The Identification and Study of Transcription factors and co-factors other than androgen receptor that could have important role in the regulation of gene expression in prostate tumor

Master's Thesis
Mukundan Pillai
Master's in Bioinformatics
BioMediTech
University of Tampere
Finland

ACKNOWLEDGEMENT

I would like to thank my supervisors, Professor Matti Nykter and Dr. Juha Kesseli, in giving me support and guidance in both the practical and writing aspects of my thesis project.

I would like to give my special thanks to my dear parents and my younger brother for their continued support and faith in me for completing the thesis successfully.

I also would thank my friends and colleagues for giving me support!!!

Mukundan Pillai

MASTER'S THESIS

Title	The Identification and Study of Transcription factors and co-factors other than androgen receptor that could have important role in the regulation of gene expression in prostate tumor
Keywords	Prostate Cancer, Androgen Receptors, Regulation of Gene Expression, Transcription Factors, Clustering
Place	University of Tampere, Finland
Author	Mukundan Pillai
Pages	54
Supervisors	Professor Matti Nykter, Dr. Juha Kesseli
Reviewers	Professor Matti Nykter, Dr. Juha Kesseli
Time	3rd December 2018

Abstract:

The overall aim of this study is to computationally identify and analyze the nature of transcription factors other than androgen receptor (AR) that play an important role in the regulation of gene expression and the subsequent progression of the castrate-resistant prostate cancer (CRPC). Anti-androgen therapy would relapse in about 20% of the treated patients leading to the development of drug resistance in CRPC. This leads to the search for other transcription factors and cofactors as suitable drug targets other than AR that have a role in proliferation of the prostate cancer cells in CRPC. The study of these transcription factors could highlight a possible role in determining an alternative treatment approach in contrast to current methods that focus primarily on AR-dependent pathways. These transcription factors also have important role in the common prostate cancer and is not displayed much due to the more active AR signaling.

CONTENTS

1. INTRODUCTION

Prostate cancer is a heterogenous and clinically diagnosed malignancy in males with high variability through progression (Shoag & Barbieri, 2016). It is one of the most known common cancer and is dependent on activation and regulation of Androgen Receptor (AR). Symptoms of prostate cancer include loss of weight, impotence, spinal cord depression, loss of mobility, and severe pain (last three symptoms are related to bone metastasis) (Hamilton, Sharp, Peters, & Round, 2006) (Bubendorf et al., 2000). The metastasis of prostate cancer affects "other organs such as liver, lungs etc" (Bubendorf et al., 2000).

AR is a ligand-activated nuclear transcription factor known to play an important role in the direct regulation of gene expression along with its coregulators in normal and tumor types of prostate tissue (Sinnesael et al., 2012). The function of AR in normal prostate cells is to act as transcription factor that controls the expression of genes associated with physiological development such as differentiation and maturation of primary and secondary male sexual phenotype, male skeletal integrity and initiation and maintenance of spermatogenesis in males. The transcriptional activity of AR is activated through the binding of androgens; testosterone and dihydrotestosterone (DHT); by nuclear translocation. The transcriptional activity is in turn influenced by the interaction and phosphorylation of AR and coregulator factors (Roy AK1, Lavrovsky Y, Song CS, Chen S, Jung MH, Velu NK, Bi BY, 1998)(Srinivas-Shankar U, 2006). In prostate cancer, the cancer cells like the normal prostate cells need AR (undergoes post-translational mutation via acetylation) and its coregulators as they play a key role in the transactivation, growth and survival of the prostate cancer cells throughout the different stages of the cancer (Bardin CW, Brown T, Isomaa VV, 1983)(Chang, 2016). One of the important aspects of AR is that it increases the ratio of cell growth proliferation and decreases the cell death (apoptosis), thus disrupting the normal cell proliferation-cell death ratio. The AR and androgen levels are also shown to be elevated in the cancer cells. The high levels of AR are maintained throughout the different stages of the cancer (Fu et al., 2002)(Denmeade SR, Lin XS, 1996). Hence, the blocking of androgens or the inhibition of the AR transcription factor by targeting the ligand-binding domain helps in the treatment of

the prostate cancer (Helsen et al., 2014). Thus, it is selected as primary drug target for anti-cancer therapy such as androgen deprivation therapy (ADT). However, recent studies have shown that prostate cancer have known to adapt and proliferate under androgen-deprived environment through the web of interactions between other transcription factors, oncogenes, tumor suppressor genes. It also leads to the abnormal reactivation of the AR (Rosa-Ribeiro et al., 2014)(R. S. S. and K. E. Knudsen, 2013).

The behavior of AR in the castrate-resistant prostate cancer (CRPC) is bit different from that of the commonly known prostate cancer. The varied nature of AR is caused to the formation of drug resistance to the anti-cancer therapeutic agents due to mutations such as gene amplifications (overexpression of AR as result of androgenic hormone suppression), gain-to-function mutations, alternate splicing, co-repressor loss of function and occurrence of intracrine androgen. Such mutations lead to uncontrolled cell-cycle progression and proliferation often leading to an incurable condition. This situation is often seen in about 10-20% of the treated patients (Eisermann, Wang, Jing, Pascal, & Wang, 2013)(Manuscript, 2011).

However, the levels of AR in the primary stages of CRPC is low when compared to the later advanced stages. This leads to a possibility of other transcription factors that act as “master regulators” in place of AR and may also modulate with the high mutations of AR. These transcription factors help in maintaining the AR levels necessary for the growth and proliferation of cancer cells through coregulation or independently and may be overshadowed by AR in normal prostate cancer and quite prominent in CRPC. They could be selected as drug targets in addition to AR for treatment of prostate cancer. Understanding these genes helps in understanding the interactions of the signaling networks that primarily initiates the development of the variant prostate cancer and thus, opening the way for better therapeutic methods.

2. REVIEW OF LITERATURE

Transcription factors are important protein molecules that regulate the rate of transcription of gene expressions; transfer of genetic data from target DNA to mRNA (messenger RNA) by binding to the specific complementary target DNA sequence; leading to all the biological and phenotype changes in all organisms (Latchman, 1993). There are around 2600 known TFs in the human genome that change the behavior of the target gene or cell by binding to their proximal or distal regulatory elements (Vaquerizas, Kummerfeld, Teichmann, & Luscombe, 2009). These process is important so that the specific gene is either turned "off" or "on"; downregulated or upregulated respectively; in order "to make sure that they are expressed in the right cell at the right time and in the right amount throughout the life of the cell and the organism". TFs either act alone or coordinate together with other TFs and proteins to form complex molecules in order to "promote or repress" the activity of the RNA polymerase enzyme to the target genes (Roeder, 1996) (Lee & Young, 2000). Transcription factors are of clinical significance because their activities; caused directly or indirectly; due to mutations could be associated with the development, cell signaling and cell cycle of specific diseases or disorders and hence potential drug targets such that acting on these specific TFs could help in regulating or combating the disease without affecting other important parts of humans and other organisms (Lee & Young, 2013). In association with cancer, TFs are either tumor suppressors (example: p53 gene) or oncogenes (example: HER2/neu gene) due to mutations or mis-regulatory activities (Dasgupta, Srinidhi, & Vishwanatha, 2012) (Zerbini, 2006). Hence it is great importance to know as many details as possible of TFs regulatory network in diseases for better chances of clinical treatment.

Androgens are sex hormones that are responsible for the development and the maintenance of the male sexual characteristics (development of primary sex organs and initiation of secondary sexual features at the onset of puberty). Testosterone and Dihydrotestosterone (DHT) are the main androgens present in males and mainly active in prostate cells. Testosterone is produced in the Leydig cells in the testis and DHT is produced with the conversion of testosterone by the enzyme 5 alpha-reductase which enables DHT to bind to AR (Debes & Tindall, 2002) . The other function of androgen includes its role as "extracellular signals for the development as well as the

functional activity of the prostate epithelial cells" (Gao et al., 2003). The activities of androgens are regulated by the transcription factor, Androgen Receptor (AR).

Androgen Receptor (AR) is a nuclear receptor and a ligand-regulated transcription factor that plays a critical role in the growth and development of male physiological characteristics particularly in the prostate gland development. AR gene is located on the X chromosome. It is expressed on prostate cells and on many tissues such as bones, muscles, adipose tissues, cardiovascular etc (Davey & Grossmann, 2016). AR is activated by binding to the male sex hormones, testosterone and dihydrotestosterone, in the cytoplasm. The binding causes the AR to undergo conformational change that initiates the dissociation of an inhibitory heat-shock protein complex and the translocation of AR complex to the nucleus and its consequent interactions for AR-target gene regulations with the help of multiple cofactors. Mutations in AR is one of the main contributor to the progression of prostate tumor (Jemal et al., 2006) (Norris et al., 2009). The mutated activation of these genes results in the progression of the prostate tumor cells and the metastasis of the prostate cancer.

Many transcription factors and co-factors along with androgen receptor plays a major role in the development of prostate cells together with AR by taking part in the AR signaling process by enabling the opening of chromatin regions to allow RNA polymerase II to target gene promoters (Wang, Carroll, & Brown, 2005). A study of the reprogramming of the AR binding sites in the prostate tumorigenesis has shown that the transcription factors; FOXA1 and HOXB13; in cooperation with AR is relevant for the transformation of normal tissue to tumor in prostate cancer (Pomerantz MM, Li F, Takeda DY, Lenci R, Chonkar A, Chabot M, Cejas P, Vazquez F, Cook J, Shivdasani RA, Bowden M, Lis R, Hahn WC, Kantoff PW, Brown M, Loda M, Long HW, 2015).

2.1. Androgen and Prostate Cancer (HDPC)

Androgen Receptor is one of the main driving force behind prostate carcinogenesis. It is responsible for the proliferative, apoptotic and angiogenic mechanisms in the tumor cells. It has been proven that different cellular mechanisms and pathways influences prostate cancer growth

in an androgen-independent environment (Debes & Tindall, 2002). AR mediates the transcription of genes such as UBE2C and cyclin D by binding to the distinct DNA sequences that result in tumor proliferation (Xu, Chen, Ross, & Balk, 2006). Pioneer factors, for example FOXA1, enable the AR to interact with chromatin and in turn interact with coactivators and corepressors which either promote or repress transcription.

AR is expressed in both Hormone-Dependent (HDPC) and Castrate-Resistant (CRPC) prostate cancers though it is present in low amounts in the primary stages of CRPC but increases its gene expression in the advanced metastasis stages. Prostate Cancer treatment focuses primarily on stopping the activity of androgen. An example consists of androgen deprivation therapy (ADT) by inhibiting the production of the androgens by “castration or using the use of androgen antagonists such as bicalutamide etc. Another example of treatment is surgical castration. 80% of the patients undergoing the treatment get cured from the tumor but there is a small chance that in the remaining 20% of the patients, the prostate tumor cells become unresponsive to the treatment and develop into hormone-independent prostate cancer when AR activity is slowly reactivated (K. E. Knudsen & Penning, 2010) (Xin Yuan & Balk, 2009) .

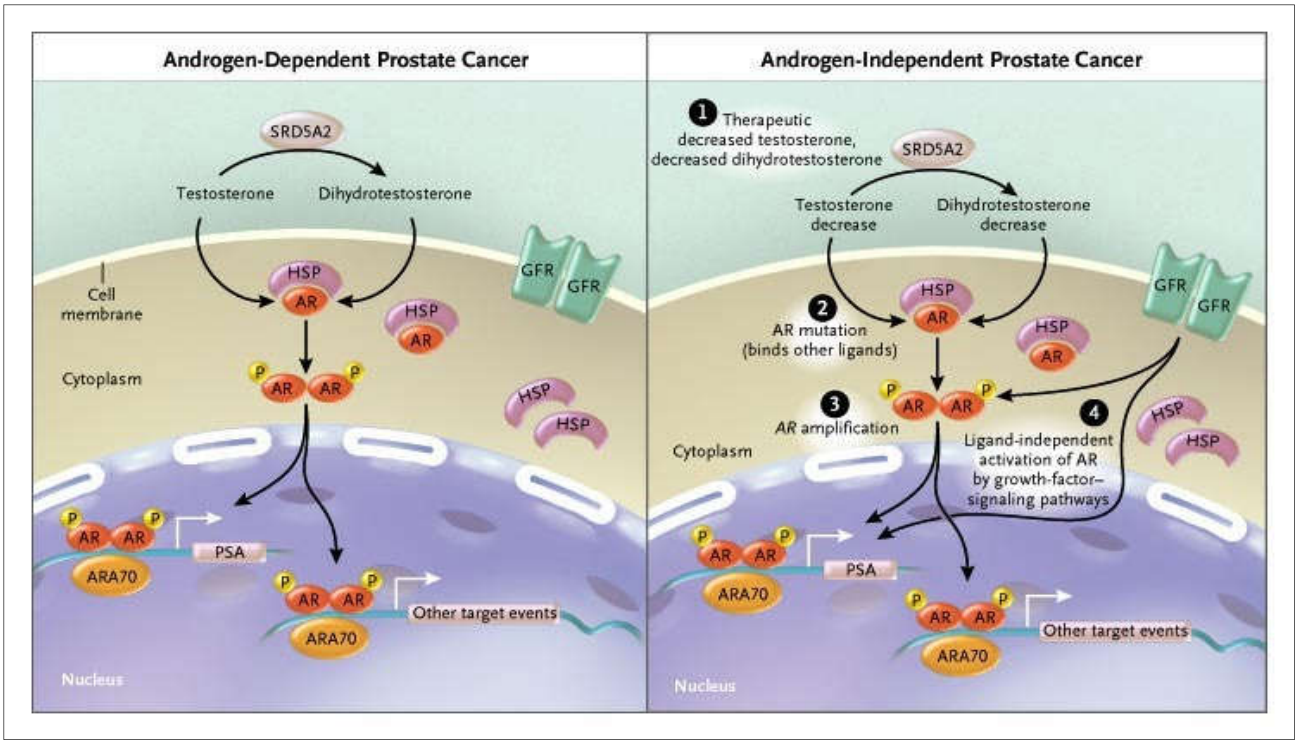


Figure 1: Visualization of Androgen Receptor interaction in the cancerous prostate cells of both Hormone-Dependent Prostate Cancer (HDPC, also known as Androgen dependent Prostate Cancer) and Castrate-Resistant Prostate Cancer (CRPC, also known as Androgen-Independent Prostate Cancer) (Nelson, De Marzo, & Isaacs, 2003)

2.2. Castrate-Resistant Prostate Cancer (CRPC)

AR is an important component of prostate cancer for its growth. The CRPC is the form of prostate cancer that arises when after the treatment methods like androgen deprivation therapy blocks AR transcription and cures the HDPC. CRPC adopts several alternative mechanisms for the tumor growth in absence or low expressions of androgens or AR. In CRPC, the AR is overexpressed through gene amplification (see Figure 1) or AR protein overexpression and allows the binding to chromatin structure in concentration lower than normal levels observed in normal prostate and HDPC (Carreira et al., 2014) (Urbanucci et al., 2011) (Cai et al., 2011). This initiates AR reactivation in the CRPC cells and enables the tumor cells to grow in low concentrations of AR. Other mechanisms involve AR mutations such as gene/RNA splicing, intra-tumoral androgen synthesis (X Yuan et al., 2014), aberrant activation of AR without androgens and mutations of AR co-regulators. AR splicing consists of expression of “C-terminal variants” called ARvs which are active

in mediating AR-regulated transcriptions and initiates tumor proliferation under castrate conditions (X. Zhang et al., 2011). In CRPC, AR follows a distinct transcription pathway different from the one in HDPC. It involves the direct activation, regulation and overexpression of genes related to M-phase of cell cycle “such as CDK1 and UBE2C” (Wang et al., 2009). This suggest an epigenetic reprogramming in CRPC in presence of low levels of AR and decrease of genes that are normally regulated by AR in normal and HDPC. This is the necessary driving force behind the need to find alternative treatment methods that do not depend on androgen and AR and hence, finding new potential drug targets. Successful targets are those that needs to be present in both forms of the prostate cancer and tend to follow same transcription program.

2.3. Gene Expression Arrays

Gene expression arrays or microarrays are one of the new technologic approaches available for modelling the expression profile of thousands of genes of interest in large amounts. This feature together with its “reproducibility of gene array studies” makes it of great practical importance in molecular biology research. More targeted gene arrays could be constructed for measuring small number of genes thus, increasing the specificity (Merker & Arber, 2010). Microarrays consist of collection of nucleic acids (DNA) attached to a solid surface such as glass slide or silicon surface and is used for detecting and measuring the expression level of target genes/nucleic acid sequences via probe hybridization. The nucleic acids present on the solid surface are called as probes which are either cDNA or oligonucleotides (Bumgarner, 2013).

Microarrays produce high-quality data and are useful for monitoring gene expression levels to help researchers to study various biological questions such as;

- Genetic interactions and study of molecular functions (Breitling, Amtmann, & Herzyk, 2004),
- Differential Expression studies
- Genome Mapping
- Mutations,
- Disease classification and pathogenicity (Miklos & Maleszka, 2004),
- Drug-Target response,

- Identification of new genes/variants (Shoemaker et al., 2001),
- Construction and analysis of metabolic/signaling pathways (Roberts et al., 2000)

Gene Expression Arrays or Microarrays depend on the manufacture of arrays, techniques implemented, probe selection, quantification of hybridization and data analysis for good reproducible results. Differentially expressed genes are detected using one- or two-color channel when comparing the experimental data with the reference sample. The channels are based on labelling the probes with fluorescent tags for identification and visualization of target genes to determine whether they are regulated positively or negatively (Liu, Bebu, & Li, 2010).

The workflow of Microarray or Gene Expression Arrays involves the following;

1. Extraction of RNA of target gene from biological sample,
2. Conversion of RNA sample to cDNA with labelling through reverse transcription,
3. Hybridization of the labeled cDNA to complementary probes,
4. Removal of unhybridized cDNA,
5. Detection and measuring expression levels of hybridized cDNA,
6. Preparing the quantitative data for data analysis and comparison with other samples ("The Basics: What is a Gene Array?," 2018) (Govindarajan, Duraiyan, Kaliyappan, & Palanisamy, 2012).

2.4. Computational Analysis of Microarray Data

Major biological breakthroughs such as the Human Genome Project helped us in reshaping our understanding of the genome of humans and other organisms. Many techniques were developed in analyzing this ever-growing biological data such microarrays to build gene expression patterns and metabolic networks based on cellular responses, phenotypes, sample conditions etc. (Quackenbush, 2001). Gene Expression Arrays produce enormous amounts of biological data and it cannot be directly interpreted due to the background differences, sample conditions, time points etc. Hence, processing the information is required and use of an algorithmic approach to make the data more understandable. The steps involved in the analysis are as follows;

- Data Preprocessing: This step, prior to analysis and result interpretation, involves the extraction of "meaningful" information from the gene expression data (raw data obtained from microarray) and its subsequent preparation for data analysis. Normalization, determination of spot intensities and background correction are some of the types of preprocessing (Tarca, Romero, & Draghici, 2006).

- Filtering: Removal of non-variant genes/samples from the data.

- Identification of differentially expressed genes (Loguinov, Mian, & Vulpe, 2004).

- Classification or Class Prediction (Supervised learning): The approach allows the assignment of data to known classes using algorithms such as K-Nearest neighbor (kNN), Artificial Neural Network etc. This step involves training of the algorithm with the samples (training set) and testing of the algorithm on a separate group of samples (test set) (Selvaraj & Natarajan, 2011)

- Clustering Methods (Unsupervised learning): The data analysis includes the discovery of new subgroups based genes/samples with common characteristics (Class Discovery) (Tarca et al., 2006). Clustering is one such method that aims on grouping the "clustering of genes or samples with similar expression profiles or relationship" together based on distance/correlation measures to allow for meaningful interpretation of the data. Clustering methods are of two types: Hierarchical (based on relationships) or Non-Hierarchical (not based on relationship).

  Hierarchical Clustering: the algorithm involves grouping together of small clusters of similar genes/samples within larger clusters of less-related genes/samples. It could be agglomerative (bottom-up approach where "each observation begins in its own clusters and merges with other clusters as it moves up the hierarchy") or divisive (to-down approach where "all observations begin in one cluster and splits into other clusters as it moves down the hierarchy). Pairwise distance measures are used for calculating objects

into clusters and includes single linkage clustering, complete linkage clustering etc. (Babu, n.d.).

Non-hierarchical Clustering: the algorithm involves determination the number of clusters required to be created and placing the observations with similar characteristics into those clusters. The two known types of non-hierarchical methods are K-means and SOM (Self Organizing Maps).

K-Means Clustering: The objects are placed in predetermined number of clusters and the centroid for each cluster is calculated. The objects are rearranged from one cluster to another "depending on which centroid is closer to the object. The centroid calculation and rearrangement is performed in an "iterative manner for fixed number of times" or until no more rearrangements are possible.

Self-Organizing Maps (SOM): The algorithm is similar in nature to K-means, but "the number and pattern of clusters" are selected with respect to each other. The SOM consists of a two-dimensional map or expression space filled with nodes (clusters). Each object is then moved to the node nearest to it. Next, a "random node is selected and the node in the locality of the object is moved closer to it". "The other nodes are moved to a small extent depending on how they are to the selected object". The process in repeated with randomly selected objects, the node positions are refined, and the region of locality becomes confined. The map is "deformed" to fit the data. SOM does not force the clusters to be equal to the number of starting nodes. SOM can produce good results even with noisy data (Esa Alhoniemi, Jaakko Hollmén, 1996) (Babu, n.d.).

- Result interpretation and data visualization (Loewe R.P., 2011).

The data profiling enables the study of gene expression patterns, regulatory networks, metabolic pathways, disease progression or mutations (Zhang Y., Szustakowski J., 2009) (Hanai, Hamada, & Okamoto, 2006). Apart from data analysis, other factors are present which is what makes computational analysis of microarray data more essential. The factors include experimental design of data (complexity of the array data) (Churchill, 2002), annotation (Barbosa-Morais et al.,

2010) and data storage. Many tools are available for microarray data analysis including Bioconductor project for the “comprehension of genomic data” based on R-Programming language. Hence, computational analysis is an important tool for biological researchers. The strategy for data analysis is determined by the purpose of the microarray experiment and biological question to be answered (Quackenbush, 2001). The results from the microarray analysis could be combined with other approaches to make new discoveries and information such as identification of transcription binding sites, protein-protein interaction networks/pathways, Gene set enrichment analysis (GSET) etc. (Selvaraj & Natarajan, 2011).

In recent years, Gene Expression Arrays plays an important role in cancer research in the form of identification of pathways, oncogenes and new therapeutic targets, molecular diagnosis and genome mapping of the tumor sample to identify signaling molecules and other factors that aid in tumor growth and differentiation (Brentani et al., 2005).

3. AIMS & OBJECTIVES

The aim of the study is focus on the data based on a previous research conducted on androgen regulation and to expand the scope, i.e., to identify and study the genes clusters that may contain transcription factors that may have important regulatory roles parallel to that of androgen receptor in the prostate cancer growth and survival.

The objectives of the study are as follows:

1) Download and process the data for analysis.
2) Filtering out non-informative genes.
3) Clustering of genes and detection of clusters with informative expression profiles (early up, early down, late up, late down, first up then down).
4) Detection of transcription factors in the gene clusters that are relevant for prostate cancer.
5) To identify transcription factors in early cluster groups that have the potential to regulate genes in the late cluster groups.

4. MATERIALS & METHODS

The primary tool used in the extraction and analysis of the data is R programming and the bioconductor packages.

4.1. GEO Dataset Source

The data (accession number: GSE18684) for the analysis was obtained from Gene Expression Omnibus (GEO), an international public genomics data repository from National Center of Biotechnology Information (NCBI). GEO stores and allows free download of microarray data, next-generation sequencing and other types of genomic data sets. NCBI is a part of the United States National Library of Medicine (NLM) and provides access to various types of information related to biotechnological, health and biomedical fields. The data was downloaded using the GEOquery, a bioconductor package used in extracting data from GEO database (Massie et al., 2011).

4.2. Experimental Design of GEO Dataset

The dataset includes the comprehensive analysis of “androgen regulated gene expression in LNCaP prostate cancer cell line”. LNCaP cells are androgen-sensitive prostate adenocarcinoma cells often used in prostate cancer research. The dataset was part of a study that was aimed to identify Androgen receptor (AR) regulated genes with the Illumina beadarray study of androgen regulated gene expression with AR ChIP-sequencing data. “The steroid-depleted growth medium in the ChIP analysis used for culture of LNCaP cells include the Roswell Park Memorial Institute (RPMI) medium which contains glucose, salts, amino acids, pH indicator and vitamins supplemented with 10% charcoal dextran stripped FBS for 72 hours prior to addition of 0.01% ethanol (vehicle control) or 1nM R1881 (treated). 1% formaldehyde was used for the cross-linking of DNA protein interactions for 10 minutes at room temperature and final addition of 125mM glycine for strengthening the interactions. 48 RNA samples were grown for 72 hours comprising of three groups: 3 time-zero samples, 10 ethanol (vehicle control) treated samples at 2h, 4h, 8h, 12h and 24h with biological replicates, 36 R1881 (synthetic androgen) treated samples for every

30 minutes for 4h followed by every hour until 24h with biological replicates at 1h, 2h, 4h, 8h, 12h, 16h, 20h and 24h. The RNA samples were extracted from ChIP using trizol and isopropanol preparation and quantification was performed using an Agilent Bioanalyser. Illumina TotalPrep RNA Amplification Kit was used in cRNA preparation and biotin labelling. Standard Illumina protocols were executed for hybridization and scanning procedures".

"The experiment involved in the construction of the target dataset used the Illumina HumanWG v2 BeadArrays. The array consists of two replicate sections that are treated as technical replicate arrays due to the small systematic shifts present between the sections that is addressed in the normalization resulting in 96 RNA samples. Data analysis was performed using beadarray software from the raw-bead level with the detection, automated removal (BASH) of spatial artefacts followed by manual curation. The dataset obtained as a result with reduced data was summarized with outliers removed to obtain a mean log-intensity and standard error for each probe/array combination. Probes that did not match to transcripts and those without any "signal above background" were removed. In the original study, a Biotrove Realtime PCR was used to confirm the AR binding sites and gene expression changes using primers, SYBRgreen chemistry (Applied Biosystems, 2x SYBRgreen master mix) in an ABI7900 instrument. The final dataset consists of 17182 probes" (Massie et al., 2011).

4.3. Autocorrelation Method to Filter Probes

The study involved in the mapping of the androgen regulated genes and construction of GEO dataset used autocorrelation at lag 1 was used as a measure to detect probes that showed a systematic smooth change over time without prescribing a form for that change. Autocorrelation is the "correlation of a time series with lags of itself". It shows how the previous state/condition (lagged observations) has an influence on the current state/condition. If the autocorrelation crosses a certain threshold it means that the specific lag is significantly correlated with the current state. It is also used to determine if the given time series is stationary or not depending on whether the value becomes zero quickly or gradually respectively. Autocorrelation in the analysis was used to identify probes where the "neighboring time-points" are more

homogeneous than the heterogeneous time points. This allowed the identification of probes that showed all smooth and systematic gene expression changes irrespective of the shapes of the expression profiles. 100 sets of observations were simulated from the known means and standard errors to calculate the autocorrelation and mean of each set, considering the variability in the measurements (Taylor, 1990). In the experiment, a cut-off value of 0.5 was suggested to have low false-discovery rate. The same cut-off value was used in the filtering of probes that showed androgen response that have > 0.5 to obtain the desired set of clusters that show the desired expression information profiles. Thus, resultant dataset consisted of 3687 probes (Massie et al., 2011).

4.4. R Programming & Bioconductor Project

R is a special programming language and free computational system for "statistical computing and graphics". It consists of a run-time environment with "graphics, a debugger, access to certain system functions and able to run programs stored in script files". The fundamental behind R is an "interpreted computer language "which allows branching, looping and modular algorithms using functions. R consists of functionality for many statistical procedures which "includes linear and generalized linear models, non-linear regression models, time-series analysis, classical parametric and nonparametric tests, clustering, smoothing and a flexible graphical environment for the implementation of data presentations. Special Add-on packages are available for additional distinct functions. R is widely used for statistical analysis, scientific computation, visualization and data mining. R is developed for Unix-based, Macs and Windows operating systems. R is available as a free software with command line interface distributed under GNU General Public License and its source code is a part of the GNU project (R Core Team, 2018). R is also available as several graphical user interfaces such as Integrated development environment and RStudio (Leonard, 2013).

Bioconductor Project is an open source development software which provides tools "for the analysis and comprehension of high-throughput genomic data. It follows the R programming language. It also includes meta-data packages consisting of pathways, organisms, microarray and

other annotations". Its packages also include "the analysis of DNA microarray, sequence, flow, SNP, and other data". There are approximately 934 packages available. It was started in 2001 and is maintained by several international institutions (Huber et al., 2015) (Gentleman et al., 2004) .

R and Bioconductor project which consists of packages for analyzing biological data is used extensively and considered fundamental by computational biologists for research. "The rich set of inbuilt functions enables for high-volume analysis, statistical simulations and high-quality graphical output at all levels of the research within R along with development of ideas for answering biological questions". In addition, "the packaging system" also makes it ideal for codes to be shared (Gentleman et al., 2004) (Eglen, 2009).

R programming with packages from Bioconductor project was used for the analysis of the microarray data. R was used for data preprocessing, gene filtering, clustering, extraction and identification of transcription factors. Some of the prominent bioconductor packages used in this study include Genefilter, ggplot2 and illuminaHumanv2.db.

4.4.1. Genefilter Package

A Bioconductor package consisting of methods for filtering genes from high-throughput experiments. Variance-based filtering (var.filter) was used to remove genes with low or no variance from the gene expression data (R. Gentleman, V. Carey, 2018). A variance cut-off value (var.cutoff) of 0.1 was used, based on multiple runs, to obtain the 3318 probes that matches approximately with the number of probes obtained in the original study.

4.4.2. ggplot2 Package

The ggplot2 package enables R to visualize the extracted information from the data for better understanding in form of line plots, scattered plots etc. It creates "publication-quality statistical graphics in an efficient, elegant, organized and coordinated manner" (Ito & Murphy, 2013).

4.4.2. illuminaHumanv2.db Package

The illuminaHumanv2.db is an annotation package used for mapping genes to specific features of interest to identify the nature of the data and is also used as means for purging data that do not have annotation during data preprocessing (Mark Dunning, 2015).

4.5. Data Preprocessing

The time-series experimental data was preprocessed by removing the biological replicates by using average means. Normalization is conducted by dividing the time-series sample with zero-time value and filtering of the data for non-variant genes was performed using autocorrelation and Genefilter package. The probes were annotated using illumina package (illuminaHumanv2.db)

4.6. K-Means Clustering

K-means clustering is a type of the unsupervised learning cluster technique that classifies together undefined data into K-clusters. Each data point of a given data set is assigned to one of the K-cluster/groups based on feature similarity; i.e. each data point in each cluster are closer to each other than to data points in other clusters. The centroid of each cluster is a collection of features which define the characteristics of each K cluster and each centroid is placed away from each other as much as possible. The data points are associated with each nearest centroid based on feature values and is completed when no more points are present. The number of k-clusters in this study is 14 which is determined after multiple simulations, final number of probes, and finally to reduce unwanted and less significant clusters to save time taken in the analysis. The “ggplot” package was used to line plot the k-clusters to identify prominent k-clusters based on the information profile determined in the original study.

4.7. ENCODE ChIP-Seq Significance Tool

ENCODE ChIP-Seq Significance Tool is a simple flexible web application that helps to analyze and identify “enriched ENCODE transcription factors from a list of genes of transcripts” by comparing with public ENCODE data. It is one of the tools that helps in analyzing data related to biological processes and phenotype expression. The significant features of the tool include as follows;

- Entering multiple gene or transcript lists at a time, thus reducing the analysis time and less errors when rerunning the tool multiple times. The results are displayed alongside their respective lists in the same table for easy comparison.
- Entering own custom-made list of background genes or transcripts and be used to search for gene or transcript lists extracted from “expression microarrays, next-generation sequencing strategies with array-capture methods and assays that do not search entire genome/transcriptome”.
- Any value between 500-5000 base pairs could be selected in the analysis window in either upstream or downstream direction.
- Customization of the analysis to create more understanding of biological processes like the usage of either the whole gene/transcript body, TSS, TTS as the center of analysis window.

The objective of the online tool is to provide more flexibility to run the analysis to suit the methodology and on answering wide range of biological questions without been held back by limitations including no limits in the addition of TFs and cell treatment combinations. It also helps in exploring the potentiality in either TSS or TTS sites.

This tool enables the identification of enriched TFs from one or multiple gene lists. “Transcription binding significance” is calculated using hypergeometric test followed by multiple hypothesis correction using Benjamini and Hochberg (False Discovery Rate - FDR) method. THE FDR method involves the identifying the rate of type I errors in null hypothesis in multiple comparisons and controlling the expected proportion of incorrect discoveries. A minimum false discovery rate is calculated as the q-value to define a transcription factor (< 0.05) as significant.

4.7.1. Hypergeometric Test

Hypergeometric distribution is a discrete probability distribution that describes "the probability of successes in random draws *without replacement* from a finite population of given size containing the objects with specific features in which each draw is either a success or a failure. This is the exact opposite of binomial distribution in which describes probability of success in draws *with replacement*". The Hypergeometric Test uses this distribution to calculate the statistical significance to identify which "sub-populations are over- or under- represented in a sample". The result of each draw is classified into one of the two mutually exclusive categories and the probability of a success changes with each draw as the population decreases (without replacement).

The formula for calculating the hypergeometric distribution is given as follows (Rice, 2007);

$$P = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

where;

- N - the size of the population
- K - the number of success in a given population
- k - the number of observed success
- n - the number of draws
- P - The probability of obtaining k successes
- $\binom{K}{k}$ - The number of combinations of K success taken k at a time

4.8. TRRUST

Development of “genome-scale regulatory circuit models” remain hindered due to the lack of all the information of the “intrinsic complexities of the human transcriptional networks” and due to technical limitations in genome mapping of such networks. Experimental procedures such as ChIP (chromatin immunoprecipitation) and genome sequencing analysis coordinated with together with research literature of cellular context helps in constructing, combining and understanding the complex TFs-target gene interactions on a genome scale.

TRRUST (Transcriptional regulatory relationships unraveled by sentence-based text mining) is a manually curated database of human and mouse TF-target regulatory interactions. A sentence-based text mining approach is used over Medline abstract for by extracting text sentences that is associated with transcriptional regulation and subjected to manual curation. There are around 8,444 TFs-target gene interactions and is currently one of the largest known public database of “literature-curated TFs regulatory interactions” to date. The significance features of the TRRUST database is as follows;

- It has annotations for mode-of-regulation in many known interactions.
- It can perform a “network analysis” of TFs-target genes interactions by incorporating functional or physical protein interaction networks to provide “various systemic context information for facilitating functional interpretations such as target modularity, TF cooperativity and TF-pathway/disease associations”.
- High enrichment of gene-pairs in top-ranked regulatory interactions inferred from high-throughput expression data.

All these features point out on how TRRUST could become an important benchmark for construction and study of complex transcriptional regulatory networks (Han et al., 2018).

In this study, this tool was used for finding transcription factors that have ability to interact and regulate target genes of late clusters (up-late and down-late information profile). Final identified TFs will be crosschecked with the TFs of early clusters for final verification.

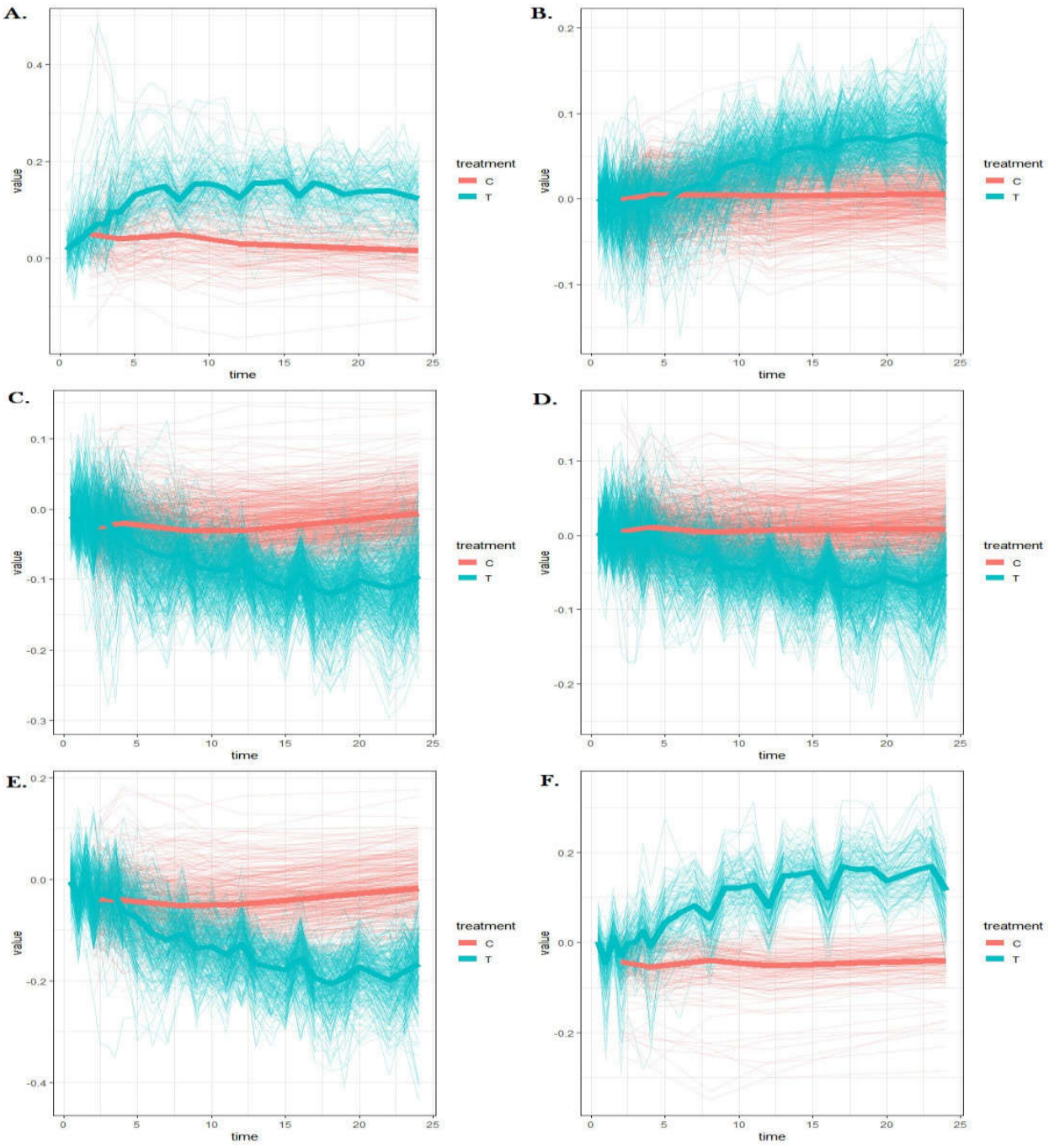
5. RESULTS

5.1. Primary Objective

14 primary clusters were obtained from the K-cluster analysis using R programming. From the primary set of clusters, 10 clusters were further selected based on the information profile obtained from the plotting of these secondary clusters using "ggplot2". The information profile of the selected clusters correlated with the desired information profiles obtained in the original research (Massie et al., 2011). The information profiles of the original study are based on the gene expression changes and were divided into two groups - up and down - based on gene response to androgen stimulation and further subdivided into two more groups - early (4.5 hrs) and late (4.5 hrs) – based on the response time (see Figure 2). These information profiles are considered as primary in this study. In addition, two more groups were constructed to pattern other gene expression changes – first up then down and first down then up. The two new groups were not present in the original study and are referred in this study as secondary information profile. Also, the "no androgen" stimulation from the original study was not considered for present study.

From the second group of 10 selected clusters, the gene list of each cluster is collected and given as input for the online tool – ENCODE ChIP-Seq Significance Tool. This tool enables the identification of transcription factors using False Discovery Rate Method. The following parameters were used for running the algorithm;

- Human genome reference - Hg19
- Regulatory elements - Protein-Coding Genes is selected for analysis
- Feature Information - Gene Symbol (Input is presented in the form of gene symbol list)
- Background Regions - Default option of using all genes/transcript that have the selected ID type as a population for hypergeometric test.
- Analysis Window Parameters - TSS of 5-prime end of the target genes is selected as the center of analysis window. The size of analysis window in base pairs (bp) in both upstream and downstream was selected to be 2000 bp. The reason for selecting 2000 bp is because



- **Figure 2:** Visualization of the gene expression data under androgen stimulation of the 6 clusters with the following information profiles: **A)** Up-Early, **B)** Up-Late, **C)** Down-Early, **D)** Down-Late, **E)** First Up then Down and **F)** First Down then Up**. Figures 2-A, B, C, D** represent primary/linear information profiles while **Figures 2-E, F** represent secondary information profiles.

it has been previously been studied that around 8.5-12.6% of AR binding sites were mapped in within 2kb of TSS (Cheng et al., 2015).

- Cell Lines - Default option of all cell lines since no other suitable ones were available in the given options.

From the output, six out of ten clusters have been identified to interact with several transcription factors (See Figure 2). These clusters also showed the information profiles as expected: up-early, up-late, down-up, down-late, first up then down and first down then up.

First focusing on the 4 clusters with primary information profiles. The first cluster (Up-Early profile) has identified 145 TFs, the second cluster (Up-Late profile) has 179 TFs having the highest number among the clusters while the third cluster (Down-Early profile) has 62 TFs with the lowest number, and finally, the fourth cluster (Down-Late profile) has 173 TFs. About 33 transcription factors were identified to be common between all the 4 clusters; HMGN3, PHF8, SIN3A, TBP, CCNT2, E2F6, POL2, MAZ, ELF1, CFOS, CREB1, E2F1, CHD2, NRSF, TAF1, MXI1, ZBTB7A, RBBP5, YY1, MAX, E2F4, SP1, CoREST, CMYC, EGR1, UBTF, GATA3, FOXA1, SMC3, USF1, ZNF143, HNF4A and BACH1.

In the case of final two clusters with secondary information profiles. The first cluster (First Up-Then Down profile) has identified 24 TFs followed by the second cluster (First Down-Then Up profile) has 79 TFs. About 12 transcription factors were identified to be common in both the clusters; POL2, ZNF263, MAZ, CTCF, RAD21, MAX, EGR1, SIN3A, TAF1, GATA3, YY1 and E2F1.

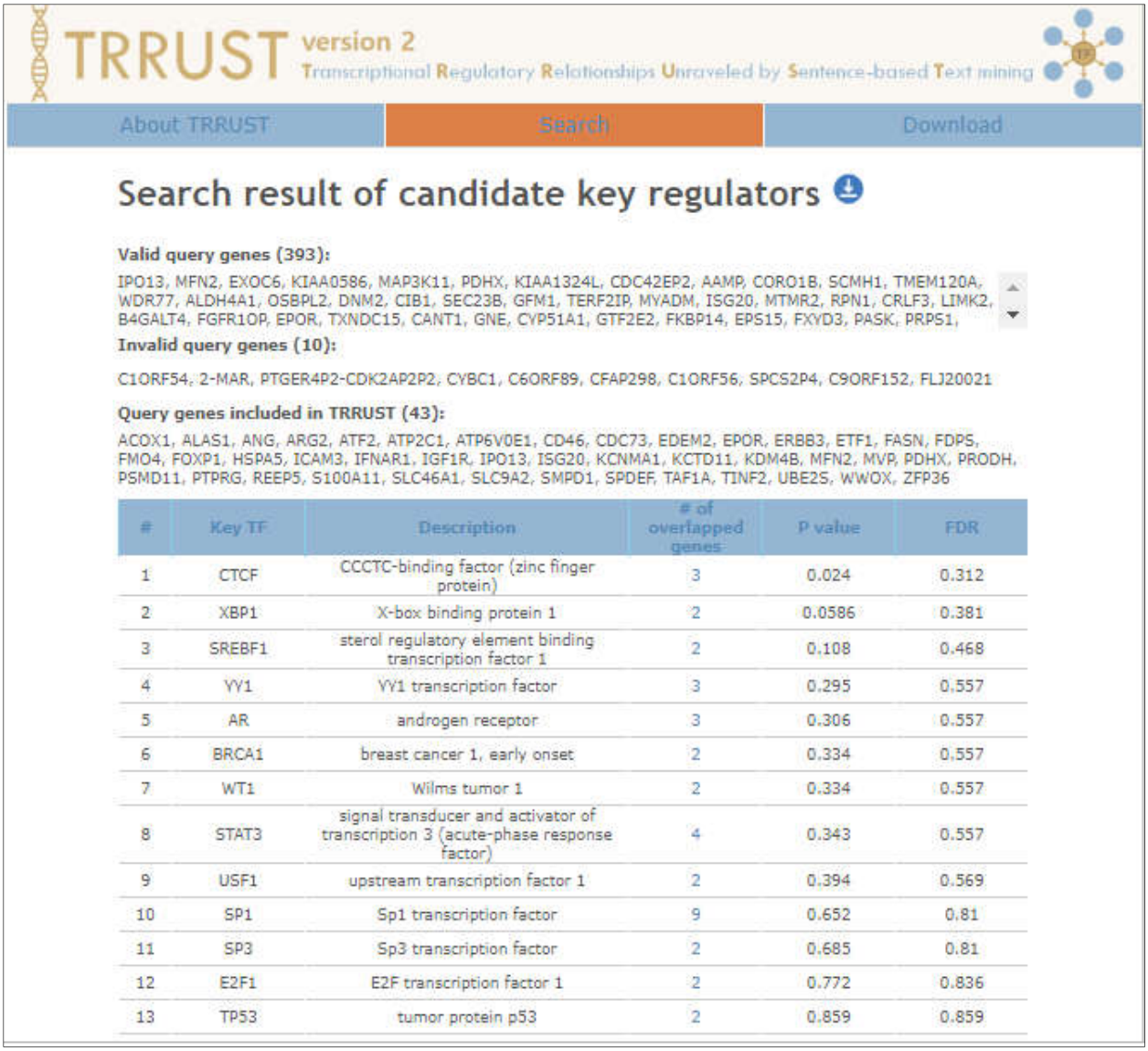
Overall, 8 transcription factors that are common to all clusters were selected due to potential ability to influence the prostate tumor growth and progression. The TFs selected are; POL2, MAZ, CTCF, EGR1, SIN3A, TAF1, GATA3, YY1 and E2F1. Pol2 (RNA Polymerase II) enzyme plays a key role in DNA transcription for almost all the genes and hence is not suitable for potential drug target in terms of specificity. MAX gene was chosen not to be discussed further due to lack of

information to support its role and effect in prostrate tumor. All these transcription factors have shown to have good significant Q-value in the ENCODE Chip-Seq Significance Tool.

Also, three additional TFs were selected; two from the cluster sets of primary information profile and one from cluster sets of secondary information profile. The TFs selected are FOXA1, SP1 and CTCF. The reason for selecting these TFs is because of their previously known relationship to androgen receptor and prostate cancer development. Another reason is due to presence or absence in the cluster sets. On one hand, FOXA1 and SP1 are shown to be active in the cluster sets of primary information profile while they are inactive in cluster sets of secondary information profile. On the other hand, CTCF is active in the cluster sets of secondary information profiles while it is inactive in the cluster sets of the primary information profiles. Both CTCF and FOXA1 are pioneer factors. Pioneer factors are transcription factors that directly bind to chromatin, thus remodeling the chromatin to exert positive or negative regulatory control on transcription by recruiting other TFs, directing histone modification enzymes on target gene enhancer or promoter region and influencing DNA methylation (Cirillo et al., 2002).

5.2. Secondary Objective

TRRUST was used in identifying transcription factors present in early cluster group (Up-Early & Down-Early profiles) that have the potential to regulate the expression of target genes in the late cluster groups (Up-Late & Down-Early profiles) (Han et al., 2018). The list of target genes of late clusters was used as input. The text-mining algorithm obtain the list of transcription factors that have regulatory relationship with the target genes. 13 TFs were found to interact with the query genes of Up-Late cluster (see Figure 3) while 28 TFs were found to interact with the query genes of Down-Late cluster. The TFs are then compared with the TFs of the early clusters obtained in the analysis of the primary objective to find positive matches. The TFs found are CTCF, YY1, USF1, E2F1, BRCA1, HDAC1, STAT1, SP1 and EGR1. Among them, YY1 is seen to be of great potential as it is present in both the late clusters and due to its multifunctional nature.



#	Key TF	Description	# of overlapped genes	P value	FDR
1	CTCF	CCCTC-binding factor (zinc finger protein)	3	0.024	0.312
2	XBP1	X-box binding protein 1	2	0.0586	0.381
3	SREBF1	sterol regulatory element binding transcription factor 1	2	0.108	0.468
4	YY1	YY1 transcription factor	3	0.295	0.557
5	AR	androgen receptor	3	0.306	0.557
6	BRCA1	breast cancer 1, early onset	2	0.334	0.557
7	WT1	Wilms tumor 1	2	0.334	0.557
8	STAT3	signal transducer and activator of transcription 3 (acute-phase response factor)	4	0.343	0.557
9	USF1	upstream transcription factor 1	2	0.394	0.569
10	SP1	Sp1 transcription factor	9	0.652	0.81
11	SP3	Sp3 transcription factor	2	0.685	0.81
12	E2F1	E2F transcription factor 1	2	0.772	0.836
13	TP53	tumor protein p53	2	0.859	0.859

Figure 3: List of transcription factors obtained by text-mining algorithm in TRRUST for gene expression data corresponding to Late Cluster Group (Up-Late Information Profile)

6. DISCUSSION

The nature and relationship of identified TFs of significance with AR in prostate cancer is explained below;

6.1. Transcription Factors of Interest

6.1.1. TAF1

Gene Location: Chromosome X, Band: Xq13.1

TAF1 (TATA binding protein-associated factor 1) is a protein-coding gene that joins with the transcription factor, TBP (TATA binding protein), to form the transcription factor complex known as TFIID (Transcription Factor II D) which in turn is a part of the preinitiation complex of RNA polymerase II. This complex is important for the "coordination of the 70 polypeptides for the initiation of transcription by RNA polymerase II"(Louder et al., 2016). TAF1 and also other members of TAF family initiate "activator-dependent transcription in a promoter and tissue specific manner" (Martel, Brown, & Berk, 2002). TAF1 gene is present on the q-arm of chromosome X and its protein has the special function(s) to act as protein kinase & histone acetyltransferase, mediate ubiquitin-activation/conjugation and phosphorylation in cells. (Dikstein, Ruppert, & Tjian, 1996) (Mizzen et al., 1996) (Pham & Sauer, 2000) . TAF1 was reported to interact and coactivate with AR in prostate cancer cells. On one hand, it represses the transactivation of AR leading to significant decrease in AR levels in the tumor cells while, on the other hand, it increases the transcriptional activity of AR by many levels by controlling c-Jun, MDM2 and cyclin D1 (Cai, Hsieh, & Shemshedini, 2007) (Gaughan, Logan, Neal, & Robson, 2005) (Narayanan, Narayanan, Davis, & Nargi, 2006). The overexpression of TAF1 in CRPC is necessary for the survival and growth of the CRPC by reducing or even ubiquitinate AR in order to prevent the potential targeting by drugs unlike its predecessor form, thus its crucial role (Peyman Tavassoli et al., 2010).

6.1.2. SIN3A

Gene Location: Chromosome 15, Band: 15q24.1

SIN3A is transcription factor with paired amphipathic helix domains necessary for protein-protein interactions (Yu, Thiesen, & Strätling, 2000). It belongs to a corepressor family of Sin3 proteins which are known for mediating the gene expression for the downregulation of transcription activity of many genes (Halleck et al., 1995). Previous research has shown that SIN3A acts as a component of transcription repression complexes and acts as a pathway for interaction between DNA repressors and HDAC (histone deacetylase) enzymes. An example includes the studies showing SIN3A interacting with its corepressors EBP1 and TGIF and in turn enhances the functional ability of both genes (Y. Zhang, Akinmade, & Hamburger, 2005) (Sharma & Sun, 2001). Both the complexes have known to inhibit the transcription of other factors like E2F1 and AR-regulated genes. Thus, SIN3A on an overall scale, is repressor of the prostate cancer growth and differentiation. In castrate-resistant prostate cancer (CRPC), SIN3A-EBP1-HDAC complex reduces the expression of AR and its target genes by decreasing the protein translation of AR. Hence, the exact mechanism is not known. In HDPC, the complex downregulates both by destabilizing mRNA of Androgen Receptor (AR) and stops its protein translation (Zhou, Zhang, & W Hamburger, 2011).

6.1.3. MAZ

Gene Location: Chromosome 16, Band: 16p11.2

MAZ (MYC-Associated Zinc Finger Protein) is a transcription factor that plays an important role in the development and survival of many types of cancers by controlling the upregulation of the gene expression of oncogenes. In prostate cancer, studies have shown that MAZ expression shows positive correlation with AR expression and in higher concentration in prostate tumor samples (W. Luo et al., 2016). The MAZ knockdown by siRNA (small interfering RNA) lead to the inhibition of cell growth, proliferation and arresting of cell cycle. It also resulted in the breakdown of the metastasis and differentiation of the cancer cells. Another interesting factor presented in the study was that when AR was knockdown by siRNA, MAZ expression decreased a bit but when MAZ knockdown occurred, AR expression levels increased. This showed the connection between

MAZ and AR. MAZ was selected as potential drug target in castrate-resistant prostate cancer (Shen, 2013).

6.1.4. YY1

Gene Location: Chromosome 14, Band: 14q32.2

YY1 (Yin Yang 1) is a transcription factor that has dual properties as both activator and repressor of many promoters by directing histone modification in both normal and cancer cells (Inouye & Seto, 1994). YY1 has been known to take part in regulation of genes by acting as co-factor to proteins that take part in cell proliferation, embryonic development, tumor mutation and apoptosis (Sui, 2009) (J. Luo et al., 2013). Upregulation of YY1 leads to disturbing the effects of its target gene and influences tumorigenesis. From in-vitro protein studies and immunoprecipitation analysis, YY1 has been reported to interact with AR and optimizes the transcription activity of AR in normal cells (Z. Deng et al., 2009) but the overexpression of YY1 in prostate cancer cells mainly downregulates AR and its cofactors by interfering the formation of the transcription complex and the subsequent "inverse effect" on target genes of AR regulation (Kashyap & Bonavida, 2014). But YY1 also has a slight positive effect on AR regulation; by suppressing the transcriptional activity of HOXB13 which is responsible for arresting the cell growth in AR-mediated tumor proliferation; by forming a complex with Histone deacetylase 4 (HDAC4) which in turn binds to the two YY1-binding sites present on HOXB13 promotor (Ren et al., 2009). YY1 also plays an important role in the initiation of five transcription-factor regulated signaling networks along with p53, AR, c-MYC and SP1 in both normal and prostate tumor (Sun et al., 2008). YY1 also regulates the transcription of miRNAs (MicroRNAs). In turn as key regulators, miRNAs influence many biological processes and act as tumor suppressors. Excess levels of miRNAs mediate prostate cancer cells "in terms of apoptosis, progression and viability by directly targeting the EGFR, RAC1, ROCK1, interleukin-1 receptor-associated kinase 1, and tumor necrosis factor-associated factor 6". YY1 downregulates the transcriptional activity of miRNA in cancer cells. This was proven in a study by downregulation YY1 and promoting miRNA activity (Huang et al., 2017). It has been proven that the processes exerted by YY1 activity are reversible in nature. The overexpression of YY1 upregulates the AR transcriptional activity in

CRPC to its pre-treated levels. Hence, YY1 has been discovered to a potential drug target in cancer treatment (Sui, 2009) (Zhiyong Deng, Cao, Wan, & Sui, 2010).

6.1.5. EGR1

Gene Location: Chromosome 5, Band: 5q31.2

EGR1 (Early Growth Response-1) is a transcriptional regulator associated with the activation of cell differentiation and mitosis trigger (Guerquin et al., 2013) (Zwang et al., 2011). In prostate cancer, EGR1 has the dual role of both suppressing and progressing tumorigenesis. It is involved in apoptosis and cell growth. EGR1 is present in high concentration (Thigpen et al., 1996) in the tumor cells than in normal cells due to the mutational loss of its repressor NAB2 (Abdulkadir et al., 2001) or mutation of B-RAF gene (Davies et al., 2002). In a hormone-dependent environment, EGR1 interacts with cascading to a series of events from AR translocation to nucleus, transcription of PSA (Prostate-Specific Antigen) and finally to tumor proliferation (S. Yang, 2003). In hormone-independent situation, EGR1 in overexpressed levels enhances the prostate cell growth by interacting with the low AR levels and downregulating AR target genes (Mulholland et al., 2011) (S.-Z. Yang, Eltoum, & Abdulkadir, 2005). Inhibition of EGR1 leads to blocking the metastasis of tumor cells (Mulholland et al., 2011). Hence EGR1 is a potential target for anti-cancer therapy.

6.1.6. E2F1

Gene Location: Chromosome 20, Band: 20q11.22

E2F1 was present in all the clusters ranging from low to high significant levels. E2F1 plays a key role in the" driving S-phase of the cell cycle during cellular proliferation" (Engelmann & Pützer, 2012). The effect of E2F1 in CRPC includes the downregulation of AR factor in prostate cancer cells which is associated with the tumor development (Davis et al., 2006). In a previous study, significant increases in E2F1 expression with its gene regulatory signatures were detected in the "metastatic prostate cancer tissue microarray". Though E2F1 downregulates AR expression, it drives the prostate cancer progression and metastasis by mediating target genes like DNMT1 gene (McCabe et al., 2006). E2F1-DNMT1 complex contributes to the metastasis by "epigenetic silencing of tumor suppressor genes (Valdez, Kunju, Daignault, Wojno, & Day, 2013).

6.1.7. GATA3

Gene Location: *Chromosome 10,* *Band:* *10p14*

GATA3 is a transcription factor encoded by the GATA3 gene which is located on the p14 of chromosome 10. It is responsible for the regulation of expression of known relevant genes (Yamashita et al., 2004), particularly the physiological and pathological immune responses and embryonic development of varied tissues (Barnes, 2018). It is important in cancer studies as an important clinical marker of known cancers particularly breast cancer (Ordóñez, 2013). The GATA3 gene codes for two variant transcripts – GATA3 variant 1 and GATA3 variant 2. There are no reported studies related to the differences between the variants. Both variants are taken into consideration in the clinically studies and laboratory research. All members of GATA transcription factors are essential in the primary stages of prostate cancer development particularly, GATA2 and GATA3 which has shown interactions with AR signaling both in coordination, correlation and regulation of gene expression. However, the removal of GATA transcription factors did not cause decrease in AR-related expression levels but instead increased tumor progression(Nguyen et al., 2013). This shows that these TFs do not actively participate in all the stages of the tumor growth (Xiao et al., 2016).

In all the final clusters, GATA3 is present in all in significant elevated levels. GATA3 is shown to be integral in AR signaling in low risk type breast cancer (Sanga, Broom, Cristini, & Edgerton, 2009). A study has shown that both GATA3 and AR show high correlation in breast cancer. A “unique high level of correlation” between the two TFs (95% confidence interval) compared to other correlations studied. It has also shown that “strong AR expression” levels in lobular and urothelial carcinoma is reported as excellent detection tool for identifying GATA3+ rich setting (Boto & Harigopal, 2018) (Kim et al., 2016). In prostate cancer, GATA3 could plays a more influential role . As explained before, GATA3 is not active in tumor development but in CRPC its role potentially become more prominent. The strong correlation could suggest that GATA3 could potentially be one of the primary factors responsible for the growth and progression of tumor in presence of low level of AR in primary stages of CRPC before the AR levels increase to levels found

in hormone-dependent prostate cancer. Another evidence of GATA3 involvement in CRPC is its interaction with transcription factor regulator – FOXA1, which is considered to be one of important factors involved in the initiation and progression of prostate cancer together with AR and HOXB13 (Pomerantz MM, Li F, Takeda DY, Lenci R, Chonkar A, Chabot M, Cejas P, Vazquez F, Cook J, Shivdasani RA, Bowden M, Lis R, Hahn WC, Kantoff PW, Brown M, Loda M, Long HW, 2015) (Albergaria et al., 2009).

The presence of GATA3 in all the clusters and the evidence of its correlation with AR suggest strongly GATA3 could take the role of AR in the initial stages. Further studies are required in fully mapping out the role and specific function of GATA3 in both the primary stages of hormone-dependent prostate cancer (PCA) and castrate-resistant prostate cancer (CRPC).

6.2. Other Significant TFs

6.2.1. FOXA1

Gene Location: *Chromosome 14,* *Band:* *14q21.1*

FOXA1 is a transcription factor that is primarily act as pioneer factor. FOXA1 (also known as forkhead box A1) by opening chromatin on the target DNA and mediates the passage of other transcription factors like Androgen Receptor (AR) to bind to its target sequence (Augello, Hickey, & Knudsen, 2011) (Sahu et al., 2011). It is also due to its pioneering trait that it is essential in the early development processes such as embryonic growth, initiation of "tissue-specific gene expression and regulation of gene expression in differentiated tissues" (Lupien et al., 2008). Thus, FOXA1 is essential in prostate growth and differentiation due to its important role in the AR activation and regulation of the target genes (Wang et al., 2007) (Gao et al., 2003). FOXA1 expression in prostate cancer are high compared to the normal prostate cells particularly in the CRPC (Gerhardt et al., 2012). A study has shown that FOXA1 regulates AR variants such that the knockdown of FOXA1 results in the reduction of tumor cell growth and thus the activity of AR and its variants. Hence, FOXA1 could be a potential drug target for cancer therapy (Jones et al., 2015).

But in this study, FOXA1 is not active all the time and probably under certain conditions, is not the prime pioneering factor which is instead replaced by CTCF as seen in the clusters of secondary information profile. This is supported by the fact that other studies had detected low levels of FOXA1 expression in CRPC tumors (Jin, Zhao, Ogden, Bergan, & Yu, 2013).

6.2.2. SP1

Gene Location: Chromosome 12, Band: 12q13.12

SP1 are proteins of zinc-finger family that mediates an important role in many cellular processes in both normal and tumor cells. SP1 is present in high concentrations in many cancers including prostate cancer and is used as cancer marker. SP1 binds to the GC rich domain in the promoter region of target genes. It regulates a range of target genes that are "involved in cell cycle, proliferation, cell growth and proliferation and apoptosis. These genes include AR, TGF-Beta, C-MET, PSA, TFIID, TBP, NF-kB etc. These features make SP1 an important therapeutic target by blocking its interactions in prostate cancer (H. Yuan, Gong, & Young, 2005) (T Sankpal, Goodison, Abdelrahim, & Basha, 2011).

In CRPC, SP1 is overexpressed. It is also found to be correlated with the PKM2 (pyruvate kinase isoenzyme type M2). PKM2 is a glycolytic enzyme that plays a key role in tumor metabolism and survival. PKM2 helps the cancer cells to obtain nutrients required for growth and proliferation. Additionally, PKM2 also plays as an "activator and protein kinase" in tumors (Dong et al., 2016). SP1 directly regulates PKM2 and in turn SP1 controls metabolism, growth and autophagy. SP1-PKM2 pathway is repressed by tumor suppressor miRNAs and focusing on this interaction is of great clinical significance for prostate cancer treatment (Ling et al., 2017).

6.1.3. CTCF

Gene Location: Chromosome 16, Band: 16q22.1

CTCF (also known as 11-zinc finger protein or CCCTC-binding factor) is a transcriptional repressor and a pioneer factor like FOXA1. The primary role of CTCF involves chromatin structure remodeling to form chromatin loops on binding to target DNA. The CTCF-DNA complex then anchors the DNA to the cells and regulates the expression of genes. CTCF also acts as barrier for

the interaction between enhancers and promoters (Phillips & Corces, 2009). Depending on the histone modification enzymes it binds to, it could act as either an activator or repressor. “When the CTCF complex binds to histone acetyltransferase (HAT), it becomes a transcriptional activator and when the complex binds to histone deacetylase (HDAC), it becomes a transcriptional repressor”(“Cancer Genetics Web - CTCF,” n.d.). In this study, CTCF has shown to be active under certain conditions in the tumor progression as seen in the data of primary information profile.

7. CONCLUSION

From the analytical results and evidence from previous studies, GATA3 is the most potential candidate for drug target in CRPC followed by YY1 and SP1. The latter two have already been proved to potential drug targets in previous studies.

Further analysis is required for considering GATA3 as a drug target. To strengthen this hypothesis, a similar approach and methodology is required by applying on a HDPC-based dataset to detect low levels of GATA3 when AR levels are high as a template for comparison with CRPC. Also, it is important to further study the presence, response and stimulation of GATA3 in CRPC and how exactly it does contribute to androgen-independent prostate cancer growth and progression to metastasis before its loss of active role.

8. LIMITATIONS

The limitations associated with this analysis is as follows;

- The identified transcription factors; alternative to Androgen Receptor (AR); are based on interpretation of the analytical results and is limited due to the incomplete understanding of the genomic network of castrate-resistant prostate cancer, possible variability of phenotype expression/mutations in tumor samples across individuals etc.
- To obtain hormone-dependent prostate cancer (HDPC) dataset of similar nature to the CRPC dataset used in both the original and current analysis to support the hypothesis of GATA3 activity and its relationship to AR in prostate cancer of both types.
- The exact significance of GATA3 in prostate cancer in order further understand its regulatory role in prostate cancer in both androgen-dependent and -independent environment is much less understood and poorly studied. Hence, more studies are required in this direction.

9. REFERENCES

Abdulkadir, S. A., Qu, Z., Garabedian, E., Song, S.-K., Peters, T. J., Svaren, J., … Milbrandt, J. (2001). Impaired prostate tumorigenesis in Egr1-deficient mice. *Nature Medicine*, *7*, 101. Retrieved from https://doi.org/10.1038/83231

Albergaria, A., Paredes, J., Sousa, B., Milanezi, F., Carneiro, V., Bastos, J., … Schmitt, F. (2009). Expression of FOXA1 and GATA-3 in breast cancer: The prognostic significance in hormone receptor-negative tumours. *Breast Cancer Research*, *11*(3), 1–15. https://doi.org/10.1186/bcr2327

Augello, M. A., Hickey, T. E., & Knudsen, K. E. (2011). FOXA1: master of steroid receptor function in cancer. *The EMBO Journal*, *30*(19), 3885–3894. https://doi.org/10.1038/emboj.2011.340

Babu, M. M. (n.d.). An Introduction to Microarray Data Analysis, 225–249.

Barbosa-Morais, N. L., Dunning, M. J., Samarajiwa, S. A., Darot, J. F. J., Ritchie, M. E., Lynch, A. G., & Tavaré, S. (2010). A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Research*, *38*(3), e17–e17. https://doi.org/10.1093/nar/gkp942

Bardin CW, Brown T, Isomaa VV, J. O. (1983). Progestins can mimic, inhibit and potentiate the actions of androgens. *Pharmacology & Therapeutics*, *23*(3), 443–459.

Barnes, P. J. (2018). Targeting cytokines to treat asthma and chronic obstructive pulmonary disease. *Nature Reviews Immunology*, *18*(7), 454–466. https://doi.org/10.1038/s41577-018-0006-6

Boto, A., & Harigopal, M. (2018). Strong androgen receptor expression can aid in distinguishing GATA3+ metastases. *Human Pathology*, *75*, 63–70. https://doi.org/10.1016/j.humpath.2018.01.024

Breitling, R., Amtmann, A., & Herzyk, P. (2004). Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, *5*, 34. https://doi.org/10.1186/1471-2105-5-34

Brentani, R. R., Carraro, D. M., Verjovski-Almeida, S., Reis, E. M., Neves, E. J., de Souza, S. J., … Reis, L. F. L. (2005). Gene expression arrays in cancer research: methods and applications. *Critical Reviews in Oncology/Hematology*, *54*(2), 95–105. https://doi.org/https://doi.org/10.1016/j.critrevonc.2004.12.006

Bubendorf, L., Schöpfer, A., Wagner, U., Sauter, G., Moch, H., Willi, N., … Mihatsch, M. J. (2000). Metastatic patterns of prostate cancer: An autopsy study of 1,589 patients. *Human Pathology*, *31*(5), 578–583. https://doi.org/https://doi.org/10.1053/hp.2000.6698

Bumgarner, R. (2013). Overview of DNA microarrays: types, applications, and their future. *Current Protocols in Molecular Biology*, *Chapter 22*, Unit-22.1. https://doi.org/10.1002/0471142727.mb2201s101

Cai, C., He, H. H., Chen, S., Coleman, I., Wang, H., Fang, Z., … Balk, S. P. (2011). Androgen receptor gene expression in prostate cancer is directly suppressed by the androgen receptor through recruitment of lysine-specific demethylase 1. *Cancer Cell*, *20*(4), 457–471. https://doi.org/10.1016/j.ccr.2011.09.001

Cai, C., Hsieh, C.-L., & Shemshedini, L. (2007). c-Jun has multiple enhancing activities in the novel cross talk between the androgen receptor and Ets variant gene 1 in prostate cancer. *Molecular Cancer Research : MCR*, *5*(7), 725–735. https://doi.org/10.1158/1541-7786.MCR-06-0430

Cancer Genetics Web - CTCF. (n.d.). Retrieved from http://www.cancer-genetics.org/CTCF.htm

Carreira, S., Romanel, A., Goodall, J., Grist, E., Ferraldeschi, R., Miranda, S., … Attard, G. (2014). Tumor clone dynamics in lethal prostate cancer. *Science Translational Medicine*, *6*(254), 254ra125-254ra125. https://doi.org/10.1126/scitranslmed.3009448

Chang, C. (2016). Androgen Receptor (AR) Coregulators : An Overview, *23*(March), 175–200.

Cheng, Y., Yu, P., Duan, X., Liu, C., Xu, S., Chen, Y., … Tao, Z. (2015). Genome-wide analysis of androgen receptor binding sites in prostate cancer cells. *Experimental and Therapeutic Medicine*, *9*(6), 2319–2324. https://doi.org/10.3892/etm.2015.2406

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays, *32*(december). https://doi.org/10.1038/ng1031

Cirillo, L. A., Lin, F. R., Cuesta, I., Friedman, D., Jarnik, M., & Zaret, K. S. (2002). Opening of

Compacted Chromatin by Early Developmental Transcription Factors HNF3 (FoxA) and GATA-4. *Molecular Cell*, *9*(2), 279–289. https://doi.org/https://doi.org/10.1016/S1097-2765(02)00459-8

Dasgupta, S., Srinidhi, S., & Vishwanatha, J. K. (2012). Oncogenic activation in prostate cancer progression and metastasis: Molecular insights and future challenges. *Journal of Carcinogenesis*, *11*, 4. https://doi.org/10.4103/1477-3163.93001

Davey, R. A., & Grossmann, M. (2016). Androgen Receptor Structure, Function and Biology: From Bench to Bedside. *The Clinical Biochemist. Reviews*, *37*(1), 3–15. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/27057074

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., … Futreal, P. A. (2002). Mutations of the BRAF gene in human cancer. *Nature*, *417*, 949. Retrieved from https://doi.org/10.1038/nature00766

Davis, J. N., Wojno, K. J., Daignault, S., Hofer, M. D., Kuefer, R., Rubin, M. A., & Day, M. L. (2006). Elevated E2F1 inhibits transcription of the androgen receptor in metastatic hormone-resistant prostate cancer. *Cancer Research*, *66*(24), 11897–11906. https://doi.org/10.1158/0008-5472.CAN-06-2497

Debes, J. D., & Tindall, D. J. (2002). The role of androgens and the androgen receptor in prostate cancer. *Cancer Letters*, *187*(1), 1–7. https://doi.org/https://doi.org/10.1016/S0304-3835(02)00413-5

Deng, Z., Cao, P., Wan, M. M., & Sui, G. (2010). Yin Yang 1: a multifaceted protein beyond a transcription factor. *Transcription*, *1*(2), 81–84. https://doi.org/10.4161/trns.1.2.12375

Deng, Z., Wan, M., Cao, P., Rao, A., Cramer, S. D., & Sui, G. (2009). Yin Yang 1 regulates the transcriptional activity of androgen receptor. *Oncogene*, *28*(42), 3746–3757. https://doi.org/10.1038/onc.2009.231

Denmeade SR, Lin XS, I. J. (1996). Role of programmed (apoptotic) cell death during the progression and therapy for prostate cancer. *The Prostate*, *28*(6), 414.

Dikstein, R., Ruppert, S., & Tjian, R. (1996). TAFII250 is a bipartite protein kinase that phosphorylates the base transcription factor RAP74. *Cell*, *84*(5), 781–790.

Dong, G., Mao, Q., Xia, W., Xu, Y., Wang, J., Xu, L., & Jiang, F. (2016). PKM2 and cancer: The

function of PKM2 beyond glycolysis. *Oncology Letters*, *11*(3), 1980–1986. https://doi.org/10.3892/ol.2016.4168

Eglen, S. J. (2009). A Quick Guide to Teaching R Programming to Computational Biology Students, *5*(8), 8–11. https://doi.org/10.1371/journal.pcbi.1000482

Eisermann, K., Wang, D., Jing, Y., Pascal, L. E., & Wang, Z. (2013). Androgen receptor gene mutation, rearrangement, polymorphism. *Translational Andrology and Urology*, *2*(3), 137–147. https://doi.org/10.3978/j.issn.2223-4683.2013.09.15

Engelmann, D., & Pützer, B. M. (2012). The Dark Side of E2F1: In Transit beyond Apoptosis. *Cancer Research*, *72*(3), 571 LP-575. Retrieved from http://cancerres.aacrjournals.org/content/72/3/571.abstract

Esa Alhoniemi, Jaakko Hollmén, O. S. and J. V. (1996). Process Monitoring and Modeling using the Self-Organizing Map.

Fu, M., Wang, C., Wang, J., Zhang, X., Sakamaki, T., Yeung, Y. G., … Pestell, R. G. (2002). Androgen receptor acetylation governs trans activation and MEKK1-induced apoptosis without affecting in vitro sumoylation and trans-repression function. *Molecular and Cellular Biology*, *22*(10), 3373–3388. https://doi.org/10.1128/MCB.22.10.3373

Gao, N., Zhang, J., Rao, M. A., Case, T. C., Mirosevich, J., Wang, Y., … Matusik, R. J. (2003). The Role of Hepatocyte Nuclear Factor-3α (Forkhead Box A1) and Androgen Receptor in Transcriptional Regulation of Prostatic Genes. *Molecular Endocrinology*, *17*(8), 1484–1507. Retrieved from http://dx.doi.org/10.1210/me.2003-0020

Gaughan, L., Logan, I. R., Neal, D. E., & Robson, C. N. (2005). Regulation of androgen receptor and histone deacetylase 1 by Mdm2-mediated ubiquitylation. *Nucleic Acids Research*, *33*(1), 13–26. https://doi.org/10.1093/nar/gki141

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., … Zhang, J. (2004). Bioconductor : open software development for computational biology and bioinformatics, (10).

Gerhardt, J., Montani, M., Wild, P., Beer, M., Huber, F., Hermanns, T., … Kristiansen, G. (2012). FOXA1 promotes tumor progression in prostate cancer and represents a novel hallmark of castration-resistant prostate cancer. *American Journal of Pathology*, *180*(2), 848–861.

https://doi.org/10.1016/j.ajpath.2011.10.021

Govindarajan, R., Duraiyan, J., Kaliyappan, K., & Palanisamy, M. (2012). Microarray and its applications. *Journal of Pharmacy & Bioallied Sciences*, *4*(Suppl 2), S310–S312. https://doi.org/10.4103/0975-7406.100283

Guerquin, M.-J., Charvet, B., Nourissat, G., Havis, E., Ronsin, O., Bonnin, M.-A., … Duprez, D. (2013). Transcription factor EGR1 directs tendon differentiation and promotes tendon repair. *The Journal of Clinical Investigation*, *123*(8), 3564–3576. https://doi.org/10.1172/JCI67521

Halleck, M. S., Pownall, S., Harder, K. W., Duncan, A. M. V, Jirik, F. R., & Schlegel, R. A. (1995). A widely distributed putative mammalian transcriptional regulator containing multiple paired amphipathic helices, with similarity to yeast SIN3. *Genomics*, *26*(2), 403–406. https://doi.org/https://doi.org/10.1016/0888-7543(95)80229-F

Hamilton, W., Sharp, D. J., Peters, T. J., & Round, A. P. (2006). Clinical features of prostate cancer before diagnosis: a population-based, case-control study. *The British Journal of General Practice : The Journal of the Royal College of General Practitioners*, *56*(531), 756–762. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/17007705

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., … Lee, I. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, *46*(D1), D380–D386. https://doi.org/10.1093/nar/gkx1013

Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., … Lee, I. (2015). TRRUST: a reference database of human transcriptional regulatory interactions. *Scientific Reports*, *5*, 11432. https://doi.org/10.1038/srep11432

Hanai, T., Hamada, H., & Okamoto, M. (2006). Application of bioinformatics for DNA microarray data to bioscience, bioengineering and medical fields. *Journal of Bioscience and Bioengineering*, *101*(5), 377–384. https://doi.org/https://doi.org/10.1263/jbb.101.377

Helsen, C., Van Den Broeck, T., Voet, A., Prekovic, S., Van Poppel, H., Joniau, S., & Claessens, F. (2014). Androgen receptor antagonists for prostate cancer therapy. *Endocrine-Related Cancer*, *21*(4). https://doi.org/10.1530/ERC-13-0545

Huang, Y., Tao, T., Liu, C., Guan, H., Zhang, G., Ling, Z., … Chen, M. (2017). Upregulation of miR-

146a by YY1 depletion correlates with delayed progression of prostate cancer. *International Journal of Oncology*, *50*(2), 421–431. https://doi.org/10.3892/ijo.2017.3840

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*, 115. Retrieved from https://doi.org/10.1038/nmeth.3252

Inouye, C. J., & Seto, E. (1994). Relief of YY1-induced transcriptional repression by protein-protein interaction with the nucleolar phosphoprotein B23. *Journal of Biological Chemistry*, *269*(9), 6506–6510.

Ito, K., & Murphy, D. (2013). Application of ggplot2 to Pharmacometric Graphics. *CPT: Pharmacometrics & Systems Pharmacology*, *2*(10), e79–e79. https://doi.org/10.1038/psp.2013.56

Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., Smigal, C., & Thun, M. J. (2006). Cancer Statistics, 2006. *CA: A Cancer Journal for Clinicians*, *56*(2), 106–130. https://doi.org/10.3322/canjclin.56.2.106

Jin, H.-J., Zhao, J. C., Ogden, I., Bergan, R. C., & Yu, J. (2013). Androgen receptor-independent function of FoxA1 in prostate cancer metastasis. *Cancer Research*, *73*(12), 3725–3736. https://doi.org/10.1158/0008-5472.CAN-12-3468

Jones, D., Wade, M., Nakjang, S., Chaytor, L., Grey, J., Robson, C. N., & Gaughan, L. (2015). FOXA1 regulates androgen receptor variant activity in models of castrate-resistant prostate cancer, *6*(30).

Kashyap, V., & Bonavida, B. (2014). Role of YY1 in the pathogenesis of prostate cancer and correlation with bioinformatic data sets of gene expression. *Genes & Cancer*, *5*(3–4), 71–83. https://doi.org/10.18632/genesandcancer.12

Kim, S., Moon, B.-I., Lim, W., Park, S., Cho, M. S., & Sung, S. H. (2016). Expression patterns of GATA3 and the androgen receptor are strongly correlated in patients with triple-negative breast cancer. *Human Pathology*, *55*, 190–195. https://doi.org/10.1016/j.humpath.2016.04.013

Knudsen, K. E., & Penning, T. M. (2010). Partners in crime: deregulation of AR activity and androgen synthesis in prostate cancer. *Trends in Endocrinology and Metabolism: TEM*,

21(5), 315–324. https://doi.org/10.1016/j.tem.2010.01.002

Knudsen, R. S. S. and K. E. (2013). Molecular Pathogenesis and Progression of Prostate Cancer R.S. *Seminars in Oncology*, *40*(3), 244–258. https://doi.org/10.1053/j.seminoncol.2013.04.001.Molecular

Latchman, D. S. (1993). Transcription factors: an overview. *International Journal of Experimental Pathology*, *74*(5), 417–422. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8217775

Lee, T. I., & Young, R. A. (2000). Transcription of Eukaryotic Protein-Coding Genes. *Annual Review of Genetics*, *34*(1), 77–137. https://doi.org/10.1146/annurev.genet.34.1.77

Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell*, *152*(6), 1237–1251. https://doi.org/10.1016/j.cell.2013.02.014

Leonard, C. (2013). *List of R Editors*. *r-dir*.

Ling, Z., Liu, D., Zhang, G., Liang, Q., Xiang, P., Xu, Y., … Tao, T. (2017). miR-361-5p modulates metabolism and autophagy via the Sp1-mediated regulation of PKM2 in prostate cancer, *38*(3), 1621–1628. https://doi.org/10.3892/or.2017.5852

Liu, H., Bebu, I., & Li, X. (2010). Microarray probes and probe sets. *Frontiers in Bioscience (Elite Edition)*, *2*, 325–338. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/20036881

Loewe R.P., N. P. J. (2011). Microarray Bioinformatics. *Biological Microarrays*, *671*, 295–320.

Loguinov, A. V, Mian, I. S., & Vulpe, C. D. (2004). Exploratory differential gene expression analysis in microarray experiments with no or limited replication. *Genome Biology*, *5*(3), R18–R18. https://doi.org/10.1186/gb-2004-5-3-r18

Louder, R. K., He, Y., Fang, J., Nogales, E., Biology, C., Division, I. B., & National, L. B. (2016). Structure of promoter-bound TFIID and insight into human PIC assembly. *Nature*, *531*(5), 604–609. https://doi.org/10.1038/nature17394.Structure

Luo, J., Zhou, X., Ge, X., Liu, P., Cao, J., Lu, X., … Zhang, S. (2013). Upregulation of Ying Yang 1 (YY1) suppresses esophageal squamous cell carcinoma development through heme oxygenase-1. *Cancer Science*, *104*(11), 1544–1551. https://doi.org/10.1111/cas.12248

Luo, W., Zhu, X., Liu, W., Ren, Y., Bei, C., & Qin, L. (2016). MYC associated zinc finger protein promotes the invasion and metastasis of hepatocellular carcinoma by inducing epithelial

mesenchymal transition, *7*(52), 86420–86432.

Lupien, M., Eeckhoute, J., Meyer, C. A., Wang, Q., Zhang, Y., Li, W., … Brown, M. (2008). FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, *132*(6), 958–970. https://doi.org/10.1016/j.cell.2008.01.018

Manuscript, A. (2011). Synthesis in Prostate Cancer, *21*(5), 315–324. https://doi.org/10.1016/j.tem.2010.01.002.Partners

Mark Dunning, A. L. and M. E. (2015). illuminaHumanv2.db: Illumina HumanWG6v2 annotation data (chip illuminaHumanv2).No Title, *R package*(1.26.0).

Martel, L. S., Brown, H. J., & Berk, A. J. (2002). Evidence that TAF-TATA box-binding protein interactions are required for activated transcription in mammalian cells. *Molecular and Cellular Biology*, *22*(8), 2788–2798. https://doi.org/10.1128/MCB.22.8.2788-2798.2002

Massie, C. E., Lynch, A., Ramos-Montoya, A., Boren, J., Stark, R., Fazli, L., … Mills, I. G. (2011). The androgen receptor fuels prostate cancer by regulating central metabolism and biosynthesis. *EMBO Journal*, *30*(13), 2719–2733. https://doi.org/10.1038/emboj.2011.158

McCabe, M. T., Low, J. A., Daignault, S., Imperiale, M. J., Wojno, K. J., & Day, M. L. (2006). Inhibition of DNA Methyltransferase Activity Prevents Tumorigenesis in a Mouse Model of Prostate Cancer. *Cancer Research*, *66*(1), 385 LP-392. Retrieved from http://cancerres.aacrjournals.org/content/66/1/385.abstract

Merker, J. D., & Arber, D. A. (2010). Chapter 21 - Molecular Diagnostics of Hematopoietic Disorders. In W. W. Grody, R. M. Nakamura, C. M. Strom, & F. L. B. T.-M. D. Kiechle (Eds.) (pp. 243–260). San Diego: Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-369428-7.00021-5

Miklos, G. L. G., & Maleszka, R. (2004). Microarray reality checks in the context of a complex disease. *Nature Biotechnology*, *22*, 615. Retrieved from https://doi.org/10.1038/nbt965

Mizzen, C. A., Yang, X. J., Kokubo, T., Brownell, J. E., Bannister, A. J., Owen-Hughes, T., … Allis, C. D. (1996). The TAF(II)250 subunit of TFIID has histone acetyltransferase activity. *Cell*, *87*(7), 1261–1270. https://doi.org/10.1016/S0092-8674(00)81821-8

Mulholland, D. J., Tran, L. M., Li, Y., Cai, H., Morim, A., Wang, S., … Wu, H. (2011). Article Cell Autonomous Role of PTEN in Regulating Castration-Resistant Prostate Cancer Growth.

Cancer Cell, *19*(6), 792–804. https://doi.org/10.1016/j.ccr.2011.05.006

Narayanan, B. a, Narayanan, N. K., Davis, L., & Nargi, D. (2006). RNA interference-mediated cyclooxygenase-2 inhibition prevents prostate cancer cell growth and induces differentiation: modulation of neuronal protein synaptophysin, cyclin D1, and androgen receptor. *Molecular Cancer Therapeutics*, *5*(5), 1117–1125. https://doi.org/10.1158/1535-7163.MCT-05-0520

Nelson, W. G., De Marzo, A. M., & Isaacs, W. B. (2003). Prostate Cancer. *New England Journal of Medicine*, *349*(4), 366–381. https://doi.org/10.1056/NEJMra021562

Nguyen, A. H. T., Tremblay, M., Haigh, K., Hervékoumakpayi, I., Paquet, M., Pandolfi, P. P., … Bouchard, M. (2013). Gata3 antagonizes cancer progression in pten-deficient prostates. *Human Molecular Genetics*, *22*(12), 2400–2410. https://doi.org/10.1093/hmg/ddt088

Norris, J. D., Chang, C. Y., Wittmann, B. M., Kunder, R. S., Cui, H., Fan, D., … McDonnell, D. P. (2009). The Homeodomain Protein HOXB13 Regulates the Cellular Response to Androgens. *Molecular Cell*, *36*(3), 405–416. https://doi.org/10.1016/j.molcel.2009.10.020

Ordóñez, N. G. (2013). Value of GATA3 Immunostaining in Tumor Diagnosis: A Review. *Advances in Anatomic Pathology*, *20*(5). Retrieved from https://journals.lww.com/anatomicpathology/Fulltext/2013/09000/Value_of_GATA3_Immunostaining_in_Tumor_Diagnosis__.6.aspx

Peyman Tavassoli, Latif A. Wafa, Helen Cheng, Amina Zoubeidi, Ladan Fazli, Martin Gleave, Robert Snoek, and P. S. R. (2010). TAF1 Differentially Enhances Androgen Receptor Transcriptional Activity via Its N-Terminal Kinase and Ubiquitin-Activating and -Conjugating Domains. *Mol Endocrinol*, *24*(4), 696–708. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5417532/

Pham, A.-D., & Sauer, F. (2000). Ubiquitin-Activating/Conjugating Activity of TAF_{II}250, a Mediator of Activation of Gene Expression in Drosophila *Science*, *289*(5488), 2357 LP-2360. Retrieved from http://science.sciencemag.org/content/289/5488/2357.abstract

Phillips, J. E., & Corces, V. G. (2009). CTCF: master weaver of the genome. *Cell*, *137*(7), 1194–1211. https://doi.org/10.1016/j.cell.2009.06.001

Pomerantz MM, Li F, Takeda DY, Lenci R, Chonkar A, Chabot M, Cejas P, Vazquez F, Cook J, Shivdasani RA, Bowden M, Lis R, Hahn WC, Kantoff PW, Brown M, Loda M, Long HW, F. M. (2015). The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nature Genetics*, *47*(11), 1346–1351. https://doi.org/10.7205/MILMED-D-14-00168.Long-chain

Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, *2*, 418. Retrieved from https://doi.org/10.1038/35076576

R. Gentleman, V. Carey, W. H. and F. H. (2018). Genefilter: methods for filtering genes from high-throughput experimentsNo Title, *R package*(version 1.62.0).

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing*. Vienna, Austria. Retrieved from https://www.r-project.org

Ren, G., Zhang, G., Dong, Z., Liu, Z., Li, L., Feng, Y., … Lu, J. (2009). Recruitment of HDAC4 by transcription factor YY1 represses HOXB13 to affect cell growth in AR-negative prostate cancers. *The International Journal of Biochemistry & Cell Biology*, *41*(5), 1094–1101. https://doi.org/https://doi.org/10.1016/j.biocel.2008.10.015

Rice, J. A. (2007). *Mathematical statistics and data analysis*. Belmont, CA: Thomson/Brooks/Cole.

Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., … Friend, † Stephen H. (2000). Signaling and Circuitry of Multiple MAPK Pathways Revealed by a Matrix of Global Gene Expression Profiles. *Science*, *287*(5454), 873 LP-880. Retrieved from http://science.sciencemag.org/content/287/5454/873.abstract

Roeder, R. G. (1996). The role of general initiation factors in transcription by RNA polymerase II. *Trends in Biochemical Sciences*, *21*(9), 327–335. https://doi.org/https://doi.org/10.1016/S0968-0004(96)10050-5

Rosa-Ribeiro, R., Nishan, U., Vidal, R. O., Barbosa, G. O., Reis, L. O., Cesar, C. L., & Carvalho, H. F. (2014). Transcription factors involved in prostate gland adaptation to androgen deprivation. *PLoS ONE*, *9*(6). https://doi.org/10.1371/journal.pone.0097080

Roy AK1, Lavrovsky Y, Song CS, Chen S, Jung MH, Velu NK, Bi BY, C. B. (1998). Regulation of androgen action. In *Vitamins & Hormones* (Vol. 55, pp. 309–352).

Sahu, B., Laakso, M., Ovaska, K., Mirtti, T., Lundin, J., Rannikko, A., … Jännne, O. A. (2011). Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *EMBO Journal*, *30*(19), 3962–3976. https://doi.org/10.1038/emboj.2011.328

Sanga, S., Broom, B. M., Cristini, V., & Edgerton, M. E. (2009). Gene expression meta-analysis supports existence of molecular apocrine breast cancer with a role for androgen receptor and implies interactions with ErbB family. *BMC Medical Genomics*, *2*. https://doi.org/10.1186/1755-8794-2-59

Selvaraj, S., & Natarajan, J. (2011). Microarray data analysis and mining tools. *Bioinformation*, *6*(3), 95–99. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/21584183

Sharma, M., & Sun, Z. (2001). 5'TG3' Interacting Factor Interacts with Sin3A and Represses AR-Mediated Transcription. *Molecular Endocrinology*, *15*(11), 1918–1928. Retrieved from http://dx.doi.org/10.1210/mend.15.11.0732

Shen, D. (2013). The prostate cancer-up-regulated Myc-associated zinc-finger protein (MAZ) modulates proliferation and metastasis through reciprocal regulation of androgen receptor The prostate cancer-up-regulated myc-associated zinc-finger protein (MAZ) modulates proliferation and metastasis through reciprocal regulation of androgen receptor, (June). https://doi.org/10.1007/s12032-013-0570-3

Shoag, J., & Barbieri, C. (2016). Clinical variability and molecular heterogeneity in prostate cancer. *Asian Journal of Andrology*, *18*(4), 543. https://doi.org/10.4103/1008-682X.178852

Shoemaker, D. D., Schadt, E. E., Armour, C. D., He, Y. D., Garrett-Engele, P., McDonagh, P. D., … Boguski, M. S. (2001). Experimental annotation of the human genome using microarray technology. *Nature*, *409*, 922. Retrieved from https://doi.org/10.1038/35057141

Sinnesael, M., Claessens, F., Laurent, M., Dubois, V., Boonen, S., Deboel, L., & Vanderschueren, D. (2012). Androgen receptor (AR) in osteocytes is important for the maintenance of male skeletal integrity: Evidence from targeted AR disruption in mouse osteocytes. In *Journal of Bone and Mineral Research* (Vol. 27, pp. 2535–2543). https://doi.org/10.1002/jbmr.1713

Srinivas-Shankar U, W. F. (2006). Drug insight: testosterone preparations. *Nature Reviews Urology*, *3*(12), 653–665.

Sui, G. (2009). The Regulation of YY1 in Tumorigenesis and its Targeting Potential in Cancer

Therapy. *Molecular and Cellular Pharmacology*, *1*(3), 157–176. https://doi.org/10.4255/mcpharmacol.09.20

Sun, A., Tang, J., Hong, Y., Song, J., Terranova, P. F., Thrasher, J. B., … Li, B. (2008). Androgen receptor-dependent regulation of Bcl-xL expression: Implication in prostate cancer progression. *The Prostate*, *68*(4), 453–461. https://doi.org/10.1002/pros.20723

T Sankpal, U., Goodison, S., Abdelrahim, M., & Basha, R. (2011). *Targeting Sp1 transcription factors in prostate cancer therapy*. *Medicinal chemistry (Shāriqah (United Arab Emirates))* (Vol. 7).

Tarca, A. L., Romero, R., & Draghici, S. (2006). Analysis of microarray experiments of gene expression profiling. *American Journal of Obstetrics and Gynecology*, *195*(2), 373–388. https://doi.org/10.1016/j.ajog.2006.07.001

Taylor, D. (1990). Time-Series Analysis: Use of Autocorrelation as an Analytic Strategy for Describing Pattern and Change. *Western Journal of Nursing Research*, *12*(2), 254–261. https://doi.org/10.1177/019394599001200210

The Basics: What is a Gene Array? (2018). Retrieved from https://www.thermofisher.com/fi/en/home/references/ambion-tech-support/rna-gene-expression/general-articles/the-basics-what-is-a-gene-array.html

Thigpen, A. E., Cala, K. M., Guileyardo, J. M., Molberg, K. H., McConnell, J. D., & Russell, D. W. (1996). Increased Expression of Early Growth Response-1 Messenger Ribonucleic Acid in Prostatic Adenocarcinoma. *The Journal of Urology*, *155*(3), 975–981. https://doi.org/10.1016/S0022-5347(01)66361-4

Urbanucci, A., Sahu, B., Seppälä, J., Larjo, A., Latonen, L. M., Waltering, K. K., … Visakorpi, T. (2011). Overexpression of androgen receptor enhances the binding of the receptor to the chromatin in prostate cancer. *Oncogene*, *31*, 2153. Retrieved from https://doi.org/10.1038/onc.2011.401

Valdez, C. D., Kunju, L., Daignault, S., Wojno, K. J., & Day, M. L. (2013). The E2F1 / DNMT1Axis Is Associated With the Development of ARNegative Castration Resistant Prostate Cancer, *1785*(February), 1776–1785. https://doi.org/10.1002/pros.22715

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*, 252. Retrieved from https://doi.org/10.1038/nrg2538

Wang, Q., Carroll, J. S., & Brown, M. (2005). Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Molecular Cell*, *19*(5), 631–642. https://doi.org/10.1016/j.molcel.2005.07.018

Wang, Q., Li, W., Liu, X. S., Carroll, J. S., Jänne, O. A., Keeton, E. K., … Brown, M. (2007). A Hierarchical Network of Transcription Factors Governs Androgen Receptor-Dependent Prostate Cancer Growth. *Molecular Cell*, *27*(3), 380–392. https://doi.org/10.1016/j.molcel.2007.05.041

Wang, Q., Li, W., Zhang, Y., Yuan, X., Xu, K., Yu, J., … Brown, M. (2009). Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer. *Cell*, *138*(2), 245–256. https://doi.org/10.1016/j.cell.2009.04.056

Xiao, L., Feng, Q., Zhang, Z., Wang, F., Lydon, J. P., Ittmann, M. M., … He, B. (2016). The essential role of GATA transcription factors in adult murine prostate. *Oncotarget*, *7*(30), 47891–47903. https://doi.org/10.18632/oncotarget.10294

Xu, Y., Chen, S., Ross, K. N., & Balk, S. P. (2006). Androgens Induce Prostate Cancer Cell Proliferation through Mammalian Target of Rapamycin Activation and Post- transcriptional Increases in Cyclin D Proteins, (15), 7783–7793. https://doi.org/10.1158/0008-5472.CAN-05-4472

Yamashita, M., Ukai-Tadenuma, M., Miyamoto, T., Sugaya, K., Hosokawa, H., Hasegawa, A., … Nakayama, T. (2004). Essential role of GATA3 for the maintenance of type 2 helper T (Th2) cytokine production and chromatin remodeling at the Th2 cytokine gene loci. *Journal of Biological Chemistry*, *279*(26), 26983–26990. https://doi.org/10.1074/jbc.M403688200

Yang, S.-Z., Eltoum, I. A., & Abdulkadir, S. A. (2005). Enhanced EGR1 activity promotes the growth of prostate cancer cells in an androgen-depleted environment. *Journal of Cellular Biochemistry*, *97*(6), 1292–1299. https://doi.org/10.1002/jcb.20736

Yang, S. (2003). Early Growth Response Gene 1 Modulates Androgen Receptor Signaling in Prostate Carcinoma Cells *, *278*(41), 39906–39911. https://doi.org/10.1074/jbc.M307250200

Yu, F., Thiesen, J., & Strätling, W. H. (2000). Histone deacetylase-independent transcriptional repression by methyl-CpG-binding protein 2. *Nucleic Acids Research*, *28*(10), 2201–2206. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10773092

Yuan, H., Gong, A., & Young, C. Y. F. (2005). Involvement of transcription factor Sp1 in quercetin-mediated inhibitory effect on the androgen receptor in human prostate cancer cells. *Carcinogenesis*, *26*(4), 793–801. Retrieved from http://dx.doi.org/10.1093/carcin/bgi021

Yuan, X., & Balk, S. P. (2009). Mechanisms mediating androgen receptor reactivation after castration. *Urologic Oncology*, *27*(1), 36–41. https://doi.org/10.1016/j.urolonc.2008.03.021

Yuan, X., Cai, C., Chen, S., Chen, S., Yu, Z., & Balk, S. P. (2014). Androgen receptor functions in castration-resistant prostate cancer and mechanisms of resistance to new agents targeting the androgen axis. *Oncogene*, *33*(22), 2815–2825. https://doi.org/10.1038/onc.2013.235

Zerbini, T. A. L. and L. F. (2006). Targeting Transcription Factors for Cancer Gene Therapy. *Current Gene Therapy*. https://doi.org/http://dx.doi.org/10.2174/156652306775515501

Zhang, X., Morrissey, C., Sun, S., Ketchandji, M., Nelson, P. S., True, L. D., … Plymate, S. R. (2011). Androgen receptor variants occur frequently in castration resistant prostate cancer metastases. *PloS One*, *6*(11), e27970–e27970. https://doi.org/10.1371/journal.pone.0027970

Zhang Y., Szustakowski J., S. M. (2009). Bioinformatics Analysis of Microarray Data. *Cardiovascular Genomics*, *573*, 259–284.

Zhang, Y., Akinmade, D., & Hamburger, A. W. (2005). The ErbB3 binding protein Ebp1 interacts with Sin3A to repress E2F1 and AR-mediated transcription. *Nucleic Acids Research*, *33*(18), 6024–6033. https://doi.org/10.1093/nar/gki903

Zhou, H., Zhang, Y., & W Hamburger, A. (2011). *EBP1 Inhibits Translation of Androgen Receptor mRNA in Castration Resistant Prostate Cancer Cells*. *Anticancer research* (Vol. 31).

Zwang, Y., Sas-Chen, A., Drier, Y., Shay, T., Avraham, R., Lauriola, M., … Yarden, Y. (2011). Two phases of mitogenic signaling unveil roles for p53 and EGR1 in elimination of inconsistent growth signals. *Molecular Cell*, *42*(4), 524–535.

https://doi.org/10.1016/j.molcel.2011.04.017