Tampere University

Noora Salokorpi

VARIANTS IN TRANSCRIPTION FACTOR BINDING SITES ALTERING GENE EXPRESSION IN PROSTATE CANCER

Faculty of Medicine and Health Technology
Master's Thesis
November 2022

ABSTRACT

Noora Salokorpi: Variants in transcription factor binding sites altering gene expression in prostate cancer
Master's thesis
Tampere University
Master's Programme in Biotechnology and Biomedical Engineering
Supervisor: Professor Matti Nykter
Examiners: Professor Matti Nykter and University Lecturer, Dr. Tech Juha Kesseli
November 2022

Prostate cancer is the 2nd most prevalent cancer and 5th most worldwide cause of death among men. There are several methods to treat prostate cancer, such as surgery, radiation therapy, hormone therapy, and chemotherapy. Non-lethal primary prostate cancer can develop into lethal castration-resistant prostate cancer. Prostate cancer development is caused by environmental and genetic factors. One promising explanation for prostate cancer development is transcription factor binding in cis-regulatory regions, which promotes or inhibits gene expression. Variants in these cis-regulatory elements can change the binding of transcription factors and, therefore, alter gene expression.

In many cases, the effects of noncoding regions of the genome on gene expression are unclear. Noncoding regions include many essential parts of gene expression regulation, such as promoters, enhancers, and silencers. ATAC-sequencing is a sequencing method used to study chromatin accessibility genome-wide. Open chromatin peaks accessed by ATAC-sequencing contain active parts of the genome, which is why it is a suitable method to study active noncoding regions.

The first aim of this Master's thesis was to perform variant calling with suitable parameters to ATAC-seq. The second aim was to discover common variants within different TFBSs. The third aim was to find out how variants affect the ability of TF to bind to its binding site. This aim was accomplished by comparing PWM scores of wild types and mutated sequences. The main objective, to discover if and which variants in TFBS can change the gene expression close to these regulatory areas, was accomplished by the three aims.

Variant calling was performed with sufficient quality, with the median percentage of ATAC-sequencing variants found from whole genome sequencing variants to be 91.4 %. The five most common transcription factor binding sites for all cell lines and prostate cell lines were CTCF, AR, ESR1, FOXA1, and MYC, and AR, FOXA1, ERG, CTCF, and E2F1, respectively. After running Wilcoxon rank-sum test and Benjamini-Hochberg multiple testing correction for each gene in samples with and without the variant, 443 genes had a p-value less than 0.05. Out of these, eight were considered significant in three transcription factors and 112 in two transcription factors. The eight genes present in three transcription factor binding sites were *ZNF195, RFXANK, PTPN3, MAP4K5, KRIT1, ITGAL, DDX17*, and *AHCY*. Previous studies of *ITGAL, DDX17*, and *AHCY* stated that these genes have a role in prostate cancer development.

To understand whether the variants in transcription factor binding sites were actually the cause of changes in gene expression, more studies would be required. These methods could be, for example, using STARR-sequencing to directly and quantitatively estimate enhancer activity.

Keywords: ATAC-sequencing, prostate cancer, transcription factor, transcription factor binding site, gene expression

TIIVISTELMÄ

Noora Salokorpi: Transkriptiofaktorien sitoutumiskohtien varianttien vaikutukset geenien ekspression muutoksiin eturauhassyövässä
Pro gradu
Tampereen yliopisto
Bioteknologian ja biolääketieteen tekniikan maisteriohjelma
Ohjaaja: Professori Matti Nykter
Tarkastajat: Professori Matti Nykter ja yliopistolehtori, tekniikan tohtori Juha Kesseli
Marraskuu 2022

Eturauhassyöpä on toiseksi yleisin tapausmäärältään ja viidenneksi yleisin kuolinsyy maailmanlaajuisesti miehillä. Eturauhassyövän hoitoon on monia menetelmiä, kuten leikkaus, sädehoito, hormonaaliset hoidot tai kemoterapia. Ei-tappava primaarinen eturauhassyöpä voi kehittyä tappavaksi kastraatioresistentiksi eturauhassyöväksi. Eturauhassyövän kehitys johtuu sekä geneettisistä että ympäristötekijöistä. Yksi lupaava selittävä tekijä eturauhassyövän kehityksessä on cis-säätelyalueen transkriptiofaktorit, jotka edistävät tai vähentävät geenien ekspressiota. Näiden cis-säätelyalueiden variantit voivat muuttaa transkriptiofaktorien sitoutumista ja täten muuttaa geenien ekspressiota.

Genomin ei-koodaavien alueiden vaikutus geenien ekspressioon on monissa tapauksissa epäselvä. Ei-koodaaviin alueisiin kuuluu monia geenien ekspression säätelyn kannalta tärkeitä alueita, kuten promoottorit sekä tehostin- ja vaimenninalueet. ATAC-sekvensointi on sekvensointimenetelmä, jonka avulla voidaan tutkia kromatiinin avoimuutta genomin laajuisesti. Avoimet kromatiinikohdat, joita ATAC-sekvensoinnilla saavutetaan, sisältävät genomin aktiiviset alueet, minkä vuoksi se on hyvä menetelmä tutkia aktiivisia ei-koodaavia alueita.

Tämän tutkielman ensimmäisenä tavoitteena oli suorittaa varianttien kutsuminen sopivilla parametreillä ATAC-sekvensoidusta datasta. Toinen tavoite oli selvittää eri transkriptiofaktorien sitoutumisalueiden yleiset variantit. Kolmas tavoite oli selvittää, kuinka variantit vaikuttavat transkriptiofaktoreiden kykyyn sitoutua sitoutumisalueelle. Tämä tavoite saavutettiin vertaamalla PWM-arvoja normaalin sekvenssin ja mutatoituneen sekvenssin välillä. Päätavoite, joka oli selvittää, jos ja mitkä variantit transkriptiofaktorien sitoutumiskohdissa muuttavat geenien ekspressiota, saavutettiin näiden tavoitteiden avulla.

Varianttien laatu oli riittävä. ATAC-sekvensoinnista saaduista varianteista mediaaniprosentiltaan 91,4 % löytyi myös koko genomin sekvensoinnin varianteista. Viisi yleisintä transkriptiofaktorin sitoutumiskohtaa kaikille solulinjoille oli CTCF, AR, ESR1, FOXA1 ja MYC ja eturauhasen solulinjoille AR, FOXA1, ERG, CTCF ja E2F1. Wilcoxonin järjestyssummatestin ja Benjamini-Hochbergin monen testin korjaamismenetelmän geenien näyteryhmille variantilla ja ilman jälkeen jäljelle jäi 443 geeniä, joiden p-arvo oli alle 0,05. Näistä geeneistä kahdeksaa pidettiin merkityksellisenä kolmessa transkriptiofaktorissa ja 112:ta kahdessa transkriptiofaktorissa. Kahdeksan geeniä, jotka löytyivät kolmesta transkriptiofaktorista, olivat *ZNF195, RFXANK, PTPN3, MAP4K5, KRIT1, ITGAL, DDX17* ja *AHCY.* Aikaisempien tutkimusten mukaan *ITGAL*, *DDX17* ja *AHCY* toimivat jonkinlaisessa roolissa eturauhassyövän kehityksessä.

Näiden transkriptiofaktorien sitoutumiskohtien varianttien merkityksen ymmärtäminen geenien ekspression säätelyssä vaatisi lisätutkimuksia. Tämä voisi tarkoittaa esimerkiksi STARR-sekvensoinnin käyttämistä tutkiakseen tehostinalueita suoraan ja määrällisesti.

Avainsanat: ATAC-sekvensointi, eturauhassyöpä, transkriptiofaktori, transkriptiofaktorin sitoutumiskohta, geenien ekspressio

PREFACE

This thesis was performed in Computational Biology research group at the Faculty of Medicine and Health Technology at Tampere University. Firstly, I would like to thank Professor Matti Nykter for giving me the topic of this thesis and answering and helping with all my questions and problems.

I would also like to thank my co-workers, friends, and family for encouraging and supporting me. My special thanks will go to Lauri and Lilja for bearing and cheering me through the work.

Tampere, 28th November 2022

Noora Salokorpi

CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

ADT	Androgen deprivation therapy
ATAC	Assay for transposase-accessible chromatin
AR	Androgen receptor
BPH	Benign prostate hyperplasia
CTCF	CCCTC-binding factor
CRPC	Castration-resistant prostate cancer
ERG	ETS-related gene
ESR1	Estrogen receptor 1
ETS	Erythroblast transformation specific
E2F1	E2F transcription factor 1
FOXA1	Forkhead box protein A1
hg38	Human genome build 38
HTS	High-throughput sequencing
PC	Prostate cancer
PFM	Position frequency matrix
PPM	Position probability matrix
PWM	Position weight matrix
QTL	Quantitative trait locus
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
TAD	Topologically associated domain
TF	Transcription factor
TFBS	Transcription factor binding site
TSS	Transcription start site
VAF	Variant allele frequency
WGS	Whole genome sequencing

1. INTRODUCTION

Prostate cancer is the most common malignancy in men. There are many ways to treat prostate cancer, such as surgery, radiation therapy, hormone therapy, and chemotherapy. One common treatment method for prostate cancer is hormone therapy called androgen deprivation therapy, which targets testosterone production or blocks it from acting on prostate cancer cells. Since some prostate cancer cells can grow in an environment of low testosterone, cancer can progress into castration-resistant prostate cancer, which can be lethal.

Transcription factors are proteins that regulate the transcription of genes by binding to specific DNA sequences. These specific sequences are called transcription factor binding sites. The binding of a transcription factor to a transcription factor binding site can be computationally modeled with *in silico* models, such as position weight matrices. These acquired scores tell whether different sequences tend to bind certain transcription factors. As everywhere in DNA, also transcription factor binding sites can contain variants. These variants can be somatic or germline mutations. Mutations may improve or weaken the binding of a transcription factor, which can then change the regulation of gene expression.

Studying noncoding regions can reveal new information explaining the development of different diseases. Multiple studies have illustrated the role of variants in disease development in noncoding regions.

This thesis aimed to perform variant calling of single nucleotide variants to 38 ATAC-sequenced samples. The effects of these variants on transcription factor binding were then studied by comparing position weight matrix scores of original and mutated sequences. Variants having scores with significant differences were matched 250 kilobases up and downstream of different genes. After this step, the gene expression scores within different samples were analyzed. The aim was to find whether called variants impacted gene expression between samples with variants and those without. The significance of the difference was tested by the Wilcoxon rank-sum test between samples containing a variant and those not in the gene window.

2. LITERATURE REVIEW

2.1 Diseases of prostate

Even though prostate cancer is a more widely spoken condition, the prostate can also develop into benign prostatic hyperplasia (BPH) and prostatitis (Motrich et al., 2018). According to studies, there is a 50 % chance of developing BPH at age 51-60 and a 70 % chance of age between 61 and 70 years in men (Miah & Catto, 2014).

2.1.1 Prostate cancer

Prostate cancer (PC) is a cancer of the prostate. It is the 2nd most prevalent cancer and 5th most worldwide cause of death among men (Bray et al., 2018). Even though most prostate cancer cases are clinically insignificant, they can develop into deadly cancer in some cases. The problem is that prostate cancer is highly heterogenous, and it can be challenging to recognize fatal cases from clinically insignificant ones. (Spans et al., 2013)

Prostate cancer development is a result of both environmental and genetic factors. Examples of genetic factors are estrogen synthesis, metabolism, and signal transduction pathways. (Y.-M. Wang et al., 2013) Examples of environmental factors are radiation and different chemicals.

Even though studies have not found a common etiology between BPH and PC, both have growth dysregulation of prostatic cells during development (Shah & Getzenberg, 2004). Chen et al., showed in their studies that there are seven hub genes among sixty differentially expressed genes that may indicate which BPH patients develop their hyperplasia into prostate cancer (Chen et al., 2022). These genes are *MYC, CXCR4, CSRP1, SNAI2, MYL9, ACTG2*, and *MYH11* (Chen et al., 2022). Due to this link and the high occurrence of BPH, it is essential to acknowledge hyperplasia samples when studying prostate cancer.

Depending on the state of prostate cancer, there are different treatment methods. These are, for example, surgery, radiation therapy, and chemotherapy. One possible method is hormone therapy since prostate cancer cells usually need testosterone to grow. Androgen deprivation therapy (ADT) is a treatment method that targets testosterone production

or blocks it from acting on prostate cancer cells. Since most prostate cancer cells die from being deprived of testosterone, ADT is a commonly used and efficient method.

The downside of ADT is that since it is not a curative treatment method, cancer can develop into lethal castration-resistant prostate cancer (CRPC) even after multimodal therapy with different treatment methods and medicine (Imamura & Sadar, 2016; G. Wang et al., 2018). This development is due to some prostate cancer cells being able to get the ability to grow in the environment of low testosterone and are therefore not affected by ADT. When there are more of these cells, and ADT cannot kill them anymore, PC has developed into CRPC. The survival rate of CRPC is much worse than in primary prostate cancer, which is why new treatment methods are needed for those cases (Kodama et al., 2020).

Studies in recent years have developed multiple agents that have strongly impacted the overall survival of CRPC cases. Examples of these agents are sipuleucel-T, radium-223, abiraterone, enzalutamide, and cabazitaxel (Komura et al., 2018). The optimization of these factors is still a work in progress.

Prostate cancer is the most common cancer in Finland. The rate of incidences has grown in number since the 1990s. However, age-standardized prostate cancer mortality has decreased since the 1980s, with 195.1 cases per 100 000 person-years in 2019. The overall mortality rate has increased, with 5245 new cases in 2019. (Pitkäniemi et al., 2021)

2.2 Gene expression

Gene can be defined in different ways. In this thesis, the gene is a DNA segment transcribed and translated into RNA or polypeptide and has some functionality (Orgogozo et al., 2016). Transcription is a process in which a part of DNA is processed into messenger RNA (mRNA) (Ganguly, 2022a). This mRNA is then used as an instruction to build a polypeptide chain (Ganguly, 2022b). Transcription happens in the nucleus, and translation occurs in ribosomes.

The transcription start site (TSS) is a DNA strand where transcription starts. TSS is located within the promoter area of the gene. A promoter is a short part of DNA in which different proteins bind and initiate the start of transcription. The promoter binds the transcription machinery, which consists of RNA Polymerase II and its associated general

transcription factors (Haberle & Stark, 2018). The promoter is typically located either at the 5’ end of the transcription start site or directly upstream (Q. Zou et al., 2019).

Enhancers are DNA sequences located remotely to promoters that can increase the transcription of genes (Andersson, 2014; Pennacchio et al., 2013). Enhancers work by forming chromatin loops that get the enhancer and target gene into close proximity (Pennacchio et al., 2013). Alternatively, silencers can also repress gene expression(Doni Jayavelu et al., 2020). Silencers are DNA sequences that bind transcription factors called repressors. Repressors prevent the binding of RNA Polymerase, which then prevents the start of transcription.

2.2.1 Regulation of gene expression

Gene expression can be regulated in many alternative ways. One key component in gene expression regulation is transcription factors (Vaquerizas et al., 2009). TFs can change the activity of cellular functions and cells’ responses to the environment. Studying TF activity and expression in different cell lines and tissues may provide information on which TFs are most active with different genes and their expression. Activating TFs are called activators and bind to enhancers while repressing TFs are called repressors and bind to silencers.

2.2.2 Gene expression matrix

A gene expression matrix is a computationally produced matrix in which rows usually represent genes and columns their gene expression scores. Gene expression matrices are typically developed from microarrays or RNA-seq data.

2.3 High-throughput sequencing

High-throughput sequencing (HTS) is also known as next-generation sequencing (NGS). HTS can be used to perform sequencing of DNA or RNA, and accessed reads can be a single-end or paired-end (Taguchi, 2018). The paired-end method means that sequenced reads are made from both ends of DNA or RNA fragments, while the single-end method means that sequencing is only done from one end of the fragment.

Even though there are different high-throughput sequencing techniques, they usually involve some same steps. These steps are template preparation, clonal amplification, and parallel sequencing (Farrar, 2019). DNA fragments are isolated, purified, and compiled in template preparation to make a DNA library. In clonal amplification, compiled libraries are copied in flow cells, and fragments are amplified into clusters. In parallel sequencing, templates are sequenced at the same time. (Farrar, 2019)

The preparation of samples makes the difference between different high-throughput sequencing methods. For example, the whole genome, RNA, or open chromatin areas of the genome are sequenced similarly, and other parts of the genome are collected in sample preparation steps (Gautam et al., 2019).

After sequencing, there are raw sequence reads. Often the next step is sequence alignment. Sequence alignment is a method in which DNA, RNA, or protein sequences are arranged to study their similarities. Mutations in sequences, such as point mutations, insertions, and deletions, are considered during alignment. Insertions and deletions are presented as gaps. (Prjibelski et al., 2019) Aligned reads can partially, completely, or not overlap with other reads.

2.3.1 Assay for Transposase-Accessible Chromatin using sequencing

Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) is used to study chromatin accessibility genome-wide (Buenrostro et al., 2015). The openness of chromatin can be split into transcriptionally active euchromatin and inactive heterochromatin (Yan et al., 2020). The difference between these structures is presented in figure 1. ATAC-seq is based on hyperactive Tn5 transposase, which utilizes the “cut and paste” mechanism (Buenrostro et al., 2015). Tn5 transposase adds sequencing primers to euchromatin areas. This step is called tagmentation (Buenrostro et al., 2015).

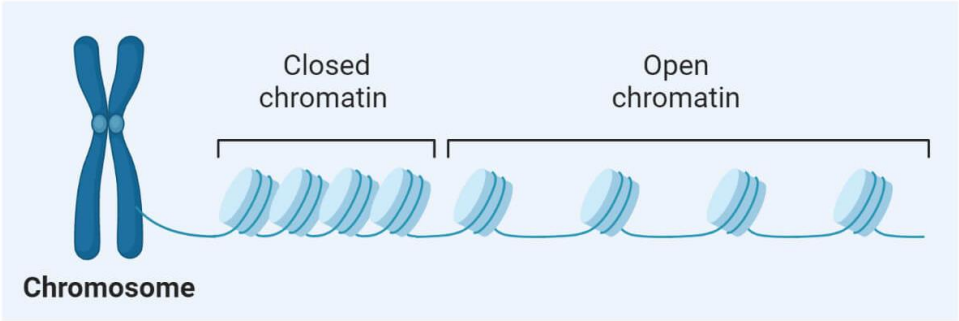

Figure 1. *Difference between closed and open chromatin, adapted from https://thebiologynotes.com/chromatin/ (accessed on 1.10.2022)*

There are three steps in ATAC-seq: nuclei preparation, transposition, and amplification. First, target cells are lysed to get the nuclei. After that, Tn5 transposase is added to tag into DNA with two adapters. Finally, collected primers are used to generate a library for sequencing. (Sun et al., 2019)

Data from ATAC-seq can be used to study peaks of open chromatin areas and, for example, their length or the genes that they are associated with (Sun et al., 2019). The biggest perks of ATAC-seq are its simpleness and time efficiency (Yan et al., 2020). ATAC-seq is also an efficient sequencing method when interested in important noncoding regions (Massarat et al., 2021). Some cell types and tissues have problems with ATAC-seq since individual optimizations are needed to sequence them properly. One way to make ATAC-seq data more comparable is to use Omni-ATAC, an improved ATAC-seq protocol (Corces et al., 2017).

2.3.2 Whole genome sequencing

In recent years, sequencing methods have become cheaper and more accessible. Therefore, sequencing of the whole genome has become a suitable method for performing genome-wide analysis. Whole genome sequencing (WGS) is a type of next-generation sequencing. (van El et al., 2013)

WGS is an excellent method since it allows for studying changes everywhere in the genome. It produces a lot of raw data to investigate further, such as identifying inherited diseases and characterizing the mutations driving cancer development. The choice of a

suitable sequencing method depends on the need for analysis. The downside of WGS is its costs when multiple individuals are sequenced (Massarat et al., 2021).

2.3.3 RNA sequencing

RNA sequencing (RNA-seq) is a next-generation sequencing method in which transcribed mRNA is converted into complementary DNA (cDNA) library and then sequenced. cDNA is more stable than RNA. RNA-seq is primarily used to study differential gene expression and alternative splicing of messenger RNAs. Nowadays, it is also possible to use RNA-seq to study, for example, single-cell gene expression and translation. (Stark et al., 2019)

RNA-seq consists of the mRNA of an individual. In practice, some parts of RNA can be left out or picked. The relative number of these RNAs also represents the expression of the corresponding genes (Finotello & di Camillo, 2015). Therefore, aligned reads of RNAs can be computationally analyzed to study gene expression.

2.4 Single nucleotide variants and single nucleotide polymorphisms

A single nucleotide variant (SNV) is a variant that takes place at a specific genomic position in a single nucleotide (H. Zou et al., 2020). Single nucleotide polymorphism (SNP) is also a variant of a single nucleotide, but it occurs in more than 1 % of a population (Børsting & Morling, 2013).

Somatic mutations can occur in any cell lineage except for the germline. Therefore, somatic mutations are not inherited. Somatic mutations are a normal part of the life cycle. They can result from stress or defects in the DNA repair system. Somatic mutations may have a role in cancer development, especially if they make a growth advantage or prevent apoptosis. (Miles & Tadi, 2022)

Germline mutations are mutations in germ cells, sperm, and egg, that are inherited by offspring. This fact means that the specific mutation occurs in each cell of the offspring's body. Germline mutations may lead to different hereditary diseases (Newkirk et al., 2017).

2.5 Transcription factors

Transcription factors (TF) are proteins that can upregulate or downregulate the transcription rate by binding to specific DNA sequences (Hombach et al., 2016). They can control gene expression and, therefore, they also control different molecular and cellular processes. TFs consist of at least two parts: a sequence-specific DNA-binding domain and a domain that acts as an activator or repressor and can depend on cofactors (Bhagwat & Vakoc, 2015).

2.5.1 Structural motifs

A structural motif means a three-dimensional structure of a protein. These structures consist of secondary protein structures, of which the most common ones are α-helix and β-sheet. In α-helix, the carbonyl of one amino acid is hydrogen bonded to the amino H of an amino acid that is four down the chain. In the β-sheet, the hydrogen bonds form between the carbonyl and amino groups of the backbone. The strands can be parallel or antiparallel. (https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure; cited 17.10.2022)

Transcription factors that need dimerization to bind into DNA sequences are typically leucine zipper factors or helix-loop-helix factors (Daniel H. Gonzalez, 2015). Leucine zipper factors consist of basic regions and leucines located seven residues apart along an α-helix. The leucine zipper factor is presented in figure 2a. Helix-loop-helix factors consist of two α-helices connected by a loop. This domain is illustrated in figure 2b. Both these classes are part of the basic domain superclass.

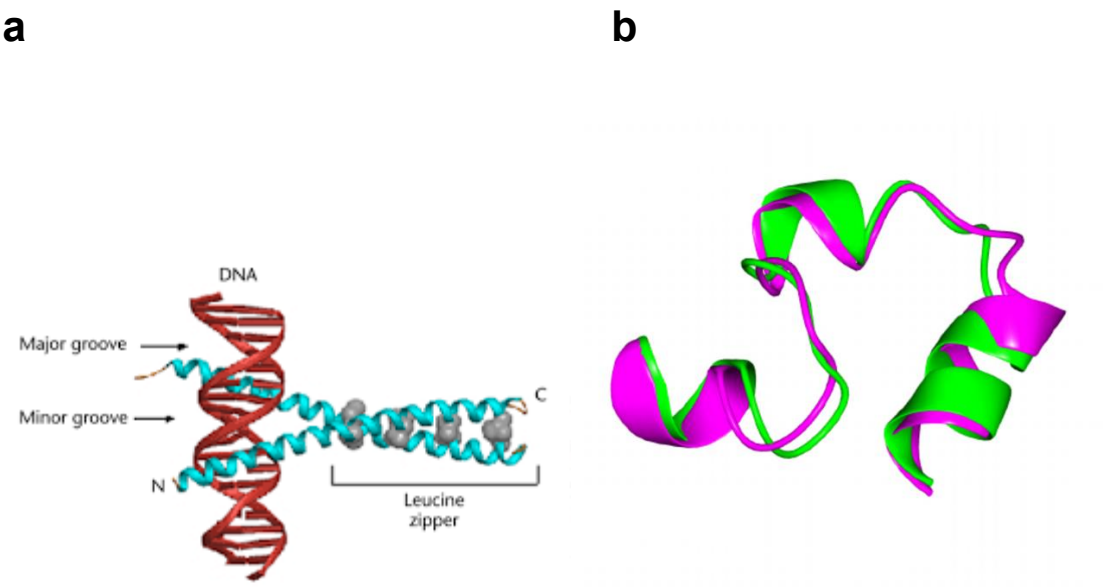

Figure 2. ***(a)*** *The structure of the leucine zipper domain. Adapted from Krylov & Vinson, 2001* ***(b)*** *The structure of the helix-loop-helix domain. Adapted from Chernodub et al., 2010.*

Besides that, there are four other superclasses of transcription factors; Zinc-coordinating DNA-binding domain, helix-turn-helix, β-Scaffold factors with minor groove contacts, and other transcription factors (Daniel H. Gonzalez, 2015). A common structure of the zinc-coordinating DNA-binding domain is the zinc finger. Zinc fingers are small protein motifs with many finger-like protrusions that then have contact with their target molecule (Klug, 1999). Zinc fingers can bind zinc, other metals, or no metal. The structure of zinc fingers is presented in figure 3a. Some transcription factors belong to the helix-turn-helix domain, consisting of two α-helices with a turn between them. A structure of the helix-turn-helix domain is presented in figure 3b.

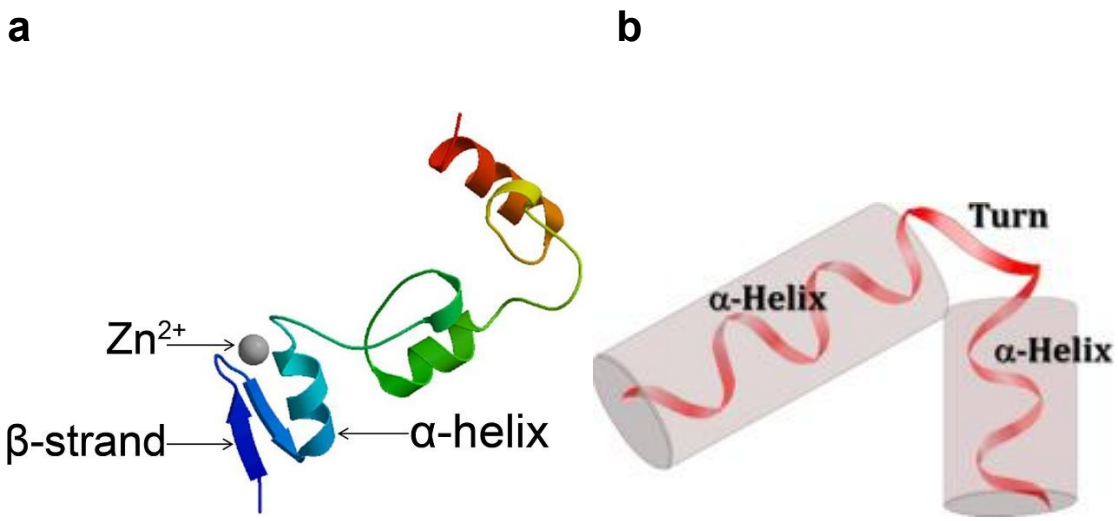

Figure 3. ***(a)*** *The structure of zinc fingers. Adapted from Han et al., 2020* ***(b)*** *The structure of the helix-turn-helix domain. Adapted from Roy & Kundu, 2021.*

2.5.2 Transcription factor binding sites

Transcription factor binding sites (TFBS) are motifs that tend to be binding sites of transcription factors of different kinds. The length of TFBSs is usually around 6-12 nucleotides (Vinson et al., 2011). TFBSs are typically located near the transcriptional start site (TSS); from there, they can activate or repress gene expression (Vinson et al., 2011).

2.5.3 Transcription factors in prostate cancer

As mentioned in 2.5 Transcription factors, many TFs have essential roles in cancer development. One significant TF in primary prostate cancer and metastases is AR (Culig & Santer, 2014). Besides prostate cancer, AR has an essential role in the normal homeostasis of the prostate, in addition to NKX3-1 and p63. ETS family members and c-MYC are important TFs in primary prostate cancer, while AR and FOXA1 are critical players in CRPC. (Labbé & Brown, 2018)

Androgen receptor (AR) is a transcription factor that acts as a regulator for androgens. AR is widely expressed in many cells and has various roles in regulating different processes, such as immune and neural systems development. AR may also have an es-

sential role in the development of cancers such as prostate and bladder. (Davey & Grossmann, 2016). The structure of AR belongs to zinc-coordinating DNA-binding domains and nuclear receptors with the C4 zinc fingers class.

ETS-related gene (ERG) is a member of the ETS family of transcription factors, and it is an oncogene. ETS is short for erythroblast transformation specific. Genes of the ETS family have roles in cell cycle control, cell proliferation, differentiation, migration, and apoptosis (Abou-Ouf et al., 2016). Gene fusion with the transmembrane protease serine 2 (TMPRSS2) is a commonly found structure in prostate cancer (Z. Wang et al., 2017). ERG belongs to the helix-turn-helix domain, specifically the tryptophan cluster factors class.

Myc is a family of proto-oncogenes that make transcription factors c-Myc (MYC), l-Myc (MYCL), and n-Myc (MYCN). MYC has a vital role in cancer formation. Therefore, it could be a potential target for cancer treatment. (Duffy et al., 2021) MYC is a basic helix-loop-helix leucine zipper transcription factor, which means it has both helix-loop-helix and leucine zipper motifs.

Forkhead box A1 (FOXA1) encodes a factor that alters the open chromatin conformation. This change allows other transcription factors to be able to bind. One example of these transcription factors is AR, which improves the growth and survival of normal prostate and prostate cancer cells. (Teng et al., 2021). FOXA1 belongs to the helix-turn-helix domain superclass and, more specifically, is a winged helix factor.

2.5.4 *In silico* models

The binding of TFs to TFBS is presented with *in silico* models, such as position frequency matrices (PFM), position probability matrices (PPM), and position weight matrices (PWM) (Hombach et al., 2016). Position frequency matrices tell how many times each nucleotide occurs in a specific position. Position probability matrices are created by dividing the number of occurrences of the position frequency matrix by the overall number of sequences. The equation for calculating PPM for nucleotide i in location j from PFM is presented in formula (1):

$$(1)\ \mathrm{PPM(i,j)} = \frac{\mathrm{PFM(i,j)}}{\sum_i(\mathrm{PFM(i,j)})},$$

where i means a nucleotide (guanine (G), adenine (A), cytosine (C), and thymine (T)) and j means the location of that nucleotide in the alignment (Fostier, 2020).

Position weight matrix scores are calculated as position-specific log-likelihoods. (Nishida et al., 2008) Calculating the position weight matrix for nucleotide i in position j from the position probability matrix is presented in formula (2):

(2) $\mathrm{PWM}(i,j) = log_2(\frac{\mathrm{PPM}(i,j)}{b_i})$,

where i means a nucleotide (A, C, G, T), j represents the location of that nucleotide in the alignment, and b_i means the corresponding background nucleotide probability (Fostier, 2020).

When the PWM score equals 0, there is an equal probability of the sequence being a functional site or random site. When the score is higher than 0, the probability of the sequence being a functional site is higher than a random site, and when less than 0, vice versa. (https://bioinformaticaupf.crg.eu/T12/MakeProfile.html; cited 12.10.2022)

There are multiple databases that have PWMs for different TFBSs. The differences in confidence between these databases differ. JASPAR and HT-SELEX-derived matrices produced more reliable results in identifying *in vivo* TFBSs than PBM-derived models (Hombach et al., 2016).

2.5.5 Variants of transcription factor binding sites in prostate cancer

Noncoding regions of the genome are unexplored in many parts. Therefore, studying these regions may reveal new mutations and changes in DNA that can explain cancer development. Variants can change the binding of proteins to better or worse, which may then influence the transcription and protein synthesis and, thereby, change the gene expression.

Cohesin is a protein complex that associates with transcription factors, especially CTCF. Cohesin is present in almost all parts of the genome, especially in locations where transcription factors are present (Katainen et al., 2015). CTCF/Cohesin-binding sites (CBSs) can alter the stability of chromatin loops. Variations in CBSs can cause multiple cancers,

including early-onset prostate cancer (Katainen et al., 2015). The known variant to cause enhancement of CTCF binding is prostate cancer-associated rs7077275, which leads to a decrease in the apoptosis of prostate cancer cells (Tseng et al., 2021).

Mutations in AR have been found in prostate cancer cases. The occurrence of these mutations may induce tumor growth. Clinical studies show the effects of variants on the development of prostate cancer. For example, a mutation Thr877Ala leads to an increased AR binding affinity. (Culig & Santer, 2014). Another variant to induce prostate cancer risk is polymorphism rs684232, which has multiple causes, such as the downregulation of AR (Tseng et al., 2021).

A few variants are identified in the binding site of ESR1, which are thought to give an advantage to prostate cancer development. According to the studies of Wang et al., different polymorphisms can cause cancer between different ethnicities and countries. For example, ESR1 PvuII (C>T) polymorphism significantly impacts prostate cancer development within the Asian population. (Y.-M. Wang et al., 2013)

Multiple studies have also found TFBS variants that can cause the development of prostate cancer. A polymorphism rs339331 found in the HOXB13-binding site has been noticed to enhance the binding of HOXB13, which then results in the upregulation of RFX6 (Tseng et al., 2021). This upregulation makes prostate cancer cells more active in dividing and growing.

According to previous studies, the altered binding of TF has been linked to various diseases, such as osteoarthritis, type-2 diabetes, and colorectal cancer (Dodd et al., 2013; Claussnitzer et al., 2015; S. Wang et al., 2015; Shi et al., 2019). According to Grishin and Gusev, multiple cancer allele-specific accessibility quantitative trait loci (as-aQTLs) alter TF binding sites and, thereby, TF binding and gene expression (Grishin & Gusev, 2022). The most extensive number of as-aQTLs were found in breast, prostate, and renal cancer. These cancers have significant heritability enrichment compared with all peaks. (Grishin & Gusev, 2022)

2.6 Background of variant calling

As mentioned before, mutations in TFBSs can be significant factors for cancer development. Therefore, an efficient and reliable variant calling pipeline is vital for this analysis. Variant calling can mean the calling of inherited SNVs, indels, somatic mutations, copy

number variants, and structural variants. (Koboldt, 2020). In this thesis, the focus is on the variant calling of SNVs. There are multiple tools for variant calling, such as BCFtools, GATK HaplotypeCaller, and FreeBayes (Koboldt, 2020).

Variants from variant calling are often filtered. For this thesis, the criteria used were read depth and quality score. Read depth means the number of reads overlapping alignments in a particular locus. (Strom, 2016). Low read depth makes it difficult to separate actual variants and sequencing errors. Another filtering criterion was a quality score. The quality score is a PHRED-scaled probability that tells whether a single base is correct (Strom, 2016). The quality score is calculated with the formula (3) below:

$$(3)\ Q = -10\, log_{10}\, P,$$

Where Q stands for PHRED-scaled probability and P for the likelihood of error.

Variant allele frequency (VAF) means the percentage of sequencing reads that have a particular variant out of overall coverage at that specific locus. Homozygous loci have a ratio of approximately 100 %, heterozygous loci approximately 50 %, and reference loci 0 %. (Strom, 2016)

2.7 Background of statistical testing

Statistical tests are used to decide whether there is enough evidence to "reject" a null hypothesis about a process. The null hypothesis assumes that two possibilities are equal, e.g., two population means or medians are equal. Another hypothesis is called the alternative hypothesis, and in that situation, it would be that two population means or medians are not equal. A common measure to analyze hypotheses of statistical tests is a p-value. The p-value is the probability of the test statistic being, at the minimum, as extreme as the one gotten if we presume the null hypothesis is true (NIST/SEMATECH, 2012). The smaller the p-value, the more likely the null hypothesis is false. In other words, if the limit of the p-value is set to be 0.05 and the p-value is less than 0.05, we can reject the null hypothesis and assume that the result can be statistically significant (Kilcoyne et al., 2013).

Statistical tests can be parametric or non-parametric. Parametric tests assume a normal distribution in the dataset (Kilcoyne et al., 2013). Examples of parametric tests are t-

tests, analysis of variance for comparing groups, and least squares regression and correlation (Kilcoyne et al., 2013). If the data does not have a normal distribution, it is wise to use a non-parametric test instead. The non-parametric test does not make any assumptions about the distribution or variance of data. Examples of non-parametric tests are Wilcoxon signed rank test, the Kruskal-Wallis test, and Wilcoxon rank-sum test (Kilcoyne et al., 2013). The selection of suitable test is based on other factors of the data. For example, Wilcoxon signed rank test assumes that compared samples are related, while Wilcoxon rank-sum test assumes the samples to be independent.

Wilcoxon rank-sum test, also known as Mann-Whitney U test, is a non-parametric statistical test for independent samples. Wilcoxon rank-sum test assumes as a null hypothesis that for two random values, A and B, from different populations, the probability of A being greater than B and B being greater than A is equal. The equation for Wilcoxon rank-sum test is presented in formula (4).

$$(4)\ \mathrm{U} = \mathrm{W} - \frac{n_2(n_2+1)}{2},$$

Where W stands for the sum of the rank and n_2 stands for the sample size for sample 2. The same equation can be proved for sample 1 with n_1. (Hogg et al., 2013)

Multiple testing correction adjusts p-values from multiple statistical tests to correct the occurrences of false positive values. Different methods exist to perform multiple testing corrections, such as Bonferroni, Bonferroni Step-Down, Westfall and Young Permutation, and Benjamini-Hochberg.

The Bonferroni correction is the most strict test. In that correction, each p-value is multiplied by the number of samples in the sample list (Silicon Genetics, 2003). The Benjamini-Hochberg procedure is the least strict of all correction methods and gives a good balance between statistically significant p-values and false positive findings (Noble, 2009; Silicon Genetics, 2003). In Benjamini-Hochberg correction, each p-value is multiplied by the total number of samples in the sample list divided by its position in the list (Silicon Genetics, 2003).

3. OBJECTIVES

The main objective is to discover if and which variants in TFBS can change the gene expression close to these regulatory areas. This objective is accessed with three aims of this thesis.

The first aim was to perform variant calling with suitable parameters to ATAC-seq. Quality control and other studies were used to achieve this goal. The second aim was to discover common variants within different TFBSs. The third aim was to find out how variants affect the ability of TF to bind to its binding site. This aim was accomplished by comparing PWM scores of wild types and mutated sequences. After all these aims have been achieved, we will get to the primary objective of this thesis.

4. MATERIALS AND METHODS

4.1 Flowchart of the workflow

The flowchart for all materials and methods can be seen in figure 4. Colored circles are materials, light grey boxes are methods, and dark grey boxes are produced dataframes.

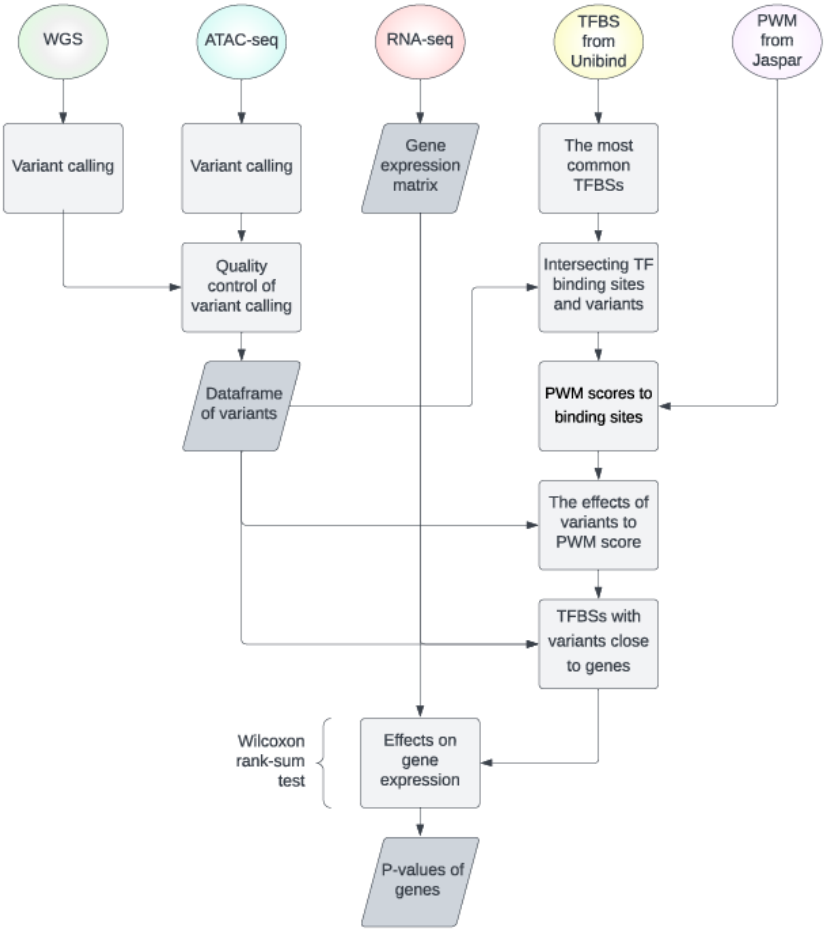

Figure 4. *Flowchart of the workflow*

4.2 Materials

This analysis was done with ATAC-sequenced data. The samples were from Tampere University Hospital and ATAC-sequenced data from Computational biology group. The

ATAC-seq data was beforehand aligned with Bowtie 2 version 2.3.4.1 against the hg38 reference genome with parameters --sensitive-local and -X 2000 (Langmead & Salzberg, 2012). After this step, filtering with parameter -q 20, sorting, and indexing was done with SAMtools version 1.8 (Danecek et al., 2021). The last step to pre-process this data set was to mark duplicates using Picard Markduplicates tool, version 2.9.2-SNAPSHOT, with parameters VALIDATION_STRINGENCY=LENIENT and REMOVE_DUPLICATES=FALSE (https://broadinstitute.github.io/picard/; cited 4.10.2022).

To perform quality control of variant calling, whole genome sequencing (WGS) data of partially the same patients as ATAC-seq data was used. The gene expression matrix of RNA-seq data was also from the same patients as ATAC-sequenced data. The bulk ATAC-seq data comprised 11 benign prostate hyperplasias, 16 untreated primary prostate cancer, and 11 castration-resistant prostate cancer samples.

4.3 Variant calling

Variant calling was performed with BCFtools version 1.9-174-g4caf1fd (Danecek et al., 2021). Genotype likelihoods at each genomic position were counted with the command BCFtools mpileup against hg38 reference genome with parameter -I. After this, variants were called with the command call from BCFools with parameter -mv and normalized with command bcftools norm with parameter -m. After this step, variants were filtered with BCFtools filter. Variants with a quality score over 35 and read depth over 10 were included for further analysis. Finally, duplicate reads were removed with Picard Markduplicates tool, version 2.9.2-SNAPSHOT, with the parameter REMOVE_DUPLICATES=TRUE. This step was done to ensure that BCFtools does not include duplicate variants for further analysis.

Variants were intersected with open peak areas of ATAC-seq. These peaks present the openness of chromatin in different areas. The openness of each peak is presented with a value starting from 0, with a bigger number indicating a more open chromatin. Peaks with a value over a threshold of 5 were collected. The percentage of peaks containing a variant out of all peaks was calculated and presented in a boxplot with R version 3.6.0.

The same steps were done for ATAC-seq data and WGS data. The difference between data sets is that variants of ATAC-seq data were only from open chromatin areas, while the

variants of WGS data were from the whole genome. Nonetheless, the data sets were similarly produced and similar enough to perform quality control as presented in 4.4 Quality control of variant calling.

4.4 Quality control of variant calling

Quality control of variant calling was performed to ensure the quality of data. The first step in analyzing variant quality was to count variant allele frequency. All known SNPs of the human genome were downloaded from the dbSNP database (Smigielski et al., 2000). These SNPs were intersected with ATAC-seq peaks with BEDtools intersect command. From these, the VAF was calculated for each SNP. The distribution of VAF was plotted with a histogram.

Besides VAF distribution and duplicates, variants of ATAC-seq data were compared to WGS data variants. If most of the ATAC-seq variants could be found from WGS data, the quality of variants would be more prominent. Variants were compared with BEDtools version 2.29.1 with the command intersect and parameter -u (Quinlan & Hall, 2010). There were nine same samples between these datasets, so these samples were studied. Intersected variants were plotted against variants of WGS data with R package ggplot2 version 3.3.6 in R version 4.1.2 (Wickham, 2016). In the plot, allele fraction against coverage was plotted.

4.5 Finding variants within TFBSs

The next step in this Master's thesis was finding variants in TFBSs. Called variants from 4.3 Variant calling were intersected with TFBSs from the Unibind database.

4.5.1 Collecting the most common transcription factors

Transcription factor binding sites were downloaded from the Unibind database (Gheorghe et al., 2019; Puig et al., 2021). These binding sites were collected from all cell lines and prostate cell lines. The 20 most common transcription factors from all cell lines and prostate cell lines were collected, and their frequency is presented in a histogram produced with R version 4.1.2. Only the 20 most common transcription factors were presented because the amount of these factors was higher than the amount of all the rest of the transcription factors

4.5.2 Intersecting TF binding sites and variants

Intersections of locations of transcription factor binding sites from Unibind and ATAC-seq variants of variant calling were analyzed with BEDtools intersect version 2.29.1 with parameters -wa and -wb. This step was done for all cell lines and prostate cell lines.

4.5.3 Calculating PWM scores to binding sites

Intersections of TFBSs and variants were downloaded to Python version 3.8.8 and used with Jupyter Notebook version 6.3.0. Python package pandas, version 1.2.4, was used to process information in dataframes in this and the next steps of methods (Mckinney, 2010; The pandas development team, 2020). If the TFBS was from the reverse strand, it was reverse complemented. The five most common transcription factor binding sites were used to further studies.

The position weight matrix was downloaded from the Jaspar database for each TFBS (Castro-Mondragon et al., 2022). Then, each wild-type TFBS got a PWM score calculated via PWM. The PWM score was calculated by collecting a value matching the nucleotide (row) and the number of nucleotides in sequence (column) and adding it to the score. The same step was repeated for each nucleotide in the sequence.

PWM score was also calculated for each TFBS with a variant. After the variant was modified to the sequence, the PWM score of each mutated TFBS was calculated.

4.5.4 The effects of variants on PWM scores

The aim was to discover which variants either improve or weaken the binding of TF. This step was done by comparing the PWM scores of the reference and mutated sequences. When the difference is positive, the variant has weakened the binding, and when negative, the variant has improved the binding.

Variants that had made a difference in PWM scores between wild-type and mutated sequence more significant than 5 or less than -5 were collected. These variants had the most significant impact on binding.

4.6 The effects of variants on gene expression

After identifying variants changing the binding affinity of TFBSs, these binding sites were compared to known genes. The changes in gene expression scores were compared between samples with and without variants in the TFBS of the particular gene.

4.6.1 TFBSs with variants close to genes

Variants of TFBSs were annotated with Homer version v4.11 with Annotatepeaks.pl with hg38 (Heinz et al., 2010). Variants annotated as 'TSS' were chosen for further studies.

The gene expression matrix was formed from RNA-sequenced data of the same samples as ATAC-sequenced data. The genomic locations of these genes were analyzed with the BiomaRt package version 2.54.0 in R (Durinck et al., 2005, 2009).

TFBSs with a variant annotated as 'TSS' that were 250 000 base pairs to up- or downstream genes were collected together. This step analyzed which TFBS is regulating each gene. This step was analyzed with the BEDtools window command with parameter -w 250 000.

4.6.2 Effects on gene expression

The gene expression matrix of RNA-seq contained gene expression scores for each gene within each sample. This matrix collected scores for genes with TSS-annotated variants in TFBS within a window of 250 kbp upstream and downstream.

TFBSs with variants and without variants were separated into groups and were then tested with Wilcoxon rank-sum test to get p-values with scipy.stats.ranksums function (Virtanen et al., 2020). After that, Benjamini-Hochberg multiple testing correction was performed for p-values with statsmodels.stats.multitest.multipletests. The Benjamini-Hochberg procedure is the least strict of all correction methods and gives a good balance between statistically significant p-values and false positive findings (Noble, 2009; Silicon Genetics, 2003).

5. RESULTS

5.1 Variant calling

Variant calling was performed for 38 ATAC-seq samples and 9 WGS samples. The number of variants of ATAC-seq after all steps is presented in table 1.

***Table 1**. Number of variants in each sample of ATAC-seq*

Sample name	Number of variants
BPH_337	5 789
BPH_456	9 844
BPH_651	6 800
BPH_652	8 154
BPH_656	2 310
BPH_659	4 050
BPH_671	1 080
BPH_677	5 329
BPH_688	7 783
BPH_689	10 383
BPH_701	9 903
PC_12517	8 152
PC_14670	6 270
PC_15420	8 378
PC_15760	9 921
PC_17163	16 784
PC_17447	8 231
PC_18307	6 480
PC_19403	5 397
PC_470	5 490
PC_4980	8 776
PC_6174	3 695
PC_6488	7 061
PC_7875	5 489
PC_8131	6 318
PC_8438	2 785
PC_9324	13 250

CRPC_261	3 062
CRPC_278	1 742
CRPC_305	11 128
CRPC_348	6 515
CRPC_435	7 515
CRPC_489	10 153
CRPC_539	3 598
CRPC_541	11 103
CRPC_542	6 700
CRPC_543	13 647
CRPC_697	8 633

The minimum value of variants in ATAC-seq is 1080, and the maximum is 16784. The difference between maximum and minimum values means there is a great variation between samples.

The number of variants of WGS after variant calling is presented in table 2. There were only nine comparable samples to ATAC-seq.

Table 2. *Number of variants in each sample of WGS*

Sample name	**Number of variants in WGS**
BPH_651	3 425 648
BPH_659	3 578 855
BPH_671	3 566 705
BPH_688	3 096 262
BPH_701	3 406 436
CRPC_278	3 422 749
CRPC_305	3 505 107
CRPC_489	3 074 471
CRPC_697	3 135 004

The minimum value of variants in WGS is 3074471, and the maximum is 3578855. There is also variation between samples in the number of variants.

5.1.1 Open chromatin areas

For each sample type, open chromatin areas of ATAC-seq were analyzed. The number of peaks in each sample type is presented in table 3.

Table 3. *The number of open chromatin areas in each sample type*

The sample type	The range of peaks of open chromatin
BPH	39 507–77 018
PC	44 138–81 806
CRPC	47 686–72 859

The percentage of open chromatin areas with variants compared to all open chromatin areas is presented in figure 5. The first boxplot represents BPH percentages, the second one PC percentages, and the third one CRPC percentages.

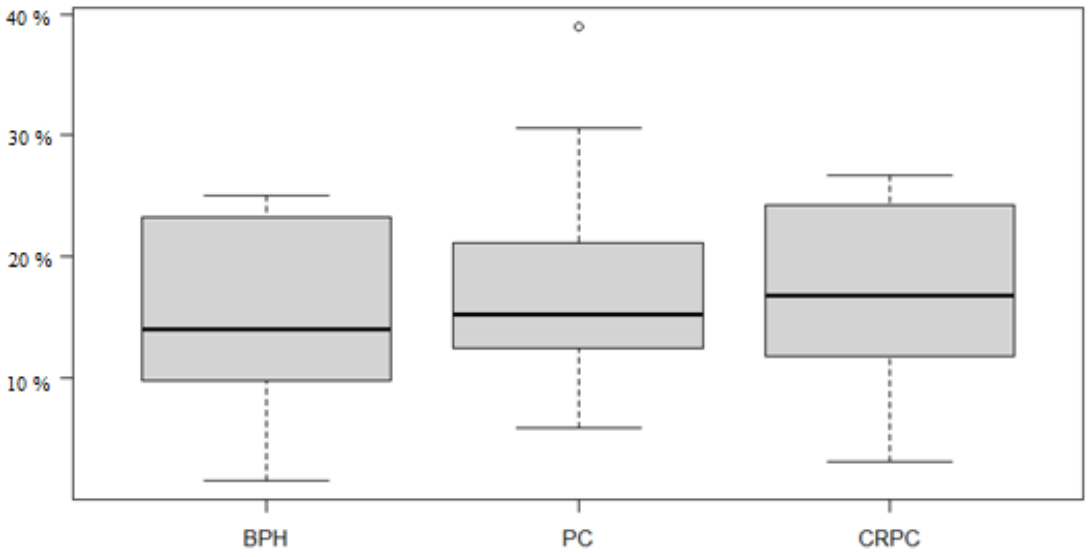

Figure 5. *Boxplot of percentages of open chromatin areas with a variant between sample types.*

The values of boxplots, minimum, 1st quarter, median, 3rd quarter, and maximum, are presented in table 4. The biggest median value is with CRPC samples, but the biggest

maximum percentage is with PC samples. The smallest minimum percentage is with BPH samples.

***Table 4**. The minimum, 1st quarter, median, 3rd quarter, and maximum values of boxplots between sample types*

Sample type	Minimum	1st quarter	Median	3rd quarter	Maximum
BPH	0.01510	0.09775	0.14040	0.23210	0.25000
PC	0.0582	0.1249	0.1519	0.2107	0.3899
CRPC	0.0309	0.1178	0.1678	0.2421	0.2668

5.2 Quality control

The quality of called variants was studied with different methods. These methods included analyzing VAF distribution and comparison of identical variants between ATAC-seq data and WGS data.

5.2.1 VAF distribution

Variant allele frequency distribution was produced with R for all samples. This histogram is presented in figure 6. The X-axis represents variants' frequency, and the Y-axis represents variant allele frequency distribution from 0 to 1.0. Most visible peaks can be seen near 100 % and 70 %. A smaller peak can also be spotted near 50 % of VAF.

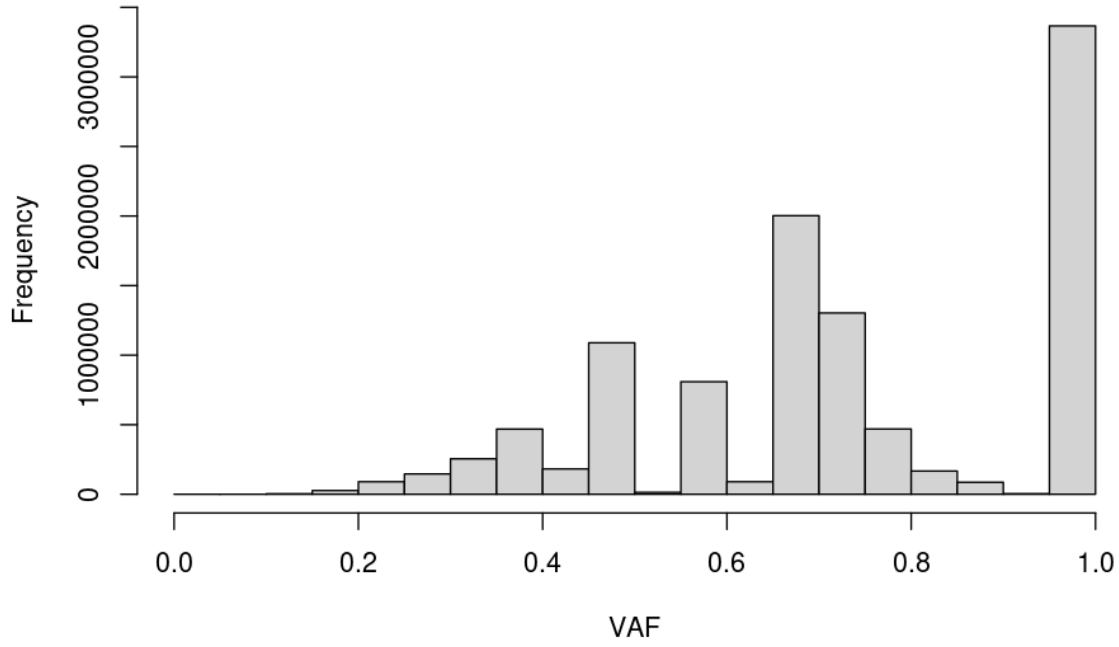

Figure 6. *Variant allele frequency distribution of all samples*

5.2.2 Variants in ATAC-seq data and WGS data

The number and percentage of variants in the same samples between ATAC-seq and WGS are represented in table 5. The number of variants in WGS data was much higher than in ATAC-seq data which was expected. The number of intersecting variants between datasets was compared to the number of variants in ATAC-seq to get the percentage of these variants.

Table 5*. Number of variants and the percentage of intersected variants*

Sample	Number of intersect variants	Percentage of intersect variants	Number of variants in ATAC-seq	Number of variants in WGS
BPH_651	6 215	91.4 %	6 800	3 425 648
BPH_659	3 861	95.3 %	4 050	3 578 855
BPH_671	978	90.6 %	10 80	3 566 705
BPH_688	6 756	86.8 %	7 783	3 096 262
BPH_701	9 250	93.4 %	9 903	3 406 436
CRPC_278	1 623	93.2 %	1 742	3 422 749
CRPC_305	10 437	93.8 %	11 128	3 505 107
CRPC_489	9 060	89.2 %	10 153	3 074 471
CRPC_697	7 722	89.4 %	8 633	3 135 004

The minimum percentage was 86.8 %, and the maximum was 95.3 %. The median percentage was 91.4 %. The distribution and occurrence of intersecting variants and variants in WGS are presented in figure 7. Green dots represent intersected variants, and red dots represent variants of WGS data. The plot consists of variants of all nine samples.

.

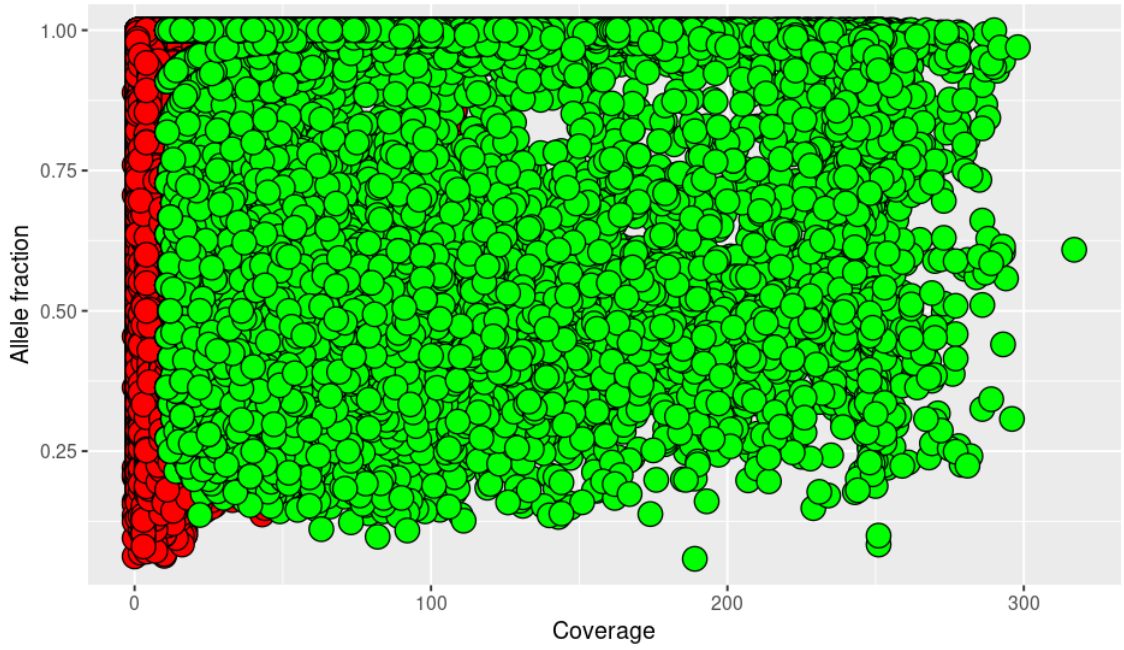

Figure 7. *Intersected variants as green plotted against variants of WGS data in red*

5.3 Transcription factor binding site analysis

Transcription factors with the highest number of binding sites in the Unibind database were collected to perform transcription factor binding site analysis. Each chosen transcription factor binding site was intersected with variants from variant calling and compared to the genes in the gene list.

5.3.1 The most common TFBSs in all cell lines

The 20 most common transcription factor binding sites in all cell lines are presented in figure 8. The last bar represents the amount of all other TFBSs. The five most common TFBSs are CTCF, AR, ESR1, FOXA1, and MYC. These transcription factors were used to study variants and gene expression.

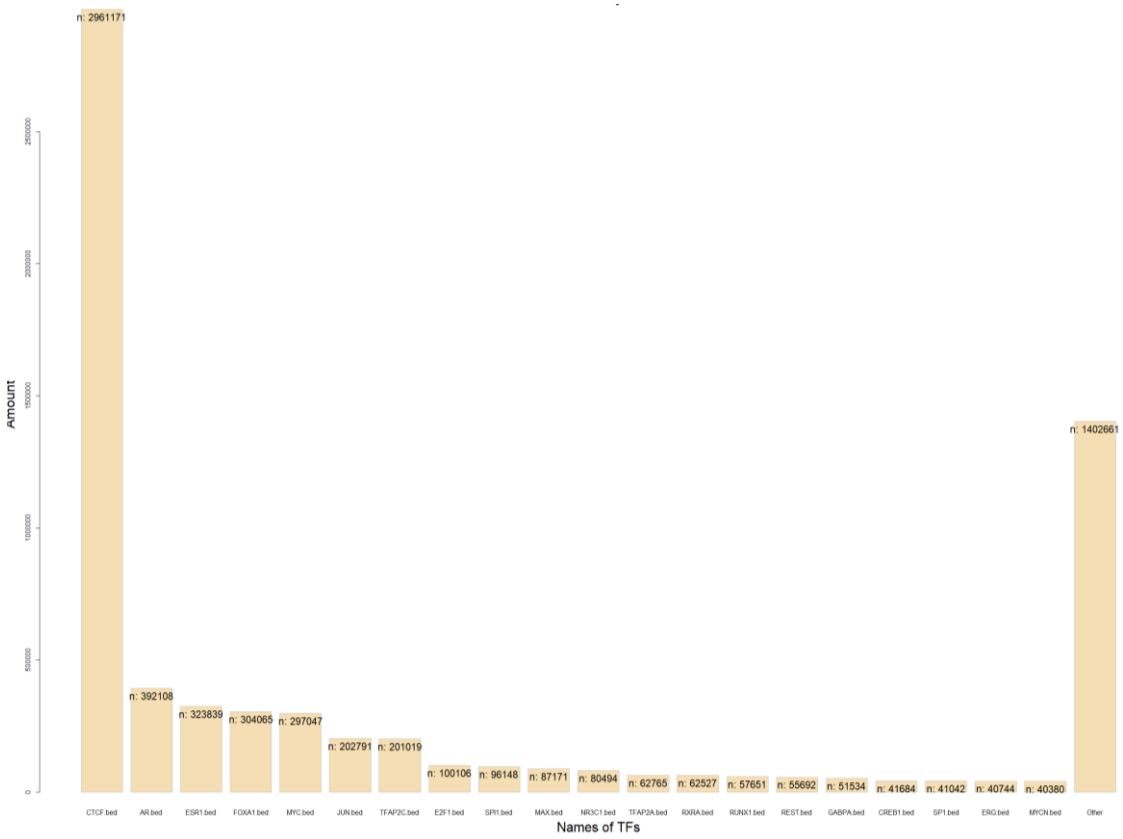

Figure 8. *20 most common transcription factors and their amount in all cell lines*

5.3.2 The most common TFs in prostate cell lines

The 20 most common transcription factor binding sites in prostate cell lines are presented in figure 9. The last bar represents the amount of all other TFBSs. The five most common TFBSs are AR, FOXA1, ERG, CTCF, and E2F1. These transcription factors were used for further studies.

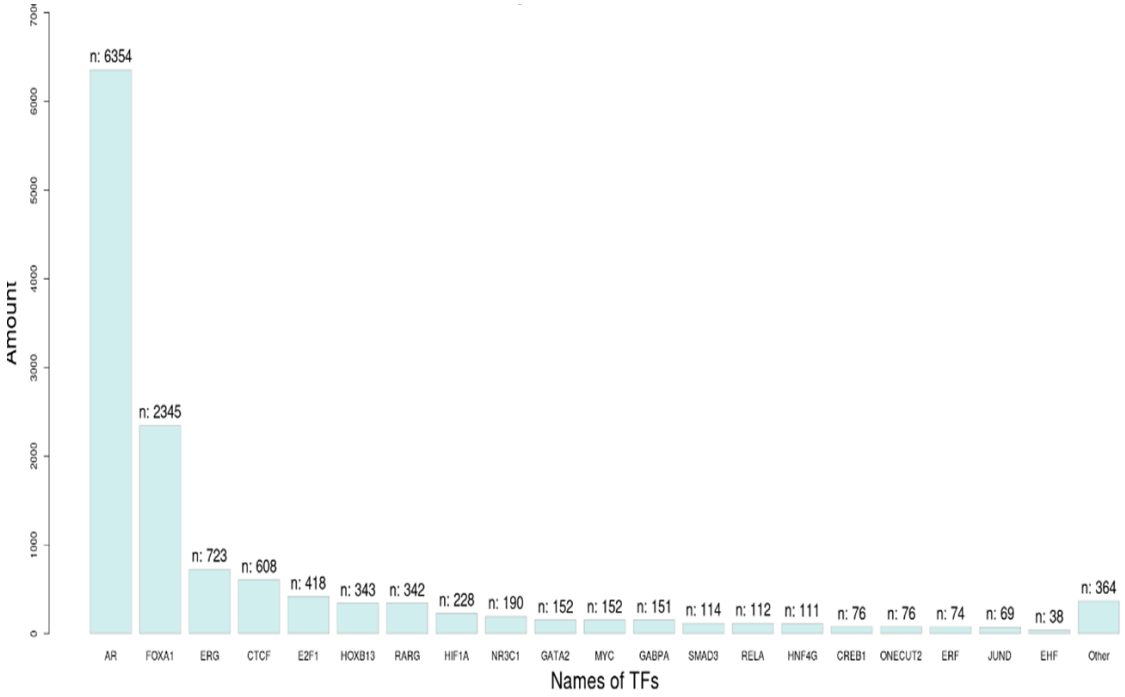

Figure 9. *20 most common transcription factors and their amount in prostate cell lines*

5.3.3 PWM scores of wild types

PWM scores of wild-type sequences were calculated for the five most common transcription factor binding sites in all cell and prostate cell lines. These transcription factors were chosen because their highest occurrence could indicate that their role in regulating gene expression could be more significant than those with less common transcription factors. After performing intersections between collected variants and binding site locations, similar binding sites from different cell lines were removed, so duplicate binding sites were not included. These binding sites were included if the same binding site and variant were discovered from different samples. The number of intersections between binding sites and variants is presented in table 6.

Table 6. *Number of intersections between TFBSs from figures 8 and 9 and ATAC-seq variants of variant calling from table 1*

Transcription factor binding site	Number of intersections with variants
CTCF (all)	18 405
AR (all)	25 471
ESR1 (all)	28 139
FOXA1 (all)	7 618
MYC (all)	31 016
AR (prostate)	23 129
FOXA1 (prostate)	6 315
ERG (prostate)	3 696
CTCF (prostate)	4 681
E2F1 (prostate)	9 131

The PWM score was calculated for each TFBS sequence with and without variants. The PWM scores were between -∞ and 30.

5.3.4 The effects of variants on PWM scores

After the PWM score was collected for wild-type and mutated sequence, the difference between these values was calculated. As an example, the distribution of differences for AR binding sites from all cell lines is presented in figure 10. Values of -∞ were excluded from histograms due to issues with plotting them.

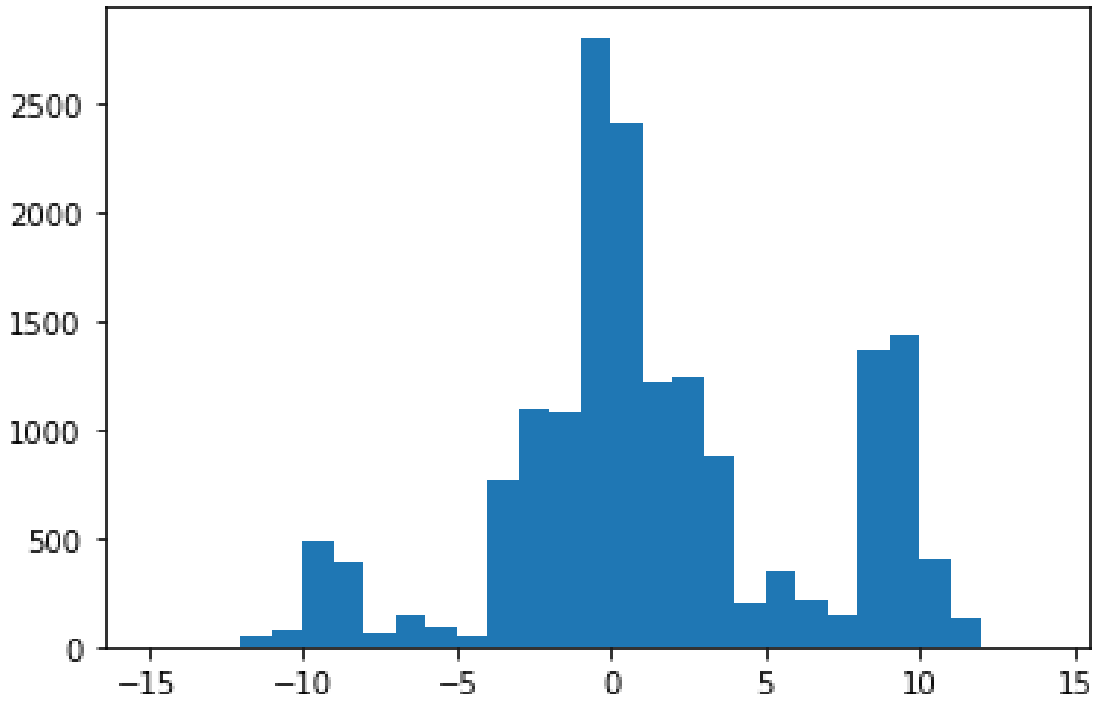

Figure 10. *Histogram of differences of wild type and mutated sequence of AR in all cell lines*

The most significant differences in better or worse binding affinity were the most interesting ones to study. The number of variants with a difference bigger than 5 or less than -5 of the PWM score between the wild-type and mutated type is presented in table 7.

Table 7. *Number of binding sites with a difference of more than 5 or less than -5 in PWM scores for each transcription factor binding site of figure 8 and 9, with an ATAC-seq variant from table 1.*

Transcription factor binding site	Number of binding sites with significant difference
CTCF (all)	1 601
AR (all)	5 355
ESR1 (all)	286
FOXA1 (all)	1 520
MYC (all)	6 218
AR (prostate)	5 022
FOXA1 (prostate)	1 300
ERG (prostate)	516
CTCF (prostate)	307
E2F1 (prostate)	1 240

5.3.5 Annotated variants in gene window

Variants of transcription factor binding sites with significant differences in PWM scores were annotated and the ones classified as transcription start sites were collected for further studies. The number of variants annotated as TSS is presented in table 8.

Table 8. *Number of ATAC-seq variants of TFBSs from table 7 annotated as TSS*

Transcription factor binding site	Number of ATAC-seq variants annotated as TSS
CTCF (all)	857
AR (all)	1 834
ESR1 (all)	177
FOXA1 (all)	254
MYC (all)	2 699
AR (prostate)	1 721
FOXA1 (prostate)	203
ERG (prostate)	189
CTCF (prostate)	91
E2F1 (prostate)	625

TFBSs with variants annotated as TSS were compared with genes of RNA-seq 250 kb upstream and downstream of gene locations. This step was done to understand which TFBS regulates each gene. The number of TFBSs within the gene window is presented in table 9.

Table 9. *Number of TFBSs with a variant annotated as TSS from table 8 in the gene window of 250 kb upstream and downstream*

Transcription factor binding site	Number of TFBSs within the gene window
CTCF (all)	8 316
AR (all)	20 841
ESR1 (all)	2 573
FOXA1 (all)	2 597
MYC (all)	29 155
AR (prostate)	19 390
FOXA1 (prostate)	1 978
ERG (prostate)	2 105
CTCF (prostate)	850
E2F1 (prostate)	7 090

5.3.6 Statistically significant genes

433 genes had a p-value less than 0.05 after Benjamini-Hochberg multiple testing correction. These genes are presented in supplementary table B.1. Out of these genes, 8 had a p-value less than 0.05 with three different TFs and 112 in 2 different TFs. Eight genes that had a statistical significance between samples with and without variants in 3 different TFs were *ZNF195* (AR, MYC, CTCF)*, RFXANK* (AR, FOXA1, CTCF), *PTPN3* (MYC, AR, FOXA1), *MAP4K5* (AR, FOXA1, CTCF), *KRIT1* (AR, FOXA1, ERG), *ITGAL* (AR, FOXA1, ERG), *DDX17* (AR, FOXA1, MYC), and *AHCY* (MYC, AR, FOXA1).

6. DISCUSSION

Transcription factors can upregulate and downregulate gene expression (Hombach et al., 2016). Transcription factor binding sites are located as cis-regulatory regions, such as promoters and enhancers. Transcription factor binding site affinity can be studied by prediction models, such as position weight matrices, or by experimental methods *in vitro* or *in vivo* (Castro-Mondragon et al., 2022). Understanding the meaning of variants to functional aspects of the genome is still a big challenge (The 1000 Genomes Project Consortium, 2012).

6.1 Variant calling

Variant calling was performed for ATAC-seq data and WGS data. Since the variants of ATAC-seq were a significant part of this thesis, the predicted results had to be reliable. The quality of variants was tested in many ways to ensure that filtering methods, read depth, and quality were well set.

The number of variants from ATAC-seq in different samples is presented in table 1, and the number of variants from WGS in different samples is presented in table 2. As we can see from table 1, the range of variants is between 1080 and 16784. This is quite a big difference since most of the variants in the human genome are inherited from parents. There is no apparent difference between the number of variants between sample types since the range is variable in the same sample types. To understand more why the range of variants is high, it would be good to link gene expression results to the variant count. This link could give an insight into why there are so many variants and the meaning of it.

The number of variants is much higher in table 2 compared to table 1. WGS data contains all variants, while ATAC-seq data contains only variants of open chromatin areas. According to studies performed with DNAse-seq and FAIRE-seq, open chromatin regions cover 1-2 % of the whole genome (Song et al., 2011). This percentage could explain the differences between the number of variants in ATAC-seq compared to WGS.

6.1.1 Reliability of variant calling

The quality control of variant calling was tested in multiple ways. This testing makes the quality of variants more reliable, but there can still be variants that are false positives. Performing variant calling with different filtering criteria could change the outcome, but it is also important not to filter out actual positive variants.

Variant allele frequency distribution was studied with all known SNPs of the human genome. These SNPs were intersected with open chromatin peaks of ATAC-seq. In figure 6, the histogram had peaks in almost all percentages, especially higher rates. This distribution means that some reads did not have clear homozygous or heterozygous SNPs in the loci, but more or less, reads had different alleles. It is usually thought that the higher the variant allele frequency, it is also more likely that the variant is from germline cells (Deleonardis et al., 2019). SNPs are germline mutations. Variant allele frequency can be different from expected, for example, due to tissue and tumor heterogeneity and copy number abnormalities (Deleonardis et al., 2019).

This thesis was only focusing on single nucleotide variants. Since insertions, deletions, and other changes in the genome were excluded, we may miss information that these would have provided about changes in gene expression.

6.1.2 Open chromatin areas

The variants were compared to open chromatin areas of ATAC-seq. Since only areas with openness scores over 5 were selected, some borderline cases may be unincluded and, therefore, some meaningful variants too.

6.2 Transcription factor binding sites

TFBSs were downloaded from the Unibind database. The five most common TFs in all cell and prostate cell lines were used for further studies. Since so many TFs were not studied at this stage, changes in those TFBSs are not noted. Therefore, we may miss some significant changes that could explain the progression of diseases. Since the thesis is limited work, some TFBSs had to be excluded.

Another way to choose used TFs could have been done by literature review. The TFs in this study could have also been the ones that are linked to prostate cancer cases. This way, different TFs could have been chosen for the analysis and changed the results.

6.2.1 The reliability of PWMs

TFBSs used in the study were obtained from the uniform processing of thousands of ChIP-seq data sets. This information makes their quality to be high in confidence. PWM matrices downloaded from Jaspar are manually curated and, therefore, can also be seen as high quality.

Other ways to study TF-DNA interactions include Markov and deep learning-based models (Castro-Mondragon et al., 2022). Using these models to study the binding affinity of TFs could change the results and how we interpret them. Another approach would be to use software designed for analyzing the impact of altered bindings of TFs on gene expression levels. One software developed for this purpose is TF2Exp, which has shown promising results in understanding the functional impact of variants (Shi et al., 2019).

6.2.2 Variants in transcription factor binding sites

After variant calling, variants were intersected with TFBSs downloaded from Unibind. The number of variants intersected with the binding site in different TFs varied between 3969 and 31016. The number of variants did not align with the number of TFBSs of each TF since the highest number of variants intersected with TFBSs was with MYC. CTCF had the highest number of binding sites overall. This number could be explained by the fact that CTCF and other TFs with an increased number of binding sites may have had many similar binding sites between different cell lines, which were then removed due to uniqueness in the analysis.

The PWM scores for these TFBSs were calculated for wild-type sequence and sequence with mutation added to it. The difference between these scores was calculated, and TFBSs with a difference bigger than 5 or less than -5 were collected. This means that the most significant changes done by variant to the binding affinity were collected. This step was done because we were interested in binding sites with possible changes due to variants. From histograms of differences, for example, in figure 10, we can see the tails on both ends of the histogram. These tails start approximately from -5 and 5, which

supports these limits. If these limits have been changed, stricter or looser, it could change a lot which TFBSs would have been chosen. But according to the histogram in figure 10, these limits seem to be a good fit.

Variants were annotated with Homer Annotatepeaks.pl. Variants that were annotated as transcription start sites were included. This inclusion means that variants of TFBSs used in gene window studies should be in transcription start sites and, therefore, have that functionality. If the annotation was not accurate enough, some TFBSs could be excluded from the analysis due to their variants being inaccurately annotated. Homer is a widely used software, so it can be assumed reliable.

6.3 Differentially expressed genes

There were altogether 433 genes that were differentially expressed between samples with a variant in regulatory and those without. A threshold of the p-value of the Wilcoxon rank-sum test and multiple testing correction was set to be 0.05 since it is a commonly accepted threshold. This p-value means that the result is consistent with the null hypothesis. Confirming the null hypothesis would require more studies. At least some of the genes had been linked to prostate cancer cases in previous studies. This finding could indicate that at least some genes can have expression changes.

6.3.1 Gene window size

The gene window size of BEDtools window command was set as 250 000 bp upstream and downstream of the gene. This window size is quite extensive, so all regulatory elements are collected with high probability. The problem with a big window is that there are also collected TFBSs that do not regulate gene expression in some parts. To avoid this incident, regulatory areas of each gene should be studied and, thereby, choose the window size for each gene independently. This task would require a lot of time and resources.

6.3.2 Genes with a significant p-value

The genes with significant p-value present in more than one different TFs are *ZNF195, RFXANK*, *PTPN3, MAP4K5, KRIT1, ITGAL, DDX17,* and *AHCY.*

According to studies, the minor alleles of rs2073917 and rs3764322 in *ITGAL* have been linked to a more considerable risk of death in men with CRPC. This risk can be explained by the fact that *ITGAL*, together with *ABL2* and *SEMA4D*, regulates T-cell motility, antigen surveillance, and T-helper cell activity, which affect cellular and humoral immunity. (Xie et al., 2019) *ITGAL* has also been associated with having an essential role in gene pathways with the number of positive lymph nodes, meaning the number of metastasized nodes in the patient's body (Zhao et al., 2019).

According to Lin et al., circular RNA *DDX17* can suppress PC cell mobility, proliferation, and invasion (Lin et al., 2020). DDX17 also has a role as a tumor suppressor in colorectal cancer (Lin et al., 2020).

In a study by Uchiyama et al., a small molecule compound, Aristeromycin (a derivative of 3-deazaneplanocin A (DZNeP)), was identified from hormone-resistant prostate cancer cells. The targeted function of aristeromycin is the inhibition of *AHCY*, which act as a catalyzer in different reactions. The inhibition of *AHCY* can lead to decreased growth of prostate cancer cells. (Uchiyama et al., 2017)

These previous studies state that *ITGAL*, *DDX17*, and *AHCY* have been linked to prostate cancer cases. Since their functionality has been studied, variants in their regulatory area could be one explaining factor for their role in prostate cancer. This result alone does not explain the role of variants, so more studies are required.

6.4 Future

More studies are needed to both validate and take forward the study. Results should be validated in the laboratory. For example, this validation could be done with STARR-seq (self-transcribing active regulatory region sequencing) data. STARR-seq can directly and quantitatively estimate enhancer activity (Arnold et al., 2013). This step could provide more insights into the activity of regulatory areas and how they are distributed in our ATAC-seq data.

This study could be performed for a more comprehensive set of transcription factors to gain more knowledge about the effects of variants on gene expression. A more enormous number of transcription factors could give more insights into what changes are actually significant.

Another thing to study in the future is the differences between sample types. Since the ATAC-seq dataset consists of BPH, PC, and CRPC samples, it could provide new information if the sample types were studied separately.

7. CONCLUSIONS

During this thesis, variant calling was performed for ATAC-seq data of 38 prostate hyperplasia or prostate cancer samples. Called variants were intersected with TFBSs from Unibind to study the effects of variants on binding affinity via position weight matrices. TFBSs that had the most significant change of PWM scores between the wild-type and the mutated type and had variants annotated as transcription start sites were compared to gene expression scores of the same samples. Wilcoxon rank-sum test and multiple testing corrections were performed for each gene to study the significance of differences.

The first aim of this Master's thesis was to perform variant calling with suitable parameters to ATAC-seq. The results revealed a wide range of variants between samples, from 1080 to 11128. Multiple methods were used to study the quality of these variants. These methods included intersecting variants of WGS data to these ATAC-seq variants and investigating variant allele frequency of known SNPs. Especially comparing variants of WGS and ATAC-seq stated that variants of ATAC-seq have good quality.

The second aim was to discover common variants within different TFBSs. These variants were found by intersecting each variant with the binding site. The range of variants intersected with TFBSs varied between 3096 and 31016, with MYC having the highest count and ERG the lowest.

The third aim was to find out how variants affect the ability of TF to bind to its binding site. PWM score was calculated for each binding site with variants for wild-type and mutated sequence, and the difference was calculated. The highest and lowest peaks of differences were collected, which meant the number of binding sites with a variant to be between 286 and 6218. The highest count was still with MYC, but the lowest was ESR1. In all TFs, many binding sites were unincluded.

The main objective was to determine if and which variants in TFBS affect gene expression close to these regulatory areas with variants. Binding sites with variants annotated as transcription start sites were intersected with the regulatory areas of each gene. The number of binding with annotated variants in the gene window area of 250 kb upstream and downstream was between 850 and 29155. MYC still had the highest number of binding sites left, while CTCF had the lowest number. The genes with these binding sites

in their regulatory areas were studied with Wilcoxon rank-sum test to find the genes with possible differences between samples with and without the variant. Our final result consisted of 433 genes whose, according to the p-value from the Wilcoxon rank-sum test and Benjamini-Hochberg multiple testing correction, gene expression could be changed due to the variants of TFBSs.

According to our results, it was possible to find which variants in TFBSs could be the explaining factor in gene expression. With these analyses, it is impossible to state that these variants affected the expression of genes. More studies, such as using STARR-sequencing, are required to confirm the results. Still, the generated lists of genes and variants can contain some vital information about the regulation of genes.

REFERENCES

Abou-Ouf, H., Zhao, L., & Bismar, T. A. (2016). ERG expression in prostate cancer: biological relevance and clinical implication. *Journal of Cancer Research and Clinical Oncology*, *142*(8), 1781–1793. https://doi.org/10.1007/s00432-015-2096-x

Andersson, R. (2014). Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays*, *37*, 314–323. https://doi.org/10.1002/bies.201400162

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., & Stark, A. (2013). Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science*, *339*(6123), 1074–1077. https://doi.org/10.1126/science.1232542

Bhagwat, A. S., & Vakoc, C. R. (2015). Targeting Transcription Factors in Cancer. *Trends in Cancer*, *1*(1), 53–65. https://doi.org/10.1016/j.trecan.2015.07.001

Børsting, C., & Morling, N. (2013). Single-Nucleotide Polymorphisms. *Encyclopedia of Forensic Sciences: Second Edition*, 233–238. https://doi.org/10.1016/B978-0-12-382165-2.00042-8

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. https://doi.org/10.3322/caac.21492

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, *109*, 21.29.1-21.29.9. https://doi.org/10.1002/0471142727.mb2129s109

Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Perez, N. M., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., … Mathelier, A. (2022). JASPAR 2022: The 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, *50*(D1), D165–D173. https://doi.org/10.1093/nar/gkab1113

Chen, X., Ma, J., Xu, C., Wang, L., Yao, Y., Wang, X., Tong Zi, ·, Cuidong Bian, ·, Denglong Wu, ·, & Wu, · Gang. (2022). Identification of hub genes predicting the development of prostate cancer from benign prostate hyperplasia and analyzing their clinical value in prostate cancer by bioinformatic analysis. *Discover Oncology*, *13*, 54. https://doi.org/10.1007/s12672-022-00508-y

Chernodub, M., Hu, S., & Niemi, A. J. (2010). Topological solitons and folded proteins. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *82*(1). https://doi.org/10.1103/PhysRevE.82.011916

Claussnitzer, M., Dankel, S. N., Kim, K., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I. S., Beaudry, J. L., Puviindran, V., Abdennur, N. A., Liu, J., Svensson, P., Hsu, Y., Drucker, D. J., Mellgren, G., Hui, C., Hauner, H., & Kellis, M. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *The New England Journal of Medicine*, *373*(10), 895–907. https://doi.org/10.1056/NEJMoa1502214

Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., Wu, B., Kathiria, A., Cho, S. W., Mumbach, M. R., Carter, A. C., Kasowski, M., Orloff, L. A., Risca, V. I., Kundaje, A., Khavari, P. A., … Chang, H. Y. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods*, *14*, 959–962. https://doi.org/10.1038/nmeth.4396

Culig, Z., & Santer, F. R. (2014). Androgen receptor signaling in prostate cancer. *Cancer and Metastasis Reviews*, *33*(2–3), 413–427. https://doi.org/10.1007/s10555-013-9474-0

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Daniel H. Gonzalez. (2015). Introduction to Transcription Factor Structure and Function. In *Plant Transcription Factors : Evolutionary, Structural and Functional Aspects* (pp. 3–11). Elsevier Science & Technology.

Davey, R. A., & Grossmann, M. (2016). Androgen Receptor Structure, Function and Biology: From Bench to Bedside. *Androgen Receptor Biology Clin Biochem Rev*, *37*(1), 3–15.

Deleonardis, K., Hogan, L., Cannistra, S. A., Rangachari, D., & Tung, N. (2019). When Should Tumor Genomic Profiling Prompt Consideration of Germline Testing? *J Oncol Pract*, *15*(9), 465–473. https://doi.org/10.

Dodd, A. W., Syddall, C. M., & Loughlin, J. (2013). A rare variant in the osteoarthritis-associated locus GDF5 is functional and reveals a site that can be manipulated to modulate GDF5 expression. *European Journal of Human Genetics*, *21*, 517–521. https://doi.org/10.1038/ejhg.2012.197

Doni Jayavelu, N., Jajodia, A., Mishra, A., & David Hawkins, R. (2020). Candidate silencer elements for the human and mouse genomes. *Nature Communications*, *11*, 1061. https://doi.org/10.1038/s41467-020-14853-5

Duffy, M. J., O'Grady, S., Tang, M., & Crown, J. (2021). MYC as a target for cancer treatment. *Cancer Treatment Reviews*, *94*, 102154. https://doi.org/10.1016/J.CTRV.2021.102154

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., de Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *BIOINFORMATICS APPLICATIONS NOTE*, *21*(16), 3439–3440. https://doi.org/10.1093/bioinformatics/bti525

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, *4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

Farrar, S. (2019, February 26). *High-throughput DNA Sequencing Techniques*.

Finotello, F., & di Camillo, B. (2015). Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis. *Briefings in Functional Genomics*, *14*(2), 130–142. https://doi.org/10.1093/bfgp/elu035

Fostier, J. (2020). BLAMM: BLAS-based algorithm for finding position weight matrix occurrences in DNA sequences on CPUs and GPUs. *BMC Bioinformatics*, *21*. https://doi.org/10.1186/s12859-020-3348-6

Ganguly, P. (2022a, September 2). *Transcription*. National Human Genome Research Institution.

Ganguly, P. (2022b, September 2). *Translation*. National Human Genome Research Institution.

Gautam, S. S., KC, R., Leong, K. W., mac Aogáin, M., & O'Toole, R. F. (2019). A step-by-step beginner's protocol for whole genome sequencing of human bacterial pathogens. *Journal of Biological Methods*, *6*(1), e110. https://doi.org/10.14440/jbm.2019.276

Gheorghe, M., Sandve, G. K., Khan, A., Chèneby, J., Ballester, B., & Mathelier, A. (2019). A map of direct TF-DNA interactions in the human genome. *Nucleic Acids Research*, *47*(4), e21. https://doi.org/10.1093/nar/gky1210

Grishin, D., & Gusev, A. (2022). Allelic imbalance of chromatin accessibility in cancer identifies candidate causal risk variants and their mechanisms. *Nat Genet.*, *54*(6), 837–849. https://doi.org/10.1038/s41588-022-01075-2

Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. In *Nature Reviews Molecular Cell Biology* (Vol. 19, Issue 10, pp. 621–637). Nature Publishing Group. https://doi.org/10.1038/s41580-018-0028-8

Han, G., Lu, C., Guo, J., Qiao, Z., Sui, N., Qiu, N., & Wang, B. (2020). C2H2 Zinc Finger Proteins: Master Regulators of Abiotic Stress Responses in Plants. *Front. Plant Sci*, *11*, 115. https://doi.org/10.3389/fpls.2020.00115

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, *38*(4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Hogg, R. v., Tanis, E. A., & Zimmerman, D. L. (2013). *Probability and statistical inference*. Pearson.

Hombach, D., Schwarz, J. M., Robinson, P. N., Schuelke, M., & Seelow, D. (2016). A systematic, large-scale comparison of transcription factor binding site models. *BMC Genomics*, *17*, 388. https://doi.org/10.1186/s12864-016-2729-8

Imamura, Y., & Sadar, M. D. (2016). Androgen receptor targeted therapies in castration-resistant prostate cancer: Bench to clinic. In *International Journal of Urology* (Vol. 23, Issue 8, pp. 654–665). Blackwell Publishing. https://doi.org/10.1111/iju.13137

Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A. E., Ristolainen, H., Hänninen, U. A., Cajuso, T., Kondelin, J., Tanskanen, T., Mecklin, J. P., Järvinen, H., Renkonen-Sinisalo, L., Lepistö, A., Kaasinen, E., Kilpivaara, O., Tuupanen, S., … Aaltonen, L. A. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature Genetics*, *47*, 818–821. https://doi.org/10.1038/ng.3335

Kilcoyne, A., O'Connor, D., & Ambery, P. (2013). Parametric and non-parametric tests. In A. Kilcoyne, P. Ambery, & D. O'Connor (Eds.), *Pharmaceutical Medicine* (pp. 274–276). Oxford University Press.

Klug, A. (1999). Zinc Finger Peptides for the Regulation of Gene Expression. *J. Mol. Biol.*, *293*, 215–218. http://www.idealibrary.com

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine*, *12*(1), 91. https://doi.org/10.1186/s13073-020-00791-w

Kodama, H., Koie, T., Oikawa, · Masaaki, Narita, T., Tanaka, T., Noro, D., Iwamura, H., Tobisawa, Y., Yoneyama, T., Hashimoto, Y., & Chikara Ohyama, ·. (2020). Castra-

tion-resistant prostate cancer without metastasis at presentation may achieve cancer-specific survival in patients who underwent prior radical prostatectomy. *International Urology and Nephrology*, *52*, 671–679. https://doi.org/10.1007/s11255-019-02339-3

Komura, K., Sweeney, C. J., Inamoto, T., Ibuki, N., Azuma, H., & Kantoff, P. W. (2018). Current treatment strategies for advanced prostate cancer. *International Journal of Urology*, *25*(3), 220–231. https://doi.org/10.1111/iju.13512

Krylov, D., & Vinson, C. R. (2001). Leucine Zipper. In *Encyclopedia of Life Sciences*. www.els.net

Labbé, D. P., & Brown, M. (2018). Transcriptional regulation in prostate cancer. *Cold Spring Harbor Perspectives in Medicine*, *8*(11), a030437. https://doi.org/10.1101/CSHPERSPECT.A030437

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

Lin, Q., Cai, J., & Wang, Q. Q. (2020). The Significance of Circular RNA DDX17 in Prostate Cancer. *BioMed Research International*, *2020*. https://doi.org/10.1155/2020/1878431

Massarat, A. R., Sen, A., Jaureguy, J., Tyndale, S. T., Fu, Y., Erikson, G., & McVicker, G. (2021). Discovering single nucleotide variants and indels from bulk and single-cell ATAC-seq. *Nucleic Acids Research*, *49*(14), 7986–7994. https://doi.org/10.1093/nar/gkab621

Mckinney, W. (2010). *Data Structures for Statistical Computing in Python*.

Miah, S., & Catto, J. (2014). BPH and prostate cancer risk. *Indian Journal of Urology*, *30*(2), 214–218. https://doi.org/10.4103/0970-1591.126909

Miles, B., & Tadi, P. (2022). Genetics, Somatic Mutation. In *StatPearls*. StatPearls Publishing.

Motrich, R. D., Salazar, F. C., Breser, M. L., Mackern-Oberti, J. P., Godoy, G. J., Olivera, C., Paira, D. A., & Rivero, V. E. (2018). Implications of prostate inflammation on male fertility. In *Andrologia* (Vol. 50, Issue 11, p. e13093). Blackwell Publishing Ltd. https://doi.org/10.1111/and.13093

Newkirk, K. M., Brannick, E. M., & Kusewitt, D. F. (2017). Chapter 6 - Neoplasia and Tumor Biology. In *Pathologic Basis of Veterinary Disease Expert Consult* (pp. 286–321). Elsevier Inc. https://doi.org/10.1016/B978-0-323-35775-3.00006-0

Nishida, K., Frith, M. C., & Nakai, K. (2008). Pseudocounts for transcription factor binding sites. *Nucleic Acids Research*, *37*(3), 939–944. https://doi.org/10.1093/nar/gkn1019

NIST/SEMATECH. (2012). What are statistical tests? In *e-Handbook of Statistical Methods*.

Noble, W. S. (2009). How does multiple testing correction work? *Nature Biotechnology*, *27*(12). https://doi.org/10.1038/nbt1209-1135

Orgogozo, V., Peluffo, A. E., & Morizot, B. (2016). The " Mendelian Gene " and the " Molecular Gene " : Two Relevant Concepts of Genetic Units. *Current Topics in Developmental Biology*, *119*, 1–26.

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: five essential questions. *Nat Rev Genet.*, *14*(4), 288–295. https://doi.org/10.1038/nrg3458

Pitkäniemi, J., Malila, N., Tanskanen, T., Degerlund, H., Heikkinen, S., & Seppä, K. (2021). Cancer in Finland 2019. *Cancer Society of Finland Publication No. 98*.

Prjibelski, A. D., Korobeynikov, A. I., & Lapidus, A. L. (2019). Sequence Analysis. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *1–3*, 292–322. https://doi.org/10.1016/B978-0-12-809633-8.20106-4

Puig, R. R., Boddie, P., Khan, A., Castro-Mondragon, J. A., & Mathelier, A. (2021). UniBind: maps of high-confidence direct TF-DNA interactions across nine species. *BMC Genomics*, *22*(1). https://doi.org/10.1186/s12864-021-07760-6

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

Roy, S., & Kundu, T. K. (2021). Chemical principles of DNA sequence recognition and gene regulation. In S. Roy & T. K. Kundu (Eds.), *Chemical Biology of the Genome* (pp. 171–223). Academic Press.

Shah, U. S., & Getzenberg, R. H. (2004). Fingerprinting the Diseased Prostate: Associations Between BPH and Prostate Cancer. *Journal of Cellular Biochemistry*, *91*, 161–169. https://doi.org/10.1002/jcb.10739

Shi, W., Fornes, O., & Wasserman, W. W. (2019). Gene expression models based on transcription factor binding events confer insight into functional cis-regulatory variants. *Bioinformatics*, *35*(15), 2610–2617. https://doi.org/10.1093/bioinformatics/bty992

Silicon Genetics. (2003). *Multiple Testing Corrections*.

Smigielski, E. M., Sirotkin, K., Ward, M., & Sherry, S. T. (2000). dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Research*, *28*(1), 352–355. http://www.ncbi.nlm.nih.gov/SNP/get_html.

Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B. K., Sheffield, N. C., Gräf, S., Huss, M., Keefe, D., Liu, Z., London, D., McDaniell, R. M., Shibata, Y., Showers, K. A., Simon, J. M., Vales, T., Wang, T., Winter, D., … Furey, T. S. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research*, *21*(10), 1757–1767. https://doi.org/10.1101/gr.121541.111

Spans, L., Clinckemalie, L., Helsen, C., Vanderschueren, D., Boonen, S., Lerut, E., Joniau, S., & Claessens, F. (2013). The genomic landscape of prostate cancer. In *International Journal of Molecular Sciences* (Vol. 14, Issue 6, pp. 10822–10851). https://doi.org/10.3390/ijms140610822

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, *20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

Strom, S. P. (2016). Current practices and guidelines for clinical next-generation sequencing oncology testing. *Cancer Biology and Medicine*, *13*(1), 3–11. https://doi.org/10.28092/j.issn.2095-3941.2016.0004

Sun, Y., Miao, N., & Sun, T. (2019). Detect accessible chromatin using ATAC-sequencing, from principle to applications. *Hereditas*, *156*, 29. https://doi.org/10.1186/s41065-019-0105-9

Taguchi, Y. H. (2018). Comparative transcriptomics analysis. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *1–3*, 814–818. https://doi.org/10.1016/B978-0-12-809633-8.20163-5

Teng, M., Zhou, S., Cai, C., Lupien, M., & Hansen He, H. (2021). REVIEW Pioneer of prostate cancer: past, present and the future of FOXA1. *Protein Cell*, *12*(1), 29–38. https://doi.org/10.1007/s13238-020-00786-8

The 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*, 56–65. https://doi.org/10.1038/nature11632

The pandas development team. (2020). *pandas-dev/pandas: Pandas*. Zenodo.

Tseng, C.-C., Wong, M.-C., Liao, W.-T., Chen, C.-J., Lee, S.-C., Yen, J.-H., Chang, S.-J., Liao, W.-T. ;, Chen, C.-J. ;, Lee, S.-C. ;, Yen, J.-H. ;, & Chang, S.-J. (2021). Genetic Variants in Transcription Factor Binding Sites in Humans: Triggered by Natural Selection and Triggers of Diseases. *Int. J. Mol. Sci*, *22*(8), 4187. https://doi.org/10.3390/ijms22084187

Uchiyama, N., Tanaka, Y., & Kawamoto, T. (2017). Aristeromycin and DZNeP cause growth inhibition of prostate cancer via induction of mir-26a. *European Journal of Pharmacology*, *812*, 138–146. https://doi.org/10.1016/j.ejphar.2017.07.023

van El, C. G., Cornel, M. C., Borry, P., Hastings, R. J., Fellmann, F., Hodgson, S. v., Howard, H. C., Cambon-Thomsen, A., Knoppers, B. M., Meijers-Heijboer, H., Scheffer, H., Tranebjaerg, L., Dondorp, W., & de Wert, G. M. W. R. (2013). Whole-genome sequencing in health care. *European Journal of Human Genetics*, *21*(6), 580–584. https://doi.org/10.1038/ejhg.2013.46

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., & Luscombe, N. M. (2009). A census of human transcription factors: Function, expression and evolution. *Nature Reviews Genetics*, *10*(4), 252–263. https://doi.org/10.1038/nrg2538

Vinson, C., Chatterjee, R., & Fitzgerald, P. (2011). Transcription factor binding sites and other features in human and drosophila proximal promoters. *Subcellular Biochemistry*, *52*, 205–222. https://doi.org/10.1007/978-90-481-9069-0_10

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., J Nelson, A. R., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang, G., Zhao, D., Spring, D. J., & Depinho, R. A. (2018). Genetics and biology of prostate cancer. *Genes and Development*, *32*(17–18), 1105–1140. https://doi.org/10.1101/gad.315739.118

Wang, S., Wu, S., Meng, Q., Li, X., Zhang, J., Chen, R., & Wang, M. (2015). FAS rs2234767 and rs1800682 polymorphisms jointly contributed to risk of colorectal cancer by affecting SP1/STAT1 complex recruitment to chromatin OPEN. *Sci Rep*, *6*, 19229. https://doi.org/10.1038/srep19229

Wang, Y.-M., Liu, Z.-W., Guo, J.-B., Wang, X.-F., Zhao, X.-X., & Zheng, X. (2013). ESR1 Gene Polymorphisms and Prostate Cancer Risk: A HuGE Review and Meta-Analysis. *PLoS ONE*, *8*(6), e66999. https://doi.org/10.1371/journal.pone.0066999

Wang, Z., Wang, Y., Zhang, J., Hu, Q., Zhi, F., Zhang, S., Mao, D., Zhang, Y., & Liang, H. (2017). Significance of the TMPRSS2: ERG gene fusion in prostate cancer. *Molecular Medicine Reports*, *16*(4), 5450–5458. https://doi.org/10.3892/mmr.2017.7281

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Xie, W., Stopsack, K. H., Drouin, S. J., Fu, H., Pomerantz, M. M., Mucci, L. A., Lee, G. S. M., & Kantoff, P. W. (2019). Association of genetic variation of the six gene prognostic model for castration-resistant prostate cancer with survival. *Prostate*, *79*(1), 73–80. https://doi.org/10.1002/pros.23712

Yan, F., Powell, D. R., Curtis, D. J., & Wong, N. C. (2020). From reads to insight: A hitchhiker's guide to ATAC-seq data analysis. *Genome Biology*, *21*, 22. https://doi.org/10.1186/s13059-020-1929-3

Zhao, X., Lei, Y., Li, G., Cheng, Y., Yang, H., Xie, L., Long, H., & Jiang, R. (2019). Integrative analysis of cancer driver genes in prostate adenocarcinoma. *Molecular Medicine Reports*, *19*(4), 2707–2715. https://doi.org/10.3892/mmr.2019.9902

Zou, H., Wu, L. X., Tan, L., Shang, F. F., & Zhou, H. H. (2020). Significance of Single-Nucleotide Variants in Long Intergenic Non-protein Coding RNAs. In *Frontiers in Cell and Developmental Biology* (Vol. 8, p. 347). Frontiers Media S.A. https://doi.org/10.3389/fcell.2020.00347

Zou, Q., Xu, X., Basu, M., Quoc, N., Le, K., Yeh, H.-Y., Kien, E., Yapp, Y., & Nagasundaram, N. (2019). Classifying Promoters by Interpreting the Hidden Information of DNA Sequences via Deep Learning and Combination of Continuous FastText N-Grams. *Frontiers in Bioengineering and Biotechnology | Www.Frontiersin.Org*, *7*, 305. https://doi.org/10.3389/fbioe.2019.00305

APPENDIX A: CODE

```
## This code is for AR in all cell lines, but a similar code can be
implemented for all TFs by changing the names and PWM from Jaspar

## Standard imports
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import os
import re
import statistics
import scipy
from scipy import stats
from Bio import motifs


def reverse_complement(s):
     """
     reverse_complement takes a reverse strand (-) and reverses it to
     be the same direction as the forward strand
     :param s: reverse strand
     """
     return ''.join([{'A':'T', 'T':'A', 'G':'C', 'C':'G',
'NaN':''}[c] for c in s])[::-1]


def calculate_score(sequence, pwm):
     """
     calculate_score goes through each base of the sequence, adds its
     score to the variable score, and finally returns the score of
     the whole sequence
     :param sequence: The sequence of TFBS
     :param pwm: Position weight matrix for the particular TFBS
     """
     score = 0
     for i, base in enumerate(sequence):
        ## i is the position of the base
        ## base is the base
           score += pwm.loc[base, i] # equal to: score = score +
           pwm.loc[]
     return score


def wilcoxon(Ensembl, samples):
     """
     Wilcoxon checks if Ensembl ID and sample can be found from the
     gene expression matrix, then goes through samples with variants
     and saves their gene expression score to variable GE_var. The
     same step is done for samples without the variant. Finally,
     their median values are calculated, and both groups are tested
```

```
    with Wilcoxon rank-sum test. The median values and statistics of
    the test are returned.
    :param Ensembl: Ensembl ID of gene
    :param samples: a list of samples with variant
    """
    GE_var = []
    GE_wild = []
    if (Ensembl in Expression_matrix.index):
       for sample in Expression_matrix.columns:
           if (sample == "BPH_337"):
               continue
           if (sample not in samples):
               value = Expression_matrix.loc[Ensembl, sample]
               GE_wild.append(value)
           else:
               value = Expression_matrix.loc[Ensembl, sample]
               GE_var.append(value)
    else:
       return None
    median_var = statistics.median(GE_var)
    median_wild = statistics.median(GE_wild)
    pval = scipy.stats.ranksums(GE_var, GE_wild)

    return pval, median_var, median_wild


def gene_name(Ensembl, index):
    """
    gene_name checks if Ensembl ID can be found from the gene ex-
    pression matrix and then returns the gene name of the ID.
    :param Ensembl: Ensembl ID of gene
    :param index: an index of Ensembl ID
    """
    Gene_matrix = pd.read_csv('RNAseq_trizolNormalized_all-
    gene_names.tsv', sep='\t')
    if (Ensembl in Expression_matrix.index):
       genename = Gene_matrix['gene_name'].loc[index]
    else:
       return None
    return genename


directory = 'Intersects_per_TF_copy'

df = pd.DataFrame()

## Going through all files in directory and adding them to the data-
frame
for filename in os.listdir(directory):
    f = os.path.join(directory, filename)
    ## Checking if it is a file
    if os.path.isfile(f):
```

```
        if df.empty:
            df = pd.read_csv(f, sep='\t', header=None)
            df['Filename'] = filename
        else:
            df2 = pd.read_csv(f, sep='\t', header=None)
            df2['Filename'] = filename
            df = df.append(df2)
            df2 = pd.DataFrame()

## Splitting a column with TF name and sequence to two new columns
df['TF'] = df[3].apply(lambda x: x.split('_')[0])
df['pre_TF_motif'] = df[3].apply(lambda x: x.split('_')[1])

## Saving the sample name of for each sequence
df['Filename'] = df['Filename'].apply(lambda x: x.split('.')[0])
df['TF_motif'] = ''

## Reverse complementing the sequence on the reverse/negative strand
df['TF_motif'][df[5] == '+'] = df['pre_TF_motif'][df[5] == '+']
df['TF_motif'][df[5] == '-'] = df['pre_TF_motif'][df[5] == '-']
df['TF_motif'][df[5] == '-'] = df['TF_motif'][df[5] == '-'].apply(re-
verse_complement)

## Collecting all TFBSs named AR
AR = df.loc[df['TF'] == 'AR', ]

## Modifying the Series of TF motifs to matrix
motifs_tmp = AR['TF_motif'].apply(lambda x: pd.Series(list(x)))
motifs_tmp.columns = ['M%s' %v for v in list(motifs_tmp.columns)]
AR_new = pd.concat([AR, motifs_tmp], axis=1)

## Downloading the PWM of AR from Jaspar
with open('MA0007.3.jaspar') as handle:
    jaspar = motifs.read(handle, "jaspar")

jaspar_df = (pd.DataFrame((jaspar.counts)))
jaspar_df_t = jaspar_df.transpose()

AR_motif_wt_pfm_prob = jaspar_df_t / jaspar_df_t.sum()

## Change to column names and make them integer
AR_motif_wt_pfm_prob.columns = AR_motif_wt_pfm_prob.col-
umns.astype(int)

## Taking a logarithm for each value and turning PFM to PWM
AR_motif_wt_pwm = np.log2(AR_motif_wt_pfm_prob/0.25)
AR_motif_wt_pwm.columns = AR_motif_wt_pwm.columns.astype(int)

AR_pos_change = pd.concat([AR_new.loc[:, 9] - AR_new.loc[:, 1] - 1,
AR_new.loc[:, 11] + '/' + AR_new.loc[:, 12]], axis=1)

## Saving sequences as lists to Series
```

```
AR_sequence_list = AR['TF_motif'].apply(lambda x: list(x)).copy()
AR_sequence_list_copy = AR_sequence_list.copy()
AR_sequence_scores = {} #key:sequence, value:score

## Calculating the PWM score for each sequence with the function cal-
culate_score
x = 0
for sequence in AR_sequence_list:
    AR_sequence_scores[''.join(sequence)] = calculate_score(sequence,
    AR_motif_wt_pwm)
    AR['TF_motif'].iloc[x] = ''.join(sequence)
    x += 1

## Creating a dataframe of sequences and PWM scores
AR_sequence_scores_df = pd.DataFrame(pd.Series(AR_sequence_scores),
columns=['score'])

AR['PWM_score'] = ''
values = pd.DataFrame()

## Saving a PWM score of each sequence to a dataframe
for i in AR_sequence_scores.keys():
    if (i in AR['TF_motif'].values) == True:
        values = AR.loc[AR['TF_motif'] == i, ].index
        for index in values:
            AR['PWM_score'].loc[index] = AR_sequence_scores[i]

## Removing NaN and infinity values of PWM scores
AR_hist = AR[~AR.isin([np.nan, np.inf, -np.inf]).any(1)]

## Plotting a histogram of PWM scores
plt.hist(AR_hist['PWM_score'])

## Collecting variants and their locations in the genome
AR_pos_change = pd.DataFrame(AR_pos_change)
variants = AR_pos_change[1].apply(lambda x: x.split('/')[1])
locations = AR_pos_change[0]

## Modifying the wild-type sequence to mutated sequence
x = 0
for seq in AR_sequence_list_copy:
    seq[locations.iloc[x]] = variants.iloc[x]
    AR_sequence_list_copy.iloc[x] = seq
    x +=1

AR['PWM_score_after_variant'] = ''
AR['TF_motif_after_variant'] = ''

## Calculating the PWM score for each mutated sequence with the func-
tion calculate_score
x = 0
AR_sequence_scores = {} #key:sequence, value:score
```

```
for sequence in AR_sequence_list_copy:
    AR_sequence_scores[''.join(sequence)] = calculate_score(sequence,
    AR_motif_wt_pwm)
    AR['TF_motif_after_variant'].iloc[x] = ''.join(sequence)
    x += 1

AR_sequence_scores_df = pd.DataFrame(pd.Series(AR_sequence_scores),
columns=['score'])

## Saving a PWM score of each mutated sequence to a dataframe
values = pd.DataFrame()
for i in AR_sequence_scores.keys():
    if (i in AR['TF_motif_after_variant'].values) == True:
        values = AR.loc[AR['TF_motif_after_variant'] == i, ].
index
        for index in values:
            AR['PWM_score_after_variant'].loc[index] = AR_se-
quence_scores[i]

## Removing NaN and infinity values of PWM scores
AR_hist = AR[~AR.isin([np.nan, np.inf, -np.inf]).any(1)]

## Plotting a histogram of PWM scores
plt.hist(AR_hist['PWM_score'])

# Calculating the difference between wild-type and mutated sequence
AR['diff'] = AR['PWM_score'] - AR['PWM_score_after_variant']
## Removing NaN and infinity values
AR_hist = AR['diff']
AR_hist.dropna(inplace = True)
AR_hist = AR_hist[~AR_hist.isin([np.nan, np.inf, -np.inf])]

## Plotting a histogram of differences with a range of -15 and 15
plt.hist(AR_hist, bins = range(-15,15))

AR_diff = AR.copy()
AR_diff['diff'].replace([np.inf, -np.inf], np.nan, inplace=True)

## Drop rows with NaN
AR_diff['diff'].dropna(inplace=True)

## Including motifs with a change over 5 or less than -5
AR_all = AR_diff[(AR_diff['diff'] > 5) | (AR_diff['diff'] < -5)]
AR_all.sort_values(by = [7, 8], inplace = True)

## Filtering by chromosome and location of variant and filename of the
sample
count_df_AR = AR_all.groupby(['7', '8', 'Filename'])

## This part is run in R for files that were previously annotated in
Homer Annotatepeaks.pl
```{r}
```
```

```
## Standard imports
library(dplyr)
library(stringr)

## Downloading files that have been annotated
files <- list.files(path="/data/projects/salokorpi/AnnPeaks", pat-
tern="*.txt", full.names=T, recursive=FALSE)

## Finding peaks that are annotated as TSS and saving them to a new
file
for (file_x in files){
  homer_peaks <- read.delim(file = file_x)
  colnames(homer_peaks)[1:4] <- c('Peak_id', 'Chr', 'Start', 'End')
  subset_homer_peaks <- homer_peaks[grepl("TSS", homer_peaks[["Annota-
tion"]]), ]
  subset_homer_peaks <- subset_homer_peaks %>% relocate("Chr",
"Start", "End", "Peak_id")
  colnames(subset_homer_peaks) <- c('chrom', 'chromStart', 'chromEnd',
'Sample')
  subset_homer_peaks$Sample <- sub("-.*", "", subset_homer_peaks$Sam-
ple)
  write.table(subset_homer_peaks[,1:4], file = paste(file_x, '.bed',
sep = ''), sep="\t", row.names=FALSE, col.names = F, quote = F)
}
```

```
## Downloading annotated file back to Python
AR_all = pd.read_csv('Homer_annotatedPeaks/AR_PWM_count_all.txt.bed',
sep='\t', header = None)
AR_all.columns = ['Chr', 'Start', 'End', 'Sample']

## Making a pivot table out of the dataframe of chromosome and start
of variant and sorting the rows
variant_counts_AR = AR_all.pivot_table(columns=['Chr', 'Start'], ag-
gfunc='size')
variant_counts_AR = variant_counts_AR.sort_values(ascending = False)

## Making a pivot table out of the dataframe at the start of the vari-
ant and sorting the rows
variant_counts_AR_2 = AR_all.pivot_table(columns=['Start'], ag-
gfunc='size')
variant_counts_AR_2 = variant_counts_AR_2.sort_values(ascending =
False)

df_AR = pd.DataFrame()

## Making a new dataframe
df_of_variant_list = pd.DataFrame(columns= ['Start of variant', 'Sam-
ples'])

## Looping through start sites of variants and making a list of sam-
ples that have the variant
```

```
for i in variant_counts_AR_2.keys():
    if (i in AR_all['Start'].values) == True:
           samples = AR_all.loc[AR_all['Start'] == i, ]
           df_of_variant_list.loc[len(df_of_variant_list.index)] = [i,
           list(samples['Sample'])]


## Saving the file to run BEDtools Window
df_AR.to_csv('Homer_annotatedPeaks\Annotated_vari-
ant_counts\AR_all_ann_counts.bed', sep ='\t', index = False, header =
False)

## Downloading file after gene window analysis
AR_all = pd.read_csv('Windows/AR_all_genes.bed', sep='\t', header =
None)

AR_all['Gene expression of sample'] = ''
scores = {}
AR = AR_all[[3, 7]]
## Looping through AR_all to get gene expression scores for each gene
for ensembl, sample in AR.itertuples(index=False):
    if ((sample in AR[7].values) & (ensembl in AR[3].values)):
           values = AR.loc[(AR[3] == ensembl) & (AR[7] == sample)].in-
           dex
           AR_all['Gene expression of sample'].loc[values] = en-
           sembl_TF(ensembl, sample)

## Removing empty rows
AR_all['Gene expression of sample'] = AR_all['Gene expression of sam-
ple'].replace('None', np.nan)
AR_all = AR_all.dropna(axis=0, subset=['Gene expression of sample'])

## Making an empty dataframe
df_of_variant_list = pd.DataFrame(columns= ['Start of variant', 'En-
sembl ID', 'Samples'])

## Collecting each gene’s samples with a variant to a list
for i in AR_all[1]:
     samples = AR_all.loc[AR_all[1] == i, ]
     df_of_variant_list.loc[len(df_of_variant_list.index)] = [i, sam-
     ples[3].iloc[0], list(samples[7])]

## Removing duplicate rows
df_of_variant_list.drop_duplicates(subset = 'Ensembl ID', inplace =
True)

## Creating a new dataframe with column names
df = pd.DataFrame(columns = ['Ensembl ID', 'Gene', "p-value", 'Median
gene expression score of samples with variant', 'Median gene expres-
sion score of wild type samples', 'Number of samples with variant'])
```

```
## Looping through the length of AR_all to get p-value and median val-
ues from function Wilcoxon. Collecting gene names matching Ensembl ID
for index in range(len(AR_all)):
      pval, median_var, median_wild = wilcoxon(AR_all['Ensembl
      ID'][index], df_of_variant_list['Samples'][index])
      gene = gene_name(AR_all['Ensembl ID'][index], index)
      df.loc[len(df.index)] = [AR_all['Ensembl ID'][index], gene,
      pval.pvalue, median_var, median_wild, len(df_of_vari-
      ant_list['Samples'][index])]

## Filtering out the rows with p-values bigger than 0.05
df = df[(df['p-value'] < 0.05)]
print(df['p-value'])

## Running multiple testing correction
df_correct = multiple_testing_correction(df['p-value'])
df['P-value after multiple testing correction'] = df_correct[1]
df = df[(df['P-value after multiple testing correction'] < 0.05)]

df.to_csv('Windows\P-values\AR_all_pvals.bed', index = False,
sep='\t')
df['Gene'].to_csv('Windows\P-values\AR_all_genes.bed', index = False,
header = False, sep = '\t')
```

APPENDIX B: DIFFERENTIALLY EXPRESSED GENES WITH THE MOST SIGNIFICANT VARIANTS AFFECTING GENE EXPRESSION

Supplementary table B.1. *Differentially expressed genes*

Ensembl ID	Gene	Chromosome of variant	Start of variant	Transcription factor	P-value	P-value after multiple testing corrections	Median gene expression score of samples with variant	Median gene expression score of wild-type samples
ENSG00000204965	AAAS	chr5	140821604	MYC	0.007095771905767007	0.048899016015033 09	4.20301843576974	2.232836202809 35
ENSG00000182919	AASS	chr11	93741591	AR	0.039945516704688794	0.049295679398841344	10.4039611526022	10.023664137305 9
ENSG00000182919	AASS	chr11	93741591	FOXA1	0.039945516704688794	0.049295679398841344	10.4039611526022	10.023664137305 9
ENSG00000165995	ABCB4	chr10	18140424	CTCF	0.045663563442 94926	0.049125641551331 5	8.96315940903224	8.096138411007 6
ENSG00000163485	ABCB5	chr1	203090654	MYC	0.035589495734 8061	0.048899016015033 09	4.055808990583785	4.883399522276 87
ENSG00000241553	ABCC8	chr3	9792495	AR	0.020242775262 6707	0.049295679398841344	11.54175037428584 9	11.316834292333 8
ENSG00000241553	ABCC8	chr3	9792495	FOXA1	0.020242775262 6707	0.049295679398841344	11.54175037428584 9	11.316834292333 8
ENSG00000230274	ABL1	chr3	40322715	MYC	0.021800904583 696122	0.048899016015033 09	0.0	0.0
ENSG00000158042	ACADVL	chr11	6680385	AR	0.000516789234 4514897	0.018675106468740 19	10.3451076370547	11.172924267325499
ENSG00000158042	ACADVL	chr11	6680385	FOXA1	0.000516789234 4514897	0.018675106468740 19	10.3451076370547	11.172924267325499
ENSG00000277462	ACD	chr1	247034637	AR	0.009959934299 948615	0.049279630886828 7	7.438048885412935	7.078980820882 63
ENSG00000128045	ACO2	chr4	52862317	MYC	0.048776777510 43249	0.049940846209705 44	7.49681961879431	6.869306678648 05
ENSG00000189334	ACOD1	chr1	153614255	MYC	0.040749180012 52758	0.048899016015033 09	8.807449663828756	7.063978005525 68
ENSG00000166337	ACOT7	chr11	6606294	AR	0.000589740204 276006	0.018675106468740 19	11.28871200198600 1	11.713484103386 3
ENSG00000166337	ACOT7	chr11	6606294	FOXA1	0.000589740204 276006	0.018675106468740 19	11.28871200198600 1	11.713484103386 3

ENSG00000130695	ACPP	chr1	26234200	AR	0.0293951624827304	0.049279630886828 7	7.65193723035406	8.015326732394 58
ENSG00000187266	ACSM3	chr19	11377207	FOXA1	0.0322547467949 9918	0.040371258891333 33	6.83717906996451	7.536352366270 03
ENSG00000103154	ADCY2	chr16	83968244	AR	0.0157948677815 83025	0.049279630886828 7	0.0	2.081238188834 59
ENSG00000103485	ADD1	chr16	29663279	AR	0.0481237110276 65966	0.049279630886828 7	10.14758687599329	8.486548176282 021
ENSG00000134627	AGBL5	chr11	94543840	AR	0.0063799171724 96606	0.048094760223435 95	6.73474187874408	5.630702482978 981
ENSG00000135702	AGK	chr16	75528530	CTCF	0.0146551107705 08534	0.049125641551331 5	0.0	0.0
ENSG00000144028	AGO1	chr2	96274338	AR	0.0033151911610 56079	0.033227028689524 63	12.625430 2929758	12.953145532786252
ENSG00000144028	AGO1	chr2	96274338	FOXA1	0.0033151911610 56079	0.033227028689524 63	12.625430 2929758	12.953145532786252
ENSG00000058404	AGPS	chr7	44210019	MYC	0.0464008903592 1505	0.049940846209705 44	7.76910927 1491205	8.908081595869 87
ENSG00000252839	AHCY	chr14	73246818	MYC	0.0355894957348 061	0.048899016015033 09	0.0	0.0
ENSG00000121413	AHCY	chr19	58083838	AR	0.0162252095700 137237	0.049295679398841 344	10.19271253213721	9.622974314153 99
ENSG00000121413	AHCY	chr19	58083838	FOXA1	0.0162252095700 137237	0.049295679398841 344	10.19271253213721	9.622974314153 99
ENSG00000142303	AK2	chr19	8580240	MYC	0.0332469190860 98039	0.048899016015033 09	8.72730873127646 6	7.174808867318 069
ENSG00000101452	AK2	chr20	38962299	CTCF	0.0244245363122 29243	0.043015503246591 896	8.47313699643665	8.105164673324 719
ENSG00000234585	AKAP8L	chr7	65038354	AR	0.0031470954188 573853	0.033227028689524 63	6.71349632844543	5.780694545524 165
ENSG00000234585	AKAP8L	chr7	65038354	FOXA1	0.0031470954188 573853	0.033227028689524 63	6.71349632844543	5.780694545524 165
ENSG00000130226	ALG1	chr7	153887097	AR	0.0456635634429 4926	0.049295679398841 344	-0.478632728312750 44	3.072567077441 8257
ENSG00000130226	ALG1	chr7	153887097	FOXA1	0.0456635634429 4926	0.049295679398841 344	-0.478632728312750 44	3.072567077441 8257
ENSG00000149150	ALKBH5	chr11	57484534	AR	0.0183581054891 73444	0.049295679398841 344	11.80275256187880 1	10.407214861404 4
ENSG00000149150	ALKBH5	chr11	57484534	FOXA1	0.0183581054891 73444	0.049295679398841 344	11.80275256187880 1	10.407214861404 4
ENSG00000170266	ALX4	chr3	32996609	AR	0.0363906100720 8641	0.049295679398841 344	10.79249644605930 1	11.053760987913 2
ENSG00000170266	ALX4	chr3	32996609	FOXA1	0.0363906100720 8641	0.049295679398841 344	10.79249644605930 1	11.053760987913 2
ENSG00000142541	ANAPC4	chr19	49487510	CTCF	0.0022135051460 282227	0.033329225999468 06	16.730963 6366908	16.280334408946 5

ENSG00000126259	ANKRD44	chr19	35855861	AR	0.017003141395691 97	0.049295679398841344	-0.363886771789227 07	0.0
ENSG00000126259	ANKRD44	chr19	35855861	FOXA1	0.017003141395691 97	0.049295679398841344	-0.363886771789227 07	0.0
ENSG00000078140	AP1S1	chr4	39698109	MYC	0.005614095037689342	0.048899016015033 09	11.661530 1023314	12.0530 37984553
ENSG00000205758	AP2S1	chr21	33589341	MYC	0.002037886283 8392143	0.048899016015033 09	9.8088523 12359313	9.28342 64588219 39
ENSG00000189306	AP5M1	chr22	42508344	E2F1	0.007438466539 466542	0.040527723565541 3	10.804079 22337545	10.4244 524446647
ENSG00000130724	APBA2	chr19	58551452	MYC	0.024030714109 758086	0.048899016015033 09	12.743524 626595399	12.0199 738420 61
ENSG00000112852	ARCN1	chr5	141094606	MYC	0.023654747527 608485	0.048899016015033 09	9.0855350 64170461	7.43128 152463390 55
ENSG00000196616	ARHGAP33	chr4	99304971	MYC	0.038062573520 36791	0.048899016015033 09	8.6670141 1998403	7.90054 800714 3925
ENSG00000133884	ARID4B	chr11	65333852	MYC	0.010152109319 221931	0.048899016015033 09	10.780834 0778652	11.1323 032983 57
ENSG00000198242	ARID4B	chr17	28719985	CTCF	0.012698559844 085624	0.049125641551331 5	10.982348 2869736	13.3657 941441 32801
ENSG00000181333	ARSF	chr11	94021354	AR	0.035288207644 12166	0.049279630886828 7	4.7512768 9873727	4.14684 662198243
ENSG00000149150	ARVCF	chr11	57484534	AR	0.018358105489 173444	0.049279630886828 7	11.802752 561878801	10.4072 148614044
ENSG00000114391	ASAP3	chr3	101681091	AR	0.046400890359 21505	0.049295679398841344	15.822952 663393849	15.3123 360260 8035
ENSG00000114391	ASAP3	chr3	101681091	FOXA1	0.046400890359 21505	0.049295679398841344	15.822952 663393849	15.3123 360260 8035
ENSG00000221662	ASNS	chr1	18897071	AR	0.025285366313 814082	0.049295679398841344	-0.064165617232853	0.0
ENSG00000221662	ASNS	chr1	18897071	FOXA1	0.025285366313 814082	0.049295679398841344	-0.064165617232853	0.0
ENSG00000150768	ASTE1	chr11	112025408	AR	0.028609752678 682827	0.049279630886828 7	9.6576912 3802302	10.4207 827969009
ENSG00000172888	ATG2B	chr3	40524878	E2F1	0.017890845597 89118	0.040527723565541 3	9.5679456 19417674	9.75680 909190 4056
ENSG00000238304	ATG4A	chr11	3781395	MYC	0.048433536389 85755	0.049940846209705 44	-0.640155055122819 9	0.0
ENSG00000181333	ATG5	chr11	94021354	AR	0.035288207644 12166	0.049295679398841344	4.7512768 9873727	4.14684 662198243
ENSG00000181333	ATG5	chr11	94021354	FOXA1	0.035288207644 12166	0.049295679398841344	4.7512768 9873727	4.14684 662198243
ENSG00000233276	ATP11A	chr3	49357176	MYC	0.042757784166 99213	0.049686196501489 62	13.523466 85409815	12.9444 724404 2915
ENSG00000185305	BAIAP3	chr5	53883942	MYC	0.049934413723 48684	0.049940846209705 44	9.9378223 75157829	10.3478 437028604

ENSG00000090238	BDKRB1	chr16	30092314	AR	0.000570869833367242	0.018675106468674019	11.38739591142330 1	10.84554503250 5902
ENSG00000090238	BDKRB1	chr16	30092314	FOXA1	0.000570869833367242	0.018675106468674019	11.38739591142330 1	10.84554503250 5902
ENSG00000235974	BPI	chr19	58012589	AR	0.042757784166 99213	0.049279630886828 7	2.188505750112435	0.372180880966 0249
ENSG00000184860	BTAF1	chr16	81988855	AR	0.031957801504 71461	0.049279630886828 7	8.86150051955847	8.408189991022 46
ENSG00000271816	BTBD7	chr10	73699151	MYC	0.018742483464 63052	0.048899016015033 09	4.04912015703945	2.711816284419 84
ENSG00000228223	BTBD7	chr6	26521709	E2F1	0.002743150187 44714	0.040527723565541 3	9.776750067199675	8.706449197437 31
ENSG00000171793	BTN3A1	chr1	40979300	AR	0.020741370796 551084	0.049279630886828 7	10.2646669626644	9.856732657518 892
ENSG00000157884	C19orf60	chr2	26581205	E2F1	0.024334695951 86201	0.040527723565541 3	1.066790558845 57	2.770944012953 67
ENSG00000141562	C20orf194	chr17	82458180	MYC	0.040749180012 52758	0.048899016015033 09	9.74175648193221	10.25020099798 585
ENSG00000131400	CA12	chr19	50358477	AR	0.035327196267 96993	0.049295679398841 344	4.858829696428191	3.897488963192 73
ENSG00000131400	CA12	chr19	50358477	FOXA1	0.035327196267 96993	0.049295679398841 344	4.858829696428191	3.897488963192 73
ENSG00000100033	CACNA1G	chr22	18912777	CTCF	0.045663563442 94926	0.045663563442949 26	3.085267172438045	4.323395899507 4655
ENSG00000241553	CACNA1G	chr3	9792495	AR	0.020242775262 6707	0.049279630886828 7	11.54175037428584 9	11.31683429233 38
ENSG00000113761	CAMTA2	chr5	177022696	MYC	0.032254746794 99918	0.048899016015033 09	8.634125220089 42	9.041452342529 276
ENSG00000186827	CARD10	chr1	1211326	AR	0.029395162482 73049	0.049295679398841 344	7.426806296836 78	6.275456310737 09
ENSG00000186827	CARD10	chr1	1211326	FOXA1	0.029395162482 73049	0.049295679398841 344	7.426806296836 78	6.275456310737 09
ENSG00000103994	CASC3	chr15	42412823	MYC	0.042953919681 27125	0.049686196501489 62	11.541206680173	11.956903526131498
ENSG00000269097	CAV2	chr19	57664280	MYC	0.024798582413 24801	0.048899016015033 09	0.966466178756775	0.0
ENSG00000175265	CBFB	chr15	34378935	MYC	0.008188646942 100473	0.048899016015033 09	8.999408565234933	6.753031621592 23
ENSG00000089159	CBX7	chr12	120210439	MYC	0.022882267210 95797	0.048899016015033 09	11.04403068686869 8	11.493823615370 5
ENSG00000159023	CCAR1	chr1	28887091	MYC	0.025878653650 157617	0.048899016015033 09	11.10798564594285	11.375416133133 25
ENSG00000067066	CCAR1	chr2	230415942	CTCF	0.010799899834 568076	0.049125641551331 5	11.51097640069454 9	11.208923235666 148
ENSG00000021776	CCDC124	chr15	34851782	MYC	0.019770905813 457654	0.048899016015033 09	10.98088766180 12	11.277007874334 13
ENSG00000265407	CCDC124	chr19	49308797	FOXA1	0.030508208148 06394	0.040371258891333 33	-0.598767523595864	0.0

ENSG00000168158	CCDC80	chr16	3355889	MYC	0.049773979632384176	0.049940846209705044	1.70362210146203	2.411134354729233
ENSG00000243297	CCDC80	chr19	36901742	AR	0.020891512093319325	0.049279630886828 7	0.0	0.955171458479 0021
ENSG00000124177	CCDC88C	chr20	41402083	AR	0.043262937541 51935	0.049279630886828 7	11.055997600687402	11.954175654525399
ENSG00000178229	CCL1	chr19	57320472	MYC	0.029174449950 973628	0.048899016015033 09	7.755637688922579 5	8.076194859045575
ENSG00000071462	CD22	chr7	73683025	MYC	0.008977942016 305054	0.048899016015033 09	11.136247949795	10.806260447850 6
ENSG00000100364	CD6	chr22	45192244	E2F1	0.022523341169 45638	0.040527723565541 3	11.5167857246366	11.137357964175902
ENSG00000024048	CD9	chr6	42564029	ESR1	0.003098121186 7389356	0.029456403692617 62	10.8702427702319	11.123870922305 3
ENSG00000204967	CDC23	chr5	140806929	MYC	0.003818153404 3456623	0.048899016015033 09	7.167594466438475	4.224032203838 3304
ENSG00000171970	CDC25B	chr19	2900928	AR	0.000831645146 3053089	0.019751572224751 088	7.59250450318568	7.187703651601 256
ENSG00000171970	CDC25B	chr19	2900928	FOXA1	0.000831645146 3053089	0.019751572224751 088	7.59250450318568	7.187703651601 256
ENSG00000254450	CDC27	chr11	111817214	AR	0.022372202881 60119	0.049295679398841 344	2.72566180713412	0.0
ENSG00000254450	CDC27	chr11	111817214	FOXA1	0.022372202881 60119	0.049295679398841 344	2.72566180713412	0.0
ENSG00000122861	CDC34	chr10	73909177	MYC	0.015155476007 282023	0.048899016015033 09	8.43328503734271	8.095609270584 65
ENSG00000204969	CDC6	chr5	140794852	MYC	0.008412922979 993676	0.048899016015033 09	4.783086372951805	2.467094573974 9097
ENSG00000138036	CDH1	chr2	43774039	AR	0.045277593291 83217	0.049279630886828 7	9.52043982909457	9.289133006970 731
ENSG00000151224	CDH7	chr10	80271820	MYC	0.032443751042 60566	0.048899016015033 09	5.156056019313059	2.879995564370 01
ENSG00000171169	CDHR2	chr9	128061233	AR	0.016690709025 18613	0.049295679398841 344	8.01233117231858	8.200746815828 671
ENSG00000171169	CDHR2	chr9	128061233	FOXA1	0.016690709025 18613	0.049295679398841 344	8.01233117231858	8.200746815828 671
ENSG00000149929	CDK14	chr16	29992330	AR	0.002795999480 998655	0.033227028689524 63	9.459224049802438	9.229065871303 4
ENSG00000149929	CDK14	chr16	29992330	FOXA1	0.002795999480 998655	0.033227028689524 63	9.459224049802438	9.229065871303 4
ENSG00000179094	CDKL5	chr17	8140472	FOXA1	0.032254746794 99918	0.040371258891333 33	11.95620458767710 1	10.756251380619 5
ENSG00000223797	CELSR3	chr3	40313802	E2F1	0.049524666270 82263	0.049524666270822 63	7.402793372620515	7.260732580858 875
ENSG00000256904	CFH	chr12	8819816	MYC	0.036290618541 40877	0.048899016015033 09	-3.446964767045805	0.0
ENSG00000134817	CFTR	chr11	57233577	FOXA1	0.005468433527 832356	0.040371258891333 33	6.07638449008865	7.656451476710 299

ENSG00000174371	CIAPIN1	chr1	241847967	FOXA1	0.041603055657 3786	0.044203246635964 77	5.413617848416109	5.942343084751 8794
ENSG00000186827	CINP	chr1	1211326	AR	0.029395162482 73049	0.049279630886828 7	7.426806296836 78	6.275456310737 09
ENSG00000124177	CLDN11	chr20	41402083	AR	0.043262937541 51935	0.049295679398841 344	11.05599760068740 2	11.954175654525399
ENSG00000124177	CLDN11	chr20	41402083	FOXA1	0.043262937541 51935	0.049295679398841 344	11.05599760068740 2	11.954175654525399
ENSG00000204789	CLDN18	chr6	27356451	CTCF	0.008212574322 89495	0.045625412904971 95	7.981804038226825	7.029230523580 19
ENSG00000267858	CLNS1A	chr19	58559125	AR	0.040349853567 755146	0.049295679398841 344	5.677418126288 12	5.131956236274 33
ENSG00000267858	CLNS1A	chr19	58559125	FOXA1	0.040349853567 755146	0.049295679398841 344	5.677418126288 12	5.131956236274 33
ENSG00000143324	CNIH1	chr1	180632022	AR	0.011483531114 935667	0.049295679398841 344	10.25895820914619 9	10.731245318554 3
ENSG00000143324	CNIH1	chr1	180632022	FOXA1	0.011483531114 935667	0.049295679398841 344	10.25895820914619 9	10.731245318554 3
ENSG00000269058	CNTN1	chr19	16479061	MYC	0.032254746794 99918	0.048899016015033 09	-2.74204154321577	0.0
ENSG00000154359	COASY	chr8	12721906	AR	0.018982308127 829865	0.049295679398841 344	9.943032023091 3	9.672269989024 97
ENSG00000154359	COASY	chr8	12721906	FOXA1	0.018982308127 829865	0.049295679398841 344	9.943032023091 3	9.672269989024 97
ENSG00000103175	COCH	chr16	84294846	AR	0.019152345574 950522	0.049295679398841 344	11.0325067935991	10.506046853815 75
ENSG00000103175	COCH	chr16	84294846	FOXA1	0.019152345574 950522	0.049295679398841 344	11.0325067935991	10.506046853815 75
ENSG00000164669	COL17A1	chr7	65141032	AR	0.040348155281 42803	0.049279630886828 7	4.868174722055 77	3.663576646997 9405
ENSG00000100580	COMP	chr14	77335029	MYC	0.029866132789 723523	0.048899016015033 09	10.09085061547239 4	10.489723779325 5
ENSG00000243742	COQ9	chr11	61615036	AR	0.027965569965 861015	0.049295679398841 344	8.271584089825879	6.990281873045 28
ENSG00000243742	COQ9	chr11	61615036	FOXA1	0.027965569965 861015	0.049295679398841 344	8.271584089825879	6.990281873045 28
ENSG00000121413	CRISPLD2	chr19	58083838	AR	0.016225209570 137237	0.049279630886828 7	10.19271253213721	9.622974314153 99
ENSG00000188295	CTCF	chr1	247099962	AR	0.045850947858 70022	0.049279630886828 7	7.941957553248 39	7.620903645261 42
ENSG00000215397	CTCF	chr20	661596	MYC	0.020307817950 580738	0.048899016015033 09	1.425072625534769 9	0.0
ENSG00000233319	CTNNA1	chr10	130131770	CTCF	0.028887526440 07262	0.049125641551331 5	0.0	0.0
ENSG00000230989	CTSA	chr16	83719311	AR	0.032443751042 60566	0.049279630886828 7	12.80670943704 01	12.385350212972 9
ENSG00000013364	CTT-NBP2	chr16	29820394	AR	0.017659161108 465212	0.049279630886828 7	11.65398115206 42	11.091106484747 801

ENSG00000163374	CYB5R4	chr1	155659443	E2F1	0.014484489735 3324	0.040527723565541 3	10.899993751036599	11.091990463771 1
ENSG00000204685	CYP26A1	chr2	96208389	AR	0.019745012130 979362	0.049295679398841 344	6.722815417783405	6.359841676425 114
ENSG00000204685	CYP26A1	chr2	96208389	FOXA1	0.019745012130 979362	0.049295679398841 344	6.722815417783405	6.359841676425 114
ENSG00000276002	CYP2W1	chr19	50258443	AR	0.045787419878 1624	0.049279630886828 7	1.229879252235060 1	0.0
ENSG00000160318	CYP46A1	chr19	51367098	MYC	0.021325424355 998226	0.048899016015033 09	6.15003301298492	5.315156385897 685
ENSG00000257108	DBF4	chr16	567005	ESR1	0.040355966069 1937	0.047751107719393 37	5.38826732402915	4.714324309873 41
ENSG00000188693	DBNDD1	chr7	92134604	CTCF	0.002609771325 520223	0.020878170604161 785	3.500161541334795	2.680855870532 21
ENSG00000182791	DCN	chr11	66590176	AR	0.042463383879 772415	0.049279630886828 7	4.896971794332625	5.504176331849 189
ENSG00000223756	DDX17	chr11	3380918	AR	0.029122162649 09046	0.049295679398841 344	4.769122726431955	3.751355706972 155
ENSG00000223756	DDX17	chr11	3380918	FOXA1	0.029122162649 09046	0.049295679398841 344	4.769122726431955	3.751355706972 155
ENSG00000150687	DDX17	chr11	86791059	MYC	0.036290618541 40877	0.048899016015033 09	11.0496644480808	12.335934895454 25
ENSG00000134627	DDX43	chr11	94543840	AR	0.006379917172 496606	0.046622471645167 5	6.734741878 74408	5.630702482978 981
ENSG00000134627	DDX43	chr11	94543840	FOXA1	0.006379917172 496606	0.046622471645167 5	6.734741878 74408	5.630702482978 981
ENSG00000186806	DESI1	chr19	51331536	AR	0.005819579884 421388	0.046071674085002 65	8.339940296404789	7.927053077788 21
ENSG00000186806	DESI1	chr19	51331536	FOXA1	0.005819579884 421388	0.046071674085002 65	8.339940296404789	7.927053077788 21
ENSG00000165392	DGKA	chr8	31033788	E2F1	0.021890751050 807285	0.040527723565541 3	9.043180476203789	8.635799833374 42
ENSG00000225285	DHX29	chr1	1430539	AR	0.011362150106 743796	0.049295679398841 344	7.588202319003130 5	6.253515820485 441
ENSG00000225285	DHX29	chr1	1430539	FOXA1	0.011362150106 743796	0.049295679398841 344	7.588202319003130 5	6.253515820485 441
ENSG00000198885	DIP2B	chr2	96325317	AR	0.039799418714 41024	0.049279630886828 7	7.016893581169844	6.587670477357 82
ENSG00000148343	DIS3	chr9	129036621	MYC	0.043475421938 80342	0.049686196501489 62	9.40892058729746	9.157036457700 169
ENSG00000150990	DLEC1	chr12	124946825	CTCF	0.049125641551 3315	0.049125641551331 5	8.944141167055461	9.150827706085 359
ENSG00000254999	DLEC1	chr3	10115675	AR	0.029072042093 867483	0.049279630886828 7	12.70752767239384 8	12.501778366366 315
ENSG00000083312	DNAH11	chr5	72816312	MYC	0.023247195148 468597	0.048899016015033 09	12.01929217577820 1	12.202185809236749
ENSG00000254450	DVL2	chr11	111817214	AR	0.022372202881 60119	0.049279630886828 7	2.72566180713412	0.0

ENSG00000180549	DYRK4	chr9	137030174	E2F1	0.011712981380090089	0.040527723565541 3	1.002548478924189 9	2.408422912909 7996
ENSG00000186806	ECM2	chr19	51331536	MYC	0.001897897023 1600747	0.048899016015033 09	8.524593721219224	7.890581578669 41
ENSG00000163121	EDN1	chr2	96497646	AR	0.045418481795 239184	0.049279630886828 7	6.154801331818565	5.486890979738 385
ENSG00000188868	EEF1A2	chr19	12317477	AR	0.029174449950 973628	0.049295679398841 344	8.077302459988651	7.762090386063 875
ENSG00000188868	EEF1A2	chr19	12317477	FOXA1	0.029174449950 973628	0.049295679398841 344	8.077302459988651	7.762090386063 875
ENSG00000127989	EIPR1	chr7	91692008	CTCF	0.022840527469 54778	0.049125641551331 5	8.618542264211 64	8.285853010970 696
ENSG00000064393	ELMO2	chr7	139561570	AR	0.024424536312 229243	0.049279630886828 7	12.052634 3306933	13.019015643583 4
ENSG00000171970	ELMO3	chr19	2900928	AR	0.000831645146 3053089	0.027167074779306 757	7.592504503185 68	7.187703651601 256
ENSG00000105254	EP300	chr19	36114289	AR	0.030508208148 06394	0.049279630886828 7	11.503947 47034880 2	11.053003194917 249
ENSG00000242028	EPN2	chr15	43796142	MYC	0.030808261005 142595	0.048899016015033 09	4.797251241421179	3.971292571125 645
ENSG00000221520	EPYC	chr7	92204015	MYC	0.027109304938 119364	0.048899016015033 09	0.841108916292019	0.0
ENSG00000234585	ERCC1	chr7	65038354	AR	0.003147095418 8573853	0.034276303279720 144	6.713496328 44543	5.780694545524 165
ENSG00000143622	ERH	chr1	155897808	MYC	0.025325223050 321977	0.048899016015033 09	10.319649 74636559 9	9.668975801328 966
ENSG00000114391	ESR1	chr3	101681091	AR	0.046400890359 21505	0.049279630886828 7	15.822952 66339384 9	15.312336026080 35
ENSG00000259516	ETV1	chr15	35181799	MYC	0.034108343191 42843	0.048899016015033 09	0.0	0.973234575248 5531
ENSG00000177842	EVI5	chr3	40477131	CTCF	0.039359508888 249725	0.049125641551331 5	8.285095411875231	7.425923152005 451
ENSG00000116525	EVX1	chr1	33145399	MYC	0.016892125802 11515	0.048899016015033 09	8.867723414829856	9.335132650076 035
ENSG00000086475	FA2H	chr10	13317428	MYC	0.043262937541 51935	0.049686196501489 62	10.764019 62446230 2	10.919338459617 3
ENSG00000261126	FAM120A	chr18	80046900	CTCF	0.003837081773 6190664	0.033329225999468 06	4.544327658056 16	3.377169759043 01
ENSG00000231584	FAM120A	chr2	96010526	AR	0.047434225098 551504	0.049279630886828 7	6.196399139905965 5	5.616932276084 116
ENSG00000257108	FAM136A	chr16	567005	CTCF	0.021268338567 301148	0.049125641551331 5	5.152132046920889	4.477715777851 56
ENSG00000142534	FAM168A	chr19	49496365	CTCF	0.003294916648 55967	0.033329225999468 06	16.411348 2175593	16.061994787498 4
ENSG00000082213	FAM76A	chr5	31532287	AR	0.026463269377 593083	0.049295679398841 344	9.739268321451 81	10.352545411371 4
ENSG00000082213	FAM76A	chr5	31532287	FOXA1	0.026463269377 593083	0.049295679398841 344	9.739268321451 81	10.352545411371 4

ENSG00000009709	FAM76B	chr1	18630846	AR	0.039359508888249725	0.049295679398841344	-2.970772948948319 7	0.0
ENSG00000009709	FAM76B	chr1	18630846	FOXA1	0.039359508888249725	0.049295679398841344	-2.970772948948319 7	0.0
ENSG00000060762	FAS	chr6	166364919	E2F1	0.043416052112 4754	0.046310455586640424	10.588871 2488476	10.2375 7350327 26
ENSG00000163421	FAT2	chr3	71771655	AR	0.030984997584820984	0.049279630886828 7	3.0837686 16717819 5	0.94504 0707789 238
ENSG00000200769	FBXW4	chr7	92202243	MYC	0.035901205609 27541	0.048899016015033 09	2.0142341 6880008	0.0
ENSG00000134962	FGF4	chr4	39406930	MYC	0.048688485592 78192	0.049940846209705 44	4.6316098 1709529	5.68568 1987260 17
ENSG00000135114	FH	chr12	121017763	MYC	0.036290618541 40877	0.048899016015033 09	4.4715243 5946972	5.95453 6557894 725
ENSG00000105656	FLY-WCH1	chr19	18442663	MYC	0.032015971082 67044	0.048899016015033 09	9.2067400 87918416	9.40153 5506273
ENSG00000177303	FMO1	chr17	75500261	AR	0.046905833086 972624	0.049279630886828 7	10.234462 3948089	10.3976 0508784 88
ENSG00000204963	FMO2	chr5	140834248	MYC	0.007729239299 174583	0.048899016015033 09	4.2317077 9706104	1.89424 6106704 64
ENSG00000244131	FNDC3A	chr12	8742428	MYC	0.025325223050 321977	0.048899016015033 09	-1.7709507 17933929 9	0.68088 1052431 9155
ENSG00000077009	FNDC3B	chr19	3933069	AR	0.026764245313 653835	0.049295679398841344	-1.4852574 23612095	2.15295 5747342 72
ENSG00000077009	FNDC3B	chr19	3933069	FOXA1	0.026764245313 653835	0.049295679398841344	-1.4852574 23612095	2.15295 5747342 72
ENSG00000105649	FOXRED2	chr19	18196784	MYC	0.020695659504 665658	0.048899016015033 09	7.7614727 35488695	8.15521 9637781 9
ENSG00000166747	FUCA2	chr16	71729000	ESR1	0.036390610072 08641	0.047751107719393 37	11.512596 361154	11.8274 1820441 77
ENSG00000164111	FUCA2	chr4	121667946	CTCF	0.043262937541 51935	0.045663563442949 26	14.630103 9669558	14.0344 0462878 93
ENSG00000005243	GABRA1	chr17	48026167	AR	0.027319202105 005273	0.049279630886828 7	9.0509254 4626707	8.53072 4416651 81
ENSG00000187051	GALC	chr22	39529093	CTCF	0.048919207273 142466	0.049125641551331 5	11.072743 4713142	10.9759 6686557 16
ENSG00000105393	GAS7	chr19	17267376	AR	0.003497581967 318382	0.034276303279720 144	8.3978936 58326009	7.89988 0390688 99
ENSG00000144028	GGA1	chr2	96274338	AR	0.003315191161 056079	0.034276303279720 144	12.625430 2929758	12.9531 4553278 6252
ENSG00000099940	GGT5	chr22	20859007	AR	0.035589495734 8061	0.049279630886828 7	10.712350 94944785	10.8596 1045086 0899
ENSG00000223476	GLP2R	chr7	64933273	CTCF	0.041611334344 40652	0.049125641551331 5	2.5093958 34627939 7	1.51540 8910083 035
ENSG00000124249	GLT8D1	chr20	44745865	CTCF	0.045663563442 94926	0.049125641551331 5	3.8707332 04259660 3	5.53192 2027396 02

ENSG00000126457	GPM6B	chr19	49675786	CTCF	0.046646807762 11384	0.049125641551331 5	12.2692449668277	12.023292628853099
ENSG00000122435	GRB10	chr1	100133150	MYC	0.039589544521 80087	0.048899016015033 09	8.48040791112921	8.071788022155 05
ENSG00000120837	GTF2IRD1	chr12	104117086	ERG	0.034731659807 5229	0.047322385249218 196	10.49885238692365 2	10.314328004070 3
ENSG00000126453	HCCS	chr19	49665142	CTCF	0.003999507119 936167	0.033329225999468 06	9.004458520322212	8.656374858687 02
ENSG00000269893	HECTD1	chr4	118278703	AR	0.038634298254 34656	0.049295679398841 344	10.78874378439180 1	11.268528912102 04
ENSG00000269893	HECTD1	chr4	118278703	FOXA1	0.038634298254 34656	0.049295679398841 344	10.78874378439180 1	11.268528912102 04
ENSG00000100316	HHAT	chr22	39312882	CTCF	0.020741370796 551084	0.049125641551331 5	16.51853833759419 7	16.232483521031 9
ENSG00000274472	HIVEP2	chr5	154705626	ERG	0.046400890359 21505	0.047322385249218 196	0.0	0.690504359460 553
ENSG00000204685	HMGXB4	chr2	96208389	AR	0.019745012130 979362	0.049279630886828 7	6.722815417783405	6.359841676425 114
ENSG00000079950	HMOX1	chr6	132445867	AR	0.036290618541 40877	0.049279630886828 7	10.41287486481075	10.906900988199
ENSG00000254911	HOOK2	chr11	93721513	AR	0.049351157209 831985	0.049351157209831 985	-0.710263359138239 9	0.0
ENSG00000152661	HPF1	chr6	121435595	AR	0.042757784166 99213	0.049279630886828 7	11.99875315028915	11.475267060220 92
ENSG00000188868	HSD17B1P1	chr19	12317477	MYC	0.029174449950 973628	0.048899016015033 09	8.077302459988651	7.762090386063 875
ENSG00000181392	HSP90AB1	chr19	36003307	AR	0.023423887704 95009	0.049295679398841 344	10.32981411514065	9.272237149935 405
ENSG00000181392	HSP90AB1	chr19	36003307	FOXA1	0.023423887704 95009	0.049295679398841 344	10.32981411514065	9.272237149935 405
ENSG00000156017	IBTK	chr9	74980790	CTCF	0.023250836647 02644	0.049125641551331 5	8.882515060 58288	9.207282836743 115
ENSG00000173992	ICA1	chr11	66593153	ESR1	0.020733486012 92209	0.037618601957463 06	9.778483180661715	9.431839576368 624
ENSG00000160325	ICA1	chr9	133459965	ERG	0.015919060573 542528	0.047322385249218 196	9.562828147008396	9.952654372232 97
ENSG00000207091	IDS	chr15	34314728	ESR1	0.043771848742 77726	0.047751107719393 37	-1.17035326394583	0.0
ENSG00000082213	IFFO1	chr5	31532287	AR	0.026463269377 593083	0.049279630886828 7	9.739268321 45181	10.352545411374
ENSG00000233369	IFT80	chr7	73154938	MYC	0.017352389190 761427	0.048899016015033 09	8.000333163736165	10.336311003428 499
ENSG00000261971	IKZF2	chr16	3037400	CTCF	0.040749180012 52758	0.049125641551331 5	8.956950763001455	7.627025713561 931
ENSG00000079950	IL11	chr6	132445867	AR	0.036290618541 40877	0.049295679398841 344	10.41287486481075	10.906900988199

ENSG00000079950	IL11	chr6	132445867	FOXA1	0.036290618541 40877	0.049295679398841344	10.41287486481075	10.906900988199
ENSG00000230274	INPP4A	chr3	40322715	CTCF	0.011362150106 743796	0.049125641551331 5	0.969181369409550 9	0.0
ENSG00000165119	INTS13	chr9	83968083	AR	0.047903189313 88099	0.049279630886828 7	14.15639078790735	14.36364387638185
ENSG00000189171	INTS6	chr1	153618787	MYC	0.040749180012 52758	0.048899016015033 09	9.281850720558118	8.047994267211376
ENSG00000199906	IPO11	chr3	40498891	AR	0.048275940961 268894	0.049295679398841344	2.41839380072851	1.8557623528056 2
ENSG00000199906	IPO11	chr3	40498891	FOXA1	0.048275940961 268894	0.049295679398841344	2.41839380072851	1.8557623528056 2
ENSG00000004777	ITGAL	chr19	35774532	AR	0.033246919086 98039	0.049295679398841344	8.538854515351876	7.242978729454576
ENSG00000004777	ITGAL	chr19	35774532	FOXA1	0.033246919086 98039	0.049295679398841344	8.538854515351876	7.242978729454576
ENSG00000155506	ITGAL	chr5	154682986	ERG	0.039359508888 249725	0.047322385249218196	12.25638053738735	12.761311787596 05
ENSG00000169439	ITIH1	chr8	96493813	CTCF	0.006658239220 719644	0.041613995129497774	12.1747645624537	11.836360626391
ENSG00000122912	ITIH4	chr10	68477998	E2F1	0.026456353486 824676	0.040527723565541 3	11.28777982727260 1	10.932811347496202
ENSG00000131400	JADE1	chr19	50358477	AR	0.035327196267 96993	0.049279630886828 7	4.858829696428191	3.897488963192 73
ENSG00000112855	KCNAB2	chr5	140691430	MYC	0.001337239556 0119928	0.048899016015033 09	9.106856991673835	8.789570157153886
ENSG00000275041	KCNK2	chr8	100897853	MYC	0.004121964761 565393	0.048899016015033 09	1.06143546723711	0.0
ENSG00000229816	KDM1A	chr2	32201600	ERG	0.047322385249 218196	0.047322385249218196	2.44592647923062	1.127234161576 25
ENSG00000119636	KIAA0100	chr14	74019349	AR	0.048776777510 43249	0.049295679398841344	7.73577382060995	8.038433131662 51
ENSG00000119636	KIAA0100	chr14	74019349	FOXA1	0.048776777510 43249	0.049295679398841344	7.73577382060995	8.038433131662 51
ENSG00000103355	KIAA0556	chr16	2783953	CTCF	0.048302379121 798675	0.049125641551331 5	1.99978117843924	0.0
ENSG00000213462	KIAA0556	chr7	64990356	AR	0.035152837353 55208	0.049279630886828 7	8.71005318640 59	9.359858938319 615
ENSG00000106686	KLC3	chr9	4553386	MYC	0.030508208148 06394	0.048899016015033 09	5.962759977498566	5.297221000699 69
ENSG00000182791	KLHL13	chr11	66590176	ESR1	0.019745012130 979362	0.037618601957463 06	5.292752636136539	5.864965164073 229
ENSG00000134817	KMT2E	chr11	57233577	AR	0.005468433527 832356	0.046071674085002 65	6.07638449008865	7.656451476710 299
ENSG00000134817	KMT2E	chr11	57233577	FOXA1	0.005468433527 832356	0.046071674085002 65	6.07638449008865	7.656451476710 299
ENSG00000072778	KRIT1	chr17	7217125	AR	0.020667876132 125937	0.049295679398841344	12.7689537107037	12.430950165231 8

ENSG00000072778	KRIT1	chr17	7217125	FOXA1	0.020667876132125937	0.049295679398841344	12.7689537107037	12.430950165231 8
ENSG00000134108	KRIT1	chr3	5122249	ERG	0.036290618541 40877	0.047322385249218196	11.326487454428	11.80845958424215
ENSG00000166133	LAMA3	chr15	40569299	E2F1	0.040749180012 52758	0.046310455586640424	7.726567651669516	8.195159434955386
ENSG00000064393	LAMC2	chr7	139561570	AR	0.024424536312229243	0.049295679398841344	12.0526343306933	13.019015643583 4
ENSG00000064393	LAMC2	chr7	139561570	FOXA1	0.024424536312229243	0.049295679398841344	12.0526343306933	13.019015643583 4
ENSG00000257108	LAMP2	chr16	567005	CTCF	0.026884689529119935	0.043015503246591896	5.510804283363991	4.714324309873 41
ENSG00000004777	LAMP2	chr19	35774532	AR	0.033246919086 98039	0.049279630886828 7	8.538854515351876	7.242978729454576
ENSG00000158553	LCP2	chr6	27285903	CTCF	0.023395491077 711397	0.049125641551331 5	0.0	0.746723399407 3895
ENSG00000215252	LIG3	chr15	34525095	ESR1	0.021002995762 711064	0.037618601957463 06	8.45437186538701	6.371380404936 92
ENSG00000143324	LIPG	chr1	180632022	AR	0.011483531114 935667	0.049279630886828 7	10.258958209146199	10.731245318554 3
ENSG00000170638	LMBR1	chr22	50185913	MYC	0.038213674575 016364	0.048899016015033 09	9.639516642008 98	9.809760525798 85
ENSG00000135097	LPCAT2	chr12	120341330	MYC	0.038197609531 63916	0.048899016015033 09	8.620826399651719	9.222205038912 131
ENSG00000267858	LRCH4	chr19	58559125	AR	0.040349853567 755146	0.049279630886828 7	5.677418126288 12	5.131956236274 33
ENSG00000167182	LRR-FIP2	chr17	47896150	AR	0.002292955061 4600126	0.033227028689524 63	10.2097637628244	10.641546655803198
ENSG00000167182	LRR-FIP2	chr17	47896150	FOXA1	0.002292955061 4600126	0.033227028689524 63	10.2097637628244	10.641546655803198
ENSG00000213462	LSG1	chr7	64990356	AR	0.035152837353 55208	0.049295679398841344	8.710053186405 9	9.359858938319615
ENSG00000213462	LSG1	chr7	64990356	FOXA1	0.035152837353 55208	0.049295679398841344	8.710053186405 9	9.359858938319615
ENSG00000101098	LTBP1	chr20	44751808	CTCF	0.019723236054 794577	0.049125641551331 5	1.693931865076620 9	6.603554821083396
ENSG00000169217	LTF	chr16	30350773	AR	0.048776777510 43249	0.049279630886828 7	11.1117214646417	10.931549344552801
ENSG00000177873	LUC7L3	chr3	40477113	MYC	0.034894859062 32599	0.048899016015033 09	7.719732026456 86	7.965434504496679
ENSG00000173080	LY75	chr1	155941638	E2F1	0.038493020681 63402	0.046310455586640424	0.0	0.0
ENSG00000182791	LYPLA2	chr11	66590176	AR	0.042463383879 772415	0.049295679398841344	4.896971794332625	5.504176331849189
ENSG00000182791	LYPLA2	chr11	66590176	FOXA1	0.042463383879 772415	0.049295679398841344	4.896971794332625	5.504176331849189
ENSG00000202533	LZTS2	chr5	132468147	MYC	0.034363873236 635034	0.048899016015033 09	2.295100485132640 4	0.960675008218 115

ENSG00000104953	MADCAM1	chr19	2977538	AR	0.001597445025766067	0.030351455489555278	5.33567429057921	4.533882128026 7805
ENSG00000104953	MADCAM1	chr19	2977538	FOXA1	0.001597445025766067	0.030351455489555278	5.33567429057921	4.533882128026 7805
ENSG00000183317	MAGEC2	chr1	37713880	AR	0.013165741067 2081	0.049279630886828 7	6.328980737116780 6	7.578958890816 5095
ENSG00000221662	MAP3K13	chr1	18897071	AR	0.025285366313 814082	0.049279630886828 7	-0.0641656 17232853	0.0
ENSG00000170458	MAP3K14	chr5	140631728	CTCF	0.037070501894 20318	0.049125641551331 5	9.95053335181501	10.45667613244 3551
ENSG00000248383	MAP3K1	chr5	140926299	MYC	0.025479830364 07347	0.048899016015033 09	4.141414009280774 5	2.704204068968 3597
ENSG00000130695	MAP4K5	chr1	26234200	AR	0.029395162482 73049	0.049295679398841 344	7.65193723035406	8.015326732394 58
ENSG00000130695	MAP4K5	chr1	26234200	FOXA1	0.029395162482 73049	0.049295679398841 344	7.65193723035406	8.015326732394 58
ENSG00000099822	MAP4K5	chr19	589881	CTCF	0.002225077752 1243773	0.033329225999468 06	4.444255307219 41	5.517891964451 16
ENSG00000188163	MAPK8IP2	chr9	137243584	E2F1	0.020393872900 58733	0.040527723565541 3	1.75078762087327	0.0
ENSG00000236296	MAST4	chr4	143559472	MYC	0.037729824121 75714	0.048899016015033 09	4.561182724322 75	3.576907707810 85
ENSG00000101104	MAT2B	chr20	44910060	CTCF	0.040749180012 52758	0.049125641551331 5	8.793219147345159	6.340493857758 435
ENSG00000184925	MATR3	chr9	136949551	E2F1	0.020242775262 6707	0.040527723565541 3	5.995978605437955	5.021917158927 38
ENSG00000034713	MBTD1	chr16	75566375	CTCF	0.005819579884 421388	0.041568427745867 054	11.789961 942511	11.336824836563 801
ENSG00000182919	MED24	chr11	93741591	AR	0.039945516704 688794	0.049279630886828 7	10.403961 1526022	10.023664137305 9
ENSG00000125878	MEGF8	chr20	604257	MYC	0.017723943735 911133	0.048899016015033 09	1.67393301632319	2.797235087532 1697
ENSG00000152256	METTL1	chr2	172555373	E2F1	0.032254746794 99918	0.043258334723474 214	10.45264552659515	9.437972021038 405
ENSG00000167182	MICALL1	chr17	47896150	AR	0.002292955061 4600126	0.034276303279720 144	10.209763 7628244	10.641546655803 198
ENSG00000106701	MKS1	chr9	105447796	CTCF	0.043648036103 188216	0.049125641551331 5	6.73920338308088	7.165271289028 86
ENSG00000184445	MOCOS	chr12	122527246	MYC	0.022372202881 60119	0.048899016015033 09	7.313080166446215	8.167539071325 67
ENSG00000010539	MPPED2	chr16	3222325	CTCF	0.045663563442 94926	0.049125641551331 5	8.681111520626665	7.977260861056 725
ENSG00000154359	MS4A12	chr8	12721906	AR	0.018982308127 829865	0.049279630886828 7	9.94303202309 13	9.672269989024 97
ENSG00000177873	MSANTD3	chr3	40477113	E2F1	0.031745341706 406895	0.043258334723474 214	7.70348415235169	7.935064874468 119
ENSG00000204961	MSH2	chr5	140847772	MYC	0.002179195053 1444846	0.048899016015033 09	3.580361602140925	0.770937709935 6245

ENSG00000152661	MSMO1	chr6	121435595	AR	0.042757784166 99213	0.049295679398841344	11.99875315028915	11.475267060 2292
ENSG00000152661	MSMO1	chr6	121435595	FOXA1	0.042757784166 99213	0.049295679398841344	11.99875315028915	11.475267060 2292
ENSG00000165119	MXD1	chr9	83968083	AR	0.047903189313 88099	0.049295679398841344	14.15639078790735	14.363643876 38185
ENSG00000165119	MXD1	chr9	83968083	FOXA1	0.047903189313 88099	0.049295679398841344	14.15639078790735	14.363643876 38185
ENSG00000173599	MYH13	chr11	66848417	ESR1	0.049524666270 82263	0.049524666270822 63	9.072396040460024	9.616656318484 878
ENSG00000197191	MYH13	chr9	137224635	E2F1	0.014149026432 152	0.040527723565541 3	5.617767124007901	5.114666550284 7695
ENSG00000254999	MYLIP	chr3	10115675	AR	0.029072042093 867483	0.049295679398841344	12.70752767239384 8	12.501778363663 15
ENSG00000254999	MYLIP	chr3	10115675	FOXA1	0.029072042093 867483	0.049295679398841344	12.70752767239384 8	12.501778363663 15
ENSG00000277462	MYLK2	chr1	247034637	AR	0.009959934299 948615	0.049295679398841344	7.438048885412935	7.078980820882 63
ENSG00000277462	MYLK2	chr1	247034637	FOXA1	0.009959934299 948615	0.049295679398841344	7.438048885412935	7.078980820882 63
ENSG00000090238	NAA10	chr16	30092314	AR	0.000570869833 3367242	0.027167074779306 757	11.387395911423301	10.845545032505902
ENSG00000160299	NANS	chr21	46324124	MYC	0.039359508888 249725	0.048899016015033 09	9.004458520322212	9.655672522449 08
ENSG00000197128	NAT9	chr19	57466663	MYC	0.025625209431 290832	0.048899016015033 09	7.995136632413 93	8.457939124198 3
ENSG00000103260	NCAPH2	chr16	715118	CTCF	0.023724375521 784902	0.049125641551331 5	9.708266404405 4	9.412173456481 119
ENSG00000184860	NECAP1	chr16	81988855	AR	0.031957801504 71461	0.049295679398841344	8.861500519558 47	8.408189991022 46
ENSG00000184860	NECAP1	chr16	81988855	FOXA1	0.031957801504 71461	0.049295679398841344	8.861500519558 47	8.408189991022 46
ENSG00000207357	NFE2L3	chr19	1021522	E2F1	0.021682531306 961156	0.040527723565541 3	2.888430740807 86	0.596547199372 2121
ENSG00000101442	NFYA	chr20	38748460	CTCF	0.033484485293 732505	0.049125641551331 5	8.293235510162 39	7.916628042303 16
ENSG00000153832	NISCH	chr2	229922302	CTCF	0.023654747527 608485	0.049125641551331 5	8.866281281086 3	8.413130157524 7
ENSG00000284154	NKAIN1	chr1	33332393	MYC	0.007553798506 537555	0.048899016015033 09	2.054497799771285	0.0
ENSG00000167112	NOD1	chr9	128305159	MYC	0.018341329646 214377	0.048899016015033 09	10.83810531441 13	11.268644479491 85
ENSG00000148341	NOMO3	chr9	129007036	MYC	0.017022329779 541212	0.048899016015033 09	11.643898992281 5	11.325542118695 65
ENSG00000186468	NOP58	chr5	82273320	CTCF	0.020393872900 58733	0.049125641551331 5	11.885207924455399	12.426419923164 1
ENSG00000055957	NOTCH3	chr3	52777595	MYC	0.033940371266 208046	0.048899016015033 09	2.589207468018 21	0.973234575248 5531

ENSG00000213918	NRIP2	chr16	3611728	MYC	0.004692093524400863	0.048899016015033309	9.651611554860815	9.013511175083991
ENSG00000105254	NRP1	chr19	36114289	AR	0.030508208148063944	0.049295679398841344	11.50394747034880	11.053003194917249
ENSG00000105254	NRP1	chr19	36114289	FOXA1	0.030508208148063944	0.049295679398841344	11.50394747034880	11.053003194917249
ENSG00000232814	NRXN3	chr13	110502575	AR	0.023247195148468597	0.049279630886828 7	0.825709553885 74	0.0
ENSG00000141622	NTSR1	chr18	46326809	MYC	0.048776777510432 49	0.049940846209705 44	7.882825849118269 4	6.893099913718 18
ENSG00000103502	NUB1	chr16	29858357	AR	0.007639545339591015	0.049279630886828 7	11.619725086470401	11.318399331656302
ENSG00000255622	NXPE1	chr5	141155996	MYC	0.031703881259114 59	0.048899016015033309	4.3232031 4603138	2.260472793710 9796
ENSG00000152223	OGFR	chr18	45800581	MYC	0.030911666145468 14	0.048899016015033309	9.4719326 8281542	9.992346953768 91
ENSG00000001629	OSBPL7	chr7	92245974	MYC	0.009796632751061707	0.048899016015033309	10.869510597332098	10.9888 14619275999
ENSG00000150625	OTUD5	chr4	175632934	MYC	0.028887526440072 62	0.048899016015033309	8.6440356 3936981	7.391754323475 35
ENSG00000158042	PAX2	chr11	6680385	AR	0.001122827445 7746993	0.027509272421480137	10.640640 2498158	11.201995422461 1
ENSG00000055957	PDE4A	chr3	52777595	AR	0.033940371266208046	0.049279630886828 7	2.5892074 6801821	0.973234575248 5531
ENSG00000122557	PDK2	chr7	35632659	FOXA1	0.024030714109758086	0.040371258891333 33	10.760487133298199	10.570647129797 32
ENSG00000127989	PDK2	chr7	91692008	CTCF	0.022840527469 54778	0.043015503246591896	8.6185422 6421164	8.285853010970696
ENSG00000172867	PDK3	chr12	52644558	AR	0.017578665106865163	0.049279630886828 7	4.9578397 9832113	2.408422912909 7996
ENSG00000215252	PDK3	chr15	34525095	MYC	0.010009221278 27532	0.048899016015033309	9.1100605 03419763	6.371380404936 92
ENSG00000134109	PDK4	chr3	5187646	ERG	0.032254746794 99918	0.047322385249218196	9.9482957 46472475	10.72938395815505
ENSG00000138036	PEX3	chr2	43774039	AR	0.045277593291 83217	0.049295679398841344	9.5204398 2909457	9.289133006970731
ENSG00000138036	PEX3	chr2	43774039	FOXA1	0.045277593291 83217	0.049295679398841344	9.5204398 2909457	9.289133006970731
ENSG00000215910	PGM3	chr1	11761787	MYC	0.048123711027 665966	0.049940846209705 44	1.318457455075739 5	0.0
ENSG00000234797	PHGDH	chr15	59768352	AR	0.025565749689 383292	0.049279630886828 7	8.5075198 5230048	8.060054092096 365
ENSG00000231584	PHKA2	chr2	96010526	AR	0.047434225098 551504	0.049295679398841344	6.196399139905965 5	5.616932276084 116
ENSG00000231584	PHKA2	chr2	96010526	FOXA1	0.047434225098 551504	0.049295679398841344	6.196399139905965 5	5.616932276084 116
ENSG00000127311	PHRF1	chr12	66302493	MYC	0.045663563442 94926	0.049940846209705 44	7.220016808403 03	7.836300587533 724

ENSG00000122359	PHTF2	chr10	80150889	MYC	0.029395162482 73049	0.048899016015033 09	13.612197 1964232	13.2345 74969305699
ENSG00000181894	PIGL	chr19	58126248	MYC	0.011362150106 743796	0.048899016015033 09	9.1184823 3352239	9.495264895501439
ENSG00000268660	PITX1	chr19	58044592	AR	0.036193346450 03881	0.049279630886828 7	-0.348108272957146 47	0.0
ENSG00000099940	PKD1	chr22	20859007	ERG	0.035589495734 8061	0.047322385249218 196	10.712350 94944785	10.859610450860899
ENSG00000065268	PKN2	chr19	984332	E2F1	0.032443751042 60566	0.043258334723474 214	10.409752 5901791	9.8683412957636 9
ENSG00000163683	PLAT	chr4	39546336	MYC	0.009317120489 652634	0.048899016015033 09	12.307548 9702284	12.730339181720 4
ENSG00000149150	PLAUR	chr11	57484534	FOXA1	0.012464873002 125357	0.040371258891333 33	11.802752 56187880 1	10.407214861406 44
ENSG00000117505	PLEKHH1	chr1	93345907	MYC	0.010799899834 568076	0.048899016015033 09	11.449475 00975915	11.056169316846102
ENSG00000126259	PLPP1	chr19	35855861	AR	0.017003141395 69197	0.049279630886828 7	-0.363886771789227 07	0.0
ENSG00000101442	PLXND1	chr20	38748460	MYC	0.033484485293 732505	0.048899016015033 09	8.2932355 1016239	7.916628042303 16
ENSG00000181392	PNPLA5	chr19	36003307	AR	0.023423887704 95009	0.049279630886828 7	10.329814 11514065	9.272237149935 405
ENSG00000225285	POLB	chr1	1430539	AR	0.011362150106 743796	0.049279630886828 7	7.5882023 190031305	6.253515820485 441
ENSG00000171169	POLD3	chr9	128061233	AR	0.016690709025 18613	0.049279630886828 7	8.0123311 7231858	8.200746815828 671
ENSG00000157992	POLR1A	chr2	27442366	AR	0.040749180012 52758	0.049279630886828 7	10.586625 79801435	9.718502196159 23
ENSG00000126460	POLR2B	chr19	49580646	CTCF	0.031470182073 70469	0.049125641551331 5	8.8116162 02405661	8.394888295064 83
ENSG00000234494	POLR2F	chr17	47897330	AR	0.031398758807 24878	0.049279630886828 7	6.1830087 32538309	5.939107669985 651
ENSG00000130165	POLR2J	chr19	11551147	FOXA1	0.022372202881 60119	0.040371258891333 33	11.515444 45731495	11.057217069730399
ENSG00000149262	PON1	chr11	77874418	ERG	0.023724375521 784902	0.047322385249218 196	9.2467185 37742131	9.458091987073 16
ENSG00000040341	POP4	chr8	73420369	MYC	0.027965569965 861015	0.048899016015033 09	10.079500 87144045	10.895807264631099
ENSG00000269893	PPIL2	chr4	118278703	AR	0.038634298254 34656	0.049279630886828 7	10.788743 78439180 1	11.268528912100 4
ENSG00000201388	PPP1R12A	chr19	32608337	E2F1	0.046180453331 71794	0.047670145374676 585	0.0	0.662626301141 214
ENSG00000186806	PPP1R16B	chr19	51331536	AR	0.005819579884 421388	0.047526569056108	8.3399402 96404789	7.927053077788 21
ENSG00000013364	PPP2R3A	chr16	29820394	AR	0.017659161108 465212	0.049295679398841 344	11.653981 1520642	11.091106484747801

ENSG00000013364	PPP2R3A	chr16	29820394	FOXA1	0.017659161108465212	0.049295679398841344	11.6539811520642	11.091106484747801
ENSG00000100033	PREX2	chr22	18912777	CTCF	0.045663563442 94926	0.049125641551331 5	3.085267172438045	4.323395899507 4655
ENSG00000106993	PRKCH	chr9	4679559	MYC	0.036193346450 03881	0.048899016015033 09	10.016640 047366849	9.484019672011 225
ENSG00000105393	PROM1	chr19	17267376	AR	0.003497581967 318382	0.033227028689524 63	8.397893658326009	7.899880390688 99
ENSG00000105393	PROM1	chr19	17267376	FOXA1	0.003497581967 318382	0.033227028689524 63	8.397893658326009	7.899880390688 99
ENSG00000104324	PRSS21	chr8	96645242	CTCF	0.044984627794 04522	0.049125641551331 5	10.768025 942090599	10.3527 9937229 0999
ENSG00000104953	PSMA3	chr19	2977538	AR	0.001597445025 766067	0.031309922505014 916	5.335674 29057921	4.533882128026 7805
ENSG00000105643	PSMB1	chr19	18001132	MYC	0.005556784285 468106	0.048899016015033 09	8.889763798558 72	8.663329293340 2
ENSG00000201113	PSMB1	chr2	32214456	ERG	0.039785700882 457885	0.047322385249218 196	0.0	0.0
ENSG00000240106	PSMC1	chr19	16539688	MYC	0.040749180012 52758	0.048899016015033 09	-0.171587743840174 97	0.0
ENSG00000176986	PSMD7	chr10	73744372	MYC	0.017782273831 103083	0.048899016015033 09	10.530198 4798343	11.121443245986 7
ENSG00000169217	PTBP1	chr16	30350773	AR	0.048776777510 43249	0.049295679398841 344	11.111721 4646417	10.931549344552801
ENSG00000169217	PTBP1	chr16	30350773	FOXA1	0.048776777510 43249	0.049295679398841 344	11.111721 4646417	10.931549344552801
ENSG00000199990	PTGR1	chr5	140711275	MYC	0.003818153404 3456623	0.048899016015033 09	1.950375396984614 7	0.0
ENSG00000168517	PTPN3	chr17	45160700	MYC	0.049669106914 99726	0.049940846209705 44	6.189313 1 4958271	5.716062287875 52
ENSG00000276002	PTPN3	chr19	50258443	AR	0.045787419878 1624	0.049295679398841 344	1.229879252235060 1	0.0
ENSG00000276002	PTPN3	chr19	50258443	FOXA1	0.045787419878 1624	0.049295679398841 344	1.229879252235060 1	0.0
ENSG00000166337	PVALB	chr11	6606294	AR	0.000589740204 276006	0.027167074779306 757	11.288712 0019860 0 1	11.713484103386 3
ENSG00000242818	PXN	chr10	119768247	AR	0.013301750912 849838	0.049279630886828 7	0.005649899650180 53	0.0
ENSG00000251369	RAB5C	chr19	57535257	MYC	0.022455783869 49659	0.048899016015033 09	8.165011886927955	8.681015850419 495
ENSG00000101452	RAD51	chr20	38962299	MYC	0.024424536312 229243	0.048899016015033 09	8.473136996436 65	8.105164673324 719
ENSG00000072778	RAD52	chr17	7217125	AR	0.020667876132 125937	0.049279630886828 7	12.768953 7107037	12.430950165231 8
ENSG00000112624	RAD52	chr6	42746958	ESR1	0.006533342920 269639	0.029456403692617 62	10.481142 0976354	10.878438119 57

ENSG00000153774	RALA	chr16	75293698	CTCF	0.035494342106862715	0.049125641551331 5	10.2359718099212	9.768367756386061
ENSG00000235974	RAN-GAP1	chr19	58012589	AR	0.042757784166992135	0.049295679398841344	2.188505750112435	0.372180880966 0249
ENSG00000235974	RAN-GAP1	chr19	58012589	FOXA1	0.042757784166992135	0.049295679398841344	2.188505750112435	0.372180880966 0249
ENSG00000171793	RB1CC1	chr1	40979300	AR	0.020741370796551084	0.049295679398841344	10.2646669626644	9.856732657518892
ENSG00000171793	RB1CC1	chr1	40979300	FOXA1	0.020741370796551084	0.049295679398841344	10.2646669626644	9.856732657518892
ENSG00000103175	RBFA	chr16	84294846	AR	0.019152345574950522	0.049279630886828 7	11.0325067935991	10.50604685381575
ENSG00000163545	RBM27	chr1	205302063	MYC	0.028759454588368368	0.048899016015033 09	7.510813398698064 5	7.981354380894226
ENSG00000199906	RCOR1	chr3	40498891	AR	0.048275940961268894	0.049279630886828 7	2.4183938 0072851	1.855762352805 62
ENSG00000162086	RCVRN	chr16	3305406	MYC	0.043424453832 98641	0.049686196501489 62	9.9749246 7814004	9.648227828232 94
ENSG00000205464	RECQL	chr5	82279462	CTCF	0.013291438780430091	0.049125641551331 5	7.5028314 9403983	6.573255698722 1695
ENSG00000142920	REV3L	chr1	33081104	MYC	0.026739586992689895	0.048899016015033 09	7.629658502163164	7.165672981996345
ENSG00000142552	REV3L	chr19	49528003	FOXA1	0.030508208148 06394	0.040371258891333 33	10.058799 84641468 5	9.498983171088664
ENSG00000122787	REXO5	chr7	138002324	MYC	0.048527948595 5361	0.049940846209705 44	2.3749659 60430390 3	0.798524625562 4896
ENSG00000055957	RFXANK	chr3	52777595	AR	0.033940371266208046	0.049295679398841344	2.5892074 6801821	0.973234575248 5531
ENSG00000055957	RFXANK	chr3	52777595	FOXA1	0.033940371266208046	0.049295679398841344	2.5892074 6801821	0.973234575248 5531
ENSG00000095209	RFXANK	chr9	105694541	CTCF	0.045605009564355425	0.049125641551331 5	9.5360965 6893718	9.194430305850 76
ENSG00000160678	RGCC	chr1	153627926	MYC	0.032254746794 99918	0.048899016015033 09	3.9848173 1038686	1.905315578989255
ENSG00000105464	RHBDD2	chr19	48393668	FOXA1	0.048720834873 56016	0.048720834873560 16	4.8496638 9899763	6.160247966734611
ENSG00000114784	RHOBTB2	chr3	40309707	CTCF	0.035759259374657175	0.049125641551331 5	10.167162 7266376	10.534640025317 3
ENSG00000230989	RIMBP2	chr16	83719311	AR	0.032443751042 60566	0.049295679398841344	12.806709 4370401	12.385350212972 9
ENSG00000230989	RIMBP2	chr16	83719311	FOXA1	0.032443751042 60566	0.049295679398841344	12.806709 4370401	12.385350212972 9
ENSG00000223756	RIMS4	chr11	3380918	AR	0.029122162649 09046	0.049279630886828 7	4.7691227 26431955	3.751355706972 155
ENSG00000064393	RNF126	chr7	139561570	MYC	0.024424536312 229243	0.048899016015033 09	12.052634 3306933	13.019015643583 4
ENSG00000243742	RNF31	chr11	61615036	AR	0.027965569965 861015	0.049279630886828 7	8.2715840 89825879	6.990281873045 28

ENSG00000213762	RNF43	chr19	57614233	MYC	0.045114199574 5356	0.049940846209705 44	9.916921798312664	10.02996666956 415
ENSG00000188536	RORA	chr16	172876	MYC	0.022716543626 886843	0.048899016015033 09	8.38901381257081	7.767505967583 975
ENSG00000130226	RPL26L1	chr7	153887097	AR	0.045663563442 94926	0.049279630886828 7	-0.478632728312750 44	3.072567077441 8257
ENSG00000199568	RPS10P5	chr15	65296051	MYC	0.010064615653 41654	0.048899016015033 09	0.841108916292019	0.0
ENSG00000237223	RPS20	chr2	108322238	ESR1	0.021944184475 186788	0.037618601957463 06	4.577664496771 72	3.123809305574 015
ENSG00000142546	RPUSD1	chr19	49555468	FOXA1	0.033246919086 98039	0.040371258891333 33	10.97609344509104 9	10.35442231309 4648
ENSG00000188693	RPUSD1	chr7	92134604	CTCF	0.002609771325 520223	0.033329225999468 06	3.500161541334795	2.680855870532 21
ENSG00000177042	RUNDC3B	chr11	695591	MYC	0.047903189313 88099	0.049940846209705 44	9.895274447241526	9.530514922515 845
ENSG00000184428	RWDD2A	chr8	143304384	FOXA1	0.036193346450 03881	0.041019125976710 65	9.155837468699215	9.565902765342 244
ENSG00000002746	SCT	chr7	43112629	MYC	0.048688485592 78192	0.049940846209705 44	7.111793131942480 5	6.522030336697 051
ENSG00000234494	SEC22C	chr17	47897330	AR	0.031398758807 24878	0.049295679398841 344	6.183008732538309	5.939107669985 651
ENSG00000234494	SEC22C	chr17	47897330	FOXA1	0.031398758807 24878	0.049295679398841 344	6.183008732538309	5.939107669985 651
ENSG00000130300	SEH1L	chr19	17351450	AR	0.045605009564 355425	0.049279630886828 7	11.5017820455039	11.692156092604 7
ENSG00000163121	SEMA3A	chr2	96497646	AR	0.045418481795 239184	0.049295679398841 344	6.154801331818565	5.486890979738 385
ENSG00000163121	SEMA3A	chr2	96497646	FOXA1	0.045418481795 239184	0.049295679398841 344	6.154801331818565	5.486890979738 385
ENSG00000109814	SFRP4	chr4	39498755	MYC	0.010716253755 469751	0.048899016015033 09	11.47238028890359 9	12.261634890909 09
ENSG00000166435	SH3GL2	chr11	74807739	MYC	0.008124016331 64379	0.048899016015033 09	9.360619883985805	8.902293104644 937
ENSG00000232814	SLC11A1	chr13	110502575	AR	0.023247195148 468597	0.049295679398841 344	0.82570955388574	0.0
ENSG00000232814	SLC11A1	chr13	110502575	FOXA1	0.023247195148 468597	0.049295679398841 344	0.82570955388574	0.0
ENSG00000198885	SLC12A2	chr2	96325317	AR	0.039799418714 41024	0.049295679398841 344	7.016893581169844	6.587670477357 82
ENSG00000198885	SLC12A2	chr2	96325317	FOXA1	0.039799418714 41024	0.049295679398841 344	7.016893581169844	6.587670477357 82
ENSG00000163421	SLC13A1	chr3	71771655	AR	0.030984997584 820984	0.049295679398841 344	3.083768616717819 5	0.945040707789 238
ENSG00000163421	SLC13A1	chr3	71771655	FOXA1	0.030984997584 820984	0.049295679398841 344	3.083768616717819 5	0.945040707789 238

ENSG00000100033	SLC16A8	chr22	18912777	MYC	0.045663563442 94926	0.0499408 46209705 44	3.0852671 72438045	4.32339 5899507 4655
ENSG00000197894	SLC22A16	chr4	99070978	MYC	0.01072 8332986 71145	0.0488990 16015033 09	12.588070 189693	12.2052 5823775 675
ENSG00000099940	SLC22A17	chr22	20859007	AR	0.03558 9495734 8061	0.0492956 79398841 344	10.712350 94944785	10.8596 1045086 0899
ENSG00000099940	SLC22A17	chr22	20859007	FOXA1	0.03558 9495734 8061	0.0492956 79398841 344	10.712350 94944785	10.8596 1045086 0899
ENSG00000109906	SLC25A14	chr11	114059041	AR	0.04074 9180012 52758	0.0492796 30886828 7	11.220406 99046409 9	9.95144 9215258 961
ENSG00000112763	SLC4A1	chr6	26457904	E2F1	0.02581 7590507 638582	0.0405277 23565541 3	9.7651385 28729821	9.33374 8370821 269
ENSG00000170296	SLC4A8	chr17	7240008	AR	0.02578 2807764 51427	0.0492956 79398841 344	3.0216801 49721450 3	2.28487 3059440 18
ENSG00000170296	SLC4A8	chr17	7240008	FOXA1	0.02578 2807764 51427	0.0492956 79398841 344	3.0216801 49721450 3	2.28487 3059440 18
ENSG00000203668	SLC7A2	chr1	241628851	FOXA1	0.02634 9276827 569677	0.0403712 58891333 33	8.1746091 4502956	8.64996 1159236 788
ENSG00000199172	SLCO1A2	chr12	95308420	MYC	0.02403 0714109 758086	0.0488990 16015033 09	2.1287233 4488417	0.87081 7714777 089
ENSG00000268660	SMARCD1	chr19	58044592	AR	0.03619 3346450 03881	0.0492956 79398841 344	-0.3481082 72957146 47	0.0
ENSG00000268660	SMARCD1	chr19	58044592	FOXA1	0.03619 3346450 03881	0.0492956 79398841 344	-0.3481082 72957146 47	0.0
ENSG00000099866	SNAPC1	chr19	489176	CTCF	0.04492 6989580 07531	0.0491256 41551331 5	3.7256209 72922990 3	4.24510 0290771 08
ENSG00000188295	SNPH	chr1	247099962	AR	0.04585 0947858 70022	0.0492956 79398841 344	7.9419575 5324839	7.62090 3645261 42
ENSG00000188295	SNPH	chr1	247099962	FOXA1	0.04585 0947858 70022	0.0492956 79398841 344	7.9419575 5324839	7.62090 3645261 42
ENSG00000207062	SNRPD3	chr7	65070538	AR	0.02459 7881870 87492	0.0492796 30886828 7	1.72675 8837585 62	0.0
ENSG00000164111	SNX11	chr4	121667946	CTCF	0.04326 2937541 51935	0.0491256 41551331 5	14.6301 0396695 58	14.034404 6287893
ENSG00000103260	SPATA20	chr16	715118	ESR1	0.02841 7650625 797382	0.0426264 75938696 076	9.61338 7527652 82	9.1557828 9055304
ENSG00000086827	SPINT3	chr11	113733187	AR	0.03629 0618541 40877	0.0492956 79398841 344	8.70957 5058813 485	9.4091023 38291456
ENSG00000086827	SPINT3	chr11	113733187	FOXA1	0.03629 0618541 40877	0.0492956 79398841 344	8.70957 5058813 485	9.4091023 38291456
ENSG00000181322	SPTLC1	chr3	138261437	MYC	0.03994 5516704 688794	0.0488990 16015033 09	4.84730 1943194 5	4.2096457 15471565
ENSG00000130300	SRCAP	chr19	17351450	AR	0.04560 5009564 355425	0.0492956 79398841 344	11.5017 8204550 39	11.692156 0926047
ENSG00000130300	SRCAP	chr19	17351450	FOXA1	0.04560 5009564 355425	0.0492956 79398841 344	11.5017 8204550 39	11.692156 0926047

ENSG00000152256	SRPK1	chr2	172555373	MYC	0.032254746794999 18	0.048899016015033 09	10.45264552659515	9.437972021038405
ENSG00000119636	ST7L	chr14	74019349	AR	0.048776777510 43249	0.049279630886828 7	7.735773820609 95	8.038433131662 51
ENSG00000089154	STAG3	chr12	120127202	MYC	0.006576193466 606716	0.048899016015033 09	11.154843643617 7	11.623720743520698
ENSG00000183317	STAU2	chr1	37713880	AR	0.013165741067 2081	0.049295679398841 344	6.328980737116 7806	7.578958890816509 5
ENSG00000183317	STAU2	chr1	37713880	FOXA1	0.013165741067 2081	0.049295679398841 344	6.328980737116 7806	7.578958890816509 5
ENSG00000009709	STK17B	chr1	18630846	AR	0.039359508888 249725	0.049279630886828 7	-2.970772948948 3197	0.0
ENSG00000150773	STX7	chr11	112063218	AR	0.045663563442 94926	0.049295679398841 344	2.585061290190 2004	5.651903605842484 5
ENSG00000150773	STX7	chr11	112063218	FOXA1	0.045663563442 94926	0.049295679398841 344	2.585061290190 2004	5.651903605842484 5
ENSG00000204962	SUCO	chr5	140841187	MYC	0.007729239299 174583	0.048899016015033 09	3.827061137462 66	1.834630165824 79
ENSG00000207062	SUPT16H	chr7	65070538	AR	0.024597881870 87492	0.049295679398841 344	1.726758837585 62	0.0
ENSG00000207062	SUPT16H	chr7	65070538	FOXA1	0.024597881870 87492	0.049295679398841 344	1.726758837585 62	0.0
ENSG00000202111	SUSD1	chr5	140718925	MYC	0.015511117156 833364	0.048899016015033 09	2.987920731238 17	0.0
ENSG00000170296	SYNE2	chr17	7240008	AR	0.025782807764 51427	0.049279630886828 7	3.021680149721 4503	2.284873059440 18
ENSG00000215397	SYNE2	chr20	661596	E2F1	0.020307817950 580738	0.040527723565541 3	1.425072625534 7699	0.0
ENSG00000240970	SYP	chr11	119003012	MYC	0.034299957321 185966	0.048899016015033 09	2.253366744011 68	1.104426949713870 1
ENSG00000184445	SYT1	chr12	122527246	AR	0.022372202881 60119	0.049279630886828 7	7.313080166446 215	8.167539077132567
ENSG00000103154	TACR2	chr16	83968244	AR	0.015794867781 583025	0.049295679398841 344	0.0	2.081238188834 59
ENSG00000103154	TACR2	chr16	83968244	FOXA1	0.015794867781 583025	0.049295679398841 344	0.0	2.081238188834 59
ENSG00000205084	TAF2	chr16	75536741	CTCF	0.027220913873 449257	0.049125641551331 5	9.106286650916 84	8.490866238089 04
ENSG00000157992	TBC1D1	chr2	27442366	AR	0.040749180012 52758	0.049295679398841 344	10.586625798014 35	9.718502196159 23
ENSG00000157992	TBC1D1	chr2	27442366	FOXA1	0.040749180012 52758	0.049295679398841 344	10.586625798014 35	9.718502196159 23
ENSG00000239827	TBCB	chr13	40882577	MYC	0.005439719036 120182	0.048899016015033 09	4.662100092561 67	4.166398858632 59
ENSG00000150768	TBPL1	chr11	112025408	AR	0.028609752678 682827	0.049295679398841 344	9.657691238023 02	10.420782796900 9

ENSG00000150768	TBPL1	chr11	112025408	FOXA1	0.028609752678682827	0.049295679398841344	9.6576912380230 2	10.4207827969009
ENSG00000215915	TBXA2R	chr1	1449689	AR	0.026596318589 88648	0.049279630886828 7	8.6178934626345 5	7.08698948202371
ENSG00000182158	TDP1	chr7	137874979	MYC	0.013953435230 67261	0.048899016015033 09	12.097608226676 43	12.61561967052385 2
ENSG00000170890	TECR	chr12	120322115	MYC	0.011906282262 233966	0.048899016015033 09	1.776044442600 9399	2.838899433408 14
ENSG00000101452	TFAP2D	chr20	38962299	CTCF	0.024424536312 229243	0.049125641551331 5	8.4731369964365	8.105164673324719
ENSG00000120798	THPO	chr12	95020229	MYC	0.032443751042 60566	0.048899016015033 09	9.8111661195642 3	9.39516145141351
ENSG00000164669	TM7SF3	chr7	65141032	AR	0.040348155281 42803	0.049295679398841344	4.8681747220557 7	3.663576646997940 5
ENSG00000164669	TM7SF3	chr7	65141032	FOXA1	0.040348155281 42803	0.049295679398841344	4.8681747220557 7	3.663576646997940 5
ENSG00000183172	TMCC3	chr22	42079691	E2F1	0.000801064407 0221912	0.025634061024710 12	10.110079197765051	9.546377514703703
ENSG00000137700	TMEM87A	chr11	119023751	MYC	0.029626434350 697173	0.048899016015033 09	10.017692829569 54	9.683417177741788
ENSG00000164880	TMEM98	chr7	1470277	FOXA1	0.028609752678 682827	0.040371258891333 33	12.217389816890 3	11.6822788238252
ENSG00000104695	TNC	chr8	30774457	E2F1	0.008949464305 029796	0.040527723565541 3	11.616755324507999	11.2918286891761
ENSG00000125878	TNK2	chr20	604257	E2F1	0.017723943735 911133	0.040527723565541 3	1.673933016323 19	2.797235087532169 7
ENSG00000206991	TNNC2	chr15	43637632	MYC	0.048094943511 2011	0.049940846209705 44	1.030321183687 4349	0.0
ENSG00000116981	TNRC6A	chr1	39651229	MYC	0.034067041229 1665	0.048899016015033 09	0.0	1.183373890980429 9
ENSG00000077009	TOLLIP	chr19	3933069	AR	0.026764245313 653835	0.049279630886828 7	-1.485257423612 095	2.1529557 4734272
ENSG00000243297	TPX2	chr19	36901742	AR	0.020891512093 319325	0.049295679398841344	0.0	0.955171458479002 1
ENSG00000243297	TPX2	chr19	36901742	FOXA1	0.020891512093 319325	0.049295679398841344	0.0	0.955171458479002 1
ENSG00000188868	TRADD	chr19	12317477	AR	0.029174449950 973628	0.049279630886828 7	8.0773024599886 51	7.762090386063875
ENSG00000131845	TRIM16L	chr19	57351271	MYC	0.015979732527 02036	0.048899016015033 09	9.2703320545306 47	9.466990563151155
ENSG00000134644	TRIT1	chr1	30931506	E2F1	0.039799418714 41024	0.046310455586640 424	12.368582944514202	12.543123 1043357
ENSG00000070047	TSPAN15	chr11	576470	MYC	0.009684053361 991634	0.048899016015033 09	7.9022768300684 39	8.649406104152039
ENSG00000121410	TSPAN6	chr19	58345178	MYC	0.035759259374 657175	0.048899016015033 09	3.6368744908895 7	2.103036069 00217

ENSG00000101442	TSPAN6	chr20	38748460	CTCF	0.033484485293732505	0.044645980391643 34	8.293235510162 39	7.91662804230316
ENSG00000082213	TSPOAP1	chr5	31532287	CTCF	0.041142235977 95097	0.049125641551331 5	10.0842 4619882 68	10.357315 277181
ENSG00000215915	TTC22	chr1	1449689	AR	0.026596318589 88648	0.049295679398841 344	8.61789 3462634 55	7.08698948202371
ENSG00000215915	TTC22	chr1	1449689	FOXA1	0.026596318589 88648	0.049295679398841 344	8.61789 3462634 55	7.08698948202371
ENSG00000168297	TTLL12	chr3	58332880	MYC	0.014172264385 036765	0.048899016015033 09	10.1869 1327538 86	9.744465659243911
ENSG00000150773	TXLNA	chr11	112063218	AR	0.045663563442 94926	0.049279630886828 7	2.58506 1290190 2004	5.651903605842484 5
ENSG00000207217	TY-ROBP	chr7	6016877	FOXA1	0.027493476186 127585	0.040371258891333 33	0.0	1.52080089890583
ENSG00000249673	UBA6	chr4	2934882	E2F1	0.026596318589 88648	0.040527723565541 3	9.38567 4924104 58	8.97376057716742
ENSG00000132950	UBTF	chr13	19823482	MYC	0.039214008292 22933	0.048899016015033 09	8.81973 6624577 565	8.632632896706895
ENSG00000137996	UGGT2	chr1	100266216	MYC	0.049940846209 70544	0.049940846209705 44	10.3492 5864030 3	10.5251087730266
ENSG00000232082	USE1	chr6	166460663	E2F1	0.008993860264 66373	0.040527723565541 3	0.0	1.556722690560370 2
ENSG00000222691	USP14	chr4	107867807	MYC	0.018200383097 1342	0.048899016015033 09	-0.062079913159 0865	0.0
ENSG00000066136	VDAC3	chr1	40691648	AR	0.010017283373 309954	0.049279630886828 7	10.7945 5380641 35	11.008713006525301
ENSG00000149929	VMP1	chr16	29992330	AR	0.002795999480 998655	0.034276303279720 144	9.45922 4049802 438	9.2290658713034
ENSG00000177728	VRK1	chr17	75441159	AR	0.013765208838 55077	0.049279630886828 7	10.9483 5239370 13	11.252103 1396268
ENSG00000132915	WAC	chr5	149857953	MYC	0.034228677227 91284	0.048899016015033 09	4.31031 9102227 83	4.895686144959558
ENSG00000242818	WDR47	chr10	119768247	AR	0.013301750912 849838	0.049295679398841 344	0.00564 9899650 18053	0.0
ENSG00000242818	WDR47	chr10	119768247	FOXA1	0.013301750912 849838	0.049295679398841 344	0.00564 9899650 18053	0.0
ENSG00000257218	WDR54	chr12	120446444	FOXA1	0.030305805288 970797	0.040371258891333 33	10.0957 241809	9.819256621971359
ENSG00000103260	WDR54	chr16	715118	CTCF	0.023724375521 784902	0.043015503246591 896	9.70826 6404405 4	9.412173456481119
ENSG00000234705	WIPI1	chr9	128663129	MYC	0.047434225098 551504	0.049940846209705 44	4.71198 1987183 42	4.28557665460278
ENSG00000103485	XPO1	chr16	29663279	AR	0.048123711027 665966	0.049295679398841 344	10.1475 8687599 329	8.486548176282021
ENSG00000103485	XPO1	chr16	29663279	FOXA1	0.048123711027 665966	0.049295679398841 344	10.1475 8687599 329	8.486548176282021

ENSG00000234797	XRN2	chr15	59768352	AR	0.025565749689383292	0.049295679398841344	8.5075198523004	8.060054092096365
ENSG00000234797	XRN2	chr15	59768352	FOXA1	0.025565749689383292	0.049295679398841344	8.5075198523004	8.060054092096365
ENSG00000204970	XYLB	chr5	140786136	MYC	0.007095771905767007	0.048899016015033 09	5.703847921469515	2.4010475392664
ENSG00000168032	YBX3	chr3	40387184	MYC	0.005556784285468106	0.048899016015033 09	8.159929412852781	7.509116077949 27
ENSG00000109906	YY1	chr11	114059041	AR	0.040749180012 52758	0.049295679398841344	11.220406990464099	9.951449215258961
ENSG00000109906	YY1	chr11	114059041	FOXA1	0.040749180012 52758	0.049295679398841344	11.220406990464099	9.951449215258961
ENSG00000103502	ZBTB32	chr16	29858357	AR	0.007639545339591015	0.049295679398841344	11.619725086470401	11.318399331656302
ENSG00000103502	ZBTB32	chr16	29858357	FOXA1	0.007639545339591015	0.049295679398841344	11.619725086470401	11.318399331656302
ENSG00000184445	ZC3H15	chr12	122527246	AR	0.022372202881 60119	0.049295679398841344	7.313080166446215	8.167539077132567
ENSG00000184445	ZC3H15	chr12	122527246	FOXA1	0.022372202881 60119	0.049295679398841344	7.313080166446215	8.167539077132567
ENSG00000254986	ZMYND10	chr11	66480013	ESR1	0.007364100923154405	0.029456403692617 62	10.107265549633949	10.568673269131 8
ENSG00000274443	ZMYND11	chr8	73241331	MYC	0.037748195385 37311	0.048899016015033 09	1.281846295061 0845	3.300561884366145
ENSG00000134817	ZNF195	chr11	57233577	AR	0.005468433527832356	0.047526569056108	6.076384490088 65	7.656451476710299
ENSG00000135409	ZNF195	chr12	53423855	MYC	0.043262937541 51935	0.049686196501489 62	5.128554866256 86	3.981565931950 45
ENSG00000228146	ZNF195	chr16	3144015	CTCF	0.040749180012 52758	0.049125641551331 5	3.964770817404095	1.806976939698065 3
ENSG00000138160	ZNF263	chr10	92574105	FOXA1	0.017441903911844983	0.040371258891333 33	7.280216081793 6645	8.334695773737 87
ENSG00000214881	ZNF275	chr10	68544489	E2F1	0.017792049974212253	0.040527723565541 3	1.359200400466 95	0.895996854024815 5
ENSG00000170266	ZNF280C	chr3	32996609	AR	0.036390610072 08641	0.049279630886828 7	10.792496446059301	11.053760 9879132
ENSG00000254911	ZNF302	chr11	93721513	AR	0.049351157209831985	0.049351157209831985	-0.710263359138 2399	0.0
ENSG00000254911	ZNF302	chr11	93721513	FOXA1	0.049351157209831985	0.049351157209831985	-0.710263359138 2399	0.0
ENSG00000086827	ZNF500	chr11	113733187	AR	0.036290618541 40877	0.049279630886828 7	8.709575058813485	9.409102338291456
ENSG00000005243	ZNF582	chr17	48026167	AR	0.027319202105005273	0.049295679398841344	9.050925446267 07	8.530724416651 81
ENSG00000005243	ZNF582	chr17	48026167	FOXA1	0.027319202105005273	0.049295679398841344	9.050925446267 07	8.530724416651 81

ENSG00000066136	ZNF638	chr1	40691648	AR	0.010017283373309954	0.049295679398841344	10.794553806413535	11.00871300652530 1
ENSG00000066136	ZNF638	chr1	40691648	FOXA1	0.010017283373309954	0.049295679398841344	10.794553806413535	11.00871300652530 1
ENSG00000167470	ZNF839	chr19	1248553	E2F1	0.043262937541 51935	0.046310455586640424	12.585917615215 55	11.66450807691980 2
ENSG00000142188	ZNF85	chr21	33432485	MYC	0.030539638893 99325	0.048899016015033 09	9.224587301374 99	8.951968735531965
ENSG00000177943	ZRANB1	chr9	136850943	E2F1	0.043400927799 53262	0.046310455586640424	7.627061035155555	6.299779507715341
ENSG00000160883	ZXDC	chr5	176880869	MYC	0.040749180012 52758	0.048899016015033 09	2.540879636068 99	4.38304652865224