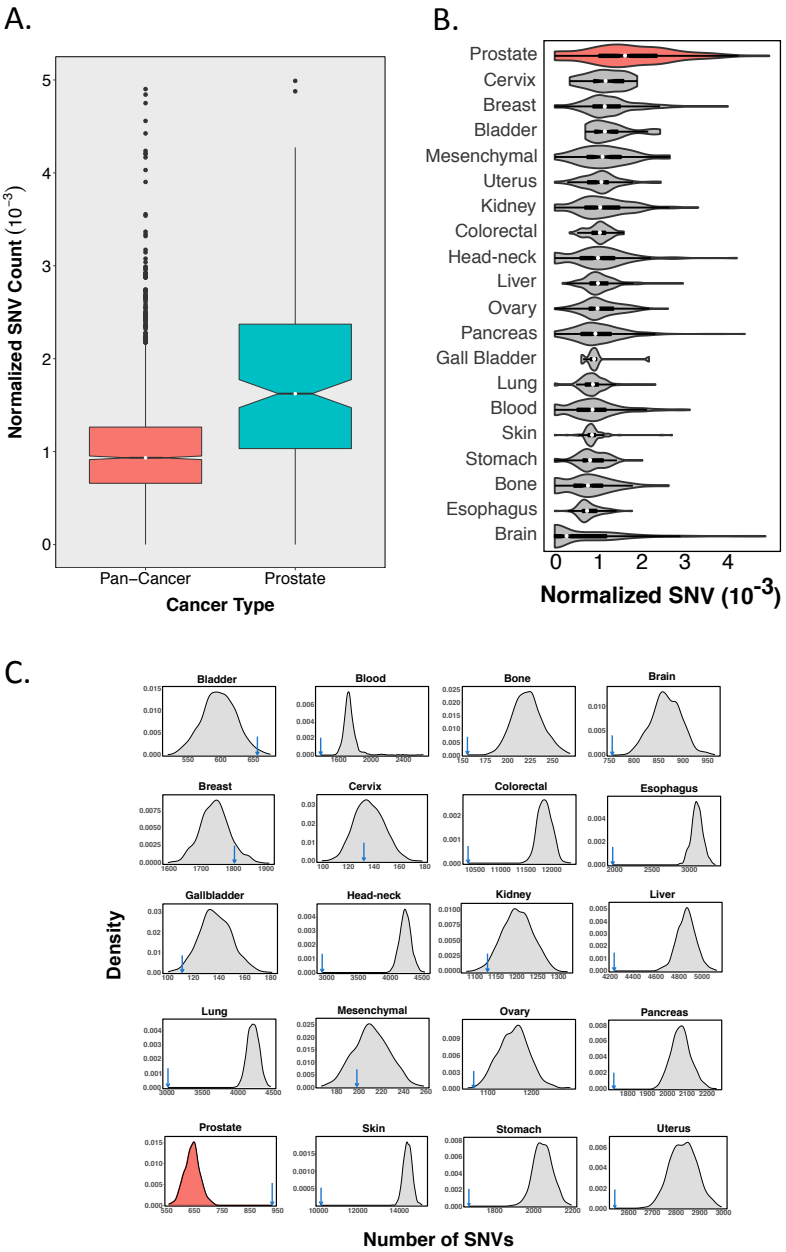

Figure 22: Mutation rate at ARBS in different cancer types. A) Normalized ARBS mutation numbers of Pan-Cancer versus and Pan-Cancer was compared. B) Prostate cancer had the highest number of ARBS mutations among Pan-Cancer cohort. C.Random mutation distribution (gray) of each cancers were compared with that cancers ARBS mutations (blue arrow). Only in prostate cancer (red) ARBS mutations were actually higher than the random. None of the 1000 randomized regions on the genome have higher rate of mutation than ARBS.

To the role of AR binding on SNVs, we investigated the mutation density at ARBS. Suggesting that the increase in mutations was due to AR occupancy, we observed a clear increase ±375bp to the AR peak. In contrast, there was no enrichment observed in Pan-cancer cohort (**Figure 23A**). To prove that this event was independent from the base composition of the ARBS, we identified those genetic regions that contain an ARE motif but no AR binding. We then quantified the mutation density at these ARE containing regions in both prostate and the Pan-Cancer cohort. Overall, there was no increase in mutations in either the Pan-Cancer or PCa patients. As mutation enrichment was only seen on those regions where AR protein binds, this suggests that AR protein is required for this mutation event and this is independent from the sequence context (**Figure 23B**). Supporting this theory, we observed a positive correlation between mutation frequency and AR ChIPseq peak height (**Figure 23C**). Using the AR peak height as a surrogate for protein occupancy, our results suggest that protein binding causes an increased rate of mutations.

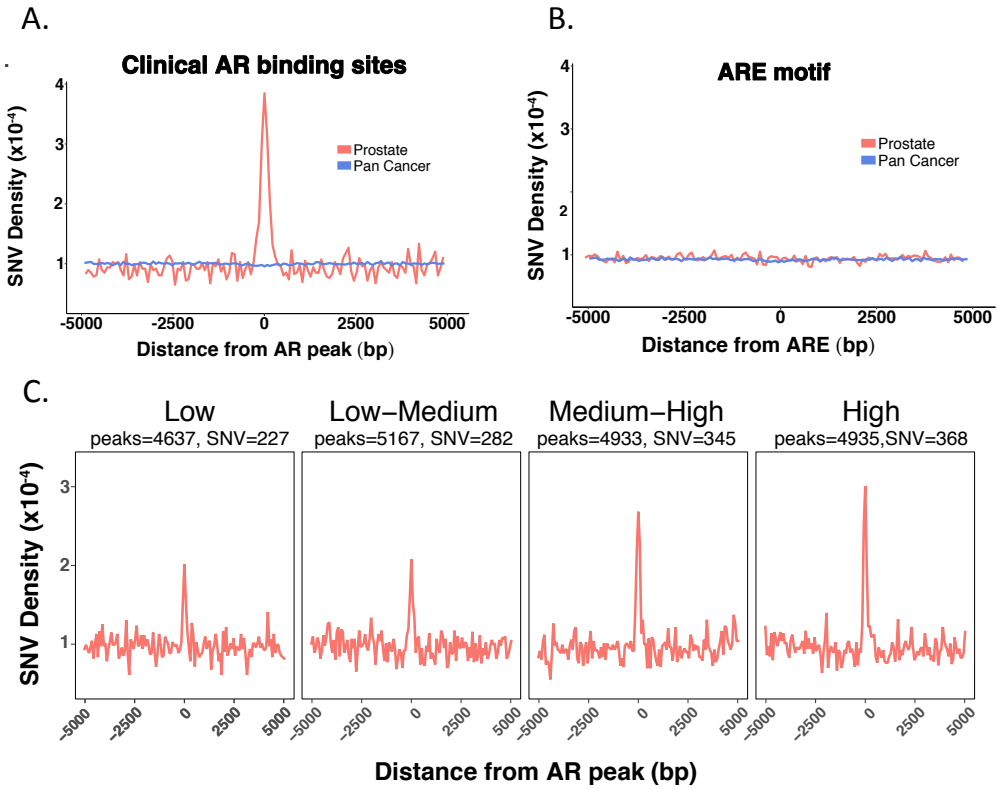

Figure 23: Distribution of mutations at AR binding sites. A) Clinical AR binding mutation density was investigated in both prostate cancer (red) and all other cancers (blue). B) ARE motif regions mutation density was investigated in prostate cancer (red) and all other cancer (blue). C) AR binding affinity is positively correlated with mutation rates.

Like prostate cancer, breast cancer is also a hormone dependent cancer. Therefore, to see if the observed increase in SNVs at cell-type specific TF binding sites is a broad phenomenon, we quantified the mutation rate at ER binding sites (ERBS) in breast cancer. We obtained ERBS of hormone dependent breast cancer cell lines MCF7 and T47D from GEO (GSM2670862) and ENCODE, (ENCFF002CNW). The mutation burden at ERBS was quantified in different clinical sub-type of breast cancer in PCAWG cohort. In agreement with our work in PCa, ER+ breast cancer had enrichment of mutation on ER binding regions. Surprisingly, ER negative breast cancer also had significant enrichment in ERBS. We then compared whole Pan-Cancer cohort and ranked them based on the mutation rate at ERBS (**Figure 24**). Similar to AR, breast cancer had the highest number of mutations at ERBS. We then obtained all of the available TF binding data for ER+ MCF7 cells from ENCODE and compared mutation frequency of each TF. Although, MCF7 derived ER and T47D derived ERBS were greater than random background mutation frequency threshold, they were not the top ranked TF (**Figure 25**). In addition, comparison of mutation distribution of breast versus rest of the Pan-Cancer cohort demonstrated that the enrichment of mutations is not as drastic as ARBS's mutation enrichment. We suggest that ER does not affect mutation events in breast cancer to the same degree of AR in prostate cancer. Although breast cancer and ER are not the main scope of this project, these results do implicate that TF mediated mutations are specific to the cell of origin (**Figure 26**).

Table 4: Table of GEO and ENCODE ids which were used in breast cancer investigation.

GEO/ENCODE ID	TF	GEO/ENCODE ID	TF
ENCFF884RAO	ARID3A	**ENCFF839EPV**	NFXL1
ENCFF353CQJ	BMI1	**ENCFF476URH**	NONO
ENCFF362XAG	CHD1	**ENCFF651PWG**	NRF1
ENCFF755DGT	COPS2	**ENCFF282SXB**	PAX8
ENCFF127RVQ	CREB1	**ENCFF922GXT**	PKNOX1
ENCFF785ZQF	CTBP1	**ENCFF002DBP**	POLR2A
ENCFF002DDK	CTCF	**ENCFF500ZLG**	RAD51
ENCFF572YVL	CUX1	**ENCFF065UFF**	RCOR1
ENCFF812MZB	DPF2	**ENCFF150TBK**	RFX1
ENCFF267FXT	ELK1	**ENCFF580NQL**	SIN3A
ENCFF040STP	ESRRA	**ENCFF278ODX**	SIX4
ENCFF580EYI	EZH2	**ENCFF351NCM**	SMARCA5
ENCFF170POB	FOS	**ENCFF193MKL**	SP1
ENCFF596OJV	FOXA1	**ENCFF870CER**	SREBF1
ENCFF671BJT	FOXK2	**ENCFF956LMS**	SUZ12
ENCFF313TUJ	GATA3	**ENCFF544CYG**	TARDBP
ENCFF231VAA	GATAD2B	**ENCFF002CZM**	TCF7L2
ENCFF456LHH	GTF2F1	**ENCFF129AXO**	YBX1
ENCFF002CZL	HA-E2F1	**ENCFF126PJF**	ZBTB11
ENCFF365WFM	HCFC1	**ENCFF447FPV**	ZBTB33
ENCFF445ANP	HDGF	**ENCFF397XTF**	ZBTB40
ENCFF902IMW	HES1	**ENCFF070WUF**	ZBTB7B
ENCFF688CIA	HSF1	**ENCFF235RIU**	ZHX2
ENCFF836SJY	JUN	**ENCFF353QLW**	ZKSCAN1
ENCFF688MSK	MAZ	**ENCFF051JAE**	ZNF207
ENCFF474AYD	MDB2	**ENCFF002CZN**	ZNF217
ENCFF686LJE	MLLT1	**ENCFF854CXA**	ZNF579
ENCFF412CLP	MNT	**ENCFF615FIR**	ZNF592
ENCFF712UDN	MTA1	**ENCFF955YVT**	ZNF687
ENCFF997RPK	MTA2	**GSM2670862**	ER
ENCFF527ALV	MTA3	**ENCFF002CNW**	ER
ENCFF002DBI	MYC	**ENCFF879VLB**	NFIB
ENCFF518OUJ	NBN	**ENCFF446FHX**	NFRKB

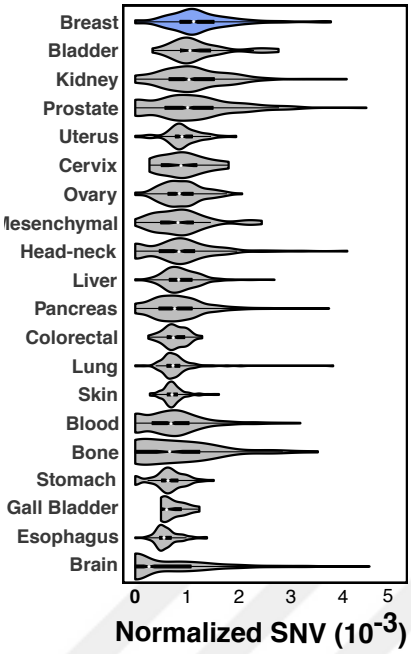

Figure 24: PCAWG cancer cohort was ranked based on the mutation numbers at ERBS. Breast cancer (blue) ranks the highest among other cancer types.

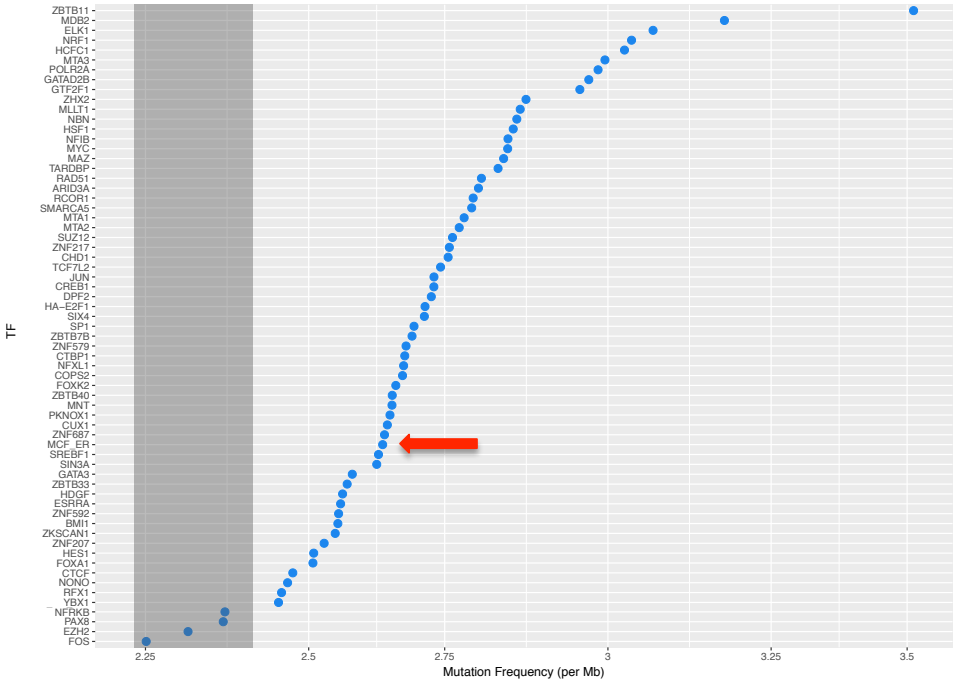

Figure 25: Mutation burden at TF binding sites in breast cancer. Red arrow indicates ER mutation frequency.

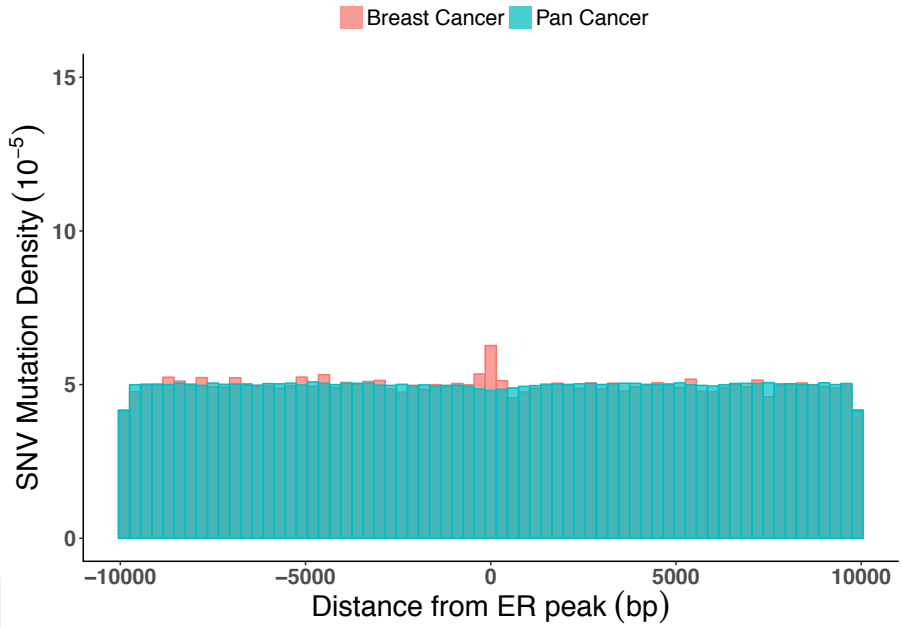

Figure 26: Minor enrichment of ER binding in breast cancer was demonstrated.

AR-mediated transcriptional is regulated by various factors such as pioneer factors, histone modifications and co-activators. Therefore, we wanted to see if any of these factors impacted the mutation frequency. To test this, we separated each ARBS regarding their mutation status. Then, we scored them if these ARBS intersected with various TF factors and histone modifications. No, clear co-binding pattern was observed that correlated with SNVs (**Figure 27**). To further study this, we implemented a random forest (RF) machine learning model to develop a model that could predict mutation occurrences based on ARBS and TF/Histone mark intersections. In this we first transformed the data to be able to run a RF algorithm on it. First, we categorically transformed overlap of ARBS to a binary score. For example, if an ARBS peak overlaps with a GATA2 peak, then that ARBS will be classified as 1 with its comparison with GATA2. As a result of transforming the data to this matrix, we classified all ARBS (n=4139) based on their overlap with 22 TF and 5 histone marks under 27 columns. However, despite extensive optimization we couldn't find any features that significantly predict ARBS mutations due to the in balanced from of data as 1010 peaks were mutated as compared to 3029 peaks are not mutated. Despite our attempts to implement under sampleing and over sampleming methods to overcome this balanced data issue, we couldn not increased our accuracy. At the end of RF predictions the error rate of our model was appeared to be %80. Therefore our current data suggests that the increase in SNVs is solely due to AR occupancy.

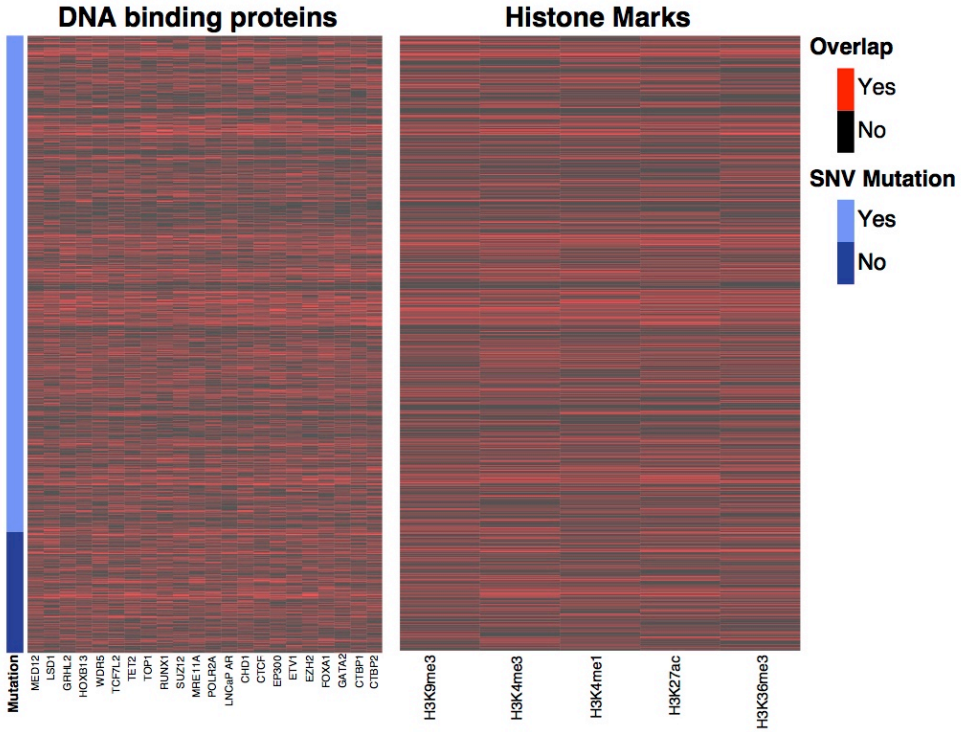

Figure 27: No clear link effect of co-localization of TF and Histone mortifications on mutation rates on ARBS.

3.3.3) Characterization of ARBS mutation types in prostate and other pan-cancer cohort

Somatic mutations can give tremendous insight into the cause of a cancer. Numerous studies have identified mutational signatures that are caused by specific etiological agents. In 2015, the COSMIC database released 30 mutation signatures that were associated with specific DNA damage. Therefore, to better understand the mutations at ARBS we characterized the mutation signatures on these regions. Strikingly, we found that mutation signature changed dramatically at ARBS compared to the remainder of the genome. Specifically, there was an increase of TpG > ApG and CpG> GpG mutations at ARBS, as compared to the remainder of genome in prostate cancer. To investigate if this enrichment was due to an ARE motif, we analyzed the mutation signature at those sites that have an ARE motif but no AR binding. We found that the signature at the ARE motif was almost identical to the whole genome signature suggesting that the mutational signature is not due to nucleotide composition but rather AR protein binding (**Figure 28A**). In addition, presence of an ARE motif on ARBS did not seem to alter the mutation signature as well (**Figure 28B**). To further test if base composition affects mutation signature, we generated 1000 randomly selected binding regions sets that were identical in size to ARBS. These random regions were separated based on similarity in nucleotide composition to ARBS. Of these, we found no difference between the whole genome mutation signatures and those sites with similar (n=159) and dissimilar (n=851)

nucleotide composition suggesting that ARBS base composition does not influence the signature (**Figure 28C**). In addition, showing the specificity none of the random regions have enriched for TpG -> ApG mutations *(***Figure 28D**).

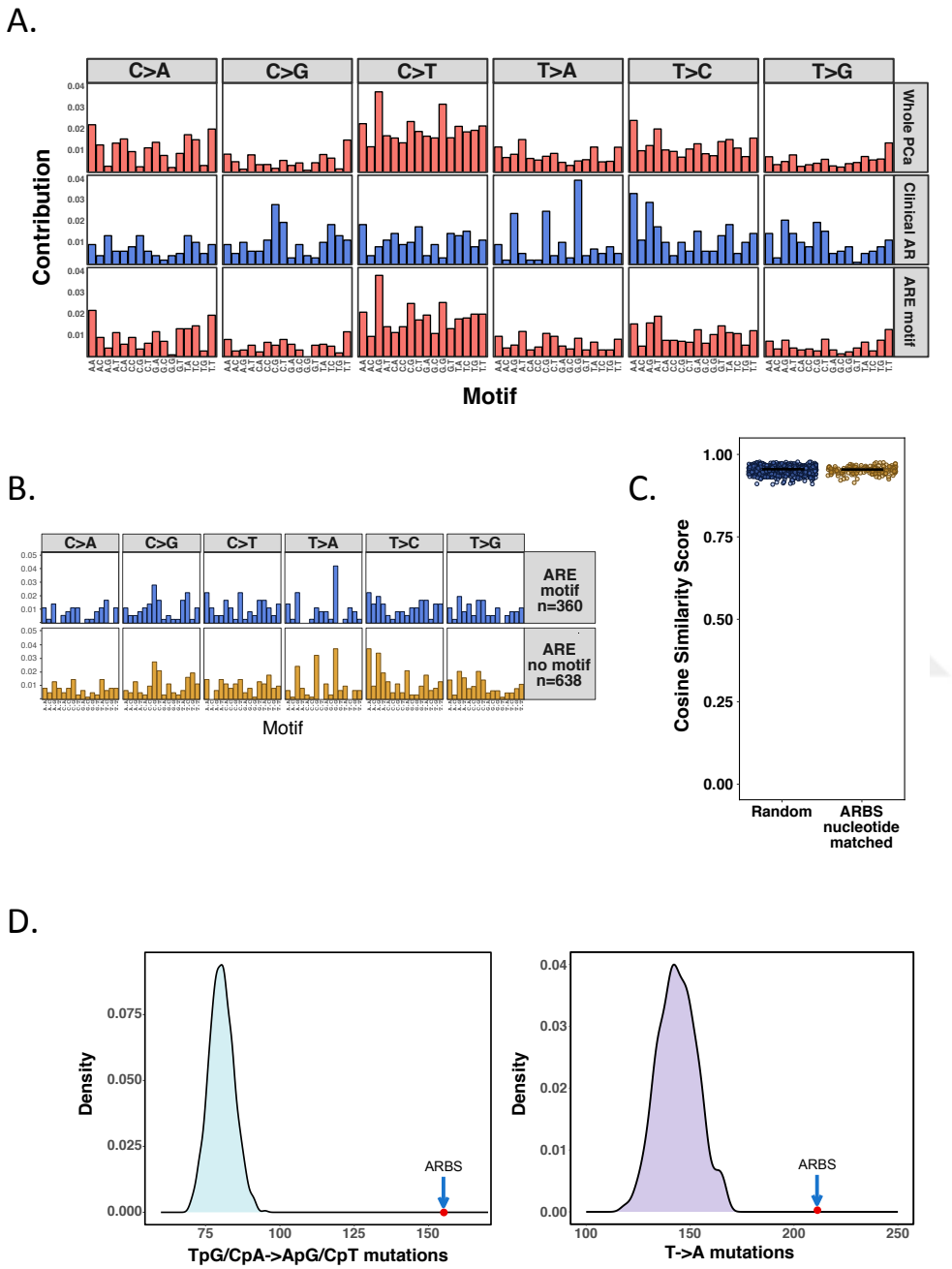

Figure 28: Characterization of the mutation signature at ARBS. A) Mutation signature of whole prostate genome, clinical ARBS and ARE motif driven regions were visualized. B) Randomly generated 1000 regions were separated based on their base composition similarity to AR regions. Then, their mutation signatures were compared to the whole genome signature. C) ARBS mutation signature was not altered significantly on peak that have ARE motif and no ARE motif. D) None of the 1000 randomly generated regions have higher TpG >ApG frequency than original ARBS regions.

AR signaling is only important in PCa. Therefore, by comparing the mutational signature in other cancers at the same chromosomal locations we can be confident that the ARBS mutation signature is caused by this specific nuclear receptor and not the regional chromosomal instability. Therefore, after removing all the cancer types that have less than 480 mutations on ARBS region (previously defined as the threshold to identify mutation signature >95% accuracy[111]), we compared ARBS mutation signature of each cancer with the remainder mutation signature. Among all of the cosine similarity values, only in prostate cancer ARBS had a different mutational signature than the remainder of the genome (**Figure 29**). Overall our results show that neither base composition nor the chromosomal location of ARBS have an effect on mutation signature.

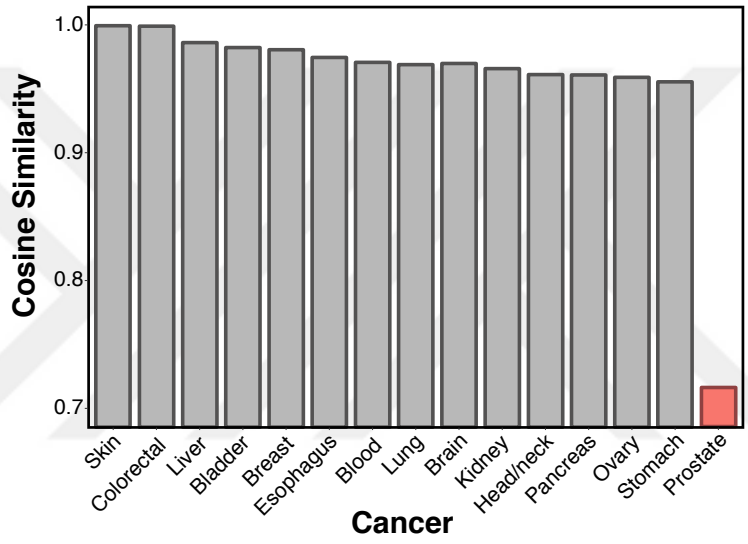

Figure 29: Comparison of mutational signature at ARBS and the remainder of the genome. Prostate cancer has the lowest similarity among all cancer types between whole genome signature and ARBS signature.

To determine if only AR caused this mutational signature, we then characterized the mutation signatures of all other TFs in PCa. Importantly, previously work had shown that you could characterize a mutational signature if there was mutation rate greater than the theoretical limit of detection (**Figure 30A**). When we compared the different mutational signatures in all TFs, we found three distinct signature subgroups. In group 1, KDM1A, GATA2 and HOXB13 were found to have an extremely similar mutation signature to AR (**Figure 30B**). As all of these proteins have been shown to be involved in AR-mediated transcription we investigated the affect of co-localization with ARBS on mutation frequency and signature. Specifically, we removed those binding sites that overlapped with ARBS. Similar to the earlier work with AR, we also characterized those sites that have a binding motif, but no protein bound. We found that mutational enrichment and signature of AR did not affect any of the TFs in this group *(***Figure 31A***)*. This suggests that the increase of mutations in GATA2 and HOXB13 sites are not due to AR binding. Also, neither pioneer factor binding motif presence affected

the signature or mutation frequency (**Figure 31B**). Similar to AR, members of group 1 only had altered mutation signatures in prostate cancer. These data suggest that AR and co-activators mutation signature are caused by the binding of the protein and it is only specific to prostate cancer (**Figure 31C**).

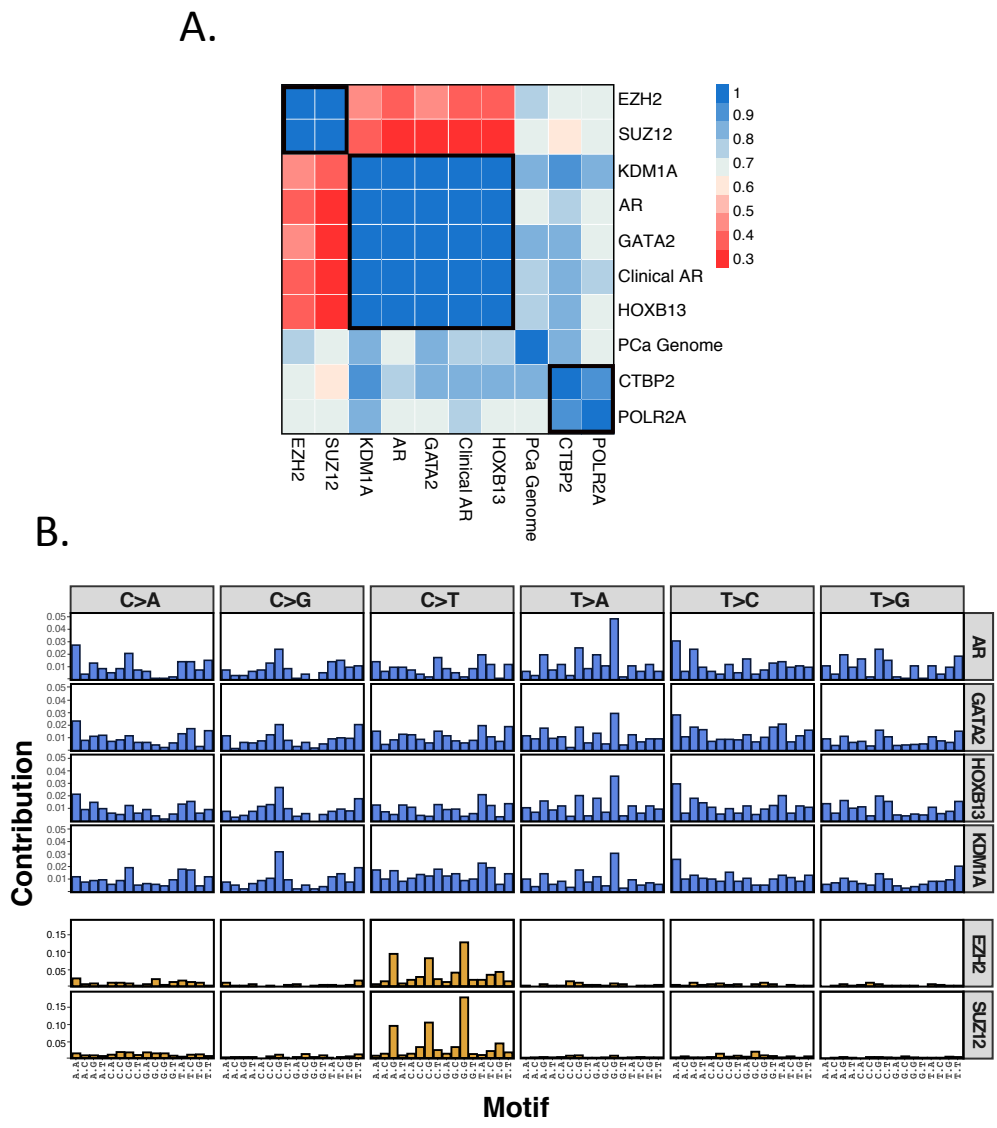

Figure 30: Characterization of mutation signature at TF binding sites. A) TF with higher frequency than background prostate cancer mutation distribution were further investigated for their mutation signature. Three groups were formed. B) Mutation signature of group1 and group 2 were shown.

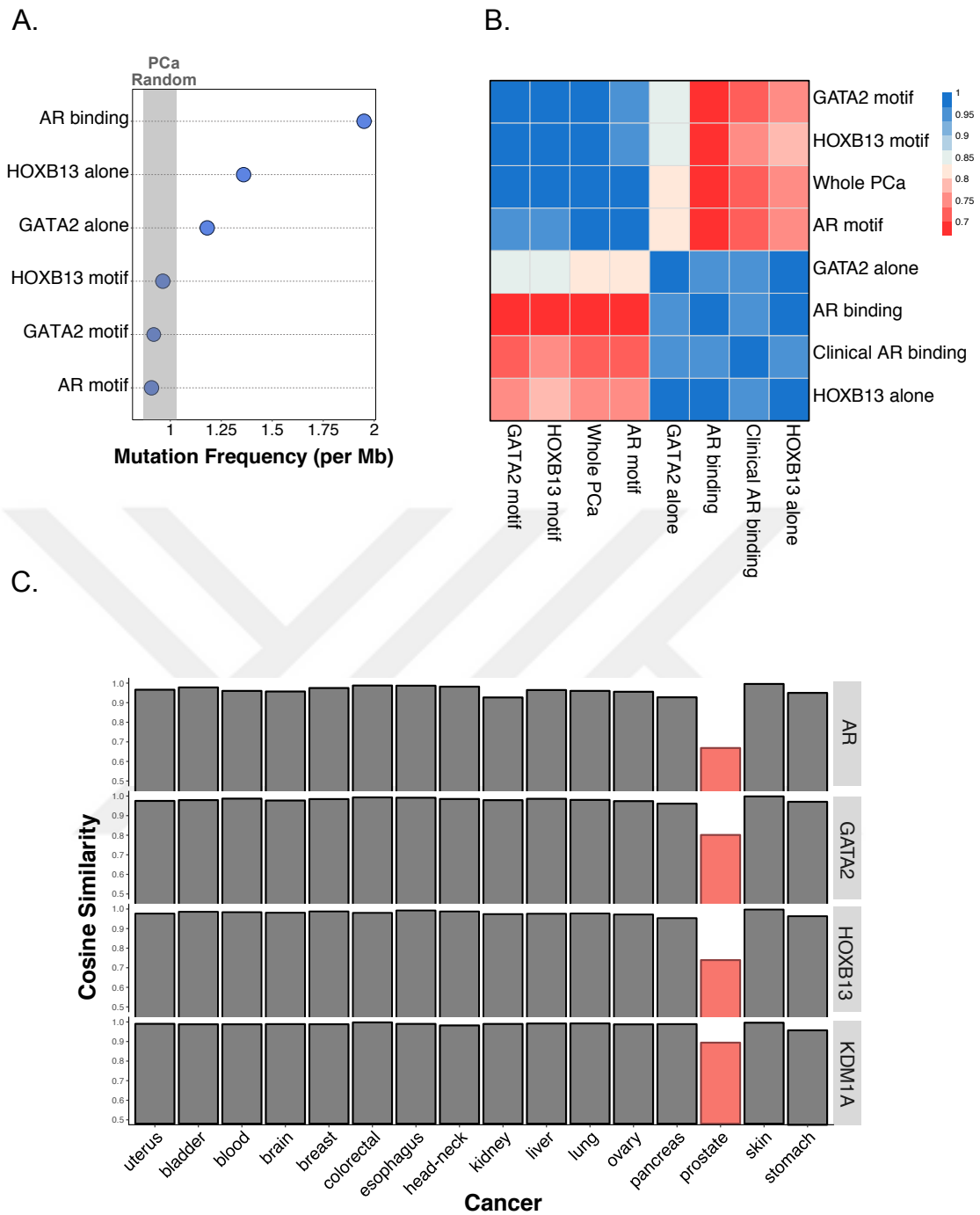

Figure 31: Characterization of mutation signature at HOXB13 and GATA2 binding sites. A) HOXB13 and GATA2 TF binding regions were excluded from ARBS. In addition, for GATA2 and HOXB13 mutation frequencies of motif driven sites were compared with protein binding sites. B) For the same sties, mutation signature similarity between whole genome and those sites were compared. C) Similar to previous comparisons, whole genome signature mutation signature of each cancer was compared with its group1 TF binding region signatures. Only in prostate cancer there is difference between whole genome and TF binding regions.

In the second group, polycomb repressive complex 2 (PRC2) members EZH2 and SUZ12 were found to have remarkably similar mutational signatures (**Figure 30B**). Specifically, there was a very high number of C>T transitions. Because of the similarity in signature and function of these two proteins, we tested if this was due to overlap between SUZ12 and

EZH2. While these proteins are both members of the PRC2, only 10% of EZH2 and 17% of SUZ12 binding sites overlapped, suggesting that the mutations signature is not simply due to overlap of these proteins (**Figure 32A**). We then examined if these regions have some chromosomal characteristic that cause this signature. We found that this signature is not prostate specific and was observed in other cancers as well (**Figure 32B**).

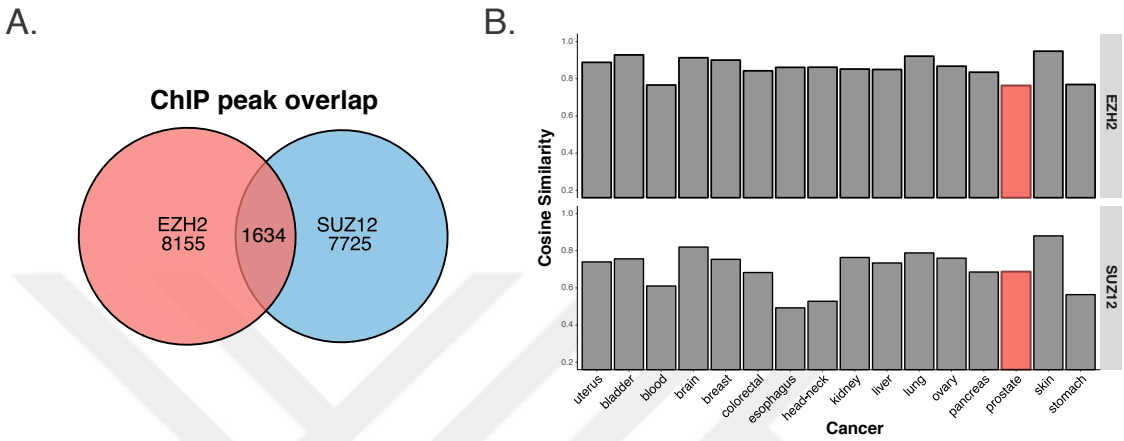

Figure 32: Characterization of SUZ12/EZH2 mutation signature. A) Intersecting regions of EZH2 and SUZ12 were calculated. ~15% of overlap was found between EZHE2 and SUZ12 binding regions. B) Unlike AR, EZH2 and SUZ12 mutation signature was not spesific to prostate cancer and seen in other cancers as well.

The third group consisted of POLR2A and CTBP1. Their signature was more similar to the whole genome signature of PCa than the other groups.

While we demonstrated that the mutation signatures are independent of their base composition with AR we wanted to confirm this with other TFs. We therefore, calculated the ATGC ratios of each TF binding regions to see if mutational signatures clustered. Our findings did not appear to be solely due to nucleotide composition. For example, POLR2A and SUZ12 had similar GC content but very different mutation signatures. Overall, base composition is not a significant factor that affects signature (**Figure 33**).

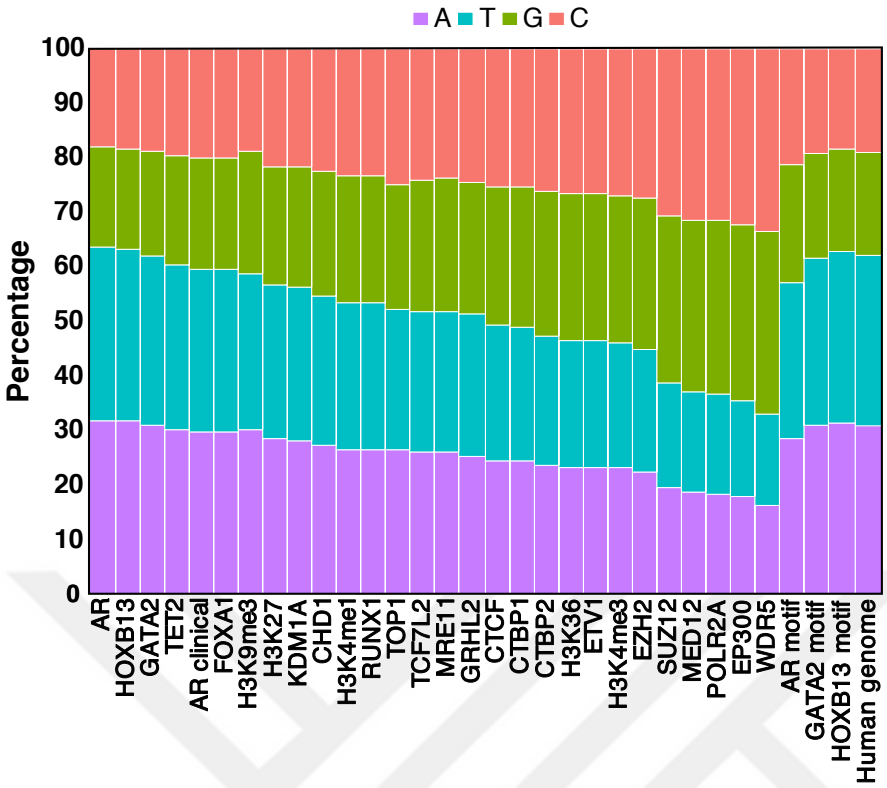

Figure 33: Nucleotide distribution of all TF and Histone marks in PCa.

To better understand the etiological cause of the increased mutation at TF binding sites we compared our results to previously published COSMIC signatures[115]. Interestingly, signature 1 was very similar to what was observed with EZH2-SUZ12 (Figure 34A). This signature has previously found in multiple cancer types and has been shown to be caused by spontaneous deamination of 5-methylcystosine (**Figure 34B**). In support of this, almost all of the C->T mutations at EZH2-SUZ12 binding sites occurred at CpG sites (**Figure 34C**). To better understand this, we investigated the methylation rates of the CpG regions with genome-wide bisulfite sequencing from LNCaP cells. We found that EZH2-SUZ12 had the one of the highest level of methylation of all TF binding sites (**Figure 34D**). Our findings demonstrate that a mutational signature can be successfully determined at a localized site.

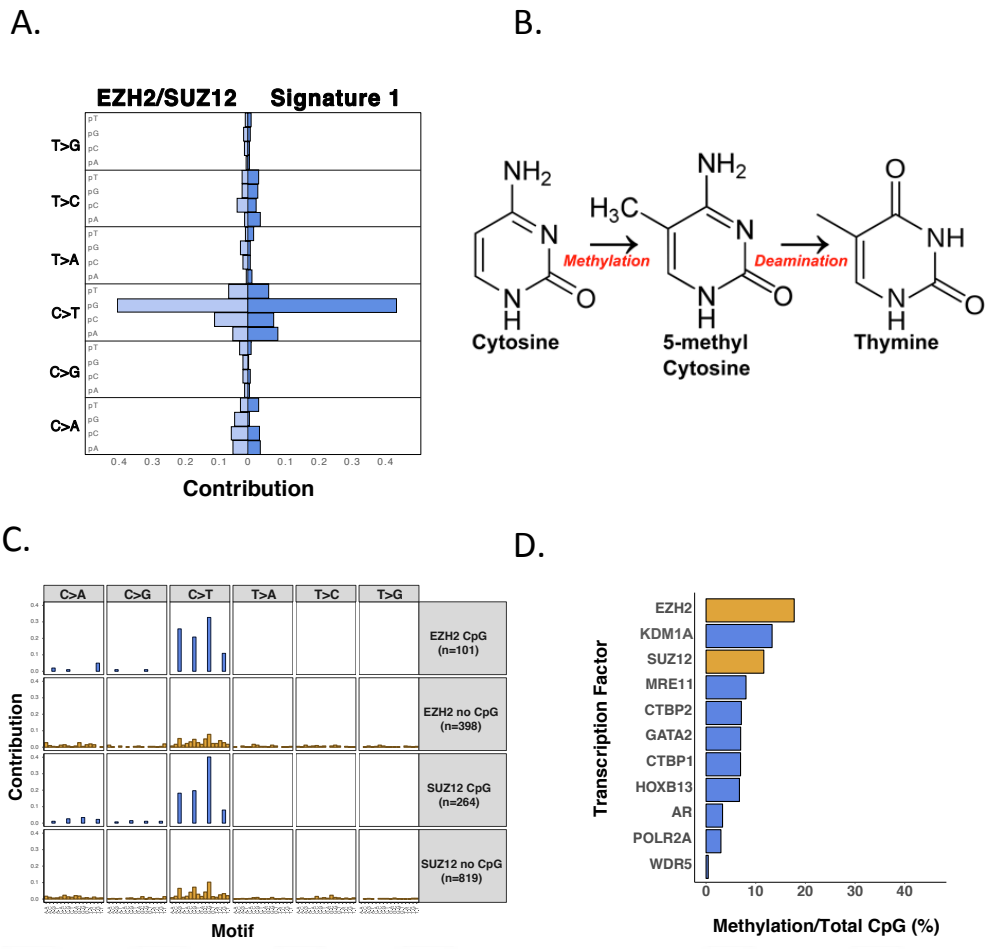

Figure 34: Characterization of EZH2/SUZ12 mutational signature. A) EZH2/SUZ12 signature similarity to Signature 1 was visualized. B) Schematic representation of spontaneous deamination was presented. C) C>T transitions of EZH2 and SUZ12 were mainly located on CpG regions. D) Among all TFs SUZ12 and EZH2 ranked top 3 on CpG methylation percentages.

Next, we characterized the mutational signature at AR/KDM1A/HOXB13/GATA2 binding sites. Interestingly, we couldn't find any COSMIC signatures similar to what we observed. The closest signature was associated with the carcinogen aristolochic acid (Signature 22)(**Figure 35A**)[115]. We decided that this was unlikely, as while the ARBS signature has an increase in T>A mutations (**Figure 35B**), it is very different than that found with the Signature 22. However, TpG -> ApG purine transversions have been previously associated with unsuccessful repair of abasic sites. This phenomenon is best shown with the carcinogen dimethylbenzantracene (DMBA). DMBA induces depurination of dexoyadenosine causing an abasic site[116–119]. When this occurs, the massive increase of these abasic sites overloads the base excision repair machinery and causes TpG>ApG transversions due to the "A-rule" whereby unpaired abasic sites are primarily fixed with adenine substitution[120]. However, it is extremely unlikely that aromatic hydrocarbon is present in the prostate gland. Yet the similarities in the mutation signature could be due to a shared failure of DNA repair

machinery. DMBA overwhelms base excision repair machinery due to the massive number, but if AR prevents access to spontaneously depurinated sites it would cause a similar mutation. Spontaneous depurination is extremely common and occurs >10,000 times/day/cell[121]. Hence, unrepaired regions will give rise to purine transversions similar to what was seen in DMBA. In support of this model, the ARBS signature observed is similar to DMBA treated animals (**Figure 35C**).

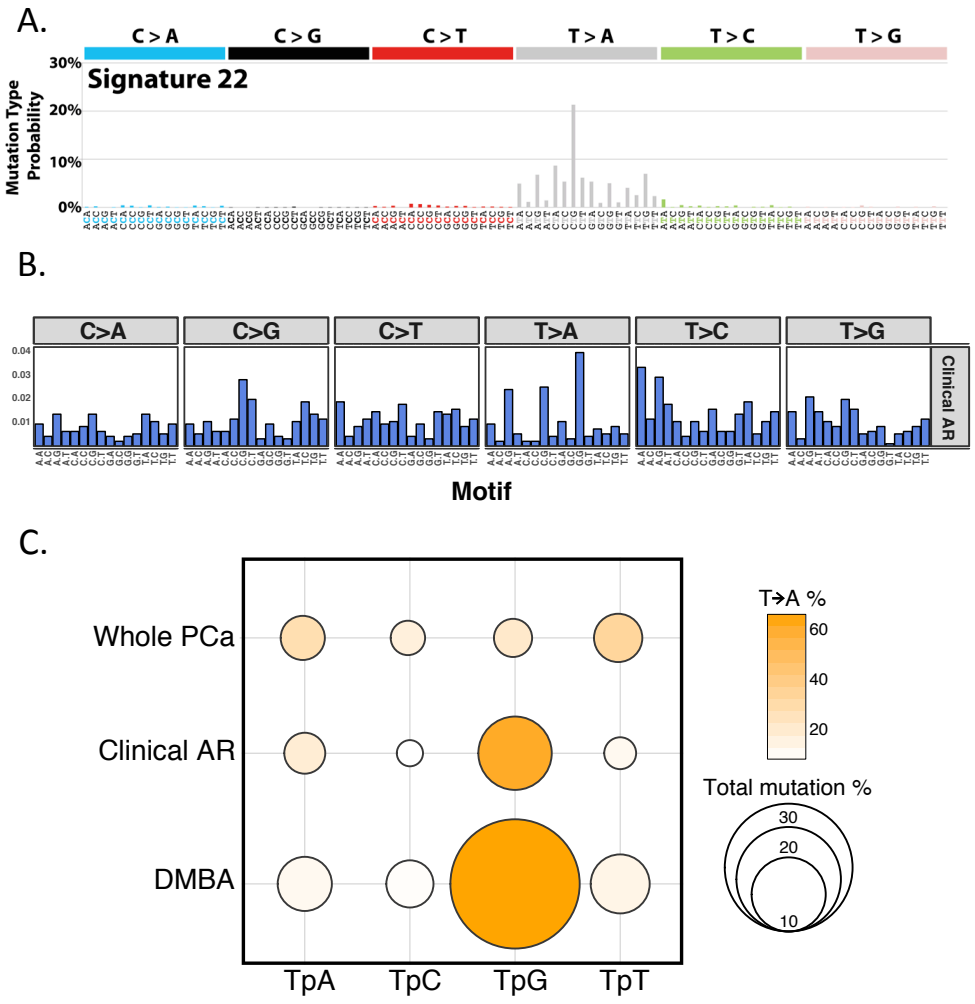

Figure 35: A) COSMIC Signature 1. B) ARBS mutation signature on prostate cancer. C) T->A transversions percentages was compared in whole PCa, ARBS and DMBA signatures with respect to their flanking regions. Regarding flanking sequence and T>A percentages, clinical AR signature and DMBA are similar to each other.

Summary

AR is critical to prostate cancer progression, however there are many different transcription factors. To be unbiased in our investigation, we examined the mutation rate at all available TF binding sites. Among all of these TF, the AR had by far the highest rate of mutations. What is more, some of the known AR regulators also had mutation frequency enrichment with respect to genome background. When we compared the mutation rate at ARBS all other cancer types only prostate cancer showed enrichment in mutations. Moreover, this enrichment

is only present in close proximity to the ARBS and depending on AR protein binding and not the motif. Considering AR's protein is crucial for mutation enrichment, we then investigated the affect of co-localization of other TF with AR on ARBS. We found no clear evidence of co-binding causes mutation enrichment on ARBS. Next, to be able propose a model for TF mediated action; we examined breast cancer that is highly affected by TF mediated action. Similar to AR in prostate cancer, ER is equally important in breast cancer. Therefore, we investigated mutations in ERBS regions. Similar to PCa, analysis of ERBS mutations in breast cancer also demonstrated that TF mediated mutation enrichment is present in breast cancer which was shown by the comparison other cancers. However, ER's contribution to mutation enrichment is not aberrant as AR's as the comparison of multiple breast cancer cell line TF showed higher rate of mutations than ERBS. Yet, there is still significant enrichment in mutation on ERBS, which supports that TF mediated mutations require more attention. So far, we have shown the increment in mutation numbers, but we have not characterized the type of the mutations. Mutation signatures give great insight about the cause of damage that lead to genomic alterations. Here in prostate cancer, we see altered mutation signature at TF binding regions for various TFs which from three distinct group of mutation signatures. Specifically, AR's mutation signature has been demonstrated that is independent from ARBS base composition and canonical ARE presence. In fact, we showed that unique AR signature is only present on the regions whereby AR protein is present. Affect of AR co-activators led us examine their affect on AR's unique signature as they have a very similar mutation signature like AR. However, removing intersected regions between AR and co-activators did not alter the phenotype, which suggest that each AR and AR co-activators have a unique signature individually. More interesti[118]ngly, AR's unique signature was only present in prostate cancer, which indicates the presence of TF mediated damage. Another interesting mutation signature was group EZH2/SUZ12 signature. Almost identical signature of this group led us investigate commonality of EZH2 and SUZ12 regions. Yet overlapping regions are low as 15% for each TF. We then used this second as a control of our methodology and showed that it is very similar to previously published spontaneous deamination caused mutation Signature 1. Having most of the C>T mutation on CpG regions and having two of the highest methylated regions, we can clearly say that our method is reliable on analyzing localized mutation signatures. Lastly, we examined the cause of AR's mutation signature and found observed T>A mutations have been erroneous abasic sites. This has been previously demonstrated with a chemical called DMBA that cause spontaneous depurination of adenosine[118]. As base excision repair mechanism is responsible for fixing these mutations, we propose that previously published TF mediated NER blockage can also be present in our case whereby, AR binding would block those abasic site to be repaired.

Chapter 4: Discussion

Cancer is caused by the accumulation of mutations. Therefore, large-scale sequencing projects such as TCGA and ICGC provide an extremely valuable resource to identify those mutations that drive the cancer. However, mutations do not occur in evenly across the genome and are affected by multiple factors. Recent work in melanoma demonstrated that TF binding can cause local mutations due to the impairment of DNA repair machinery. Yet, this has been only studied in those cancers that have an extremely high mutation rate as a small number of whole-genome datasets can give sufficient statistical power. It is unclear if the TF mediated damage is specific to only these cancers with high mutation rates or if it is a broad phenomenon. Therefore, the goal of this work was to investigate the mutation rates at TF binding sites in prostate cancer. By focusing on AR, a TF that is only important in prostate cancer, we can determine how much the nucleotide composition or chromosomal location influences the observed mutations.

Comparison of the mutation rate at TF binding sites demonstrated that both ARBS from cell lines and clinical samples have the highest mutation rates of all TFs. Although AR-mediated transcription is dependent on protein interactions, the drastic enrichment of mutation at ARBS was striking. Therefore, we investigated various aspects of these mutations in ARBS. Importantly, we found that this enrichment was independent of ARBS chromosomal locations or base composition and was dependent on AR protein. Moreover, other cancer types, which do not require AR, showed no increase in mutations at ARBS, thereby clearly showing that TF binding mutations are cell-type specific. In addition, among AR's co-regulators, HOXB13 and GATA2 had significant enrichment over background mutation distribution. Contrary to previous studies in melanoma and colorectal cancers, we did not observe increased mutations at CTCF binding sites in prostate cancer. Interestingly, another studied investigated CTCF mutations of various cancer and found that given CTCF binding regions are not always highly mutated in all cancers[122]. Cancers such as in gastric and colorectal are highly mutated where as breast, liver, lung, pancreas and lymphoma have low mutation enrichment in CTCF binding region. This gives further evidence that TF binding site mutations are cell-type specific[122].

To find the potential causes of these localized mutations, we characterized the mutation signature at ARBS and other TF binding sites. We demonstrated that the ARBS mutation signature were different than the rest of the cancer genome. Specifically, we found a high frequency of TpG->ApG and CpG->CpG purine transversions at ARBS and the binding sites of AR regulators HOXB13, GATA2 and KDM1A. These mutation types were independent

from various factors including; ARE presence, co-localization of co-activators with AR and base composition of binding regions. This was critical because AR mediated transcription is heavily dependent on co-localization of pioneer factors. We also examined all possible intersections between AR, GATA2 and HOXB13 but due to the very few number of mutation numbers it was not possible to accurately quantify the mutation signature. (Data not shown). The novel mutation signature of AR-related proteins, are only observed at protein binding sites as the mutation rate and signature was not altered at those sites that had a known binding site motif but no protein. Moreover, none of the other cancers from PCAWG had an altered mutation signature at either AR or AR coregulator binding sites. This underlines the importance of cell type specific enrichment of TF mediated mutations. In the second group, the mutation signature at SUZ12 and EZH2 binding sites were very similar. This led us to examine if these regions have any overlap. As both proteins are members of the PCR2 complex, surprisingly we found little overlap between SUZ12 and EZH2 binding regions. The C>T dominated signature at EZH2/SUZ12 sites showed striking similarity to COSMIC signature 1. This mutational signature is associated with spontaneous deamination of 5-methylcytosine in multiple cancer types and is therefore correlated with high methylation at CpG regions. Supporting this theory, EZH2/SUZ12 binding sites had some of the highest rates of DNA methylation.

There could be two possible mechanisms for observed mutation events at ARBS. During AR mediated transcription, AR could directly cause DNA damage as previously proposed[53]. However, as the mutation signatures of AR co-regulators were nearly identical to ARBS, this suggests the AR itself cannot be the cause as those sites where is not co-bound still have the same mutational signature. It is therefore more likely that binding of the TF blocks DNA replication machinery, which therefore gives rise, to the specific type of mutation. Supporting this model, previous work demonstrated that TF binding blocks nucleotide NER machinery in melanoma patients[94]. From our study, we propose an expanded mechanism whereby TF binding also prevents the repair of DNA damage by BER in addition to NER. This is supported by previous work that demonstrated the TpG>ApG signature is associated incorrect repair of abasic sites with BER. Mice with dysfunctional BER have an increase in endogenous T->A and C->G purine transversions similar to what was observed at ARBS[123]. Surprisingly, the deamination-related signature of EZH2/SUZ12 has also been previously associated with the malfunction of BER[124]. Therefore, we prose that blockage of DNA repair machinery by TF could be a more general event that impairs BER in addition to NER.

Breast and prostate cancers are similar in terms of the critical role hormone induced transcription plays in the growth of the cancer. Like AR in prostate cancer, ER can initiate transcription upon steroid hormone induction. Therefore, we investigated the mutation rate at ERBS in ICGC's breast cancer cohort and found that there is significant enrichment at ERBS. In comparison of SNV numbers, breast cancer had the highest rate of mutations at ERBS of all cancers. We then examined clinical subtypes in breast cancer. Given that the prostate cancer tumours in this work were all primary cancer, they are almost always hormone dependent and such a classification of hormone dependent and independent is not necessary. However, in breast cancer case there are hormone dependent and independent subtypes included in ICGC cohort. Therefore, investigated the different cancer types on ERBS mutations. Interestingly, in all subtypes of breast cancer ER mutation numbers were significantly enriched. This suggests that even in triple negative breast cancer the ER may be important in the development of the cancer. However, these results require deeper investigation with more patients. Although they were not as striking as AR and prostate, this still supports our theory about TF mediated mutations.

Numerous proteins are needed to initiate transcription. In addition to transcription factors, histone modifications on chromatin can also influences mutation rates[125]. We therefore investigated if any histone marks correlated with ARBS mutations. Previously random forest methodologies have been implemented in various genomic projects to predict output come of multi-variable events[126]. However, when we tested this and other machine learning methodologies, we could fine no correlation between any specific TF or histone marks with ARBS. However, scarcity, sample size and balance of data are very critical for accurate prediction. All of these parameters are linked. When we classified our peaks we described them in a binary manner. This therefore, limits our ability to understand the number of mutations at a binding site, as we cannot distinguish those sites that have multiple mutations. While we lose this information, it was necessary as there are few peaks that had multiple mutations which would affect the training. Further, balanced data is crucial in a random forest predictive model. Therefore, the number of mutated peaks and non-mutated peaks should have been equal. However, in our case we had 1-3 ratio of mutated peaks to non-mutated peaks. This caused a skewed distribution in the random sampling stage, which decreased the predictor accuracy. Overall, more patients are needed to improve the resolution of such a model.

Herein, we demonstrated that TF binding sites have an increase in mutations that are cancer specific as not all of TFs have equal amount of mutation enrichment and altered signature.

Considering the critical importance of AR in prostate cancer growth, some of these ARBS mutations may have a crucial role in cancer progression. However, experimental validations of ARBS regions are required to demonstrate the affect of these mutation transcription and other cellular processes.

4.1) Future work

While novel, further studies are needed to understand the impact of ARBS mutations. One of the largest outstanding questions is if these non-coding mutations can act as drivers. When we investigated the ARBS mutation recurrences we found only one binding site that has more mutations than expected (*p=0.015)*. However, due to our low SNV number overall, potential driver mutations may still occur but may be "long-tail" mutations. Thus, more patient samples are required to investigate these relatively rare SNVs. However, to truly understand the impact of these mutations we will need to characterize how they impact gene expression. ICGC currently has released RNAseq data from only 21 patients. However, to conduct expression quantative loci experiments analysis, we need far more RNAseq samples. Complicating the analysis somatic mutations typically have lower VAF thereby limiting the signal strength in a heterogeneous tumour. Further, enhancer-promoter interactions are difficult to define due to chromatin looping. While several HiC datasets are available for LNCaP, the low resolution of HiC (i.e. 20Kb window) makes it too noisy to analyze specific TF binding regions. Therefore, we will need ChIA-pet or HiCHip data to gives genome wide interactions of only specific TF binding regions and define the enhancer-promoter interactions. Future studies should also incorporate genome wide association studies to better understand the somatic mutations identified in this work.

Chapter 5: References

1. Prostate Cancer: Statistics. at <https://www.cancer.net/cancer-types/prostate-cancer/statistics>
2. Agoulnik, I. U. & Weigel, N. L. Androgen Receptor Action in Hormone-Dependent and

Recurrent Prostate Cancer. **372,** 362–372 (2006).

3. Debes, J. D. & Tindall, D. J. The role of androgens and the androgen receptor in prostate cancer. *Cancer Lett.* **187,** 1–7 (2002).
4. Bruchovsky, N. *et al.* Intermittent androgen suppression for prostate cancer: Canadian Prospective Trial and related observations. *Mol. Urol.* **4,** 191–9;discussion 201 (2000).
5. Grino, P. B., Griffin, J. E. & Wilson, J. D. Testosterone at high concentrations interacts with the human androgen receptor similarly to dihydrotestosterone. *Endocrinology* **126,** 1165–1172 (1990).
6. Davey, R. A. & Grossmann, M. Androgen Receptor Structure, Function and Biology: From Bench to Bedside. *The Clinical Biochemist Reviews* **37,** 3–15 (2016).
7. He, B., Kemppainen, J. A., Voegel, J. J., Gronemeyer, H. & Wilson, E. M. Activation function 2 in the human androgen receptor ligand binding domain mediates interdomain communication with the NH(2)-terminal domain. *J. Biol. Chem.* **274,** 37219–37225 (1999).
8. Tyagi, R. K. *et al.* Dynamics of intracellular movement and nucleocytoplasmic recycling of the ligand-activated androgen receptor in living cells. *Mol. Endocrinol.* **14,** 1162–1174 (2000).
9. Saporita, A. J. *et al.* Identification and characterization of a ligand-regulated nuclear export signal in androgen receptor. *J. Biol. Chem.* **278,** 41998–42005 (2003).
10. Quigley, C. A. *et al.* Androgen receptor defects: historical, clinical, and molecular perspectives. *Endocr. Rev.* **16,** 271–321 (1995).
11. Harris, W. P., Mostaghel, E. A., Nelson, P. S. & Montgomery, B. Androgen deprivation therapy: progress in understanding mechanisms of resistance and optimizing androgen depletion. *Nature clinical practice. Urology* **6,** 76–85 (2009).
12. Horie-Inoue, K., Bono, H., Okazaki, Y. & Inoue, S. Identification and functional analysis of consensus androgen response elements in human prostate cancer cells. *Biochem. Biophys. Res. Commun.* **325,** 1312–7 (2004).
13. Wang, Q. *et al.* A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth. *Mol. Cell* **27,** 380–392 (2007).
14. Urbanucci, A. *et al.* Overexpression of androgen receptor enhances the binding of the receptor to the chromatin in prostate cancer. *Oncogene* **31,** 2153 (2011).
15. CHENG, Y. U. E. *et al.* Genome-wide analysis of androgen receptor binding sites in prostate cancer cells. *Experimental and Therapeutic Medicine* **9,** 2319–2324 (2015).
16. Rodriguez-Bravo, V. *et al.* The role of GATA2 in lethal prostate cancer aggressiveness. *Nature reviews. Urology* **14,** 38–48 (2017).
17. Zhao, J. C. *et al.* FOXA1 acts upstream of GATA2 and AR in hormonal regulation of gene expression. *Oncogene* **35,** 4335–4344 (2016).
18. Norris, J. D. *et al.* The homeodomain protein HOXB13 regulates the cellular response to androgens. *Mol. Cell* **36,** 405–416 (2009).
19. Gao, N. *et al.* The role of hepatocyte nuclear factor-3 alpha (Forkhead Box A1) and androgen receptor in transcriptional regulation of prostatic genes. *Mol. Endocrinol.* **17,** 1484–1507 (2003).

20. Foley, C. & Mitsiades, N. Moving Beyond the Androgen Receptor (AR): Targeting AR-Interacting Proteins to Treat Prostate Cancer. *Hormones & cancer* **7,** 84–103 (2016).
21. Jin, H.-J., Zhao, J. C., Wu, L., Kim, J. & Yu, J. Cooperativity and equilibrium with FOXA1 define the androgen receptor transcriptional program. *Nat. Commun.* **5,** 3972 (2014).
22. Wu, D. *et al.* Three-tiered role of the pioneer factor GATA2 in promoting androgen-dependent gene expression in prostate cancer. *Nucleic Acids Research* **42,** 3607–3622 (2014).
23. Jung, C., Kim, R.-S., Zhang, H.-J., Lee, S.-J. & Jeng, M.-H. HOXB13 induces growth suppression of prostate cancer cells as a repressor of hormone-activated androgen receptor signaling. *Cancer Res.* **64,** 9185–9192 (2004).
24. Glass, C. K. & Rosenfeld, M. G. The coregulator exchange in transcriptional functions of nuclear receptors. *Genes Dev.* **14,** 121–141 (2000).
25. Aalfs, J. D. & Kingston, R. E. What does 'chromatin remodeling' mean? *Trends Biochem. Sci.* **25,** 548–555 (2000).
26. Shang, Y., Myers, M. & Brown, M. Formation of the androgen receptor transcription complex. *Mol. Cell* **9,** 601–610 (2002).
27. Brady, M. E. *et al.* Tip60 is a nuclear hormone receptor coactivator. *J. Biol. Chem.* **274,** 17599–17604 (1999).
28. Fu, M. *et al.* Hormonal control of androgen receptor function through SIRT1. *Mol. Cell. Biol.* **26,** 8122–8135 (2006).
29. Gaughan, L., Logan, I. R., Cook, S., Neal, D. E. & Robson, C. N. Tip60 and histone deacetylase 1 regulate androgen receptor activity through changes to the acetylation status of the receptor. *J. Biol. Chem.* **277,** 25904–25913 (2002).
30. Di Lorenzo, A. & Bedford, M. T. Histone arginine methylation. *FEBS Lett.* **585,** 2024–2031 (2011).
31. Daujat, S. *et al.* Crosstalk between CARM1 methylation and CBP acetylation on histone H3. *Curr. Biol.* **12,** 2090–2097 (2002).
32. Mellor, J. Dynamic nucleosomes and gene transcription. *Trends Genet.* **22,** 320–329 (2006).
33. Majumder, S., Liu, Y., Ford, O. H. 3rd, Mohler, J. L. & Whang, Y. E. Involvement of arginine methyltransferase CARM1 in androgen receptor function and prostate cancer cell viability. *Prostate* **66,** 1292–1301 (2006).
34. Wang, H. *et al.* Methylation of histone H4 at arginine 3 facilitating transcriptional activation by nuclear hormone receptor. *Science* **293,** 853–857 (2001).
35. Zhao, J. C. *et al.* Cooperation between Polycomb and androgen receptor during oncogenic transformation. *Genome Res.* **22,** 322–331 (2012).
36. Xu, K. *et al.* EZH2 Oncogenic Activity in Castration Resistant Prostate Cancer Cells is Polycomb-Independent. *Science (New York, N.Y.)* **338,** 1465–1469 (2012).
37. de la Cruz, C. C. *et al.* The Polycomb Group Protein SUZ12 regulates histone H3 lysine 9 methylation and HP1α distribution. *Chromosom. Res.* **15,** 299–314 (2007).
38. Metzger, E. *et al.* LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription. *Nature* **437,** 436–439 (2005).

39. Paltoglou, S. *et al.* Novel Androgen Receptor Coregulator GRHL2 Exerts Both Oncogenic and Antimetastatic Functions in Prostate Cancer. *Cancer Res.* **77,** 3417 LP-3430 (2017).
40. Lodish H, Berk A, Zipursky SL, et al. Proto-Oncogenes and Tumor-Suppressor Genes. *Molecular Cell Biology. 4th edition* (2000). at <https://www.ncbi.nlm.nih.gov/books/NBK21662/>
41. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314,** 268–274 (2006).
42. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).
43. Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321,** 1807–1812 (2008).
44. McFarland, C. D. *et al.* The Damaging Effect of Passenger Mutations on Cancer Progression. *Cancer Res.* **77,** 4763 LP-4772 (2017).
45. Merid, S. K., Goranskaya, D. & Alexeyenko, A. Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. *BMC Bioinformatics* **15,** (2014).
46. Rheinbay, E. *et al.* Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. *bioRxiv* (2017). at <http://biorxiv.org/content/early/2017/12/23/237313.abstract>
47. Sunkel, B. & Wang, Q. Mapping mutations in prostate cancer exomes. *Asian Journal of Andrology* **14,** 801–802 (2012).
48. Blattner, M. *et al.* SPOP Mutation Drives Prostate Tumorigenesis In Vivo through Coordinate Regulation of PI3K/mTOR and AR Signaling. *Cancer Cell* **31,** 436–451 (2017).
49. Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161,** 1215–1228 (2015).
50. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163,** 1011–1025 (2015).
51. Yang, L. *et al.* Molecular classification of prostate adenocarcinoma by the integrated somatic mutation profiles and molecular network. *Sci. Rep.* **7,** 738 (2017).
52. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310,** 644–648 (2005).
53. Haffner, M. C. *et al.* Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. *Nat. Genet.* **42,** 668–675 (2010).
54. Demichelis, F. *et al.* TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort. *Oncogene* **26,** 4596–4599 (2007).
55. Perner, S. *et al.* TMPRSS2:ERG fusion-associated deletions provide insight into the heterogeneity of prostate cancer. *Cancer Res.* **66,** 8337–8341 (2006).
56. Nam, R. K. *et al.* Expression of the TMPRSS2:ERG fusion gene predicts cancer recurrence after surgery for localised prostate cancer. *British Journal of Cancer* **97,** 1690–1695 (2007).
57. Yoshimoto, M. *et al.* Three-Color FISH Analysis of TMPRSS2/ERG Fusions in Prostate Cancer Indicates That Genomic Microdeletion of Chromosome 21 Is Associated with

Rearrangement. *Neoplasia (New York, N.Y.)* **8,** 465–469 (2006).

58. Lapointe, J. *et al.* A variant TMPRSS2 isoform and ERG fusion product in prostate cancer with implications for molecular diagnosis. *Mod. Pathol. an Off. J. United States Can. Acad. Pathol. Inc* **20,** 467–473 (2007).
59. Huang, F. W. *et al.* Highly Recurrent TERT Promoter Mutations in Human Melanoma. *Science (80-.).* (2013). at <http://science.sciencemag.org/content/early/2013/01/23/science.1229259.abstract>
60. Mansour, M. R. *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346,** 1373–1377 (2014).
61. Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526,** 519–524 (2015).
62. Khurana, E. *et al.* Integrative Annotation of Variants from 1092 Humans : Application to Cancer Genomics Integrative Annotation of Variants. **342,** (2013).
63. Stirnimann, C. U., Petsalaki, E., Russell, R. B. & Muller, C. W. WD40 proteins propel cellular networks. *Trends Biochem. Sci.* **35,** 565–574 (2010).
64. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47,** 818–821 (2015).
65. Rieber, N. Coverage Bias and Sensitivity of Variant Calling for Four Whole-genome Sequencing Technologies. **8,** (2013).
66. Griffith, M. *et al.* Optimizing cancer genome sequencing and analysis. *Cell systems* **1,** 210–223 (2015).
67. Lee, H. & Schatz, M. C. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics* **28,** 2097–2105 (2012).
68. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
69. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
70. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
71. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
72. Xu, F. *et al.* A fast and accurate SNP detection algorithm for next-generation sequencing data. *Nat. Commun.* **3,** 1258 (2012).
73. Alioto, T. S. *et al.* A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* **6,** 10001 (2015).
74. Xu, C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput. Struct. Biotechnol. J.* **16,** 15–24 (2018).
75. Smith, N. G. C., Webster, M. T. & Ellegren, H. Deterministic mutation rate variation in the human genome. *Genome Res.* **12,** 1350–1356 (2002).
76. Duret, L., Semon, M., Piganeau, G., Mouchiroud, D. & Galtier, N. Vanishing GC-rich

isochores in mammalian genomes. *Genetics* **162,** 1837–1847 (2002).

77. Brown, T. C. & Jiricny, J. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* **54,** 705–711 (1988).
78. Morton, B. R., Oberholzer, V. M. & Clegg, M. T. The Influence of Specific Neighboring Bases on Substitution Bias in Noncoding Regions of the Plant Chloroplast Genome. *J. Mol. Evol.* **45,** 227–231 (1997).
79. Shibutani, S., Suzuki, N., Tan, X., Johnson, F. & Grollman, A. P. Influence of Flanking Sequence Context on the Mutagenicity of Acetylaminofluorene-Derived DNA Adducts in Mammalian Cells,. *Biochemistry* **40,** 3717–3722 (2001).
80. Meunier, J. & Duret, L. Recombination Drives the Evolution of GC-Content in the Human Genome. *Mol. Biol. Evol.* **21,** 984–990 (2004).
81. Xia, J., Han, L. & Zhao, Z. Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome. *BMC Genomics* **13 Suppl 8,** S7 (2012).
82. Spontaneus Deamination. at <https://upload.wikimedia.org/wikipedia/commons/5/5b/Cytosine_becomes_thymine.png>
83. Blackburn, E. H., Greider, C. W. & Szostak, J. W. Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. *Nat. Med.* **12,** 1133–1138 (2006).
84. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518,** 360 (2015).
85. Hellmann, I. *et al.* Why do human diversity levels vary at a megabase scale? *Genome Research* **15,** 1222–1231 (2005).
86. Tyekucheva, S. *et al.* Human-macaque comparisons illuminate variation in neutral substitution rates. *Genome Biology* **9,** R76 (2008).
87. Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41,** 393 (2009).
88. Cairns, B. R. Chromatin remodeling: insights and intrigue from single-molecule studies. *Nat. Struct. Mol. Biol.* **14,** 989–996 (2007).
89. Melton, C., Reuter, J. A., Spacek, D. V & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* **47,** 710 (2015).
90. Maurano, M. T. *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.* **47,** 1393 (2015).
91. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295,** 1306–1311 (2002).
92. Curtin, P. *et al.* A distant gene deletion affects beta-globin gene function in an atypical gamma delta beta-thalassemia. *Journal of Clinical Investigation* **76,** 1554–1558 (1985).
93. Kaiser, V. B., Taylor, M. S. & Semple, C. A. Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLOS Genet.* **12,** e1006207 (2016).
94. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532,** 264 (2016).

95. Sharma, A. *et al.* The retinoblastoma tumor suppressor controls androgen signaling and human prostate cancer progression. *J. Clin. Invest.* **120,** 4478–4492 (2010).
96. Guseva, N. V, Rokhlin, O. W., Glover, R. A. & Cohen, M. B. P53 and the proteasome regulate androgen receptor activity. *Cancer Biol. Ther.* **13,** 553–558 (2012).
97. Goodwin, J. F. *et al.* A hormone-DNA repair circuit governs the response to genotoxic insult. *Cancer Discov.* **3,** 1254–1271 (2013).
98. Asagoshi, K. *et al.* FEN1 Functions in Long Patch Base Excision Repair Under Conditions of Oxidative Stress in Vertebrate Cells. *Molecular cancer research : MCR* **8,** 204–215 (2010).
99. Al-Ubaidi, F. L. T. *et al.* Castration therapy results in decreased Ku70 levels in prostate cancer. *Clin. Cancer Res.* **19,** 1547–1556 (2013).
100. Mayeur, G. L. *et al.* Ku is a novel transcriptional recycling coactivator of the androgen receptor in prostate cancer cells. *J. Biol. Chem.* **280,** 10827–10833 (2005).
101. Champoux, J. J. DNA topoisomerases: structure, function, and mechanism. *Annu. Rev. Biochem.* **70,** 369–413 (2001).
102. Lin, C. *et al.* Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. *Cell* **139,** 1069–1083 (2009).
103. Kuzminov, A. Single-strand interruptions in replicating chromosomes cause double-strand breaks. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 8241–8246 (2001).
104. Campbell, P. J., Getz, G., Stuart, J. M., Korbel, J. O. & Stein, L. D. Pan-cancer analysis of whole genomes. *bioRxiv* (2017). at <http://biorxiv.org/content/early/2017/07/12/162784.abstract>
105. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38,** 576–589 (2010).
106. Ambrosini, G., Groux, R. & Bucher, P. PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics* bty127-bty127 (2018). at <http://dx.doi.org/10.1093/bioinformatics/bty127>
107. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46,** D260–D266 (2018).
108. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
109. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28,** 1919–1920 (2012).
110. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: Inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31,** 3673–3675 (2015).
111. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3,** 246–59 (2013).
112. Pomerantz, M. M. *et al.* The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nat. Genet.* **47,** 1346–1351 (2015).

113. Chapman, B. Validated, scalable, community developed variant calling, RNA-seq and small RNA analysis. at <https://github.com/bcbio/bcbio-nextgen>
114. Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nature biotechnology* **32,** 71–75 (2014).
115. Alexandrov, L. B. Clock-like mutational processes in human somatic cells. **47,** 1402–1407 (2015).
116. Chakravarti, D., Pelling, J. C., Cavalieri, E. L. & Rogan, E. G. Relating aromatic hydrocarbon-induced DNA adducts and c-H-ras mutations in mouse skin papillomas: the role of apurinic sites. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 10422–10426 (1995).
117. RamaKrishna, N. V *et al.* Mechanism of metabolic activation of the potent carcinogen 7,12-dimethylbenz[a]anthracene. *Chem. Res. Toxicol.* **5,** 220–226 (1992).
118. Nassar, D., Latil, M., Boeckx, B., Lambrechts, D. & Blanpain, C. Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma. *Nat. Med.* **21,** 946–954 (2015).
119. McCreery, M. Q. *et al.* Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers. *Nat. Med.* **21,** 1514–1520 (2015).
120. Takeshita, M. & Eisenberg, W. Mechanism of mutation on DNA templates containing synthetic abasic sites: study with a double strand vector. *Nucleic Acids Res.* **22,** 1897–1902 (1994).
121. Atamna, H., Cheung, I. & Ames, B. N. A method for detecting abasic sites in living cells: age-dependent changes in base excision repair. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 686–691 (2000).
122. Guo, Y. A. *et al.* Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nat. Commun.* **9,** 1520 (2018).
123. Sobol, R. W. *et al.* Mutations associated with base excision repair deficiency and methylation-induced genotoxic stress. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 6860–6865 (2002).
124. Sciences, M. L. Base excision repair : the long and short of it. **66,** 981–993 (2009).
125. Jones, B. Chromatin influence on cancer mutation rate. *Nat. Rev. Genet.* **13,** 596 (2012).
126. Degenhardt, F., Seifert, S. & Szymczak, S. Evaluation of variable selection methods for random forests and omics data sets. *Brief. Bioinform.* bbx124-bbx124 (2017). at <http://dx.doi.org/10.1093/bib/bbx124>