FUNCTIONAL ANALYSIS OF MIRNA REGULATED GENES IN PROSTATE CANCER AS POTENTIAL DIAGNOSTIC MOLECULES

UNIVERSITY of the
WESTERN CAPE

Gadija Abdullah

Student no: 3575135

MSc Biotechnology

The Faculty of Natural Sciences

Department of Biotechnology, University of the Western Cape

Supervisor: Dr Ashley Pretorius

Co-supervisor: Dr Firdous Khan

DECLARATION

I declare that this thesis is a presentation of my original research work. I have written the enclosed Master Thesis, *Functional Analysis of miRNA regulation gene in prostate cancer as potential diagnostic molecules* completely by myself and have documented all sources and material used. This thesis was not previously presented to another examination board in any other university and has not been published.

Gadija Abdullah December 2016

……

Signature

ACKNOWLEGDEMENTS

I thank Associate Professor. A Pretorius, Chairperson of the Science Faculty, Assessment Committee and Principle Investigator of the Bioinformatics Research Group on Disease Diagnostics at the University of the Western Cape, South Africa. I am grateful and appreciative to him for the constant assistance and guidance. I would also like to thank my co-Supervisor, Dr Firdous Khan.

I wish to thank, Shuaib Abdullah, for his support as a loving husband. I would not have survived these years of study without his constant encouragement and unconditional love. He has taught me that I could be anything in spite of whatever circumstances in which I found myself. He has in every sense earned this honour more than me. And my wonderful kids, Yaghya, Yusuf, Noorah and baby Nuhaa, who throughout my studies somehow understood and complained far less. To my mom, thank you for all the support with the kids.

This work was funded by IOP funding (1st year) and I thank them for this opportunity to further my studies. Funds were generously provided by CSIR in the second year of this study. I also thank the University of the Western Cape for this study opportunity.

I would like to thank my fellow students at the Bioinformatics Research Group (BRG) of the University of the Western Cape for their insights, suggestions and training on various databases which assisted me in this research study.

ABSTRACT

Prostate Cancer is the leading cause of cancer-related death in males in the Western world. It is a common biological disease originating from the reproductive system of the male namely, the prostate gland, usually in older patients (over the age of 50) and with a family history of this disease. The disease shows clinical aggressiveness due to genetic alterations of gene expression in prostate epithelial cells. Prostate cancer is currently diagnosed by biopsy and prostate cancer screening via the Prostate-Specific Antigen (PSA) blood test. Early detection is critical and although PSA was discovered to aid in the diagnoses of this cancer at its early stages, it has a disadvantage due to its low specificity thus causing unnecessary biopsies of healthy individuals and over-treatment of patients. Although various studies and efforts have been made to identify the ideal biomarker for prostate cancer and many even being applied to clinical use, it is still challenging and has not replaced the best-known biomarker PSA. PSA test has minimal invasive characteristics, at relatively low cost together with high sensitivity but low specificity. Biomarker discovery is a challenging process and a good biomarker has to be sensitive, specific and its test highly standardized and reproducible as well as identify risk for or diagnose a disease, assess disease severity or progression, predict prognosis or guide treatment. Computational biology plays a significant role in the discovery of new biomarkers, the analyses of disease states and the validation of potential biomarkers. Bioinformatic approaches are effective for the detection of potential micro ribonucleic acid (miRNA) in cancer. Altered miRNA expression may serve as a biomarker for cancer diagnosis and treatment. Small non-protein coding RNA, miRNA are small regulatory RNA molecules that modulate the expression of their target genes. miRNAs influence numerous cancer-relevant processes such as proliferation, cell cycle control, apoptosis, differentiation, migration and metabolism. Discovery and

existence of extracellular miRNAs that circulate in the blood of cancer patients has raised the possibility that miRNAs may serve as novel diagnostic markers. Since a single miRNA is said to be able to target several mRNAs, aberrant miRNA expression is capable of disrupting the expression of several mRNAs and proteins. Biomarker discovery for prostate cancer of mRNA and miRNA expression are strongly needed to enable more accurate detection of prostate cancer, improve prediction of tumour aggressiveness and facilitate diagnosis.

The aim of this project was to focus on functional analyses of genes and their protein products regulated by previously identified miRNA in prostate cancer using bioinformatics as a tool. Most proteins function in collaboration with other proteins and therefore this study further aims to identify these protein-protein interactions and the biological relevance of these interactions as it relates to Prostate cancer. Various computational databases were used such as STRING, DAVID and GeneHub-GEPIS for functional analyses of these miRNA regulated genes. The main focus was on the 21 genes regulated by several miRNAs identified in a previous study. Results from this study identified six genes; ERP44, GP1BA, IFNG, SEPT2, TNFRSF13C and TNFSF4, as possible diagnostic biomarkers for prostate cancer. These results are promising, since the targeted biomarkers would be easily detectable in bodily fluids with the Gene Ontology (GO) analysis of these gene products showing enrichment for cell surface expression. The six genes identified *in silico* were associated to transcription factors (TFs) to confirm regulatory control of these TFs in cancer promoting processes and more specifically prostate cancer. The CREB, E2F, Nkx3-1 and p53 TFs were discovered to be linked to the genes IFNG, GP1BA, SEPT2 and TNFRSF13C respectively. The expression of these TFs show strong association with cancer and cancer related pathways specifically prostate cancer and thus demonstrates

that these genes can be assessed as possible biomarkers for prostate cancer. The prognostic and predictive values of the candidate genes were evaluated to assess their relationship to prognosis of this disease by means of several *in silico* prognostic databases. The results revealed expression differences for the majority of the candidate genes were not significantly sufficient to be distinguished as strong prognostic biomarkers in several prostate cancer populations. Although one marker, GP1BA was supported as having prognostic value for prostate cancer based on it's statistical p-value in one of the prostate cancer patient datasets used. Another candidate gene SEPT2 showed promise as it has some prognostic value in the early stages of the disease. Although the results yielded, based on the *in silico* analysis, were not the discovery of an ideal diagnostic marker based on the set criteria in this study, further analysis using a molecular approach qRT-PCR can be considered for a detailed follow-up study on selected candidate genes to evaluate their roles in disease initiation and progression of prostate cancer using cell lines as well as patient samples.

Keywords: miRNA, gene expression, biomarker, STRING, Prostate-Specific Antigen, Bioinformatics

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ACVR1C	:	Activin A receptor, type IC
ADNP	:	Activity-dependent neuroprotector homeobox
AJCC	:	American Joint Committee on Cancer
AMACR	:	alpha-methylacyl-coA racemase
AP-2	:	activator protein-2
ATM	:	Ataxia telangiectasia mutated
BFAR	:	Bifunctional Apoptosis regulator
bHLHZ	:	basic helix-loop-helix leucine zipper
BP	:	Biological process
BPH	:	benign prostatic hyperplasia
BTG2	:	BTG family member 2
C/EBP	:	CCAAT/enhancer binding protein
CAT	:	Computerized axial tomography
CC	:	Cellular components
Cdk	:	cyclin dependent kinase
cDNA	:	Complementary DNA
CFLAR	:	CASP8 and FADD-like apoptosis regulator
CLN8	:	Ceroid-lipofuscinosis, neuronal 8
Cpsa	:	complexed prostate specific antigen
CRE	:	cAMP response element
CRM	:	cis-regulatory module
CSRNP3	:	Cysteine-serine-rich nuclear protein 3
CT	:	Computerized tomography

CTC	:	circulating tumour cells.
DAVID	:	Database for Annotation, Visualization and Integrated Discovery
DE	:	Differentially expressed
DEG	:	Differentially expressed genes
DNA	:	Deoxyribonucleic acid
DRE	:	digital rectum examination
DU145	:	prostate cancer cell lines
EBI	:	European Bioinformatics Institute
EBI	:	European Bioinformatics Institute
EGFR	:	Epidermal growth factor receptor
ERP44	:	TXNDC4 thioredoxin domain containing 4 (endoplasmic reticulum)
EST	:	Expressed Sequence Tags
FDA	:	Food and Drug Administration
FOXC1	:	Forkhead box C1
G0	:	Rest phase of cell cycle
GBM	:	glioblastoma
GEO	:	Gene Expression Omnibus
GO	:	Gene Ontology
GP1BA	:	glycoprotein Ib (platelet), alpha polypeptide
GSTP1	:	glutathione s-transferase pi 1
HFH-1	:	Hepatocyte nuclear factor-3 homologue 1
HSPGs	:	Heparan sulfate proteoglycans
IARC	:	International Agency for Research on Cancer
IFNG	:	interferon, gamma

ING4	:	Inhibitor of growth family, member 4
KDD	:	Knowledge Discovery in Databases
KEGG	:	Kyoto Encyclopedia of Genes and Genomes
KLK2	:	kallikrein and kallikrein-related peptidase 2
LIG4	:	Ligase IV, DNA, ATP-dependent
LNCaP	:	Lymph Node Carcinoma of the Prostate cell lines
MF	:	Molecular function
miRNA	:	Micro ribonucleic acid
MNT	:	MAX binding protein
MPF	:	Maturation Promoting Factor
MRI	:	Magnetic resonance imaging
MSKCC		Memorial Sloan Kettering Cancer Center.
NCBI	:	National Center for Bioinformatics
NCI	:	National Cancer Institute
NGS	:	Next Generation Sequencing
p27	:	kinase inhibitory protein 1
P53	:	Tumour protein P53
PAK7	:	p21 protein (Cdc42/Rac)-activated kinase 7
PANTHER	:	protein annotation through evolutionary relationship
PAP	:	Prostate acid phosphatase
PCA3	:	prostate cancer antigen 3
PC3	:	prostate cancer cell lines
PIA	:	Proliferative inflammatory atrophy
PIN	:	Prostatic Intraepithelial Neoplasia

PPI	:	protein–protein interactions
PRAD	:	Prostate Adenocarcinoma
PRKCI	:	Protein kinase C, iota
PSA	:	Prostate-Specific Antigen,
PSMA	:	prostate-specific membrane antigen
PTPRC	:	Protein tyrosine phosphatase, receptor type, C
RAB27A	:	RAB27A, member RAS oncogene family
RRP	:	radical retropubic prostatectomy
RT-PCR	:	Real-time polymerase chain reaction
SEER	:	Surveillance, Epidemiology, and End Results
SEPT2	:	septin 2
SH3RF1	:	SH3 domain containing Ring Finger
SNP	:	single nucleotide polymorphism
STAT	:	Signal transducers and activators of transcription
STRING	:	Search Tool for the Retrieval of Interacting Genes/Proteins
TCGA	:	The Cancer Genome Atlas
TF	:	transcription factor
TiGER	:	Tissue-specific Gene Expression and Regulation
TMX1	:	Theoredoxin-related transmembrane protein
TNFRSF13C	:	tumour necrosis factor receptor superfamily, member 13C
TNFSF13B	:	Tumour necrosis factor
TNFSF15	:	Tumour necrosis factor (ligand) superfamily, member 15
TNFSF4	:	tumour necrosis factor (ligand) superfamily, member 4

TNM	:	Tumour node metastasis
TPC	:	tumour-propagating cell
TRUS	:	transrectal ultrasound
TSG	:	tissue-specific gene
WHO	:	World Health Organization
YWHAZ	:	Tyrosine 3-monooxygenase/ tryptophan 5-monooxygenase activation protein, zeta polypeptide

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

LIST OF TABLES

LIST OF APPENDICES

Chapter 1

Literature Review

1.1. Introduction

Cancer is defined as a disease in which a group of abnormal cells grow uncontrollably by disregarding the normal rules of cell division. The foundation of modern cancer biology rests on a simple principle that virtually all mammalian cells share similar molecular networks that control cell proliferation, differentiation and cell death (Hejmadi, 2009). Cancer is a disease that involves changes or mutations in the cell genome and normal cells are transformed into cancers as a result of changes in networks at the molecular, biochemical and cellular level. Phenomenal advances in cancer research in the past 50 years have given insight into how cancer cells develop (Hejmadi, 2009).

Omics is the biomedical field describing the study in biology such as genomics and proteomics. Many of the emerging fields of large-scale data-rich biology are designated by adding the suffix '-omics' onto previously used terms. Genomics is a term referring to all the genes and the interactions of those genes with each other and the environment whereas proteomics refers to the study of proteins and its modifications made by an organism (Cho, 2007). Proteomics would not be possible without genomics. Therefore, in order to understand cancer biology there should be some understanding of molecular basis dealing with that of disease and proteomics for the discovery of biomarkers in complex diseases such as cancer (Cho, 2007).

In Bioinformatics, computational tools are applied on molecular data (genes and proteins) in a way to analyse and discover new outcomes by sequence recognition of genes and the prediction thereof. Genes are only the guidelines of the cell, while the proteins encoded by the genes are ultimately the functional drive for both normal and disease physiology. Cancer-related proteins and altered gene expression of miRNA, specifically in prostate cancer was the focus of this research, in order to identify biomarkers involved in prostate cancer development and progression thereof. MiRNA molecules are already entering the clinic as diagnostic and prognostic biomarkers for patient stratification and also as therapeutic targets and agents (Jansson and Lund, 2012).

1.2. Biology of cancer

Cancer is a large group of diseases involving the uncontrolled growth and spread of abnormal cells that do not die, known as malignant tumours (Bashyam, 2002). Cancer is a multi-gene, multi-step disease originating from a single abnormal cell (clonal origin) with an altered DNA sequence (mutation). Uncontrolled proliferation of these abnormal cells is due to successive rounds of mutation and selective expansion of these cells results in the formation of a tumour mass. These changes (DNA mutations) produce proteins that disrupt the delicate cellular balance between cell division resulting in cells that keep dividing to form cancers (Hejmadi, 2009).

Normal cells in the body have an orderly path of growth, division and death. Proteins within the cell control the cell cycle however cancer cells have an abnormal cell cycle and divide repeatedly out of control even though they are not needed; they crowd out normal cells and function abnormally (Bashyam, 2002).

1.3. Classification of cancer

Cancer can be classified according to the type of cell that the tumour cells resemble which is likely to be the origin of the tumour (i) Carcinomas; are derived from epithelial cells such as breast, prostate and lung, (ii) Sarcomas; arising from connective tissue such as bone, cartilage and fat, (iii) Lymphomas; begin in the lymph nodes and immune system tissues and (iv) Leukaemia; begin in the bone marrow and accumulate in the blood stream. Cancers are often referred to by terms that contain a prefix which is related to cell type the cancer originated from and a suffix such as –sarcoma,-carcinoma or –oma. Some types of cancers are named according to the shape and size of the cells under a microscope such as small-cell carcinoma (International Classification of Diseases for Oncology, Third Edition - ICD-O-3).

1.4. The Cell cycle

Most eukaryotic cells follow a process of growth and division called the cell cycle. These stages include; (1) a growth stage (2) mitosis or nuclear division and (3) cytokinesis or division of the cytoplasm. Throughout interphase the cell is engaged in growth and metabolic activities. Interphase can be broken down into three phases: G1, S and G2 (as seen in figure 1.1). During the G1 or first growth phase, normal cell function occurs as well as cell growth. S phase, DNA replicates producing two copies of each chromosome. G2 phase, the cell continues to prepare for mitosis and cell division. Mitosis or M phase has four stages. These phases are sequentially known as; prophase, metaphase, anaphase and telophase (Campbell and Reece, 2002).

During prophase the chromosomes become visible and condense, becoming shorter and thicker. Each identical copy of a single chromosome is called a sister chromatid.

The nuclear envelope breaks down and spindle fibers form as microtubules grow out of the centrioles that move to opposite poles of the cell. During metaphase, the double-stranded chromosomes line up along the equator of the cell. The microtubules attach to each sister chromatid (Campbell and Reece, 2002). Anaphase begins when the sister chromatids of each chromosome begin to separate (Campbell and Reece, 2002). The centromeres that hold the sister chromatids together divide and the chromosomes move away from each other along its spindle fiber. During telophase, the two groups of chromosomes reach the opposite ends of the cell, as a new nuclear envelope starts to form, the chromosomes uncoil and the spindle disappears. Cytokinesis or the C phase, involves the division of the cytoplasm and organelles. Two genetically identical cells are formed as a result of mitosis and cytokinesis (Campbell and Reece, 2002)

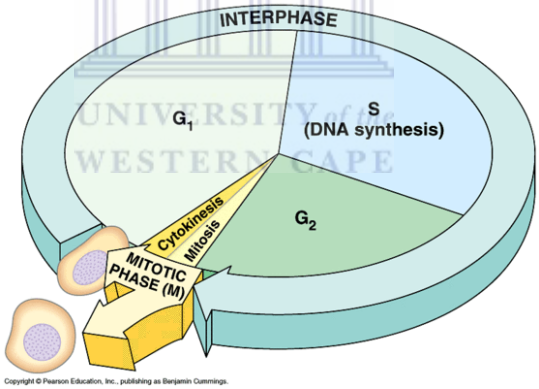


Figure 1.1: The cell cycle (Taken from Campbell and Reece, 2002)

1.5. Role of the cell cycle in cancer

The study of these phases, the proteins that regulate them, and the complex biochemical interactions that stop or start DNA replication and cell division (cytokinesis) are the primary concerns of cell cycle biologists. Many genes and proteins, that influence the passage from one phase of the cell cycle to another have

been identified and when their expression is altered by mutation or aberrant regulation, they are usually classed as oncogenes. Other proteins act to hold the cell at distinct points in the cycle (checkpoints) and are known as tumour suppressor genes. Most cancers are the result of inappropriate cell division, often stemming from aberrations in normal cell cycle regulation (Kastan and Bartek, 2004).

1.6. Regulation of the cell cycle

How cell division (and thus tissue growth) is controlled is very complex. These regulatory molecules exist largely in two varieties: protein kinases, enzymes that serve to activate or inactivate other proteins through phosphorylation, and cyclins. Cdk (cyclin dependent kinase, adds phosphate to a protein), along with cyclins, are major control switches for the cell cycle, causing the cell to move from G1 to S or G2 to M phase (Carleton *et al.,* 2007). Maturation Promoting Factor (MPF) includes the CdK and cyclins that triggers progression through the cell cycle. p53 is a protein that functions to block the cell cycle if the DNA is damaged. If the damage is severe this protein can cause apoptosis (cell death). p53 levels are increased in damaged cells (Carleton *et al.,* 2007). This allows time to repair DNA by blocking the cell cycle. p53 is frequently mutated in a number of cancers as an early genetic event. An extreme case of this is Li Fraumeni syndrome, where a genetic defect in p53 leads to a high frequency of cancer in affected individuals. p27 is a protein that binds to cyclin and Cdk blocking entry into S phase. Alteration of miRNA levels can also contribute to pathological conditions, including tumourgenesis, that are associated with loss of cell cycle control (Carleton *et al.,* 2007).

Cell cycle checkpoints are regulatory pathways that control the order and timing of cell cycle transitions and ensure that critical events such as DNA replication and chromosome segregation are completed. Checkpoints respond to damage by arresting the cell cycle to provide time for repair and by inducing transcription of genes that facilitate repair. Checkpoint loss, results in genomic instability and has been implicated in the evolution of normal cells into cancer cells (Kastan and Bartek, 2004).

Cyclins are among the most important components of the core cell cycle control system. Cyclins are a group of related proteins and there are four basic types: G1-cyclins, G1/S-cyclins, S-cyclins and M-cyclins and each cyclin is associated with a particular phase, transition or set of phases in the cell cycle and helps drive the events of that phase or period (Kastan and Bartek, 2004).

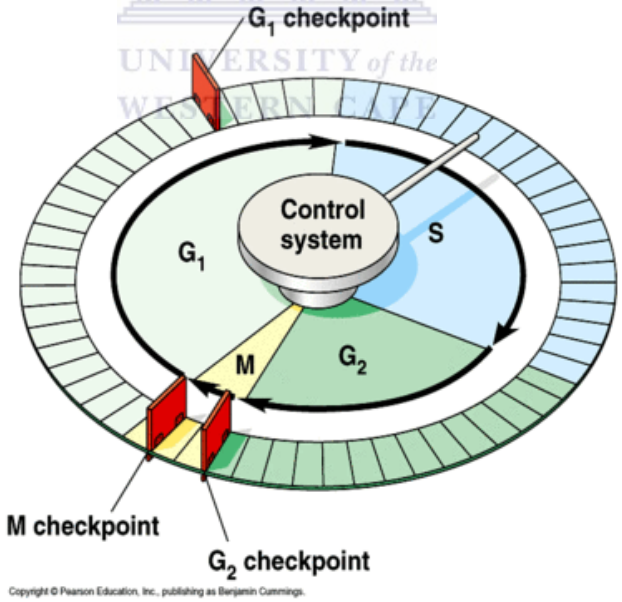


Figure 1.2: Mechanical analogy for the cell cycle control system (Taken from Campbell and Reece, 2002).

Checkpoints prevent cell cycle progression at specific points, allowing verification of necessary phase processes and repair of DNA damage. The cell cannot proceed to the next phase until checkpoint requirements have been met. Checkpoints typically consist of a network of regulatory proteins that monitor and dictate the progression of the cell through the different stages of the cell cycle. Several checkpoints are there to ensure that damaged or incomplete DNA is not passed on to daughter cells (Kastan and Bartek, 2004).

1.7. Hallmarks of cancer

All cancers share six common traits ("hallmarks") that govern the transformation of normal cells to cancer (malignant or tumour) cells. The traits ("hallmarks") are (1) Cancer cells stimulate their own growth (self-sufficiency in growth signals), (2) they resist inhibitory signals that might otherwise stop their growth (insensitivity to anti-growth signals), (3) they resist programmed cell death (evading apoptosis), (4) they can multiply indefinitely (limitless replicative potential), (5) they stimulate the growth of blood vessels to supply nutrients to tumours (sustained angiogenesis) and (6) they invade local tissue and spread to distant sites (tissue invasion and metastasis) (Hanahan and Weinberg, 2000).

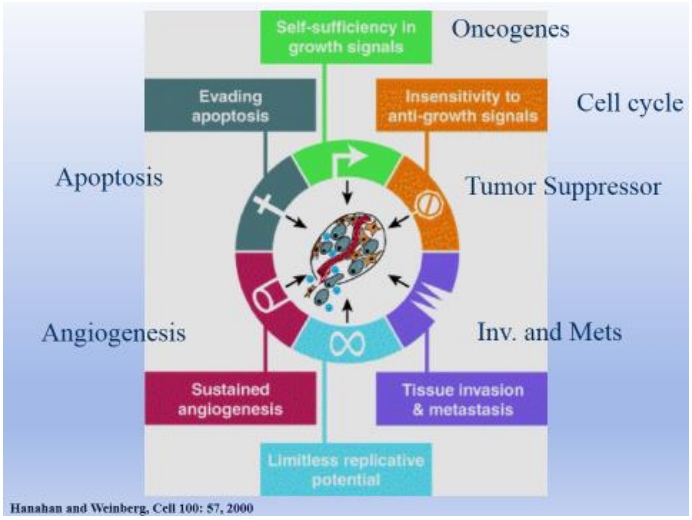


Figure 1.3: The hallmarks of cancer (Taken from Hanahan and Weinberg, 2000)

In an update published in 2011 ("Hallmarks of cancer: the next generation"), Weinberg and Hanahan proposed two new hallmarks: (1) abnormal metabolic pathways where most cancer cells use abnormal metabolic pathways to generate energy with the capability to modify, or reprogram, cellular metabolism in order to most effectively support neoplastic proliferation and (2) evading the immune system where cancer cells appear to be invisible to the body's immune system allowing cancer cells to evade immunological destruction, in particular by T and B lymphocytes, macrophages, and natural killer cells (Hanahan and Weinberg, 2011).

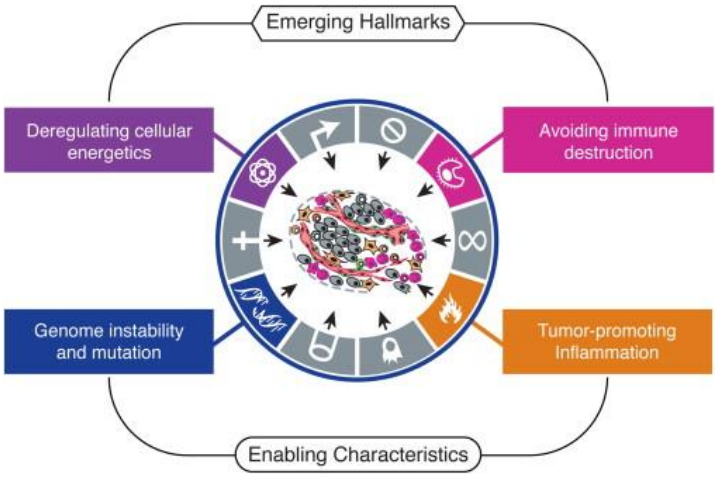


Figure 1.4: Two additional hallmarks of cancer are involved in the pathogenesis of some and perhaps all cancers (Hanahan and Weinberg, 2011).

All normal cells have mechanisms that ensure errors or damages are detected within their own control systems. The tumour cells created by multiple mutations are able to push their way through the epithelial tissue's basement membrane, which are proteins that normally create a barrier (Hanahan and Weinberg, 2000). During the development of earlier stages of the tumour, angiogenesis takes place. Angiogenesis is the development of new blood vessels forming from the pre-existing vessels. With the new blood supply the growth of the tumour accelerates and individual cells from the tumour enter into the network of newly formed blood vessels using these blood vessels to move to other parts of the body. Most tumours are lethal due to their ability to metastasize (to establish new tumour sites at other locations throughout the body) (Hanahan and Weinberg, 2000).

Metastasis is now underway, as tumour cells from the original cancer growth travel throughout the body (Bashyam, 2002). Invasion and metastasis are a dynamic multi-step process. Most of these cells will die soon after entering the blood or lymph circulation. The tumour cell leaves the blood vessels and invades tissue. While the

primary tumour may result from mutations in the growth control genes, metastasis probably results mainly from changes in gene expression patterns in the cell (Bashyam, 2002).

1.8. Genetics of cancer

Genetic changes occur at different levels and by different mechanisms. Most common mutations are changes in the nucleotide sequence of genomic DNA. Mutations include point mutations, deletions and insertions affecting the gene's coding sequence and altering the function of the genes protein product. Although complex error correction and prevention is built into each cell and tries to safeguard the cell against cancer, errors do occur and the control process fails, the mutation will survive and passed on to the daughter cells. Errors of mutations include (i) mutation in an oncogene causing the cell to reproduce faster, (ii) mutations might cause a loss of a tumour suppressor gene resulting in disruption of the apoptotic signalling pathway causing an immortal cell, (iii) mutations in the DNA damage and repair mechanisms will cause more errors in the daughter cells. The transformation of a normal cell into cancer is a chain reaction caused by initial errors and mutations allowing the cell to escape the normal control mechanisms for normal cell growth. Once cancer has begun it is an on-going process of progression to more invasive stages (Merlo *et al.,* 2006).

Four types of genes are responsible for the cell division process playing a role in cancer development: (i) Oncogenes instructs the cell when to divide - promote cell growth and reproduction, (ii) Tumour suppressor genes instruct the cell when not to divide - codes for proteins that inhibit cell division (Croce, 2008). These are stated to be the two major genes that play a role in triggering cancer (Knudson, 2001), (iii)

Suicide genes or self-destruction genes control apoptosis (programmed cell death) and instruct the cell to kill itself if something goes wrong, (iv) The DNA in every cell in our body is constantly in danger of being damaged. DNA repair genes instruct the cell to repair damaged DNA.

Several mutations need to occur to give rise to cancer. Cancer is caused by failure to regulate tissue growth, when the genes that regulate cell growth are altered (Croce, 2008). Furthermore, epigenetic alterations are frequent in the DNA sequences coding for small RNAs called miRNAs. When oncogenes are expressed they are regulated by miRNAs. MiRNAs are small non-coding, single-stranded RNAs 21-25 nucleotides in length that control gene expression by downregulating them. Mutations in such miRNAs (known as oncomirs) can lead to activation of oncogenes. These alterations are caused by both DNA sequence mutation in oncogenes and tumour suppressor genes (Esquela-Kerscher *et al*., 2006).

1.8.1 Epigenetic factors

Although mutations are in the tumour suppressor genes, oncogenes and even chromosomal abnormalities, it has been found that cancer can also be driven by epigenetic alterations (Baylin *et al.,* 2006). Epigenetics alterations are any functional modifications to the genome that does not involve any change to the nucleotide sequence (Kanwal *et al.,* 2012). The epigenetic alterations serve to regulate gene expression without any change to the DNA sequence. According to various research studies, a large number of epigenetic alterations were found in cancers such as epigenetic alterations in DNA repair genes causing a reduced or silence expression of DNA repair proteins which results in deficient DNA repair (Jacinto *et al.,* 2007, Lahtz

et al., 2011, Bernstein *et al.,* 2013 and Bernstein *et al.,* 2013). Like other cancers, prostate cancer occurs due to various genetic and epigenetic changes.

1.9. Prostate cancer

1.9.1. The prostate gland

The prostate gland is divided into three zones: peripheral, transition and central zones (refer to figure 1.5).

Peripheral zone

This zone is the largest and closest to the rectum and therefore easily found during a digital rectum examination (DRE). The majority (about 70-80%) of prostate tumours originate in the peripheral zone (Basic principles 2010).

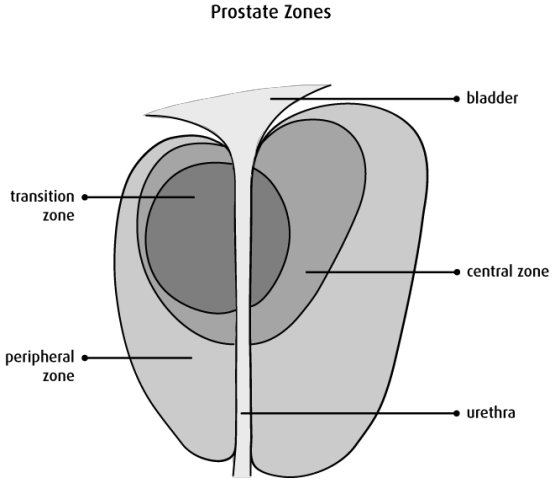


Figure 1.5: Zones of the prostate gland affected by cancer

(http://www.cancer.ca/~/media/CCE/1158/dc5c29f2337751c9e3038122fd0b3ed9.png)

Transition zone

This is the middle area of the prostate gland, between the peripheral and central zones. When men age the transition zone enlarges and eventually becomes the largest area of the prostate and it is responsible for the disease of benign prostatic hyperplasia (BPH) (Basic principles 2010).

Central zone

Found in front of the transition zone and farthest from the rectum surrounding the ejaculatory ducts, accounting for about 2.5% of prostate cancers (Cohen *et al.,* 2008).

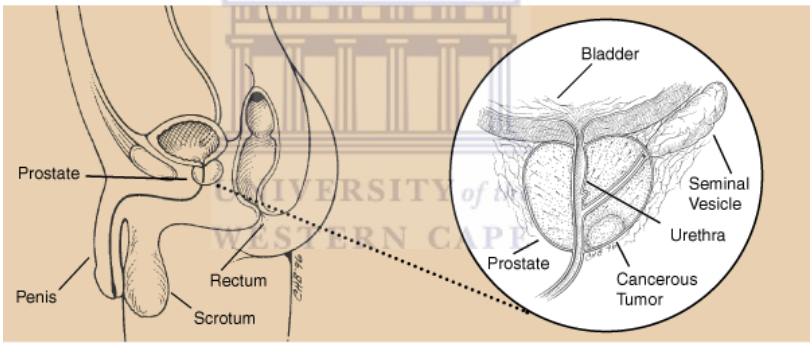


Figure 1.6: Anatomy of the prostate gland (American Cancer Society, 2015)

1.9.2. Types of Prostate Cancer

Carcinoma of the prostate (or prostate cancer) is the development of cancer in the prostate gland found in the reproductive system of the male. A healthy male prostate is a walnut-sized gland located below the urinary bladder surrounding the urethra, the tube that carries urine from the bladder during urination and semen during ejaculation (Moore *et al.,* 1999). In prostate cancer, it's the prostate cells of the gland that mutate

into cancer cells. The prostate contains cells that make some of the fluid (semen) that protects and nourishes the sperm.

Most prostate cancers are a type called adenocarcinomas, starting from the gland cells in the prostate. The type of cancer is found by sampling the cells from the prostate during a biopsy. Some types of prostate cancer include: (i) Ductal adenocarcinoma (ii) Transitional cell (urothelial) cancer (iii) Squamous cell cancer (iv) Carcinoid (v) Small cell cancer (vi) Sarcomas and sarcomatoid cancer (American Cancer Society, 2015).

There are various conditions of the prostate that are non-cancerous:

Benign Prostatic Hyperplasia (BPH) is a condition that causes the prostate to get larger as men age. BPH is not cancer and does not change into cancer. But it can cause problems urinating if the prostate gets larger. BPH is often treated with drugs that shrink the prostate or relax the muscles in it, which can help urine flow (Verhamme *et al.,* 2002).

In Prostatic Intraepithelial Neoplasia (PIN), there are changes in how the prostate gland cells look under the microscope, but the cells don't metastasize, like cancer cells (American Cancer Society, 2015).

Proliferative inflammatory atrophy (PIA) is another possible finding on a prostate biopsy. In PIA, the prostate cells look smaller than normal, and there are signs of inflammation. PIA is not cancer, but sometimes leads to high-grade PIN, or perhaps to prostate cancer (American Cancer Society, 2015).

1.10. Grading and staging of prostate cancer

The system used most often for grading prostate cancer is called the Gleason scoring system. Samples from two areas of the prostate are each graded from 1 to 5, and the number grades are added to give a Gleason score or sum of between 2 and 10. Most biopsies have a Gleason score of at least 6. A higher score means the cells look less normal and the cancer is likely to grow more progressively. The stage (extent) of a cancer is one of the most important factors in determining treatment options and the outlook of recovery. The stage is based on the prostate biopsy results (including the Gleason score), the PSA level, and any other exams or tests that were done to assess how far the cancer has spread (BMJ Group, 2009).

The TNM Staging System was developed and is maintained by the American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (UICC) (Edge & Compton 2010). According to AJCC, TNM staging system is based on 5 key pieces of information:

1. The extent of the main tumour (T category)
2. Has the cancer spread to nearby lymph nodes (N category)
3. Has the cancer metastasized (spread) to other parts of the body (M category)
4. The PSA level at the time of diagnosis
5. The Gleason score, based on the prostate biopsy (or surgery) (American Cancer Society, 2015)

These factors are combined to determine an overall stage, using Roman numerals i to iv (1 - 4). The lower the number, the less the cancer has spread. A high number, such

as stage IV (4), means a more advanced cancer (i) Stage I cancer is found only in the prostate and usually grows slowly (ii) Stage II cancer has not spread beyond the prostate gland, but involves more than one part of the prostate, and may tend to grow more quickly (iii) Stage III cancer has spread beyond the outer layer of the prostate into nearby tissues or to the seminal vesicles, the glands that help produce semen (iv) Stage IV cancer has spread to other areas of the body such as the bladder, rectum, bone, liver, lungs, or lymph nodes (AJCC).

1.11. Epidemiology of prostate cancer

Prostate cancer is the most common malignant tumour in men (Gronberg, 2003) and androgen-ablation therapy, the current management of advanced prostate cancer, reduces symptoms in about 70–80% of patients but do not provide a cure and most tumours relapse within 2 years resulting in and ultimately being responsible for prostate cancer mortality (Damber and Aus, 2008). Since 2012, prostate cancer is the most frequently diagnosed cancer (15% of all male cancers) and with an estimated 307,000 deaths in 2012, prostate cancer is the fifth leading cause of death from cancer in men (6.6% of the total deaths in men). PSA testing has a much greater effect on incidence than on mortality, there is less variation in mortality rates worldwide (ten-fold from approximately 3 to 30 per 100,000) than is observed for incidence, [GLOBOCAN 2012 (IARC), Section of Cancer Surveillance (26/6/2016)] (Ferlay *et al.,* 2014).

Geography: Prostate cancer is most common in North America, North Western Europe, Australia, and the Caribbean, and it is less common in Asia, Africa, and Central and South America (Breslow *et al.,* 1977).

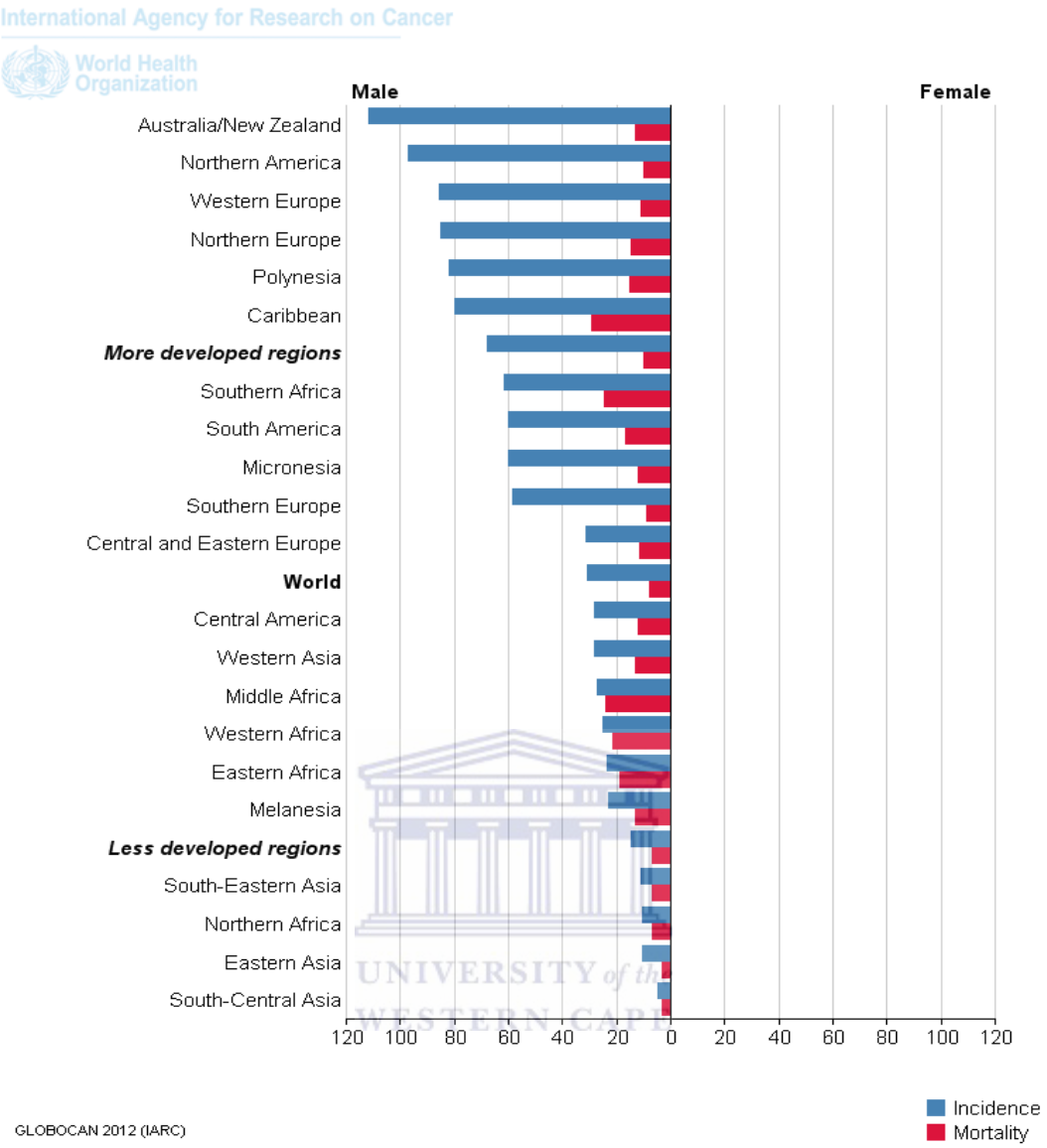


Figure 1.7: Mortality and incidence rates of prostate cancer based on geographical location in males estimated age-standardised rates (World) per 100,000.

Taken from: GLOBOCAN 2012 (IARC), Section of Cancer Surveillance (26/6/2016); World Cancer Report 2014).

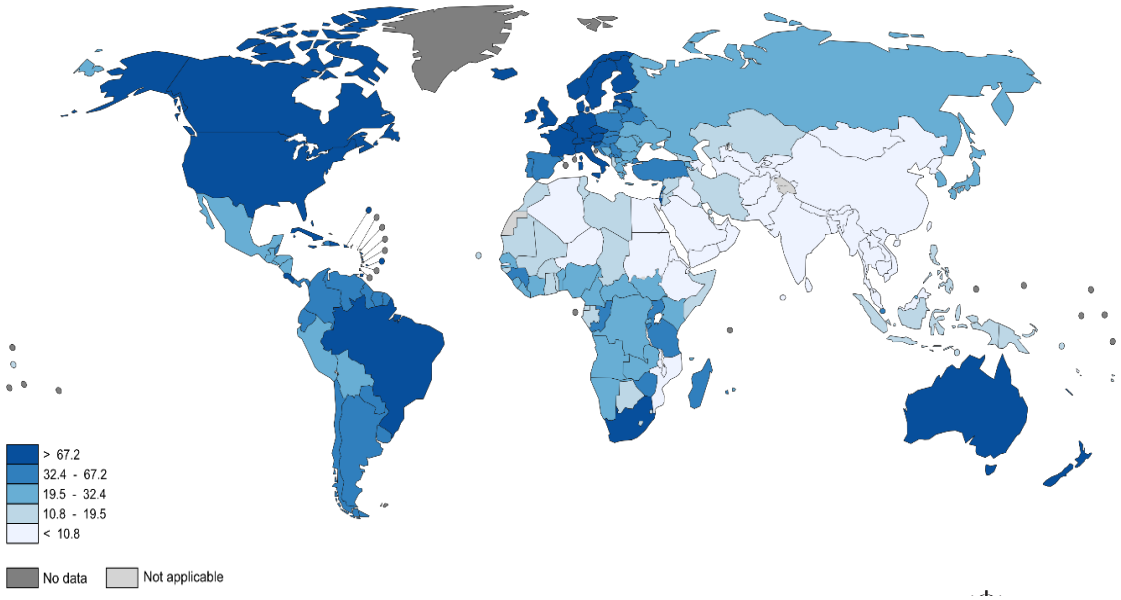


Figure 1.8: Estimated Prostate Cancer Incidence Worldwide in 2012 Estimated age-standardised rates (World) per 100,000 (GLOBOCAN 2012 (IARC) International Agency for Research on Cancer), WHO World Health Organization, 2015).

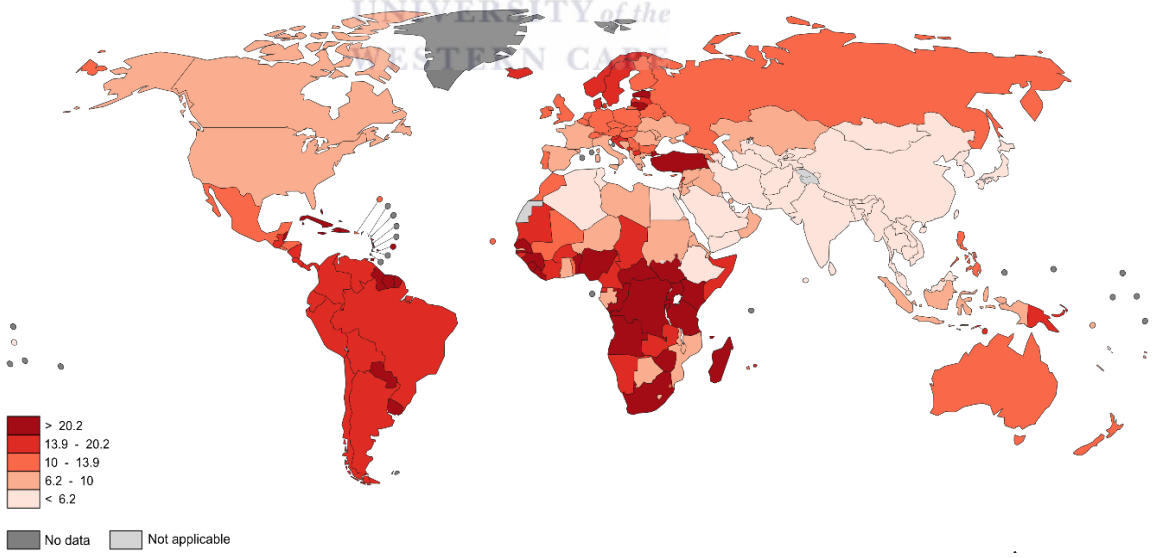


Figure 1.9: Estimated Prostate Cancer Mortality Worldwide in 2012 Estimated age-standardised rates (World) per 100,000(GLOBOCAN 2012(IARC) International Agency for Research on Cancer), WHO. World Health Organization, 2015). Risk factors of prostate cancer.

The global and region-specific estimates presented here and in more detail online (http://globocan.iarc.fr) provide a key resource for cancer researchers on the burden of cancer in 2012, and the cancer-specific patterns in 184 countries worldwide (Ferlay *et al.*, 2014).

1.12. Factors causing Prostate Cancer

Epidemiology includes endogenous factors [family history, hormones, race, aging and oxidative stress] and exogenous factors [diet, environmental agents, occupation and other factors, including lifestyle factors] (Bostwick *et al*., 2004). Epidemiologic studies have provided the greatest amount of information to date regarding risk of prostate cancer. However, most of these studies have significant problems with exposure and disease characterization (Bostwick *et al*., 2004).

Age: Prostate cancer risk goes up as men get older (Hankey *et al*., 1999). Almost 6 cases in 10 are diagnosed at the age of 65 years or later. Age, especially 55 years and above had almost 17-fold higher risk of developing prostate cancer as compared to age less than 55 years (Bashir, 2015).

Race: In the US, African-American men are more likely to get prostate cancer and die of it than in men of other races. Differences in prostate cancer risk by race may reflect three factors: differences in exposure, such as dietary differences (exogenous factors); differences in detection (reflecting exogenous factors); and genetic differences (endogenous factors) (Bostwick *et al*., 2004). African-Americans have the highest rates of prostate cancer in the world (223.0 per 100,000 men) (Bashir, 2015).

Family history: Men with close family members (father or brother) who have had prostate cancer are more likely to get it themselves (Zeegers *et al*., 2003).

Men with a father or brother affected were twice as likely to develop prostate cancer as men with no relatives affected. Increasing risk with increasing number of affected family members such that men with two or three first degree relatives affected had a five and 11–fold increased risk of developing prostate cancer (Steinberg *et al.*, 1990, Bashir, 2015).

Diet: Men who eat a lot of red meat or high-fat dairy products seem to have a greater chance of getting prostate cancer. Fat consumption, especially polyunsaturated fat, shows a strong, positive correlation with prostate cancer incidence and mortality (Bostwick *et al.*, 2004, Hayes *et al.,* 1999).

A small increased risk with lack of exercise is found in few cases of prostate cancer patients (Friedenreich *et al*., 2010). Obese men had almost six-fold higher risk of developing prostate cancer as compared to non-obese men (Bashir, 2015).

1.13. Diagnosis of prostate cancer

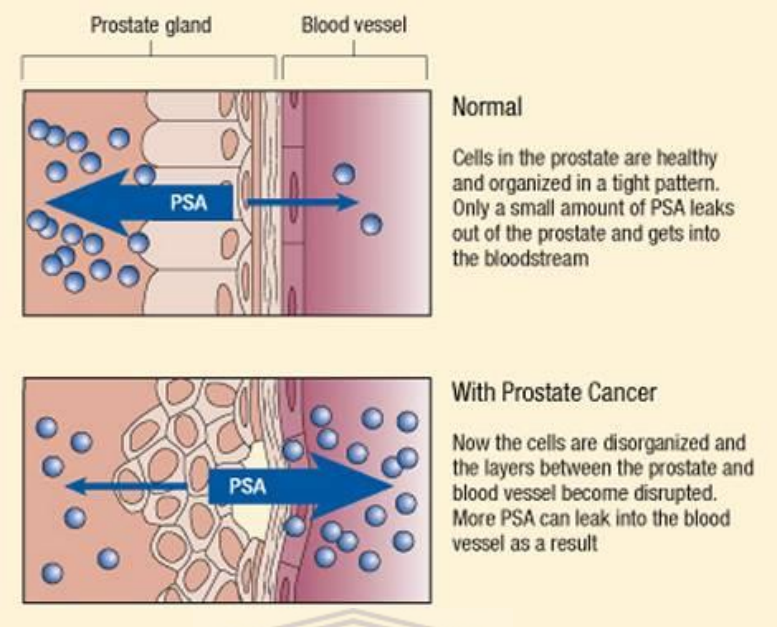


Figure 1.10: Progress of cancer in the prostate gland (http://www.prostateuk.org/psa/psa.htm)

The widespread use of PSA screening has led to a dramatic downstaging of prostate cancer at diagnosis. Prostate cancer is currently characterized by its clinical TNM stage, Gleason grade, and PSA serum level. Imaging is becoming increasingly important in the assessment of prostate cancer since it can guide treatment selection, as well as treatment planning (Hricak *et al.,* 2007). Imaging tests can determine if the cancer has spread. Microscopic spread of disease and early macroscopic invasion cannot be reliably shown using current technology (Heenan, 2004). There is no reliable and accurate test or imaging modality that can confidently diagnose and stage prostate cancer. A combination of digital rectal examination (DRE), prostate-specific antigen (PSA) level and transrectal ultrasound (TRUS) with systematic core biopsy and assessment of Gleason grade are used for diagnosis (Hricak *et al.,* 2007).

- Magnetic resonance imaging (MRI) offers the single most accurate assessment of local disease and regional metastatic spread of tumours. Numerous studies have reported varying accuracies indicating that MRI is not the perfect imaging modality (Hricak *et al.,* 2007). Currently MRI is not advocated for routine staging but it offers advantages over other imaging techniques in selected patients. The use of MRI concluded that prostate cancer could not be differentiated from benign prostatic hyperplasia (BPH) and that its usefulness in staging was comparable to digital rectal examination (DRE) (Heenan, 2004). MR imaging, with high-resolution T2-weighted scans, MR spectroscopy, and dynamic contrast enhancement, is increasingly seen as a method that can improve prostate cancer detection, characterization, staging, and treatment follow-up (Hricak *et al.,* 2007).

- Computerized axial tomography scan (CAT scan) continues to be widely used in patients with newly diagnosed prostate cancer; it has virtually no role in prostate cancer detection or primary tumour staging. On CT scans, the separation between the prostate and the levatorani muscle is poorly defined, and intraprostatic anatomy is not well demonstrated. The major role of CT is in the nodal staging of prostate cancer, for which it is limited (Hricak *et al.,* 2007).
- Prostatic ultrasound/ transrectal ultrasound (TRUS) is the most widely used clinical imaging method and it is the essential imaging tool for prostate cancer biopsy guidance. When prostate cancer is suspected, the diagnostic test of choice is a systematic needle biopsy with ultra sound guidance. Even with such systematic sampling, underdiagnosis of the extent of prostate cancer can occur with transrectal ultra sound-guided biopsy (Hricak *et al*., 2007).

Cancer screening is an examination to detect cancer before a person has symptoms. The American Cancer Society recommends screening for early detection, particularly for high-risk people or people with symptoms. Diagnosis is confirmed by various methods - clinical (physician) examination, laboratory testing (PSA test), scans (MRI, CAT), DRE (digital rectal examination) and transrectal ultrasound (TRUS). Precise indications for and sensitivity and specificity of conventional imaging methods such as radionuclide bone scanning, computed tomography (CT), magnetic resonance (MR) imaging, ultrasonography (US), and combined positron emission tomography (PET)/CT remain under debate (Hricak *et al.,* 2007).

1.14. Prostate-specific antigen

PSA is a protein produced by the cells of the prostate gland. PSA is found in the blood and measured in nanograms of PSA per millilitre (ng/mL) of blood. As seen in figure 1.10, increased PSA levels are found in the blood of men with prostate cancer. The test is used to monitor the progression of prostate cancer. PSA is one of the best-known biomarkers in medicine but due to its insufficient specificity, researchers are looking at tests based on several biomarkers. PSA testing increases cancer detection but does not decrease mortality (Djulbegovic *et al*., 2010). Small cell carcinoma is a rare type of prostate cancer that is serious and spreads quickly to other parts of the body and cannot be detected using the PSA test thus various other biomarkers are necessary to detect these kinds of rare and hidden cancers (Nutting et al., 1997; Wei et al., 2009).

PSA level is a strong indicator of stage and prognosis and is helpful in monitoring response to therapy. However, absolute PSA serum levels must be interpreted carefully with regard to the age of the patient, the size of the gland, and the presence of infection

(Hricak *et al.,* 2007). PSA was first used by forensic scientists as a marker for human semen. In the late 1970s, the initial laboratory studies were conducted to evaluate the relationship between PSA and prostatic disease (Loeba and Catalonab, 2007).

In 1986, the U.S. Food and Drug Administration (FDA) approved the PSA test for prostate cancer screening. In the late 1980s the first clinical studies were initiated to examine the role of PSA in prostate cancer screening. Previously, most prostate cancer cases were diagnosed either through a suspicious digital rectal examination (DRE) or as an incidental finding in the prostate chips from transurethral resection for presumed benign prostatic hyperplasia (BPH) (Loeba and Catalonab, 2007).

There are several problems that complicate its use in daily practice. Firstly, PSA can also be elevated in benign prostatic conditions, limiting its specificity for prostate cancer. Even certain medications (e.g., Finasteride), ejaculation, and prostate manipulation (e.g., catheterization, cystoscopy, and prostatic massage) can alter PSA levels. The concerns about overdiagnosis of prostate cancer have become increasingly more resonant (Loeba and Catalonab, 2007) and it's unlikely that PSA by itself will be an effective screening tool for the early diagnosis of prostate cancer. However, if combined with digital rectal examination and/or transrectal ultrasound it is a vital part of any early detection program. Conditions such as bacterial prostatitis and acute urinary retention can also falsely elevate the serum PSA level. Approximately 25% of the patients with BPH will have an elevated serum PSA concentration. Prostatic intraepithelial neoplasia may also be associated with moderately elevated serum PSA levels (Oesterling, 1991).

1.15. Biomarkers

1.15.1. What are Biomarkers?

A biomarker is a gene, protein/peptide or metabolite present in a biological system, used to indicate a physiological or pathological state that can be recognized or monitored (Azuaje, 2010).

Biomarkers are used to measure the progress of disease or the physiological effects of therapeutic intervention in the treatment of disease. The utility of a biomarker lies in its ability to provide an early indication of the disease, to monitor disease progression, to provide ease of detection and to provide a factor measurable across populations (Srinivas *et al*., 2002). A practical serological biomarker should have certain characteristics, *i.e.* it is a secreted or shed protein and has the ability to diffuse into the circulation during tumour development and progression, through either angiogenesis or invasion of surrounding tissues and vasculature by cancer cells (Diamandis, 2004).

Biomarkers can be proteins, metabolites, RNA transcripts, DNA, or epigenetic modifications of DNA, among other alterations. They can be detected through patient tissue samples, obtained either by biopsy or surgical resection, or non-invasively through the isolation of cells and/or molecules from bodily fluids, such as blood or urine (Prensner *et al*., 2012).

1.15.2. Types of Biomarkers

Promising biomarkers should be overexpressed proteins but this is not generally true for some of the best known cancer biomarkers such as PSA (Diamandis, 2004).

On the basis of their application to the detection of disease, three main classes of biomarkers may be specified: screening, diagnostic and prognostic biomarkers. Screening biomarkers are used to predict the potential occurrence of a disease in asymptomatic patients. Diagnostic biomarkers are used to make predictions on patients suspected of having the disease. Prognostic biomarkers are applied to predict the outcome of a patient suffering from a disease. Molecular biomarkers are measured in biological samples: solid tissues, blood or other bodily fluids (Azuaje, 2010).

1.15.3. Methods to discover Biomarkers

Biomarker discovery is a challenging process and a good biomarker has to be sensitive, specific and its test highly standardized and reproducible as well as identify risk for or diagnose a disease, assess disease severity or progression, predict prognosis or guide treatment. Biomarker discovery is a major research field of differential omics. It is a process of discovery, verification and validation. During the discovery and validation process there should be insight into the understanding of the molecular mechanism of the diseases (Azuaje, 2010).

Biomarker verification is done in two ways:

I. Wet-lab verification - Sample processing molecular ("omic") data generation hypotheses.
II. Computer-based techniques (Bioinformatics) - Data pre-processing, feature extraction and selection, statistical analysis, predictive modelling and refined / new hypotheses (Azuaje, 2010).

1.16. Biomarkers currently used for prostate cancer diagnosis

For early diagnosis, it is best to find a non-invasive cancer biomarker to monitor molecular differences in tumours thus assisting in better treatment for cancer patients. A cancer biomarker, prostate-specific antigen (PSA), is a protein found in the blood of adult men with prostate cancer (Makarov *et al*., 2009).

Prostate acid phosphatase (PAP) was the first reported biomarker to be elevated in the serum of patients with prostate cancer in 1930s (Gutman AB *et al*., 1938). However, PAP proved to be insensitive to detect localized lesion of prostate cancer and was replaced by PSA that was discovered in 1970 (Makarov *et al*., 2009). The low specificity of PSA in detecting prostate cancer caused relative high false- positive rate in screening (Makarov *et al*., 2009).

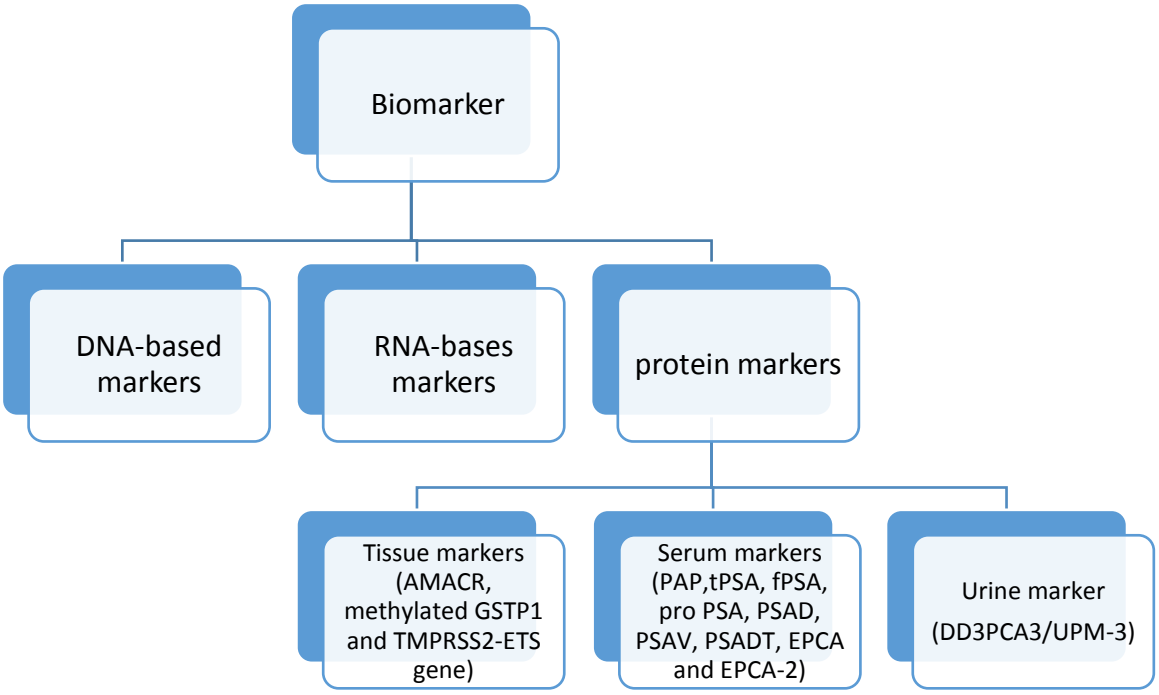


Figure 1.11: Classification of biomarkers in prostate cancer (Ludwig JA *et al*., 2005).

As new biomarkers are discovered, the following are needed in prostate cancer biomarker discovery; validating the many existing prostate biomarkers already discovered; (i) developing markers to minimize the number of unnecessary prostate biopsies. (ii) developing markers to identify men with indolent prostate cancer who will not be affected by disease in their lifetimes and do not need treatment. (iii) developing markers to identify men with aggressive disease who will benefit from local therapy and those who are likely to fail local therapy and require adjuvant intervention and (iv) developing markers that may serve as surrogate endpoints for clinical progression or survival (Makarov *et al*., 2009).

Future research can focus on validation of already existing biomarkers and the discovery of new markers to identify men with aggressive prostate cancer (Baumgartner *et al.*, 2011). miRNAs has come to the fore to that it necessitates closer inspections as prostate cancer biomarkers.

1.17. MiRNAs activity

MiRNA are regulatory RNA which play a role in gene expression. MiRNAs do not code for proteins, but can "target" protein-coding genes and reduce their expression. In mutations, the expression of the miRNAs of the genes it regulates is misexpressed resulting in cancer. Altered miRNA expression or the target genes thereof may serve as a biomarker for cancer diagnosis and treatment. Circulating miRNAs found in the blood of cancer patients has raised the possibility that miRNAs can serve as a novel diagnostic biomarker (Kosaka *et al*., 2010).

The miRNAs play a role in physiological and pathological processes, such as development, cell proliferation, apoptosis and stress responses (Bartel, 2004). A single miRNA is said to target several mRNAs and in this sense if there are mutations in the miRNA expression then the expression of several mRNAs and proteins are also disrupted. In various studies, researchers have shown that measuring different miRNA levels in serum demonstrated that they could distinguish patients with cancer from healthy individuals (Lawrie *et al*., 2008, Mitchell *et al*., 2008).

1.18. MiRNAs and its target genes/proteins as diagnostic biomarkers

MiRNA suppress the expression of oncogenes, growth promoting, survival and angiogenic genes (low in tumours). MiRNA suppress expression of tumour suppressor, growth inhibitory and proapoptotic genes. A select number of miRNAs may serve as diagnostic markers or even potential therapeutic targets for different tumour types as several miRNAs are up-or down-regulated in multiple tumours (Sevli *et al*., 2010).

MiRNAs has critical functions in gene expression and dysregulation may cause tumour formation and progression. When miRNAs are up-regulated it inhibits tumour suppressor genes in tumour cells (Sevli *et al*., 2010) and are known as oncogenic miRNAs or oncomirs whereas the miRNAs that are down-regulated causing tumour progression are tumour suppressor miRNAs. The reduced expression is a result of increased levels of oncogene expression (Sevli *et al*., 2010).

miRNAs regulate gene expression by direct cleavage of the targeted mRNAs or inhibiting translation through perfect or nearly perfect complementarity to targeted mRNAs. These targeted genes control multiple biological processes, including stem cell division, apoptosis and cancer (Zhang *et al*., 2007).

miRNAs with high influence on protein networks are valuable biomarkers that can be used in clinical investigations for cancer treatment. Over- or under-expression of specific miRNAs in different tumours makes them potential diagnostic or prognostic biomarkers, however, miRNAs that are differentially expressed and influence their targets and target partners are important regulators and thus are more promising for diagnostics, prognostics or therapy (Alshalalfa *et al*., 2012).

1.19. Bioinformatics – a search tool for biomarker discovery

Computational approaches are effective for the detection of potential miRNAs in cancer. Biomarker discovery for prostate cancer of mRNA and microRNA expression are strongly needed to enable more accurate detection of prostate cancer, improve prediction of tumour aggressiveness and facilitate diagnosis. Computational biology (bioinformatics) plays a significant role in the discovery of new biomarkers, the analyses of disease states and the validation of potential biomarkers (Srinivas *et al*., 2002).

Bioinformatics is crucial in the analysis of functional genomics data such as data clustering or principal component analysis. Functional enrichment analysis is used to determine the extent of over- or under-expression of functional categories relative to a background sets (Subramanian *et al.,* 2005). Functional genomics describes the use of large scale data produced by high throughput technologies or databases to understand the function of genes and other parts of the genome (Narayanan, 2007).

Bioinformatics, such as digital differential display and *in silico* Northern blotting is used to compare gene expression between normal and cancerous tissues to identify overexpressed genes (Diamandis, 2004).

This technology is the key to understanding and management of biological information and is essential to use the genomic information in understanding human diseases. There is growth in biological data due to the numerous amounts of research being done in the molecular field. Large databases are available, accessible and analysed for biological research and education. Prediction of a sequence (what we know), we can predict the structure and function (what we don't know). Target proteins structure is compared with related proteins and proteins with similar sequences (Baumgartner *et al.*, 2011).

In general, the search, verification, biological and biochemical interpretation and independent validation of disease biomarkers require new innovations in high-throughput technologies, biostatistics and bioinformatics and thus make necessary the interdisciplinary expertise and teamwork of clinicians, biologists, analytical- and biochemists and bioinformaticians to carry out all steps of a biomarker cohort study with professional planning, implementation, and control. Bioinformatics plays a key role in the biomarker discovery process, bridging the gap between initial discovery

phases such as experimental design, clinical study execution, and bioanalytics, including sample preparation, separation and high-throughput profiling and independent validation of identified candidate biomarkers (Baumgartner *et al*., 2011).

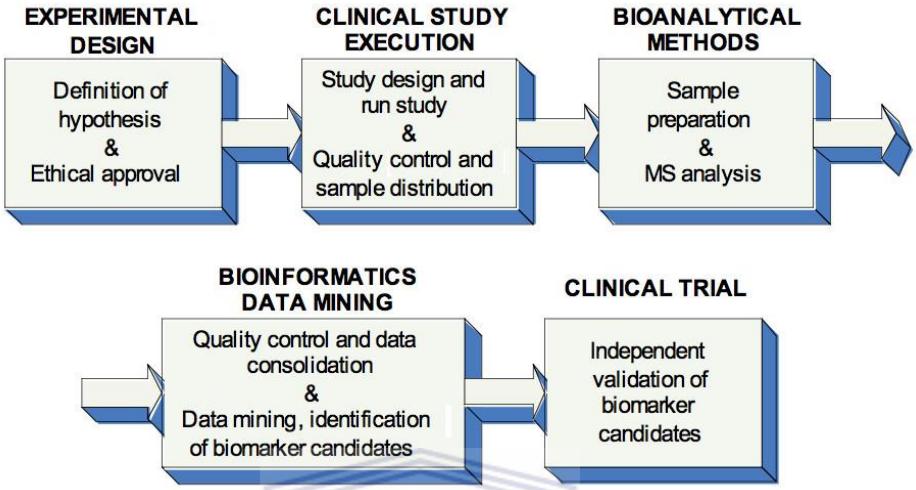


Figure 1.12: Tools for discovery and validation of biomarkers (Baumgartner *et al.*, 2011).

1.20. Conclusion

There is a need to develop new biomarkers for prostate cancer and this project is an attempt to identify effective biomarkers for prostate cancer using bioinformatics with the goal of finding more specific biomarkers to avoid overdiagnosis and overtreatment associated with PSA screening of prostate cancer. No individual marker is ideal but further validation of promising markers and continued discovery of biomarkers is needed. Even a combination of biomarkers should improve the predictive accuracy and using protein-protein interaction network for prostate cancer should hopefully help improvement in clinical outcome.

1.21. Aims and objectives

The aim of this study was to identify prostate specific biomarkers for early diagnosis of the disease using *in silico* methods.

Objectives:

1. Identifying co-expression genes/proteins of 21 genes regulated by 13 miRNAs in prostate cancer.
2. Extracting and refining a gene list from various databases for identifying potential biomarkers as diagnostic molecules for prostate cancer.
3. Associate the candidate gene list to transcription factors to confirm their regulatory control in cancer processes and specifically prostate cancer.
4. Assess the association of the candidate genes to pathways that contribute to the outcome of prostate cancer for better and supportive diagnostic evaluation.
5. Evaluate the Prognostic and Predictive value of the candidate genes to assess their usefullness in the prognosis of prostate cancer.

1.22. References

ALSHALALFA, M., BADER, G. D., GOLDENBERG, A., MORRIS, Q. & ALHAJJ, R. 2012. Detecting microRNAs of high influence on protein functional interaction networks: a prostate cancer case study. *BMC systems biology*, 6, 1.

AMBS, S., PRUEITT, R. L., YI, M., HUDSON, R. S., HOWE, T. M., PETROCCA, F., WALLACE, T. A., LIU, C. G., VOLINIA, S., CALIN, G. A., YFANTIS, H. G., STEPHENS, R. M. & CROCE, C. M. 2008. Genomic profiling of microRNA and messenger RNA reveals deregulated microRNA expression in prostate cancer. *Cancer Res,* 68**,** 6162-70.

BALDASSARRE, G., BATTISTA, S., BELLETTI, B., THAKUR, S., PENTIMALLI, F., TRAPASSO, F., FEDELE, M., PIERANTONI, G., CROCE, C. M. & FUSCO, A. 2003. Negative regulation of BRCA1 gene expression by HMGA1 proteins accounts for the reduced BRCA1 protein levels in sporadic breast carcinoma. *Mol Cell Biol,* 23**,** 2225-38.

BARTEL, D. P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell,* 116**,** 281-97.

BASHIR, M. N. 2015. Epidemiology of Prostate Cancer. *Asian Pac J Cancer Prev,* 16**,** 5137-41.

BASHYAM, M. D. 2002. Understanding cancer metastasis: an urgent need for using differential gene expression analysis. *Cancer,* 94**,** 1821-9.

BAUMGARTNER, C., OSL, M., NETZER, M. & BAUMGARTNER, D. 2011. Bioinformatic-driven search for metabolic biomarkers in disease. *J Clin Bioinforma,* 1**,** 2.

BAYLIN, S. B. & OHM, J. E. 2006. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? *Nat Rev Cancer,* 6**,** 107-16.

BERNSTEIN, C., NFONSAM, V., PRASAD, A. R. & BERNSTEIN, H. 2013a. Epigenetic field defects in progression to cancer. *World J Gastrointest Oncol,* 5**,** 43-9.

BERNSTEIN, C., R, A., NFONSAM, V. & BERNSTEI, H. 2013b. DNA Damage, DNA Repair and Cancer.

BLAKE, J. A., DOLAN, M., DRABKIN, H., HILL, D. P., NI, L., SITNIKOV, D., BURGESS, S., BUZA, T., GRESHAM, C., MCCARTHY, F., PILLAI, L., WANG, H., CARBON, S., LEWIS, S. E., MUNGALL, C. J., GAUDET, P., CHISHOLM, R. L., FEY, P., KIBBE, W. A., BASU, S., SIEGELE, D. A., MCINTOSH, B. K., RENFRO, D. P., ZWEIFEL, A. E., HU, J. C., BROWN, N. H., TWEEDIE, S., ALAM-FARUQUE, Y., APWEILER, R., AUCHINCHLOSS, A., AXELSEN, K., ARGOUD-PUY, G., BELY, B., BLATTER, M. C., BOUGUELERET, L., BOUTET, E., BRANCONI-QUINTAJE, S., BREUZA, L., BRIDGE, A., BROWNE, P., CHAN, W. M., COUDERT, E., CUSIN, I., DIMMER, E., DUEK-ROGGLI, P., EBERHARDT, R., ESTREICHER, A., FAMIGLIETTI, L., FERRO-ROJAS, S., FEUERMANN, M., GARDNER, M., GOS, A., GRUAZ-GUMOWSKI, N., HINZ, U., HULO, C., HUNTLEY, R., JAMES, J., JIMENEZ, S., JUNGO, F., KELLER, G., LAIHO, K., LEGGE, D., LEMERCIER, P., LIEBERHERR, D., MAGRANE, M., MARTIN, M. J., MASSON, P., MOINAT, M., O'DONOVAN, C., PEDRUZZI, I., PICHLER, K., POGGIOLI, D., MILLAN, P. P., POUX, S., RIVOIRE, C., ROECHERT, B., SAWFORD, T., SCHNEIDER, M., SEHRA, H., STANLEY, E., STUTZ, A., SUNDARAM, S., TOGNOLLI, M., XENARIOS, I., FOULGER, R., LOMAX, J., RONCAGLIA, P., CAMON, E., KHODIYAR, V. K., LOVERING, R. C., TALMUD, P. J., CHIBUCOS, M., GIGLIO, M. G., DOLINSKI, K., HEINICKE, S., LIVSTONE, M. S., STEPHAN, R., HARRIS, M. A., OLIVER, S. G., RUTHERFORD, K., et al. 2012. The Gene Ontology: enhancements for 2011. *Nucleic Acids Research,* 40**,** D559-D564.

BOSTWICK, D. G., BURKE, H. B., DJAKIEW, D., EULING, S., HO, S. M., LANDOLPH, J., MORRISON, H., SONAWANE, B., SHIFFLETT, T., WATERS, D. J. & TIMMS, B. 2004. Human prostate cancer risk factors. *Cancer,* 101**,** 2371-490.

BOURDOUMIS, A., PAPATSORIS, A. G., CHRISOFOS, M., EFSTATHIOU, E., SKOLARIKOS, A. & DELIVELIOTIS, C. 2010. The novel prostate cancer antigen 3 (PCA3) biomarker. *Int Braz J Urol,* 36**,** 665-8; discussion 669.

BRESLOW, N., CHAN, C. W., DHOM, G., DRURY, R. A., FRANKS, L. M., GELLEI, B., LEE, Y. S., LUNDBERG, S., SPARKE, B., STERNBY, N. H. & TULINIUS, H. 1977. Latent carcinoma of prostate at autopsy in seven areas. The International Agency for Research on Cancer, Lyons, France. *Int J Cancer,* 20**,** 680-8.

CANNELL, I. G., KONG, Y. W. & BUSHELL, M. 2008. How do microRNAs regulate gene expression? *Biochem Soc Trans,* 36**,** 1224-31.

CARLETON, M., CLEARY, M. A. & LINSLEY, P. S. 2007. MicroRNAs and cell cycle regulation. *Cell Cycle,* 6**,** 2127-32.

CARY, K. C. & COOPERBERG, M. R. 2013. Biomarkers in prostate cancer surveillance and screening: past, present, and future. *Ther Adv Urol,* 5**,** 318-29.

CHO, W. C. S. 2007. Contribution of oncoproteomics to cancer biomarker discovery. *Molecular Cancer,* 6, 25. http://doi.org/10.1186/1476-4598-6-25

CHO, W. C. 2007. Proteomics technologies and challenges. Genomics Proteomics *Bioinformatics*, 5, 77-85.

COHEN, R. J., SHANNON, B. A., PHILLIPS, M., MOORIN, R. E., WHEELER, T. M. & GARRETT, K. L. 2008. Central zone carcinoma of the prostate gland: a distinct tumor type with poor prognostic features. *J Urol,* 179**,** 1762-7; discussion 1767.

COPPOLA, V., DE MARIA, R. & BONCI, D. 2010. MicroRNAs and prostate cancer. *Endocr Relat Cancer,* 17**,** F1-17.

CROCE, C. M. 2008. Oncogenes and cancer. *N Engl J Med,* 358**,** 502-11.

DAMBER, J. E. & AUS, G. 2008. Prostate cancer. *Lancet,* 371**,** 1710-21.

'Defining Cancer'. *National Cancer Institute.* Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer [10 June 2014].

DIAMANDIS, E. P. 2004. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol Cell Proteomics,* 3**,** 367-78.

DJULBEGOVIC, M., BEYTH, R. J., NEUBERGER, M. M., STOFFS, T. L., VIEWEG, J., DJULBEGOVIC, B. & DAHM, P. 2010. Screening for prostate cancer: systematic review and meta-analysis of randomised controlled trials. *BMJ,* 341**,** c4543.

EDGE, S. B. & COMPTON, C. C. 2010. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann Surg Oncol,* 17**,** 1471-4.

ESQUELA-KERSCHER, A. & SLACK, F. J. 2006. Oncomirs - microRNAs with a role in cancer. *Nat Rev Cancer,* 6**,** 259-69.

FERLAY, J., SOERJOMATARAM, I., DIKSHIT, R., ESER, S., MATHERS, C., REBELO, M., PARKIN, D. M., FORMAN, D. & BRAY, F. 2015. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer,* 136**,** E359-86.

FRIEDENREICH, C. M., NEILSON, H. K. & LYNCH, B. M. 2010. State of the epidemiological evidence on physical activity and cancer prevention. *Eur J Cancer,* 46**,** 2593-604.

FRITZ, A., PERCY, C., JACK, A., SHANMUGARATNAM, K., SOBIN, L., PARKIN, D. M. & WHELAN, S. 2000. International classification of diseases for oncology (No. Ed. 3 pp.vii + 240 pp.) World Health Organization.

GALLAGHER, R. P. & FLESHNER, N. 1998. Prostate cancer: 3. Individual risk factors. *CMAJ,* 159**,** 807-13.

GOLDSTEIN, A. S., HUANG, J., GUO, C., GARRAWAY, I. P. & WITTE, O. N. 2010. Identification of a cell of origin for human prostate cancer. *Science,* 329**,** 568-71.

GRONBERG, H. 2003. Prostate cancer epidemiology. *Lancet,* 361**,** 859-64.

GROUP, BMP. 2009 Retrieved 2010. Prostate cancer: How far has your cancer spread? The TNM system.

GUTMAN, A. B. & GUTMAN, E. B. 1938. An " Acid " Phosphatase Occurring in the Serum of Patients with Metastasizing Carcinoma of the Prostate Gland. *J Clin Invest,* 17**,** 473-8.

HANAHAN, D. & WEINBERG, R. A. 2000. The hallmarks of cancer. *Cell,* 100**,** 57-70.

HANAHAN, D. & WEINBERG, R. A. 2011. Hallmarks of Cancer: The Next Generation. *Cell,* 144**,** 646-674.

HANKEY, B. F., FEUER, E. J., CLEGG, L. X., HAYES, R. B., LEGLER, J. M., PROROK, P. C., RIES, L. A., MERRILL, R. M. & KAPLAN, R. S. 1999. Cancer surveillance series: interpreting trends in prostate cancer--part I: Evidence of the effects of screening in recent prostate cancer incidence, mortality, and survival rates. *J Natl Cancer Inst,* 91**,** 1017-24.

HANSEL, D. E., DEMARZO, A. M., PLATZ, E. A., JADALLAH, S., HICKS, J., EPSTEIN, J. I., PARTIN, A. W. & NETTO, G. J. 2007. Early prostate cancer antigen expression in predicting presence of prostate cancer in men with histologically negative biopsies. *J Urol,* 177**,** 1736-40.

HAYES, R. B., ZIEGLER, R. G., GRIDLEY, G., SWANSON, C., GREENBERG, R. S., SWANSON, G. M., SCHOENBERG, J. B., SILVERMAN, D. T., BROWN, L. M., POTTERN, L. M., LIFF, J., SCHWARTZ, A. G., FRAUMENI, J. F., JR. & HOOVER,

R. N. 1999. Dietary factors and risks for prostate cancer among blacks and whites in the United States. *Cancer Epidemiol Biomarkers Prev,* 8**,** 25-34.

HEENAN, S. D. 2004. Magnetic resonance imaging in prostate cancer. *Prostate Cancer Prostatic Dis,* 7**,** 282-8.

HEJMADI, M. 2009. *Introduction to cancer biology*, Bookboon. Holstebro, Denmark: Ventus Publishing.

HILL, D. P., SMITH, B., MCANDREWS-HILL, M. S. & BLAKE, J. A. 2008. Gene Ontology annotations: what they mean and where they come from. *BMC Bioinformatics,* 9 Suppl 5**,** S2.

HRICAK, H., CHOYKE, P. L., EBERHARDT, S. C., LEIBEL, S. A. & SCARDINO, P. T. 2007. Imaging prostate cancer: a multidisciplinary perspective. *Radiology,* 243**,** 28-53.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res,* 37**,** 1-13.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc,* 4**,** 44-57.

JACINTO, F. V. & ESTELLER, M. 2007. Mutator pathways unleashed by epigenetic silencing in human cancer. *Mutagenesis,* 22**,** 247-53.

JACKSON, R. J. & STANDART, N. 2007. How do microRNAs regulate gene expression? *Sci STKE,* 367 (re1).

JANSSON, M. D. & LUND, A. H. 2012. MicroRNA and cancer. *Mol Oncol,* 6**,** 590-610.

JEMAL, A., BRAY, F., CENTER, M. M., FERLAY, J., WARD, E. & FORMAN, D. 2011. Global cancer statistics. *CA Cancer J Clin,* 61**,** 69-90.

KANWAL, R. & GUPTA, S. 2012. Epigenetic modifications in cancer. *Clin Genet,* 81**,** 303-11.

KAPUSHESKY, M., ADAMUSIAK, T., BURDETT, T., CULHANE, A., FARNE, A., FILIPPOV, A., HOLLOWAY, E., KLEBANOV, A., KRYVYCH, N., KURBATOVA, N., KURNOSOV, P., MALONE, J., MELNICHUK, O., PETRYSZAK, R., PULTSIN, N., RUSTICI, G., TIKHONOV, A., TRAVILLIAN, R. S., WILLIAMS, E., ZORIN, A., PARKINSON, H. & BRAZMA, A. 2012. Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res,* 40**,** D1077-81.

KASTAN, M. B. & BARTEK, J. 2004. Cell-cycle checkpoints and cancer. *Nature,* 432**,** 316-23.

KNUDSON, A. G. 2001. Two genetic hits (more or less) to cancer. *Nat Rev Cancer,* 1**,** 157-62.

KOSAKA, N., IGUCHI, H. & OCHIYA, T. 2010. Circulating microRNA in body fluid: a new potential biomarker for cancer diagnosis and prognosis. *Cancer Sci,* 101**,** 2087-92.

LAHTZ, C. & PFEIFER, G. P. 2011. Epigenetic changes of DNA repair genes in cancer. *J Mol Cell Biol,* 3**,** 51-8.

LAWRIE, C. H., GAL, S., DUNLOP, H. M., PUSHKARAN, B., LIGGINS, A. P., PULFORD, K., BANHAM, A. H., PEZZELLA, F., BOULTWOOD, J., WAINSCOAT, J. S., HATTON, C. S. & HARRIS, A. L. 2008. Detection of elevated levels of tumour-associated microRNAs in serum of patients with diffuse large B-cell lymphoma. *Br J Haematol,* 141**,** 672-5.

LIU, X., YU, X., ZACK, D. J., ZHU, H. & QIAN, J. 2008. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics,* 9**,** 271.

LOEB, S. & CATALONA, W. J. 2008. What to do with an abnormal PSA test. *Oncologist*, 13, 299-305.

LUDWIG, J. A. & WEINSTEIN, J. N. 2005. Biomarkers in cancer staging, prognosis and treatment selection. *Nat Rev Cancer,* 5**,** 845-56.

LUO, J. H., YU, Y. P., CIEPLY, K., LIN, F., DEFLAVIA, P., DHIR, R., FINKELSTEIN, S., MICHALOPOULOS, G. & BECICH, M. 2002. Gene expression analysis of prostate cancers. *Mol Carcinog,* 33**,** 25-35.

MADU, C. O. & LU, Y. 2010. Novel diagnostic biomarkers for prostate cancer. *J Cancer,* 1**,** 150-77.

MAKAROV, D. V., LOEB, S., GETZENBERG, R. H. & PARTIN, A. W. 2009. Biomarkers for prostate cancer. *Annu Rev Med,* 60**,** 139-51.

MARTIN, R. M., VATTEN, L., GUNNELL, D. & ROMUNDSTAD, P. 2010. Blood pressure and risk of prostate cancer: Cohort Norway (CONOR). *Cancer Causes Control,* 21**,** 463-72.

MERLO, L. M., PEPPER, J. W., REID, B. J. & MALEY, C. C. 2006. Cancer as an evolutionary and ecological process. *Nat Rev Cancer,* 6**,** 924-35.

MITCHELL, P. S., PARKIN, R. K., KROH, E. M., FRITZ, B. R., WYMAN, S. K., POGOSOVA-AGADJANYAN, E. L., PETERSON, A., NOTEBOOM, J., O'BRIANT, K. C., ALLEN, A., LIN, D. W., URBAN, N., DRESCHER, C. W., KNUDSEN, B. S., STIREWALT, D. L., GENTLEMAN, R., VESSELLA, R. L., NELSON, P. S., MARTIN, D. B. & TEWARI, M. 2008. Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci U S A,* 105**,** 10513-8.

MOORE, K. L., DALLEY, A. F. & AGUR, A. M. 2013. Clinically oriented anatomy, Lippincott Williams & Wilkins.

MORGAN, R., BOXALL, A., BHATT, A., BAILEY, M., HINDLEY, R., LANGLEY, S., WHITAKER, H. C., NEAL, D. E., ISMAIL, M., WHITAKER, H., ANNELS, N., MICHAEL, A. & PANDHA, H. 2011. Engrailed-2 (EN2): a tumor specific urinary biomarker for the early diagnosis of prostate cancer. *Clin Cancer Res,* 17**,** 1090-8.

NARAYANAN, R. 2007. Bioinformatics approaches to cancer gene discovery. *Methods Mol Biol,* 360**,** 13-31.

NILSSON, B. O., CARLSSON, L., LARSSON, A. & RONQUIST, G. 2001. Autoantibodies to prostasomes as new markers for prostate cancer. *Ups J Med Sci,* 106**,** 43-9.

NUTTING, C., HORWICH, A., FISHER, C., PARSONS, C. & DEARNALEY, D. P. 1997. Small-cell carcinoma of the prostate. *J R Soc Med,* 90**,** 340-1.

OESTERLING, J. E. 1991. Prostate specific antigen: a critical assessment of the most useful tumor marker for adenocarcinoma of the prostate. *J Urol,* 145**,** 907-23.

PANG, Y., YOUNG, C. Y. & YUAN, H. 2010. MicroRNAs and prostate cancer. *Acta Biochim Biophys Sin (Shanghai),* 42**,** 363-9.

PRENSNER, J. R., RUBIN, M. A., WEI, J. T. & CHINNAIYAN, A. M. 2012. Beyond PSA: the next generation of prostate cancer biomarkers. Science translational medicine, 4**,** 127rv3-127rv3.

PETRYSZAK, R., BURDETT, T., FIORELLI, B., FONSECA, N. A., GONZALEZ-PORTA, M., HASTINGS, E., HUBER, W., JUPP, S., KEAYS, M., KRYVYCH, N., MCMURRY, J., MARIONI, J. C., MALONE, J., MEGY, K., RUSTICI, G., TANG, A. Y., TAUBERT, J., WILLIAMS, E., MANNION, O., PARKINSON, H. E. & BRAZMA, A. 2014. Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res,* 42**,** D926-32.

PORKKA, K. P., PFEIFFER, M. J., WALTERING, K. K., VESSELLA, R. L., TAMMELA, T. L. & VISAKORPI, T. 2007. MicroRNA expression profiling in prostate cancer. *Cancer Res,* 67**,** 6130-5.

PROSTATE CANCER TREATMENT GUIDE. Prostate Cancer Information from the Foundation of the Prostate Gland. Available: http://www.prostate-cancer.com/prostate-cancer-treatment-overview/overview-prostate-anatomy.html. [14 June 2010].

QU, M., REN, S. C. & SUN, Y. H. 2014. Current early diagnostic biomarkers of prostate cancer. *Asian J Androl,* 16**,** 549-54.

SCHAEFER, A., JUNG, M., MOLLENKOPF, H. J., WAGNER, I., STEPHAN, C., JENTZMIK, F., MILLER, K., LEIN, M., KRISTIANSEN, G. & JUNG, K. 2010. Diagnostic and prognostic implications of microRNA profiling in prostate carcinoma. *Int J Cancer,* 126**,** 1166-76.

SEVLI, S., UZUMCU, A., SOLAK, M., ITTMANN, M. & OZEN, M. 2010. The function of microRNAs, small but potent molecules, in human prostate cancer. *Prostate Cancer Prostatic Dis,* 13**,** 208-17.

SHEN, M. M. & ABATE-SHEN, C. 2010. Molecular genetics of prostate cancer: new prospects for old challenges. *Genes Dev,* 24**,** 1967-2000.

STAHLHUT ESPINOSA, C. E. & SLACK, F. J. 2006. The role of microRNAs in cancer. *Yale J Biol Med,* 79**,** 131-40.

STAHLHUT ESPINOSA, C. E. & SLACK, F. J. 2006. The role of microRNAs in cancer. *Yale J Biol Med,* 79**,** 131-40.

STEINBERG, G. D., CARTER, B. S., BEATY, T. H., CHILDS, B. & WALSH, P. C. 1990. Family history and the risk of prostate cancer. *Prostate,* 17**,** 337-47.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER,

E. S. & MESIROV, J. P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A,* 102**,** 15545-50.

TONG, A. W., FULGHAM, P., JAY, C., CHEN, P., KHALIL, I., LIU, S., SENZER, N., EKLUND, A. C., HAN, J. & NEMUNAITIS, J. 2009. MicroRNA profile analysis of human prostate cancers. *Cancer Gene Ther,* 16**,** 206-16.

VELONAS, V. M., WOO, H. H., DOS REMEDIOS, C. G. & ASSINDER, S. J. 2013. Current status of biomarkers for prostate cancer. *Int J Mol Sci,* 14**,** 11034-60.

VERHAMME, K. M., DIELEMAN, J. P., BLEUMINK, G. S., VAN DER LEI, J., STURKENBOOM, M. C., ARTIBANI, W., BEGAUD, B., BERGES, R., BORKOWSKI, A., CHAPPEL, C. R., COSTELLO, A., DOBRONSKI, P., FARMER, R. D., JIMENEZ CRUZ, F., JONAS, U., MACRAE, K., PIENTKA, L., RUTTEN, F. F., VAN SCHAYCK, C. P., SPEAKMAN, M. J., STURKENBOOM, M. C., TIELLAC, P., TUBARO, A., VALLENCIEN, G., VELA NAVARRETE, R. & TRIUMPH PAN EUROPEAN EXPERT, P. 2002. Incidence and prevalence of lower urinary tract symptoms suggestive of benign prostatic hyperplasia in primary care--the Triumph project. *Eur Urol,* 42**,** 323-8.

VON MERING, C., HUYNEN, M., JAEGGI, D., SCHMIDT, S., BORK, P. & SNEL, B. 2003. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research,* 31**,** 258-261.

WILD, C. P., STEWART, B. W. & WILD, C. 2014. World Cancer Report 2014, IARC Press.

WEI, Z. F., XU, H., WANG, H., WEI, W., CHENG, W., ZHOU, W. Q., GE, J. P., ZHANG, Z. Y., GAO, J. P. & YIN, H. L. 2009. [Clinicopathological characterization of

prostatic small cell carcinoma: a case report and review of the literature]. *Zhonghua Nan Ke Xue,* 15**,** 829-32.

WEN, X., DENG, F. M. & WANG, J. 2014. MicroRNAs as predictive biomarkers and therapeutic targets in prostate cancer. *Am J Clin Exp Urol,* 2**,** 219-30.

ZEEGERS, M. P., JELLEMA, A. & OSTRER, H. 2003. Empiric risk of prostate carcinoma for relatives of patients with prostate carcinoma: a meta-analysis. *Cancer,* 97**,** 1894-903.

ZHANG, B., PAN, X., COBB, G. P. & ANDERSON, T. A. 2007a. microRNAs as oncogenes and tumor suppressors. *Dev Biol,* 302**,** 1-12.

ZHANG, Y., LUOH, S. M., HON, L. S., BAERTSCH, R., WOOD, W. I. & ZHANG, Z. 2007b. GeneHub-GEPIS: digital expression profiling for normal and cancer tissues based on an integrated gene database. *Nucleic Acids Res,* 35**,** W152-8.

Chapter 2

Generation of a putative gene list for the early diagnosis of prostate cancer using a bioinformatics approach

2.1. Introduction

The National Center for Biotechnology Information defines Bioinformatics as:

Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. There are three important sub-disciplines within Bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains and protein structures and the development and implementation of tools that enable efficient access and management of different types of information (NCBI, 2001).

As a complex disease, cancer is related to a large number of genes and proteins. Biomedical researchers are interested in data-mining literature based on cancer-related genes and proteins to study cancer diagnostics, treatment, and prevention (Zhu *et al.,* 2012).

Existing scientific literature represents a rich source of knowledge based on the associations between genes, diseases and cellular processes. The biomedical literature is an important source of knowledge on the function of genes and on the mechanisms by which these genes regulate cellular processes (Frijters *et al*., 2010).

Owing to the increasing body of text and the open-access policies of many journals, literature mining is also becoming useful for both hypothesis generation and biological discovery. However, the biological discovery will require the integration of literature and high-

throughput data, which should encourage close collaborations between biologists and computational linguists (Jensen *et al*., 2006).

Cancer remains a major public health challenge despite progress in detection and therapy (Srinivas *et al.,* 2002). Among the important tools critical to detection, diagnosis, treatment, monitoring, and prognosis are biomarkers (Srinivas *et al*., 2001). Bioinformatics tools are essential in the discovery of sensitive and specific biomarkers in cancer research (Srinivas *et al*., 2002).

Unlike benign prostate hyperplasia (BPH) or prostatitis, prostate cancer may not give symptoms in its early, curable stage, and therefore it is often diagnosed in the advanced stages of the disease. An intervention at an early stage may reduce the progression of small localized carcinoma to a large metastatic lesion, thereby reducing prostate cancer-related deaths. Discovery of other biomarkers are needed, if improvements in diagnosis and prognosis of prostate cancer are to be realized (Adam *et al*., 2001). One of the best ways to diagnose cancer early, aid in its prognosis, or predict therapeutic response, is to use serum or tissue biomarkers (Kulasingam, 2008). Cancer biomarkers can be DNA, mRNA, proteins, metabolites, or processes such as apoptosis, angiogenesis or proliferation (Hayes *et al*., 1996). Powerful bioinformatics tools, has a direct and major impact on the way the search for cancer biomarkers is conducted. The modern technologies are capable of performing parallel rather than serial analyses and they can help to identify distinguishing patterns and multiple markers rather than just a single marker; such strategies represent a central component and a paradigm shift in the search for novel biomarkers (Kulasingam, 2008).

Prostatic carcinomas most often arise in the glandular epithelium of the prostate periphery. Although *PSA* (*KLK3*) gene transcription is down-regulated in prostate cancer, PSA protein levels in the circulation of patients with prostate cancer increase through disruption of the

anatomic barriers between the glandular lumen and capillaries. Concomitant to early-stage prostate cancer is the loss of basal cells, disruption of cell attachment, degradation of the basement membrane, initiation of lymphangiogenesis. A goal of finding more specific biomarkers to avoid over-diagnosis and overtreatment associated with PSA screening of prostate cancer (Stacker *et al*., 2002).

Ideally, a cancer biomarker should be detectable in the blood or other body fluids that can be accessed in a non-invasive manner and none has met the original goal of discovering cancer at an early stage. One reason for the low sensitivity and specificity is the presence of these markers in the serum of individuals without cancer or with non-malignant disease. Many potentially valuable biomarkers are expressed at very low levels and are difficult to detect. Finding new and better methods for detecting and identifying these low-abundant proteins represents a new challenge for routine diagnostics (Seibert *et al*., 2005).

With the Human Genome Project's completion at hand and with increasing amounts of expression data becoming available, growing attention is being paid to *in silico* biology. Broadly speaking, the term *in silico* biology refers to the use of computers to perform biological studies and a high-throughput experimental technology which is generating biological data at unprecedented rates and the pace will only accelerate. The bioinformatics infrastructure that tabulates, curates and makes these data retrievable is developing in parallel (e.g., WIT, EcoCyc, MIPS, KEGG, Biology WorkBench, EMP, and Swiss-Prot) (Palsson, 2000).

2.2. Computational approaches to predict a novel biomarker

2.2.1. Data mining

The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining is the analysis step of the "Knowledge Discovery in Databases" process, or KDD (Fayyad *et al*., 1996).

A particular active area of research in bioinformatics is the application and development of data mining and techniques to solve biological problems. Analysing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. Examples of this type of analysis include protein structure prediction, gene classification, cancer classification based on microarray data, clustering of gene expression data, statistical modelling of protein-protein interaction, etc. (Raza 2012).

2.2.2. Biomedical text mining

Biomedical text mining (also known as BioNLP) refers to text mining applied to texts and literature of the biomedical and molecular biology domain. It is a rather recent research field on the edge of natural language processing, bioinformatics, medical informatics and computational linguistics. The main developments in this area have been related to the identification of biological entities (named entity recognition), such as protein and gene names as well as chemical compounds and drugs. Information extraction and text mining methods have been explored to extract information related to biological processes and diseases (Krallinger *et al*., 2010).

There have been an enormous number of publications on cancer research. This integrated but unstructured biomedical text is of great value for cancer diagnostics, treatment, and prevention. The immense body and rapid growth of biomedical text on cancer has led to the appearance of a large number of text mining techniques aimed at extracting novel knowledge

from scientific text. Biomedical text mining on cancer research is computationally automatic and high-throughput in nature (Zhu *et al.,* 2012). Text mining can aid in uncovering information and knowledge from a mountain of text and it is now widely applied in biomedical research (Zhu *et al.,* 2012).

Text Mining is an information retrieval task aimed at discovering new, previously unknown information, by automatically extracting it from different text resources. Research and development in the analysis of bioinformatics literature aims to provide bioinformaticians with effective means to access and exploit the knowledge contained in scientific publications (Manconi *et al*., 2012).

2.2.2.1. PubMed

Hands-on literature mining currently means a keyword search in PubMed (https://www.ncbi.nlm.nih.gov/pubmed) or any literature database. In terms of information retrieval systems, PubMed is one of the best known biomedical databases and it contains more than 20 million citations on biomedical articles from MEDLINE and life science journals, which provides a convenient web-based search portal for users as well as an application program interface for developers (McEntyre *et al*., 2001).

PubMed is a database developed by NCBI National Library of Medicine (NLM), it works as a part of the NCBI Entrez retrieval system, primarily designed to provide access to references and abstracts from biomedical and life sciences journals. PubMed provides links that allow access to the full-text journal articles of participating publishers (Lindberg, 2000). PubMed remains an optimal tool in biomedical electronic research (Falagas, 2008).

2.2.2.2. Google scholar

Google Scholar (http://scholar.google.com) enables the user to search specifically for scholarly literature, including peer-reviewed papers, theses, books, preprints, abstracts and

technical reports from all broad areas of research. Google Scholar is used to find articles from a wide variety of academic publishers, professional societies and universities and available across the web. Google Scholar provides a new method of locating potentially relevant articles on a given subject by identifying subsequent articles that cite a previously published article (Noruzi, 2007).

2.2.3. Biological databases

Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology and computational analysis (Attwood *et al.,* 2012).

Biological databases represent an invaluable resource in support of biological research. Much can be learned about a particular molecule by searching various databases and using available analysis tools. A large number of databases are available for that task and some databases are general while some are much specialised. For best results one often need to access multiple databases. Biological database design, development and long-term management are a core area of the discipline of Bioinformatics (Bourne, 2005).

2.2.3.1. STRING

Interactions between proteins are very important to understand their functions and biological processes. Several approaches and tools have been defined to deal with this challenge (Manconi *et al*., 2012). The STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database (http://string-db.org) aims to provide a critical assessment and integration of protein–protein interactions, including direct (physical) as well as indirect (functional) associations. The new version 10.0 of STRING covers more than 2000 organisms, which has necessitated novel, scalable algorithms for transferring interaction information between organisms (Szklarczyk *et al.,* 2014). STRING characterises functional

links between proteins and the protein network, i.e. the summary of all known or predicted protein interactions in an organism.

STRING specializes in three ways: (i) it provides uniquely comprehensive coverage, with >2000 organisms, 5 million proteins and >200 million interactions stored; (ii) it is one of very few sites to hold experimental, predicted and inferred interactions, together with interactions obtained through text mining; and (iii) it includes a wealth of accessory information, such as protein domains and protein structures, improving its day-to-day value for users. Protein–protein associations have proven to be a useful concept, by which to group and organize all protein-coding genes in a genome (Franceschini, 2013). The confidence score is the approximate probability that a predicted link exists between two enzymes in the same metabolic map in the KEGG (Kyoto Encyclopedia of Genes and Genomes) database. Confidence limits are as follows: low confidence - 20% (or better), medium confidence - 50%, high confidence - 75%, highest confidence - 95% (Szklarczyk *et al*., 2010). STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms. The database currently covers 9'643'763 proteins from 2'031 organisms (Szklarczyk *et al*., 2015).

2.2.3.2. DAVID

DAVID **(**Database for Annotation, Visualization, and Integrated Discovery) (http://david.abcc.ncifcrf.gov/) is a web-accessible program that integrates functional genomic annotations with intuitive graphical summaries. Lists of gene or protein identifiers are rapidly annotated and summarized according to shared categorical data for Gene Ontology, protein domain, and biochemical pathway membership (Dennis, 2003). The DAVID database aims to provide functional interpretation of large lists of genes and to

analyse these gene lists derived from high-throughput genomic experiments (Huang *et al*., 2009).

2.2.3.3. Gene Ontology

The Gene Ontology tool (GO) (http://www.geneontology.org) consists of a collaborative effort to 'annotate' gene products (e.g. proteins) with terms that describe their functions and cellular location of action (Gene Ontology Consortium, 2012).

The discipline of ontology allows experimental data to be stored in such a way that it constitutes a formal, structured representation of the reality captured by the underlying biological science. Ontology of a given domain represents types and the relations between them, and is designed to support computational reasoning about the instances of these types (Hill, 2008). GO resources include biomedical ontologies that cover molecular domains of all life forms as well as extensive compilations of gene product annotations to these ontologies that provide largely species-neutral, comprehensive statements about what gene products do (Blake, 2013).

An enrichment analysis is performed directly from the home page of the GOC website. This service connects to the analysis tool from the PANTHER Classification System, which is maintained up to date with GO annotations. The PANTHER (protein annotation through evolutionary relationship) classification system (http://www.pantherdb.org/) is a comprehensive system that combines gene function, ontology, pathways and statistical analysis tools that enable biologists to analyse large-scale, genome-wide data from sequencing, proteomics or gene expression experiments (Mi *et al*., 2013).

2.2.3.4. TiGER

Most of currently available biological databases do not focus on tissue-specific gene regulation but a web database called TiGER (Tissue-specific Gene Expression and

Regulation) (http://bioinfo.wilmer.jhu.edu/) house data on several genes across various tissues. The database contains three types of data including tissue-specific gene expression profiles, combinatorial gene regulations, and cis-regulatory module (CRM) detections. The database provides three views (gene view, transcription factor view, and tissue view) to allow users to conveniently retrieve information about genes, TFs or tissues of interest. TiGER provides visualizations of the gene expression profiles, TF interactions and CRM detections (Liu *et al*., 2008).

2.2.3.5. GeneHub-GEPIS

GeneHub-GEPIS (http://research-public.gene.com/Research/genentech/genehub-gepis/index.html) is a web server useful tool for performing gene expression analysis across many normal and cancer tissues for both mouse and human genes (Zhang, 2007). One distinguishing characteristic of this tool is that ESTs are mapped to pre-defined gene structures along the genome.

GeneHub-GEPIS can report estimated expression levels in about 40 different types of normal and cancerous tissues for a given gene or a list of genes. This tool is built upon the previous GeneHub database, which integrates gene and protein information from several databases (Zhang, 2007).

2.3. Previous Study

Using a bioinformatics framework as well as a molecular approach, 21 genes regulated by 13 miRNAs were identified in a previous study (Khan, PhD thesis, 2015) as potential Prostate cancer miRNA targets.

The 13 miRNAs were analysed using Real-Time PCR (qPCR) using a panel of prostate cancer cell lines as well as a panel of various cancer cell lines to determine the specificity of these miRNAs for prostate cancer as potential diagnostic biomarkers. This work focused on

the miRNA regulated genes to create a candidate gene list for prostate cancer diagnosis as well as prognosis. Thus a combination of miRNAs and protein biomarkers can be used to specifically and sensitively diagnose prostate cancer at its early stages as well as monitor outcome to treatment (prognosis).

Table 2.1: 21 miRNA target genes that were identified by computational prediction methods as candidate biomarkers for detection of prostate cancer.

Target	Gene description	MiRNA related with target
TMX1	Theoredoxin-related transmembrane protein	MiR 10
TNFSF15	Tumor necrosis factor (ligand) superfamily, member 15	MiR 11
LIG4	Ligase IV, DNA, ATP-dependent	MiR 12
FOXC1	Forkhead box C1	MiR 13
MNT	MAX binding protein	MiR 13
YWHAZ	Tyrosine 3-monooxygenase/ tryptophan 5-monooxygenase activation protein, zeta polypeptide	MiR 9
TNFSF13B	Tumour necrosis factor	MiR 9
ADNP	Activity-dependent neuroprotector homeobox	MiR1
PAK7	p21 protein (Cdc42/Rac)-activated kinase 7	MiR1
RAB27A	RAB27A, member RAS oncogene family	MiR1
ACVR1C	Activin A receptor, type IC	MiR1
BTG2	BTG family member 2	MiR2
SH3RF1	SH3 domain containing Ring Finger	MiR3
BFAR	Bifunctional Apoptosis regulator	miR3
PRKCI	Protein kinase C, iota	MiR4
PTPRC	Protein tyrosine phosphatase, receptor type, C	MiR4
ING4	Inhibitor of growth family, member 4	MiR5
ATM	Ataxia telangiectasia mutated	MiR6
CLN8	Ceroid-lipofuscinosis, neuronal 8	MiR7
CFLAR	CASP8 and FADD-like apoptosis regulator	MiR8
CSRNP3	Cysteine-serine-rich nuclear protein 3	MiR8

2.4. Aims

1. Expand on the current list of 21 genes.
2. Extract and refine a list of genes as possible potential biomarkers for early diagnosis of prostate cancer, using the various databases outlined above.
3. Confirm the candidate genes as novel to prostate cancer and generate a manageable candidate gene list.

2.5. Materials and methods

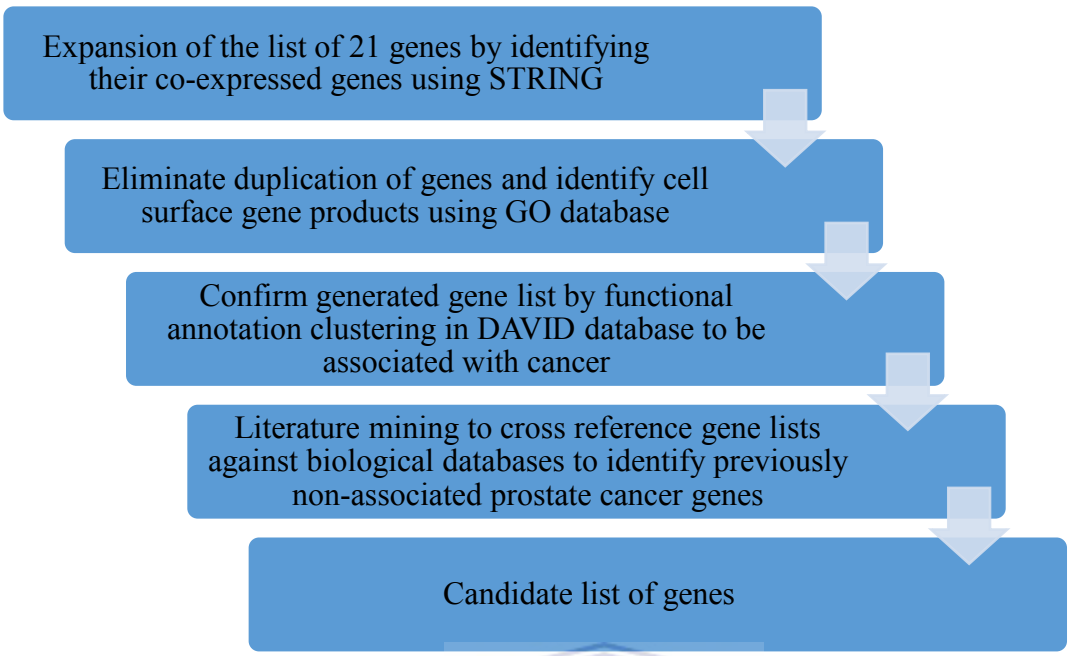


Figure 2.1: Outline of Methodology for generation of a candidate gene list

2.5.1. Expansion of candidate gene list

2.5.1.1. Identification of co-expressed genes using the STRING database

The STRING database was accessed at (http://string-db.org) and the 21 miRNA target genes were individually used in STRING to generate a list for each gene of co-expressed known and predicted protein-protein interactions. Input query of gene/protein name included organism of *Homo sapiens* as the species of interest. Predicted functional partners were derived from four sources: Genomic context, high-throughput experiments, co-expression and previous knowledge (PubMed etc.). Active Prediction Methods included neighborhood, gene fusion, co-occurrence, co-expression, experiments, databases and text mining as parameters.

The default settings and additional criteria used were as follows: High confidence score (0.700) of 20 protein-protein interactions and a medium confidence score (0.400) with 10

interactions. These settings were used as a stringency measure to filter the number of interactions that will be produced at a lower confidence score. Interactions found for each gene equals the amount of genes/proteins identified.

Further refinement of the identified co-expressed genes by elimination of duplication and overlapping of genes was performed; this gene list was used further to generate a manageable candidate gene list using additional bioinformatics tools.

2.5.1.2. Gene Ontology based enrichment analysis

Enrichment analysis was done on the gene set generated in the previous section, by inserting all gene names one per row and selecting the species; *Homo sapiens*. The ontology, cellular component was used to calculate the enrichment for genes corrospondng to a particlar term. Criteria: GO terms relating to the cellular component (CC) parts of a cell or its extracellular environment (e.g. cytoplasm, integral to plasma membrane) were used to further refine the list of genes. Ontology criteria included cell surface and displaying only the results with $P<0.05$.

Analysis Summary of GO

Analysis Type: PANTHER Overrepresentation Text

Annotation Version and Release Date: GO Ontology database

Analyzed List: upload_1 (Homo sapiens)

Reference List: Homo sapiens (all genes in database)

Annotation Data Set: GO cellular component complete

The Bonferroni correction for multiple testing was used. The Bonferroni correction is important because many statistical tests are performed (one for each ontology term) at the

same time. This correction multiplies the single-test P-value by the number of independent tests to obtain an expected error rate (Thomas *et al*., 2016).

2.5.1.3. Functional Annotation clustering using DAVID

DAVID database was accessed as a confirmation of results obtained from GO analysis. All genes were submitted to DAVID for functional annotation clustering by inserting the gene list and then choosing "OFFICIAL_GENE_SYMBOL" as the unique gene identifier followed by "gene list" as list type. '*Homo sapiens*' was selected as species to limit annotations.

The uploaded gene list was analysed using, "Functional Annotation Clustering", selected from DAVID's functional annotation tools. Class stringency was set to medium, and options were defaulted to display Benjamini analysis. Selections of clusters were queried using the following terms: cell surface; regulation of cell death; regulation of cell proliferation; apoptosis and cell cycle process as onset progression of prostate cancer. Enrichment scores of 1.3 or more was selected. A list of genes was created and further refined by removal of duplication resulting in the same genes generated from GO thus confirming results.

2.5.1.4. Gene expression profiling using GeneHub-GEPIS

The number of genes resulting following GO and DAVID analysis were then inserted into GeneHub-GEPIS database to eliminate genes already experimentally linked to prostate cancer through their expression within prostate cancer as annotated by this database, thus ensuring that genes selected as potential biomarkers would be novel. This database was specifically used due to the gene expression profiling of a large panel of normal and cancer tissues based on human EST sequence abundance. Either expressed or non-expressed genes in prostate cancer would be included within the new gene list generated. The following criteria were used: target species: Human; Search by Accession/Gene Symbol: e.g. ITGB1.

Bar graphs (figures 2.3 and 2.4) showing tumour (yellow bar) versus normal (blue bar) expression of various tissues for each gene was generated.

2.5.1.5. Gene expression analysis using TiGER

The same genes inserted in GeneHub-GEPIS were used in the TIGER database. Bar graphs (figures 2.5 and 2.6) showing results (red bar) of Expressed Sequence Tag (EST) Profile enrichment scores, with high enrichment being overly expressed in that specific tissue. The gene view allowed information to be retrieved through a simple search engine by entering the gene symbol.

ESTs may be used to identify gene transcripts, and are instrumental in gene discovery and in gene-sequence determination (Adams *et al*., 1991). A gene is considered as tissue-specific gene if it satisfies the two criteria: the enrichment score is greater than 5 and the P-value is smaller than $10^{-3.5}$.

2.5.2. Literature mining of the Candidate genes

The databases used for literature mining were: PubMed and Google Scholar. Genes were searched for by entering search words in the query box and the "search" button pressed. The gene name with the Boolean term "AND" and the cancer of interest e.g. <prostate cancer> AND <gene name> was entered. All relevant literature, abstracts and journal articles were searched for information linking the genes as biomarkers for prostate cancer. All genes within the list generated that had been previously studied as candidate cancer or benign disease serum biomarkers in prostate cancer were eliminated and a final candidate gene list was generated.

2.6. Results and Discussion

2.6.1. Identifying possible candidate biomarkers

For the purpose of this study 21 genes known to be regulated by several miRNAs as identified in a previous study '*Identification of miRNA's as specific biomarkers in prostate cancer diagnostics: a combined in silico and molecular approach'* (Khan, PhD thesis, 2015) were the main focus. As data mining can only uncover patterns present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined. 13 miRNAs were identified as potential prostate cancer miRNA targets and a gene/protein list of 21 miRNA targeted genes were generated and linked to prostate cancer (Khan, 2015).

The STRING database was used for prediction of protein-protein networks. The 21 genes were used individually as driver genes and a total of 300 additional genes/proteins were identified and extracted from the STRING database.. Through expansion of the initial 21 miRNA targeted genes, intermediary proteins functioning in concert with these genes are identified through protein-protein interactions. Identification and characterization of protein–protein interactions (PPIs) is one of the key aims in biological research. More accurate PPI detection will also improve the ability to extract experimental data related to PPIs and provide evidence for each interaction (Niu *et al*., 2010).

An example of an interaction network produced for TMX1 using STRING is shown in figure 2.2.

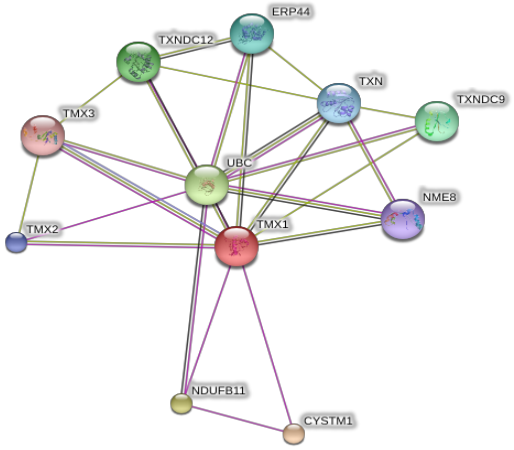


Figure 2.2: TMX1 showing interactions of its 10 predicted functional partners using STRING. Different line colours between the proteins indicate the various types of interaction evidence for generating a specific interaction (Szklarczyk *et al*., 2015).

Table 2.2: Number of genes extracted from STRING database for each miRNA targeted gene:

miRNA	**Target**	**Confidence Score**	**Identified genes/interactions**
MiR1	ADNP	0.400	10
MiR1	PAK7	0.700	20
MiR1	RAB27A	0.700	20
MiR1	ACVR1C	0.700	20
MiR2	BTG2	0.400	10
MiR3	SH3RF1	0.400	10
MiR3	BFAR	0.400	10
MiR4	PRKCI	0.700	20
MiR4	PTPRC	0.700	20

MiR5	ING4	0.400	10
MiR6	ATM	0.700	20
MiR7	CLN8	0.400	10
MiR8	CFLAR	0.700	20
MiR8	CSRNP3	0.400	10
MiR 9	YWHAZ	0.700	20
MiR 9	TNFSF13B	0.700	20
MiR 10	TMX1	0.400	10
MiR 11	TNFSF15	0.400	10
MiR 12	LIG4	0.400	10
MiR 13	FOXC1	0.400	10
MiR 13	MNT	0.400	10

The gene list was refined and duplicates were eliminated using a Perl script, resulting in a total of 231 genes from the STRING database.

GO analysis

The identification of cell surface markers is critical to the development of new diagnostic and therapeutic modalities for the management of prostate cancer. To identify potential markers for the diagnosis and treatment of prostate cancer, membrane-bound tumour antigens were searched for in that they may provide insights into the biology of prostate cancer progression (Reiter *et al*., 1998). Since these markers can be shed into the surrounding fluid and thus easily detected in various bodily fluids.

The database GO was searched using the 231 genes identified and included the cellular component (CC) as part of the analysis GO terms generated a number of mapped

identifications of 26 genes in relation to their expression on the cell surface with a P –value of <0.05.

The smaller the p-value, the larger the significance since it indicates that the hypothesis under consideration may not adequately explain the observation. A small p-value (≤ 0.05) indicates strong evidence as the p-value is widely used in statistical hypothesis testing (Nuzzo, 2014).

Cellular component included terms such as 'ribosome' or 'proteasome', specifying where multiple gene products would be found. It also includes terms such as 'nuclear membrane' or 'Golgi apparatus' (Ashburner *et al*., 2000). The GO concept is intended to make possible, in a flexible and dynamic way, the annotation of homologous gene and protein sequences in multiple organisms using a common vocabulary that results in the ability to query and retrieve genes and proteins based on their shared biology (Ashburner *et al*., 2000).

Results obtained from GO is displayed in table 2.3 below. Column 1 indicates the annotation data category. Column 2 indicates the number of reference genes linked to the annotation i.e. all genes within the database linked to this term. Column 3 shows the number of uploaded genes mapped to the annotation i.e. the number of genes within the list of 231 genes corresponding to this GO term. Column 4 shows the expected value, which is the number of genes expected in the list for this category, based on the reference list. Column 5 shows the Fold Enrichment of the genes observed in the uploaded list over the expected (number in the list divided by the expected number). The number is greater than 1, indicating that the category is overrepresented for this subset of genes in this study. Column 6 has either a + or - sign. The plus sign indicates over-representation of GO category. Column 7 has the p-value. A small p-value indicates that the number observed is significant and potentially interesting. A cut-off of 0.05 is recommended as a starting point (Thomas *et al*., 2016). Results obtained from GO and DAVID is shown in table 2.4 and displays the comparison thereof.

Table 2.3: Genes related to the GO term cellular component extracted from GO database

GO cellular component complete	Reference List	Input List	expected	Fold Enrichment	+/-	P value
external side of plasma membrane - cell surface	779	26	8.46	3.07	+	6.27E-04

Functional annotation using DAVID

Two comparisons were performed, DAVID was mined for the exact same genes extracted from GO to confirm the GO term cellular component would result in the same output using DAVID.

Genes extracted from DAVID database resulted in 20 clusters generated that was queried using the GO terms:

Cluster 1 cell surface - enrichment score 6.25; cluster 4 regulation of cell death - enrichment score 4.19; cluster 6 regulation of cell proliferation - enrichment score 3.2; cluster 11 apoptosis - enrichment score 2.11 and cluster 18 cell cycle process - enrichment score 0.86. Combined clusters list and removal of duplication resulted in 26 genes. The confirmed 26 genes from GO and DAVID analysis were then used to be further analysed for their expression in prostate cancer.

Table 2.4: Genes extracted from GO and DAVID

Cell Surface - 26 genes		
Gene Ontology	**DAVID**	**Name of gene**
CD22	CD22	hypothetical protein FLJ22814
CD3E	CD3E	CD3e molecule, epsilon (CD3-TCR complex)
CD4	CD4	CD4 molecule
EGFR	EGFR	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)
ERP44	ERP44	**Gene Name:** TXNDC4 thioredoxin domain containing 4 (endoplasmic reticulum)
FAS	FAS	Fas (TNF receptor superfamily, member 6)
FASLG	FASLG	Fas ligand (TNF superfamily, member 6)
GP1BA	GP1BA	glycoprotein Ib (platelet), alpha polypeptide
IFNG	IFNG	interferon, gamma
IL13	IL13	interleukin 13
IL6	IL6	interleukin 6 (interferon, beta 2)
ITGA4	ITGA4	integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)
ITGAM	ITGAM	integrin, alpha M (complement component 3 receptor 3 subunit)
ITGB1	ITGB1	integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)
LAMP1	LAMP1	lysosomal-associated membrane protein 1
LGALS1	LGALS1	lectin, galactoside-binding, soluble, 1 (galectin 1)
MMP7	MMP7	matrix metallopeptidase 7 (matrilysin, uterine
SEPT2	SEPT2	septin 2
SERPINE2	SERPINE2	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 2,
SHH	SHH	sonic hedgehog homolog (Drosophila
TDGF1	TDGF1	teratocarcinoma-derived growth factor 1
TNFRSF10A	TNFRSF10A	tumor necrosis factor receptor superfamily, member 10a
TNFRSF10B	TNFRSF10B	tumor necrosis factor receptor superfamily, member 10b
TNFRSF13C	TNFRSF13C	tumor necrosis factor receptor superfamily, member 13C
TNFRSF1A	TNFRSF1A	tumor necrosis factor receptor superfamily, member 1A
TNFSF4	TNFSF4	tumor necrosis factor (ligand) superfamily, member 4 (tax-transcriptionally activated glycoprotein 1, 34kDa)

2.6.2. Expression analysis

TiGER and GeneHub-GEPIS are databases containing tissue specific enriched genes and expression analysis and these were used to eliminate genes already experimentally linked to prostate cancer. All 26 genes were then mined in GeneHub-GEPIS, TIGER, PubMed and Google Scholar.

Summary of expression analysis:

GeneHub	TIGER
8 non-expressed genes	5 non-expressed genes
3 genes no EST hits	3 genes not found in databases

In GeneHub-GEPIS, 8 genes showed no expression distribution in prostate tissue and 3 genes had no EST hits in the database. Outputs of 5 genes were non-expressed in the prostate in the TIGER database and results for 3 genes were not found in this database. Of the non-expressed genes in these two databases, only 3 were common to both and only 1 gene, were not found in either database.

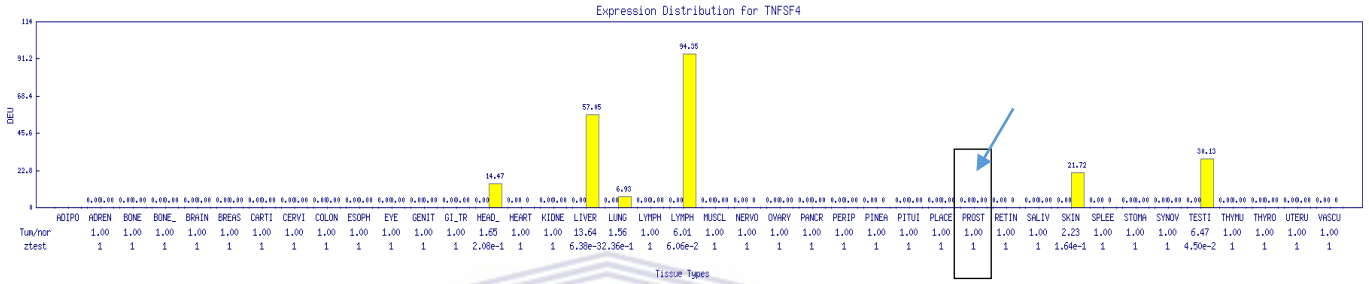


Figure 2.3: Gene TNFSF4 showing no expression in both normal and tumour tissue as seen in the GeneHub-GEPIS database.

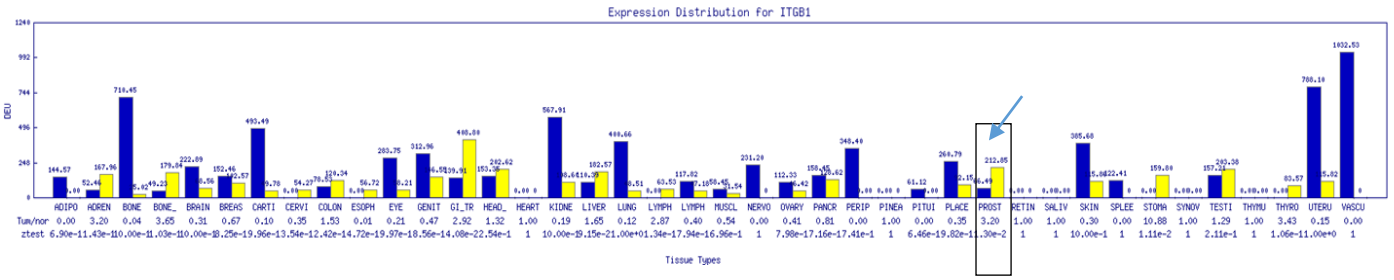


Figure 2.4: Gene ITGB1 showing expression in both normal and tumour tissue, but more expression in tumour of the prostate in GeneHub-GEPIS database.

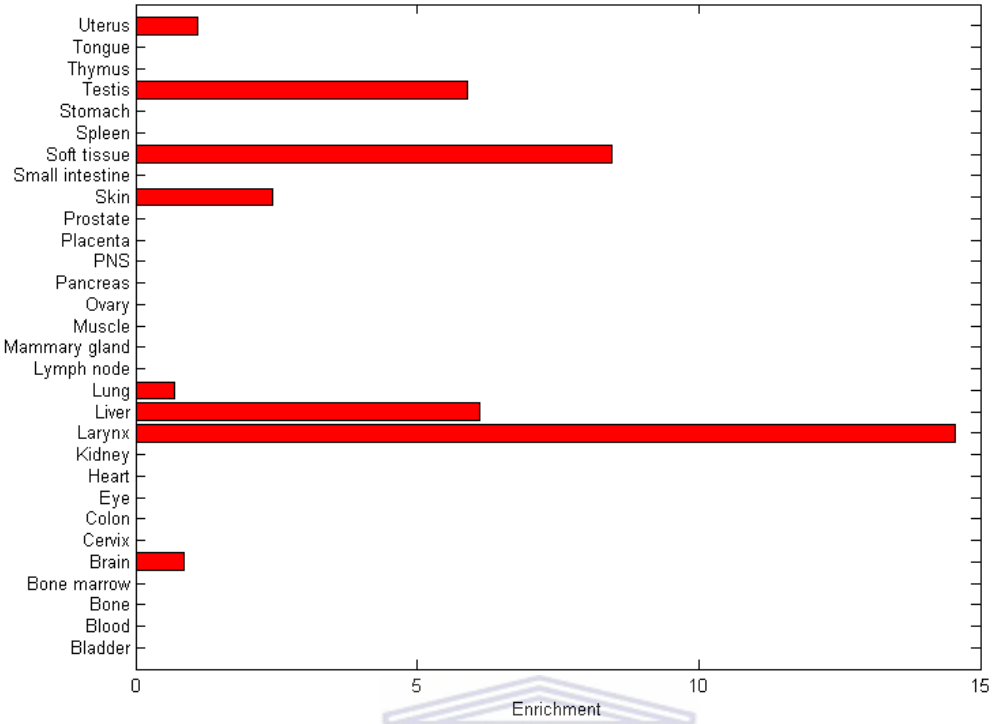


Figure 2.5: Gene TNFSF4 displaying no expression in prostate tissue in the TIGER database.

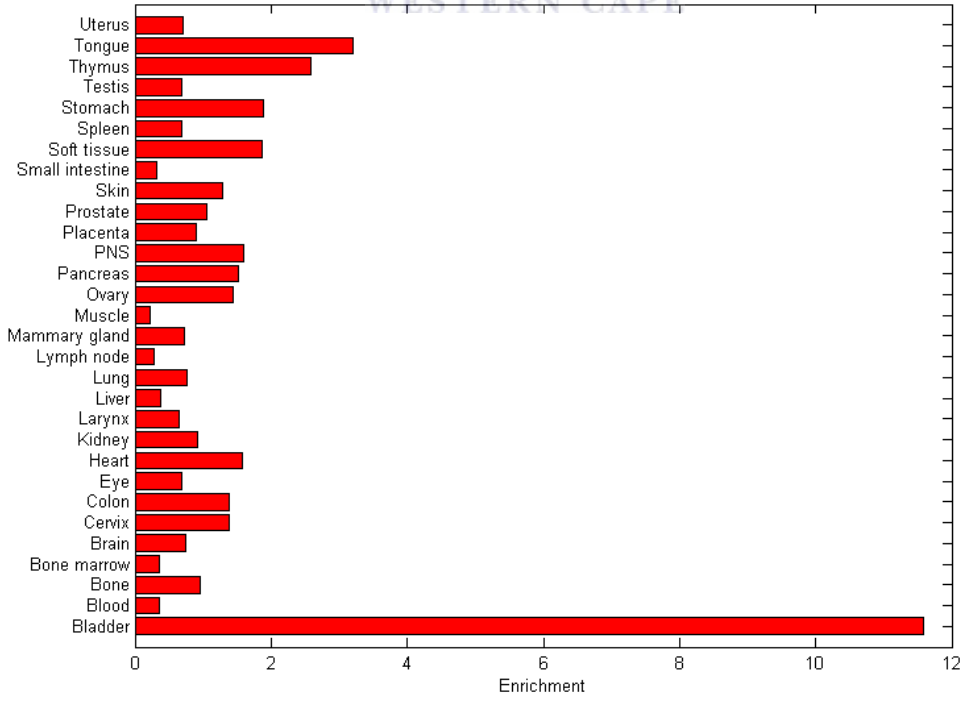


Figure 2.6: Gene LAMP1 displaying expression in various tissues in TIGER database.

Literature mining was then used, in order to obtain genes of greater relevance to be validated as novel biomarkers for prostate cancer.

2.6.3. Literature mining of the Candidate genes

Of the 26 genes, only 6 showed no related articles to prostate cancer in PubMed and in Google Scholar of that 6, only 5 were not linked to prostate cancer. The resulting 6 genes from PubMed were chosen. Since Google Scholar offers results of inconsistent accuracy (Falagas, 2008).

The final candidate gene list was narrowed down to the following 6 genes: ERP44 (Gene Name: TXNDC4 thioredoxin domain containing 4 (endoplasmic reticulum), GP1BA (glycoprotein Ib (platelet), alpha polypeptide), IFNG (interferon, gamma), SEPT2 (septin 2), TNFRSF13C (tumour necrosis factor receptor superfamily, member 13C) and TNFSF4 (tumour necrosis factor (ligand) superfamily, member 4 (tax-transcriptionally activated glycoprotein 1, 34 kDa).

2.7. Discussion and Conclusion

In this chapter the focus was on the diagnostic application of biomarkers using a bioinformatics approach. Using bioinformatics, the aim was to identify biomarkers for diagnosis of prostate cancer with a high sensitivity, specificity and accuracy. The biomarker should preferably be tissue specific, such that a change in serum level can be directly attributed to disease (for example, cancer) of that tissue (Diamandis, 2010). Currently the most widely used serological biomarker include prostate-specific antigen (PSA, also known as kallikrein-related peptidase (KLK) 3) in prostate cancer (Bostwick, 1994). This biomarker have been proven to be ineffective for the accurate diagnosis of prostate cancer hence the need to identify additional biomarkers for this disease which was the aim of this study

STRING database was used to look at known and predicted protein-protein associations (direct or indirect) for prostate cancer of the 21-miRNA target genes previously identified. The protein-protein interactions identified were based on experimental data, databases and literature and from predictions of genomic content analysis. The basic interaction unit in STRING is the *functional association*, i.e. a specific and productive functional relationship between two proteins, likely contributing to a common biological purpose (Szklarczyk *et al*., 2015). From STRING analysis 231 genes were generated which was further refined using GO, DAVID and expression analysis through TIGER and GeneHub-GEPIS. Gene Ontology (GO) analysis showed 26 out of the 231 genes being expressed on the cell surface with functional analysis using DAVID, showing the same number of genes to be linked to processes such as apoptosis, cell cycle and cancer onset and progression.

A PubMed search revealed that 20 of the 26 genes identified had been previously studied in relation to prostate cancer. As seen in figure 2.3 and figure 2.5 the gene TNFSF4 were not expressed in prostate tissue in either TiGER or GeneHub-GEPIS database and is thus one of the six candidate genes identified. The coverage of the genes in TiGER and GeneHub-GEPIS databases with no association to prostate cancer was linked to and confirmed in literature mining to prostate cancer.

These computational approaches led to six identified genes; ERP44, GP1BA, IFNG, SEPT2, TNFRSF13C and TNFSF4, as possible diagnostic biomarkers for prostate cancer. These results are promising, since the targeted biomarkers would be easily detectable in bodily fluids as the GO analysis of these genes are enriched for cell surface. The identification of biomarkers for clinical applications remains an important issue for improving diagnostics, prognostics and therapy in many diseases, including prostate cancer. Investigation of these candidates should be prioritized for further verification and validation studies. The genes identified using an *in silico* approach within this chapter will be associated to transcription

factors and disease causing pathways in Chapter 3 to identify the underlying regulation of these genes and prognostic and predictive validation of these genes in prostate cancer (Chapter 4).

2.8. References

ADAM, B. L., VLAHOU, A., SEMMES, O. J. & WRIGHT, G. L. 2001. Proteomic approaches to biomarker discovery in prostate and bladder cancers. *Proteomics,* 1**,** 1264-1270.

ADAMS, M. D., KELLEY, J. M., GOCAYNE, J. D., DUBNICK, M., POLYMEROPOULOS, M. H., XIAO, H., MERRIL, C. R., WU, A., OLDE, B., MORENO, R. F., KERLAVAGE, A. R., MCCOMBIE, W. R. & VENTER, J. C. 1991. Complementary-DNA Sequencing - Expressed Sequence Tags and Human Genome Project. *Science,* 252**,** 1651-1656.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

ATTWOOD, T. K., GISEL, A., ERIKSSON, N-E. AND BONGCAM-RUDLOFF, E. 2011. Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective. *Bioinformatics – Trends and Methodologies*

BLAKE, J. A. 2013. Ten Quick Tips for Using the Gene Ontology. *Plos Computational Biology,* 9.

BOSTWICK, D. G. 1994. Prostate-specific antigen. Current role in diagnostic pathology of prostate cancer. *Am J Clin Pathol,* 102**,** S31-7.

BOURNE, P. 2005. Will a biological database be different from a biological journal? *PLoS Comput Biol,* 1**,** 179-81.

DATABASE DATABASE FOR ANNOTATION, VISUALIZATION AND INTEGRATIVE DISCOVERY (DAVID). Available: http://david.abcc.ncifcrf.gov/. [2014, August,12].

DENNIS, G., JR., SHERMAN, B. T., HOSACK, D. A., YANG, J., GAO, W., LANE, H. C. & LEMPICKI, R. A. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol,* 4**,** P3.

DIAMANDIS, E. P. 2010. Cancer biomarkers: can we turn recent failures into success? *J Natl Cancer Inst,* 102**,** 1462-7.

FALAGAS, M. E., PITSOUNI, E. I., MALIETZIS, G. A. & PAPPAS, G. 2008. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB J,* 22**,** 338-42.

FAYYAD, U., PIATETSKY-SHAPIRO, G. & SMYTH, P. 1996. From data mining to knowledge discovery in databases. *AI Magazine*. 37-54.

FRANCESCHINI, A., SZKLARCZYK, D., FRANKILD, S., KUHN, M., SIMONOVIC, M., ROTH, A., LIN, J., MINGUEZ, P., BORK, P., VON MERING, C. & JENSEN, L. J. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res,* 41**,** D808-15.

FRIJTERS, R., VAN VUGT, M., SMEETS, R., VAN SCHAIK, R., DE VLIEG, J. & ALKEMA, W. 2010. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput Biol,* 6.

GENEHUB-GEPIS. Available: http://share.gene.com/Research/genentech/genehub-gepis/index.html.

GENE ONTOLOGY, C. 2012. The Gene Ontology: enhancements for 2011. *Nucleic Acids Res,* 40**,** D559-64.

GOOGLE SCHOLAR. Available: http://scholar.google.com.

HAYES, D. F., BAST, R. C., DESCH, C. E., FRITSCHE, H., KEMENY, N. E., JESSUP, J. M., LOCKER, G. Y., MACDONALD, J. S., MENNEL, R. G., NORTON, L., RAVDIN, P., TAUBE, S. & WINN, R. J. 1996. Tumor marker utility grading system: A framework to evaluate clinical utility of tumor markers. *Journal of the National Cancer Institute,* 88**,** 1456-1466.

HILL, D. P., SMITH, B., MCANDREWS-HILL, M. S. & BLAKE, J. A. 2008. Gene Ontology annotations: what they mean and where they come from. *BMC Bioinformatics,* 9 Suppl 5**,** S2.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc,* 4**,** 44-57.

JENSEN, L. J., SARIC, J. & BORK, P. 2006. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet,* 7**,** 119-29.

KHAN, F. 2015. *Identification of miRNAs as specific biomarkers in prostate cancer diagnostics: a combined in silico and molecular approach.* PhD, University of the Western Cape

KRALLINGER, M., LEITNER, F. & VALENCIA, A. 2010. Analysis of biological processes and diseases using text mining approaches. *Methods Mol Biol,* 593**,** 341-82.

KULASINGAM, V. & DIAMANDIS, E. P. 2008. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nature Clinical Practice Oncology,* 5**,** 588-599.

LINDBERG, D. A. 2000. Internet access to the National Library of Medicine. *Eff Clin Pract,* 3**,** 256-60.

LIU, X., YU, X., ZACK, D. J., ZHU, H. & QIAN, J. 2008. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics,* 9**,** 271.

MANCONI, A., VARGIU, E., ARMANO, G. & MILANESI, L. 2012. Literature retrieval and mining in bioinformatics: state of the art and challenges. *Adv Bioinformatics,* 2012**,** 573846.

MCENTYRE, J. & LIPMAN, D. 2001. PubMed: bridging the information gap. *CMAJ,* 164**,** 1317-9.

MI, H., MURUGANUJAN, A., CASAGRANDE, J. T. & THOMAS, P. D. 2013. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc,* 8**,** 1551-66.

MI, H., POUDEL, S., MURUGANUJAN, A., CASAGRANDE, J. T. & THOMAS, P. D. 2016. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res,* 44**,** D336-42.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (NCBI). Available: https://www.ncbi.nlm.nih.gov/ [2001].

NIU, Y., OTASEK, D. & JURISICA, I. 2010. Evaluation of linguistic features useful in extraction of interactions from PubMed; application to annotating known, high-throughput and predicted interactions in I2D. *Bioinformatics,* 26**,** 111-9.

NORUZI, A. 2007. Google Scholar: The New Generation of Citations Indexes. *Libri,* 55, 170-180.

NUZZO, R. 2014. Statistical Errors. *Nature,* 506**,** 150-152.

OMNIBUS, G. E. National center for biotechnology information.

PALSSON, B. 2000. The challenges of in silico biology. *Nat Biotechnol,* 18**,** 1147-50.

PRASSAS, I., CHRYSTOJA, C. C., MAKAWITA, S. & DIAMANDIS, E. P. 2012. Bioinformatic identification of proteins with tissue-specific expression for biomarker discovery. *BMC Med,* 10**,** 39.

PUBMED. Available: https://www.ncbi.nlm.nih.gov/pubmed/.

RAZA, K. 2012. Application of data mining in bioinformatics. *Indian Journal of Computer Science and Engineering*. 1(2):114-118.

REITER, R. E., GU, Z., WATABE, T., THOMAS, G., SZIGETI, K., DAVIS, E., WAHL, M., NISITANI, S., YAMASHIRO, J., LE BEAU, M. M., LODA, M. & WITTE, O. N. 1998. Prostate stem cell antigen: a cell surface marker overexpressed in prostate cancer. *Proc Natl Acad Sci U S A,* 95**,** 1735-40.

SEARCH TOOL FOR THE RETRIEVAL OF INTERACTING GENES/PROTEINS (STRING) database. Available: http://string-db.org.

SEIBERT, V., EBERT, M. P. & BUSCHMANN, T. 2005. Advances in clinical cancer proteomics: SELDI-ToF-mass spectrometry and biomarker discovery. *Brief Funct Genomic Proteomic,* 4**,** 16-26.

SRINIVAS, P. R., KRAMER, B. S. & SRIVASTAVA, S. 2001. Trends in biomarker research for cancer detection. *Lancet Oncology,* 2**,** 698-704.

SRINIVAS, P. R., VERMA, M., ZHAO, Y. M. & SRIVASTAVA, S. 2002. Proteomics for cancer biomarker discovery. *Clinical Chemistry,* 48**,** 1160-1169.

STACKER, S. A., ACHEN, M. G., JUSSILA, L., BALDWIN, M. E. & ALITALO, K. 2002. Lymphangiogenesis and cancer metastasis. *Nature Reviews Cancer,* 2**,** 573-583.

SZKLARCZYK, D., FRANCESCHINI, A., KUHN, M., SIMONOVIC, M., ROTH, A., MINGUEZ, P., DOERKS, T., STARK, M., MULLER, J., BORK, P., JENSEN, L. J. & VON MERING, C. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res,* 39**,** D561-8.

SZKLARCZYK, D., FRANCESCHINI, A., WYDER, S., FORSLUND, K., HELLER, D., HUERTA-CEPAS, J., SIMONOVIC, M., ROTH, A., SANTOS, A., TSAFOU, K. P., KUHN, M., BORK, P., JENSEN, L. J. & VON MERING, C. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res,* 43**,** D447-52.

TISSUE -SPECIFIC GENE EXPRESSION AND REGULATION (TiGER). Available: http://bioinfo.wilmer.jhu.edu/. [2015, October, 18].

THE GENE ONTOLOGY (GO). Available: http://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz/. [2015, November, 20].

ZHANG, Y., LUOH, S. M., HON, L. S., BAERTSCH, R., WOOD, W. I. & ZHANG, Z. 2007. GeneHub-GEPIS: digital expression profiling for normal and cancer tissues based on an integrated gene database. *Nucleic Acids Res,* 35**,** W152-8.

ZHU, F., PATUMCHAROENPOL, P., ZHANG, C., YANG, Y., CHAN, J., MEECHAI, A., VONGSANGNAK, W. & SHEN, B. 2013. Biomedical text mining and its applications in cancer research. *J Biomed Inform,* 46**,** 200-11.

Chapter 3

Identification of Transcription factors and disease associated pathways of the identified biomarkers

3.1. Introduction

Promoters assist in controlling gene expression, since it is at this site that the RNA polymerase binds for initiation of transcription. Multiple functional sites are involved in the binding of the polymerase and elements such as the TATA box, GC box, and CAAT box serve as binding sites for transcription factors (TFs). The proteins that mediate transcriptional regulation are called TFs. TFs play a central role in gene regulation but are not solely responsible. TFs influence the transcription of specific genes, essentially determining whether a particular gene will be turned "on" or "off" in an organism. Each TF can regulate multiple genes (Phillips and Hoopes, 2008).

Transcriptional activation of a gene requires the binding of specific transcription factors to regulatory DNA elements and the assembly of the preinitiation complex with RNA polymerase and RNA synthesis initiation. Many human oncogenes encode for transcription factors and some of the most prominent tumour suppressors (e.g. p53) are TFs. TF families further increase the level of genetic complexity in eukaryotes and many TFs within the same family often work together to affect transcription of a single gene (Delgado and León, 2006).

The identification of TF targets in prostate cancer provides potential new targets for therapeutic intervention of prostate cancer such as treatment strategies, disease progression and metastases. In this chapter TFs were identified and characterized with respect to their expression in the prostate. Overexpression of TFs in prostate cancer implicates the development of carcinomas and the molecular diagnosis and treatment of prostate cancer. In

order to facilitate the discovery of potential biomarkers, candidate genes, were investigated using TF annotation to identify linkage to prostate cancer.

3.2. GeneCards

GeneCards (www.genecards.org) is a readily available Web resource database that integrates data for 152 704 human genes from 125 sources (Fishilevich *et al.,* 2016). GeneCards is accessible for searchable human gene annotations. Data are automatically mined from ~120 sources and presented as an integrated web card for every human gene (Fishilevich *et al.,* 2016).

It contains comprehensive information about human genes, including data about the cellular functions of their products, their involvement in diseases and genomic, transcriptomic, proteomic, genetic, clinical and functional information. The presented information aims at giving immediate insight into current knowledge about the respective gene, including a focus on its functions in health and disease (Rebhan *et al.,* 1998).

3.3. KEGG Pathway database

The KEGG (Kyoto Encyclopedia of Genes and Genomes) (http://www.genome.jp/kegg/) database provides a systematic analysis of gene functions, linking genomic information with higher order functional information in terms of the network of interacting molecules (Kanehisa and Goto, 2000).

KEGG is an integrated database resource consisting of seventeen main databases and KEGG pathway database (http://www.genome.jp/kegg/pathway.html) is one of the main databases and is part of the research projects of the Kanehisa Laboratories used for pathway-based information (Kanehisa *et al.,* 2010). KEGG PATHWAY contains scientific literatures on the biological molecular interaction and reaction networks, including protein-protein interaction,

protein-DNA binding, protein-ligand interaction and enzyme-mediated biomass reactions. Interactions within one specific biological process or function are drawn manually to pathway maps. Currently, there are 365 pathway maps collected from 113,760 references (Kanehisa *et al*., 2010).

3.4. Aims

1. Extract transcription factors associated to the six candidate genes
2. Functionally annotate the identified TFs and their relation to prostate cancer using GeneCards
3. Identify the TFs that are linked to pathways associated in prostate cancer using the KEGG database

3.5. Materials and methods

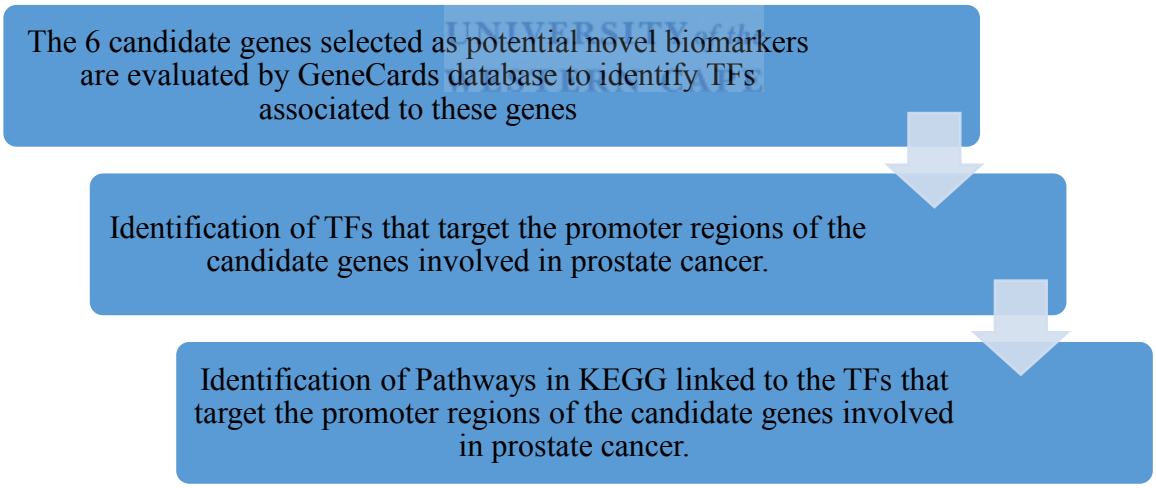


Figure 3.1: Flow chart of the methodology used for identification of transcription factors of the candidate genes

3.5.1. Identification of TFs using GeneCards

The GeneCards database was launched at www.genecards.org and used to extract the TFs that are associated to the six genes of interest. These six identified genes; ERP44, GP1BA, IFNG, SEPT2, TNFRSF13C and TNFSF4 were entered individually in the 'Explore a gene' box and selected for further exploration. The GeneCard showed all the data concerning the genes used as input. GeneCards sections and sub-headings include: Aliases, Disorders, Domains, Drugs, Expression, Function, Genomics, Localization, Orthologs, Paralogs, Pathways, Products, Proteins, Publications, Sources, Summaries, Transcripts and Variants of the gene of interest.

The Genomics section displays the chromosome, cytogenetic band and map location of the GeneCards gene as extracted from GeneLoc, HGNC, Entrez Gene, Nature and miRBase, as well as genomic views from UCSC and Ensembl, RefSeq DNA sequence links and TF binding sites from Qiagen. The section, under the sub-heading 'Regulatory Elements for Gene' is where the TFs of the candidate genes are stored. The TF binding sites by Qiagen in the gene promoter displayed all related regulatory elements data, thus providing insight into the transcriptional regulation of the gene and cell proliferation, differentiation and tumourigenesis.

3.5.2. Pathway discovery of the TFs associated to prostate cancer linked to the candidate genes using KEGG Pathway analysis

Pathways in KEGG linked to the TFs that target the promoter regions of the candidate genes involved in prostate cancer were investigated and related to known prostate cancer pathways. Within the KEGG database, the KEGG Pathway database presented with pathway maps to molecular interaction and reaction networks of the prostate cancer pathway (See figure 3.2). The pathway map for prostate cancer was found under the sub-heading 'Cancers: Specific types' of the Human Diseases. Based on the results or extraction of the TFs associated to

prostate cancer from GeneCards, these TFs (Table 3.2) were used in the KEGG Pathway as keyword entries.

3.6. Results and Discussion

Table 3.1: All TFs associated with the candidate genes identified in Chapter 2 extracted from GeneCards.

GENE	Regulatory Transcription factor
ERP44	NF-kappaB1, NF-kappaB, C/EBPalpha, HTF, HFH-1, HNF-4alpha2, POU3F1
GP1BA	AP-1,c-Jun, ATF-2, NF-kappaB1, E2F, E2F-1, E2F-2, E2F-3a, E2F-4
IFNG	STAT3, deltaCREB, CREB, POU2F1c, POU2F1b, POU2F1a, POU2F1, Ik-1, ATF-2, Hlf
SEPT2	Max1, Nkx3-1v4, c-Myc, Nkx3-1, Nkx3-1v1, Nkx3-1v2, Nkx3-1v3, Cdc5
TNFRSF13C	AP-1, c-Jun, ATF-2, Sp1, p53, E47, USF-1, USF1, SEF-1(1)
TNFSF4	AP-1, c-Jun, ATF-2, POU3F2, C/EBPalpha, POU6F1(c2), aMEF-2, MEF-2, MEF-2A, AML1a

The common TFs are highlighted amongst this subset of genes and those associated with a particular gene.

3.6.1. Transcription factors that target the promoter regions of the candidate genes that are involved in or showing a direct link to prostate cancer

NF-kappaB

Nuclear factor-κB (NF-κB) is a TF family that consists of five members in mammalian cells: NF-κB1 (p50), NF-κB2 (p52), RelA (p65), RelB and c-Rel. NF-κB is implicated in multiple physiological and pathological processes, including cell proliferation and differentiation, inflammatory and immune response, cell survival and apoptosis, cellular stress reactions and tumourigenesis (Yu *et al.,* 2009; Beinke and Ley 2004).

NF-κB1 is produced as a precursor protein, NF-κB1 p105, which is processed by the proteasome to produce the mature transcription factor NF-κB1 p50. As a cleaved product of p105, p50 only has a DNA binding domain and must form a heterodimer with RelA, RelB or C-Rel to act as a transcription factor to regulate its target gene transcription (Yu *et al*., 2009). Overexpression of p50 has been demonstrated in a large percentage of non-small cell carcinomas (Bours *et al.,* 1994; Mukhopadhyay *et al*., 1995). p50 involvement leads to the development of inflammation-associated cancers, including hepatoma, some breast cancers and colitis-associated cancers (Yu *et al.,* 2009). Wang *et al.,* 2006 reported that a peptide designed to bind with p50 can inhibit NF-κB activation and abate local acute inflammation.

There are reports demonstrating that TNF-α (Tumour necrosis factor alpha) also induce cell anti-apoptosis associated with NF-κB activation in prostate cancer cells (Sumitomo, 1999). The combination of TNF-alpha and NF-kappaB inhibitors could be constituted an effective therapy to TNF-alpha-resistant human prostate cancer cells (Sumitomo, 1999).

Transcription factor NF-kappaB1, showing expression in various cancers was annotated to regulate genes ERP44 and GP1BA (as seen in table 3.1).

C/EBPalpha.

The CCAAT/enhancer binding protein (C/EBP) family of TFs control the differentiation of a range of cell types, having key roles in regulating cellular proliferation through interaction with cell cycle proteins. C/EBPs have been described as both tumour promoters and tumour suppressors (Nerlov, 2007). Common to gene ERP44 and gene TNFSF4, C/EBPα is distinctive in that in addition to its transcriptional activity, it inhibits cell proliferation by several non-genomic mechanisms, causing it to be a tumour suppressor. In a study by Zhang *et al.,* 2008, they found that in contrast to other tumour suppressors, C/EBPα was frequently highly expressed in prostate tumours. A significant observation is the frequent co-expression of C/EBPα with androgen receptor (AR) in prostate cancer, particularly in the more aggressive metastatic tumours, which are both major regulators of transcription. This strongly suggests a combined role for them in determining the molecular phenotype of prostate tumours (Zhang *et al.,* 2008).

C/EBPα plays a role in hepatocytes and adipocytes mediating cell cycle arrest and transcription of hepatic-specific and adipose-specific genes respectively. In normal prostate, C/EBPα is expressed in the basal layer of the epithelium. Overexpression of C/EBPα inhibited epigenetically PSA expression and was accompanied by the loss of expression of AR (Yin *et al.,* 2006).

POU3F1

POU domain class 3, transcription factor I Oct-6 (Pou3f1, SCIP, Tst-1) is a member of the Pit-Oct-Unc (POU) family of transcription factors (Hofmann *et al.,* 2010).

In another study, candidate genes identified showed a differential methylation pattern between normal and prostate cancer cell lines (LNCaP, PC3, DU145) and observed additional expression differences in a gene family belonging to transcription factors (STAT1, POU3F1,

MYOD) (Mishra *et al.,* 2007). New insight into Oct-6 gene regulation was uncovered, with a potential impact on the control of nerve myelination (Hofmann *et al.,* 2010). POU3F1 was linked to gene ERP44 (table 3.1).

AP-1

Common to gene GP1BA, TNFRSF13C and TNFSF4, AP-1 nuclear transcriptional factors regulate expression of multiple genes involved in tumour growth, metastasis and angiogenesis. Activation of AP-1 has been implicated in prostate cancer development and growth, and therefore may represent promising therapeutic targets for cancer prevention and treatment (Uzzo *et al.,* 2006).

AP-1 is a heterodimer composed of proteins belonging to the c-Fos, c-Jun, ATF and JDP families. AP-1 controls a number of cellular processes including differentiation, proliferation, and apoptosis (Ameyar *et al.,* 2003). In addition to being activated by oncogenic signal transduction cascades, AP-1 is itself strongly oncogenic (Ozanne *et al.,* 2007).

c-Jun

In human prostate cancer, up-regulation of c-Jun proteins occurs in advanced disease and high levels of c-Jun expression are associated with disease recurrence. In a study by Ouyang *et al.*, 2008, the study revealed that an unappreciated role for AP-1 transcription factors in prostate cancer progression and identified c-Jun as a marker of high-risk prostate cancer (Ouyang *et al.,* 2008). Associated to genes GP1BA, TNFRSF13C and TNFSF4, c-Jun TF shows expression in prostate cancer.

E2F family

The E2F family of TFs was found to be associated to only gene GP1BA of the candidate genes. The E2F family are involved in the cell cycle regulation and synthesis of DNA. E2F1,

E2F2 and E2F3 function to promote the expression of cell cycle regulated genes and promote cell cycle progression, even if this promotion by deregulation of these E2F activities leads to defective cell cycling and apoptosis (DeGregori and Johnson, 2006). E2F-1, a regulator of cell proliferation and viability, reportedly plays a role in the development of hormone-refractory prostate cancer. Its ability to repress AR transcription, elevated levels of E2F-1 may contribute to the progression of hormone-refractory prostate cancer (Davis *et al.,* 2006).

STAT3

Signal transducers and activators of transcription, STAT as TFs are critical in mediating virtually all cytokine driven signaling. Activation of one STAT family member, Stat3, associated to gene IFNG, in human prostate cancer cell lines and primary prostate tumours revealed that elevated Stat3 activity was localized primarily in the tumour cells of prostate carcinoma specimens (Mora *et al*., 2002). Stat3 expression occurs frequently in primary prostate adenocarcinomas and is critical for the growth and survival of prostate cancer cells. These studies further suggest that Stat3 signaling represents a potentially novel molecular target for prostate cancer therapy (Mora *et al*., 2002). The inhibition of Stat3 signaling blocks the growth of prostate cancer cells and this suggest that targeting Stat3 signaling may yield a potential therapeutic intervention for prostate cancer (Ni *et al.,* 2000).

CREB

Akt is an antiapoptotic serine-threonine kinase that regulates a number of critical cellular pathways including those leading to cellular proliferation and inhibition of apoptosis. One of the downstream targets of Akt is CREB and has been shown to phosphorylate CREB; it is possible that Akt-mediated activation of CREB plays an important role in prostate carcinogenesis (Garcia *et al*., 2006). The transcription factor CREB regulates genes involved in various cellular processes by binding to cAMP response element (CRE) sequences present

in their promoter regions (Kim *et al.,* 2005). Defects in apoptotic signaling pathways are often associated with uncontrolled cell proliferation, high mutation rate and malignant transformation. Transcription factors, such as the mammalian ATF/CREB family of transcriptional regulators, have diverse functions in controlling cell proliferation and apoptosis (Persengiev and Green, 2003). AIbZIP is a novel member of the CREB/ATF family of transcription factors that is highly expressed in prostate tumours and of which the expression is up-regulated by androgen in LNCaP cells (Qi *et al*., 2002).

POU2F1

POU2F1 (also known as Oct-1) is the transcription factor associated to gene IFNG and Obinata *et al*., 2012 demonstrated that Oct1 can be a prognostic factor in prostate cancer and as a coregulator of AR and may lead to the development of a new therapeutic intervention for prostate cancer. Oct1 is involved in the proliferation of LNCaP cells (Obinata *et al*., 2012).

NKX3-1

NKX3-1 is a prostatic tumour suppressor gene. It is an androgen-regulated, prostate-specific gene whose expression is predominantly localized to prostate epithelium. It acts as a transcription factor that has critical function in prostate development and tumour suppression. Based on this, NKX3.1 is a candidate gene for playing a role in the opposing processes of androgen-driven differentiation of prostatic tissue and loss of that differentiation during the progression of prostate cancer and early stages of prostate carcinogenesis. Loss of NKX3A protein expression is a common finding in human prostate carcinomas and prostatic intraepithelial neoplasia (He *et al.,* 1997). Nkx3.1 is expressed early in the development of the prostate gland and is likely to play an important role in the differentiation of prostatic epithelia. Four splice variants (NKX3.1v1-4) (Korkmaz *et al.,* 2000) are associated with gene SEPT2.

Sp1

Associated with gene TNFRSF13C, Sp1 is elevated in several malignancies including prostate cancer and is associated with the prognosis of patients of this disease (Sankpal *et al*., 2011). Studies have shown that in prostate cancer, Sp1 regulates important genes like androgen receptor, TGF-β, c-Met, fatty acid synthase, matrix metalloprotein (MT1-MMP), PSA, and α-integrin. These results highlight the importance of Sp1 in prostate cancer and emphasize the potential therapeutic value of targeting Sp1 (Sankpal *et al*., 2011).

p53

Annotation of this TF was shown to gene TNFRSF13C, p53 mutations are detected in at least 20% of advanced prostate carcinomas and the overall p53 mutation rate is lower in prostate cancers than in many other cancers such as colon, lung, brain, breast, and bladder where p53 is mutated in over 50% of the cases studied (Surget *et al.,* 2014). The p53 tumour suppressor protein is widely known for its role as a transcription factor that regulates the expression of stress response genes and mediates a variety of anti-proliferative processes. It is the most commonly mutated tumour suppressor in human cancers (Surget *et al.,* 2014). Reported by Isaacs *et al.,* 1991, p53 gene mutations in prostate cancer cells suggested a functional role for the p53 gene in suppressing prostatic tumourigenesis.

E47

Transcription factor 3 (E2A immunoglobulin enhancer-binding factors E12/E47), also known as TCF3 is a multifunctional basic helix loop helix (bHLH), transcription factor, regulating transcription of target genes by homo- or heterodimerization with cell specific bHLH proteins. In general, E2A promotes cell differentiation, acts as a negative regulator of cell proliferation in normal cells and cancer cell lines. E2A, considered as a tumour suppressor is

highly expressed in prostate cancer. Loss of E2A promotes doxorubicin dependent apoptosis in prostate cancer cells (Patela and Chaudhary, 2012).

c-Myc

c-Myc affects the transcription of genes which participate in apoptosis. It is a key molecular integrator of cell cycle machinery and cellular metabolism and determines the common and divergent patterns of c-Myc target gene expression in a variety of physiological and neoplastic conditions (Dang, 1999). The transcriptionally active Max/Myc dimer promotes cell proliferation as well as apoptosis. The protein Max is a member of the basic helix-loop-helix leucine zipper (bHLHZ) family of transcription factors. It is able to form homodimers and heterodimers with other family members, which include Mad, Mxl1 and Myc. Myc is an oncoprotein implicated in cell proliferation, differentiation and apoptosis (Amati and Land, 1994).

Mutations of genes of the Myc family have been shown to be among the most frequently affected in the majority of human malignancies (Nair and Burley, 2006). Max1 and c-Myc are two TFs found to regulate the candidate gene IFNG only, of which Max1 showed no expression in prostate cancer from results obtained but numerous studies of human prostate cancer have demonstrated increased c-Myc expression in the prostate as well as increased Myc gene copy number in up to 30% of tumours, even at the preneoplastic stage called prostate intraepithelial neoplasia (PIN) (Ellwood-Yen *et al.,* 2003). Two commonly used prostate tumour cell lines, LNCaP and PC-3, have significant c-Myc amplification and overexpression (Zhong *et al*., 2000).

Table 3.2: Summary of TFs that target the promoter regions of the candidate genes involved in prostate cancer.

GENE	Prostate cancer associated Transcription factors
ERP44	NF-kappaB1, NF-kappaB, C/EBPalpha, POU3F1
GP1BA	AP-1,c-Jun, NF-kappaB1, E2F, E2F-1, E2F-2, E2F-3a, E2F-4
IFNG	STAT3, deltaCREB, CREB, POU2F1c, POU2F1b, POU2F1a, POU2F1
SEPT2	Nkx3-1v4, c-Myc, Nkx3-1, Nkx3-1v1, Nkx3-1v2, Nkx3-1v3,
TNFRSF13C	AP-1, c-Jun, Sp1, p53, E47
TNFSF4	AP-1, c-Jun, C/EBPalpha,

3.6.2. Transcription factors that target the promoter regions of the candidate genes that have shown no direct link to prostate cancer

The following transcription factors; HTF, HFH-1, HNF-4alpha2, ATF-2, Ik-1, HLF, Max1, POU3F2, POU6F1 (c2), Cdc5, USF-1, USF1, SEF-1(1), and aMEF-2, MEF-2, MEF-2A and AML1a showed no direct link to prostate cancer.

HTF

A TF, linked only to gene ERP44, HTF (HER2 transcription factor) from human epidermal growth factor receptor 2 has been identified as an AP-2 (activator protein-2) transcription factor. Vernimmen *et al.,* 2003 showed in their study that HTF is identical to AP-2 (activator protein-2). Overexpression of HTF has been shown to play an important role in the development and progression of certain aggressive types of breast cancer. In recent years the protein has become an important biomarker and target of therapy for approximately 30% of breast cancer patients (Mitri *et al.,* 2012).

HFH-1

Hepatocyte nuclear factor-3 homologue 1 (HFH-1) belongs to the family of the FOX transcription factors (previously called HNF-3/forkhead transcription factors) that play a role in tissue-specific and development gene regulation (Clevidence *et al.,* 1993).

HFH-1, linked only to gene ERP44 play important roles in early embryonic development and during development of many organ systems and represses the transcription of smooth muscle-specific genes (Hoggatt *et al.,* 2000).

HNF-4alpha2

Transcription factors, such as nuclear receptors activate transcription through interaction with coactivators. Nuclear receptors comprise a large superfamily of relatively conserved transcription modulators that play a role in nearly every aspect of growth, differentiation, and development (Erdmann *et al*., 2007). Hepatocyte nuclear factor 4alpha (HNF4alpha) is a tissue-specific transcription factor expressed in many cell types, including pancreatic beta-cells. HNF4alpha2 is a spliced form of HNF4alpha (Erdmann *et al.,* 2007). This TF was associated to gene ERP44.

ATF-2

Common to genes GP1BA, IFNG, TNFRSF13C and TNFSF4, ATF2 responds to stress-related stimuli and may thereby influence cell proliferation, inflammation, apoptosis, oncogenesis, neurological development and function, and skeletal remodelling. A study, implicated abnormal activation of ATF-2 in growth and progression of mammalian skin tumours (Leslie and Bar-Eli, 2005).

Heparan sulfate proteoglycans (HSPGs) are cell surface molecules that are extremely important in both development and cancer progression due to their regulation of cellular

processes such as angiogenesis, tumour growth, proliferation, tumour invasion and metastasis. HS2ST1 (heparan sulfate 2-O-sulfotransferase, 2OST) is essential for maximal proliferation and invasion of prostate cancer cells. HS2ST1 is upregulated by ATF2 (Ferguson and Datta, 2011).

POU family

POU homologs are divided into six major classes, including common names:

POU1 - POU1F1 (Pit-1); POU2 - POU2F1 (Oct-1), POU2F2 (Oct-2), POU2F3 (Oct-11); POU3 - POU3F1 (Oct-6; SCIP), POU3F2 (Oct-7; Brn-2), POU3F3 (Oct-8; Brn-1), POU3F4 (Oct-9; Brn-4; DFN3); POU4 - POU4F1 (Brn-3a; RDC-1; Oct-T1), POU4F2 (Brn-3b; Brn-3.2), POU4F3 (Brn-3c; Brn-3.1; DFNA15); POU5 - POU5F1(Oct-3; Oct-4), POU5F2 (SPRM-1), Pou2/V; POU6 - POU6f1 (Brn-5; mPOU), POU6f2 (Emb; RPF-1) (Gold *et al.,* 2014).

The transcription factors, POU3F2 and POU6F1 (c2), were linked to gene TNFSF4 which showed no expression in prostate cancer. In glioblastoma (GBM), a subset of stem-like tumour-propagating cells (TPCs) appears to drive tumour progression and underlie therapeutic resistance. A core set of neurodevelopmental TFs (POU3F2, SOX2, SALL2, and OLIG2) co-ordinately bind and activate TPC-specific regulatory elements and are essential for GBM propagation (Suvà1 *et al*., 2014).

POU6F1 (c2), referred to as brain-5 (Brn-5), is widely expressed with highest levels in the developing brain and spinal cord from embryonic day 12.5. In the adult, Brn-5 mRNA is most abundant in the brain, where it is diffusely expressed with the exception of enrichment in layer IV of the neocortex (Andersen *et al.,* 1993).

Ik-1

Linked to gene IFNG, the Ikaros family are zinc finger transcription factors. They participate in a complex network of interactions with gene regulatory elements, other family members and other transcriptional regulators to control gene expression. Ikaros family members regulate important cell-fate decisions during haematopoiesis. Mutation of several family members results in haematological malignancies, especially those of a lymphoid nature (John and Ward, 2011).

HLF

HLF is a member of the PAR (proline and acidic amino acid-rich region) subfamily of b/ZIP (basic region leucine zipper) transcription factors in acute lymphoblastic leukaemia (Hunger *et al*., 1992). Hepatic leukaemia factor inhibits cell death and is a candidate circadian factor (Watersa *et al*., 2013). Linked only to gene IFNG, this TF did not show any direct link to prostate cancer.

CDC5

The cell division cycle 5 (CDC5) is considered a putative transcription factor, as it is a MYB (a transcription factor)-related protein. In human and yeast, CDC5 has been shown to act as a component of the spliceosome to participate in mRNA splicing (Zhang *et al.,* 2013).

Polo-like kinases play critical roles during multiple stages of cell cycle progression. Polo-like kinase 1 (Plk1), a well-characterized member of serine/threonine kinases Plk family, has been shown to play pivotal roles in mitosis and cytokinesis in eukaryotic cells and plays critical roles in DNA replication and Pten null prostate cancer initiation. Cdc5 is a Plk homolog known to play important roles in cell cycle regulation (Luo and Liu, 2012), also only linked to gene SEPT2.

USF-1

USF is a family of transcription factors characterized by a highly conserved basic-helix-loop-helix-leucine zipper (bHLH-zip) DNA-binding domain. Upstream stimulatory factor 1(USF-1, USF1), can activate transcription through pyrimidine-rich initiator (Inr) elements and E-box motifs. This gene has been linked to familial combined hyperlipidaemia (FCHL) (Lee *et al.,* 2006). This TF was associated to gene TNFRSF13C.

SEF-1(1), aMEF-2, MEF-2, MEF-2A and AML1a

aMEF-2, MEF-2 myocyte enhancer factor-2 (Mef2) proteins are a family of transcription factors which through control of gene expression are important regulators of cellular differentiation and consequently play a critical role in embryonic development (Potthoff and Olson, 2007). Studies have revealed a central role for the myocyte enhancer factor-2 (MEF2) family of transcription factors in linking calcium-dependent signaling pathways to the genes responsible for cell division, differentiation and death (McKinsey *et al.,* 2002). The AML1 is associated with myelogenous leukaemias and encodes a DNA-binding protein. From AML1 gene, two representative forms of proteins, AML1a and AML1b, are produced by alternative splicing (Tanaka *et al*., 1997). SEF-1(1), TF showing no expression in prostate cancer, associated with gene TNFRSF13C. TFs (aMEF-2, MEF-2, MEF-2A and AML1a) showed no expression in prostate cancer and were linked to gene TNFSF4, as seen in table 3.1.

3.6.3. KEGG Pathway Analysis

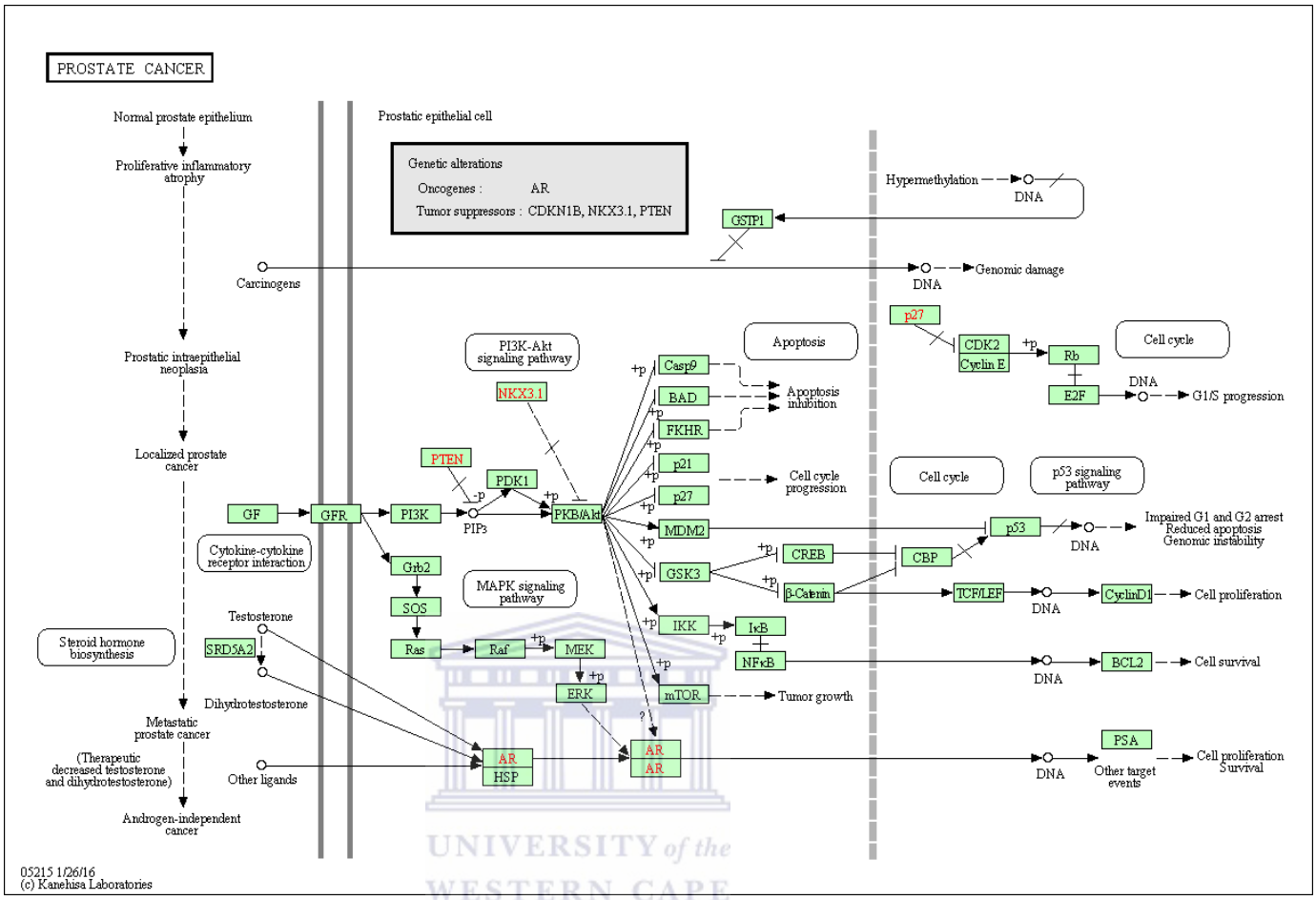


Figure 3.2: Extracted from the KEGG Pathway database showing gene and TF links in Prostate Cancer (Kanehisa Laboratories©). NKX3.1, PTEN, and p27 (shown in red) regulate the growth and survival of prostate cells in the normal prostate. Inadequate levels of PTEN and NKX3.1 lead to a reduction in p27 levels and to increased proliferation and decreased apoptosis (Kanehisa Laboratories©).

Searching the KEGG pathway database with the TFs shown in table 3.2, associated to the six genes, resulted in various TFs with some links to cancer however only 4 (NKX3.1, p53, E2F and CREB), showed association to the Prostate Cancer pathway.

NKX3-1 TF is seen in the Prostate Cancer pathway and related to the gene of interest SEPT2 as seen in table 3.2. p53 activation is induced by a number of stress signals, including DNA

damage, oxidative stress and was also found in the Prostate Cancer pathway linked to candidate gene TNFRSF13C. E2F, one of the TFs seen in the Prostate Cancer pathway was also found in Pathways in Cancer, p53 Signaling pathway and a number of other cancer pathways. E2F was linked to the gene of interest GP1BA. Another TF in the Prostate Cancer pathway, CREB (linked to gene of interest, IFNG) is also seen in other KEGG pathways such as the TNF Signaling pathway and Pathways in Cancer.

The KEGG pathway database also revealed various TFs with links to cancer pathways. The TF AP-1 showed linkage to pathways such as TNF signaling pathway and Apoptosis pathway (Appendix B) and Cell Survival pathway as well as Inflammation and Immunity pathways.

c-Jun, was found to be linked to the Apoptosis pathway, TNF Signaling pathway, Pathways in Cancer (Appendix C) and Colorectal Cancer pathway. The TFs, NF-kappaB involved in AGE-RAGE Signaling pathway and STAT3 TF is seen in Pathways in Cancer. c-Myc is seen in the Transcriptional Misregulation in Cancer pathway and Pathways in Cancer in the KEGG Pathway Database. Sp1 and E47 TFs were found in the Transcriptional Misregulation in Cancer pathway. The remaining TFs as seen in table 3.2 were not found within the KEGG Pathway databases revealing no results.

3.7. Conclusion

This chapter focused only on the TFs and the link of these TFs to pathways of the candidate biomarkers to prostate cancer.

The set of 44 TFs found to be associated to the candidate genes; there are 15 TFs out of the 44 that showed no experimental information about the genes they regulate with any direct links to prostate cancer.

The TFs that target the promoter region of the candidate genes have been reported to be involved in prostate cancer initiation, repression or progression. The remaining TFs linked to prostate cancer showing association to the candidate prostate cancer biomarkers (table 3.2); suggest that these genes could be potential diagnostic or prognostic markers as well as markers for monitoring therapeutic outcomes in prostate cancer.

To assess the association of pathways and TFs provides new clues to genes, pathways, and TFs that contribute to the outcome of prostate cancer and might be exploited in designing new biomarkers thus demonstrating that effectively incorporating pathway information with TFs can provide better diagnostic evaluation.

CREB, E2F, Nkx3-1 and p53 TFs were discovered to be linked to the genes IFNG, GP1BA, SEPT2 and TNFRSF13C respectively in prostate cancer and therefore shows that these genes can be assessed as possible biomarkers for prostate cancer.

3.8. References

AMATI, B. & LAND, H. 1994. Myc-Max-Mad: a transcription factor network controlling cell cycle progression, differentiation and death. *Curr Opin Genet Dev,* 4**,** 102-8.

AMEYAR, M., WISNIEWSKA, M. & WEITZMAN, J. B. 2003. A role for AP-1 in apoptosis: the case for and against. *Biochimie,* 85**,** 747-52.

ANDERSEN, B., SCHONEMANN, M. D., PEARSE, R. V., 2ND, JENNE, K., SUGARMAN, J. & ROSENFELD, M. G. 1993. Brn-5 is a divergent POU domain factor highly expressed in layer IV of the neocortex. *J Biol Chem,* 268**,** 23390-8.

BEINKE, S. & LEY, S. C. 2004. Functions of NF-kappaB1 and NF-kappaB2 in immune cell biology. *Biochem J,* 382**,** 393-409.

BOURS, V., DEJARDIN, E., GOUJON-LETAWE, F., MERVILLE, M. P. & CASTRONOVO, V. 1994. The NF-kappa B transcription factor and cancer: high expression of NF-kappa B- and I kappa B-related proteins in tumor cell lines. *Biochem Pharmacol,* 47**,** 145-9.

CLEVIDENCE, D. E., OVERDIER, D. G., TAO, W., QIAN, X., PANI, L., LAI, E. & COSTA, R. H. 1993. Identification of nine tissue-specific transcription factors of the hepatocyte nuclear factor 3/forkhead DNA-binding-domain family. *Proc Natl Acad Sci U S A,* 90**,** 3948-52.

DANG, C. V. 1999. c-Myc target genes involved in cell growth, apoptosis, and metabolism. *Mol Cell Biol,* 19**,** 1-11.

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 Res,* 66**,** 11897-906.

DEGREGORI, J. & JOHNSON, D. G. 2006. Distinct and Overlapping Roles for E2F Family Members in Transcription, Proliferation and Apoptosis. *Curr Mol Med,* 6**,** 739-48.

DELGADO, M. D. & LEON, J. 2006. Gene expression regulation and cancer. *Clin Transl Oncol,* 8**,** 780-7.

ELLWOOD-YEN, K., GRAEBER, T. G., WONGVIPAT, J., IRUELA-ARISPE, M. L., ZHANG, J., MATUSIK, R., THOMAS, G. V. & SAWYERS, C. L. 2003. Myc-driven murine prostate cancer shares molecular features with human prostate tumors. *Cancer Cell,* 4**,** 223-38.

ERDMANN, S., SENKEL, S., ARNDT, T., LUCAS, B., LAUSEN, J., KLEIN-HITPASS, L., RYFFEL, G. U. & THOMAS, H. 2007. Tissue-specific transcription factor HNF4alpha inhibits cell proliferation and induces apoptosis in the pancreatic INS-1 beta-cell line. *Biol Chem,* 388**,** 91-106.

FERGUSON, B. W. & DATTA, S. 2011. Role of heparan sulfate 2-o-sulfotransferase in prostate cancer cell proliferation, invasion, and growth factor signaling. *Prostate Cancer,* 2011**,** 893208.

FISHILEVICH, S., ZIMMERMAN, S., KOHN, A., INY STEIN, T., OLENDER, T., KOLKER, E., SAFRAN, M. & LANCET, D. 2016. Genic insights from integrated human proteomics in GeneCards. *Database (Oxford),* 2016.

GARCIA, G. E., NICOLE, A., BHASKARAN, S., GUPTA, A., KYPRIANOU, N. & KUMAR, A. P. 2006. Akt-and CREB-mediated prostate cancer cell proliferation inhibition by Nexrutine, a Phellodendron amurense extract. *Neoplasia,* 8**,** 523-33.

GENECARDS. Available: www.genecards.org.

GOLD, D. A., GATES, R. D. & JACOBS, D. K. 2014. The early expansion and evolutionary dynamics of POU class genes. *Mol Biol Evol,* 31**,** 3136-47.

HE, W. W., SCIAVOLINO, P. J., WING, J., AUGUSTUS, M., HUDSON, P., MEISSNER, P. S., CURTIS, R. T., SHELL, B. K., BOSTWICK, D. G., TINDALL, D. J., GELMANN, E. P., ABATE-SHEN, C. & CARTER, K. C. 1997. A novel human prostate-specific, androgen-regulated homeobox gene (NKX3.1) that maps to 8p21, a region frequently deleted in prostate cancer. *Genomics,* 43**,** 69-77.

HOFMANN, E., REICHART, U., GAUSTERER, C., GUELLY, C., MEIJER, D., MULLER, M. & STROBL, B. 2010. Octamer-binding factor 6 (Oct-6/Pou3f1) is induced by interferon and contributes to dsRNA-mediated transcriptional responses. *BMC Cell Biol,* 11**,** 61.

HOGGATT, A. M., KRIEGEL, A. M., SMITH, A. F. & HERRING, B. P. 2000. Hepatocyte nuclear factor-3 homologue 1 (HFH-1) represses transcription of smooth muscle-specific genes. *J Biol Chem,* 275**,** 31162-70.

HUNGER, S. P., OHYASHIKI, K., TOYAMA, K. & CLEARY, M. L. 1992. Hlf, a novel hepatic bZIP protein, shows altered DNA-binding properties following fusion to E2A in t(17;19) acute lymphoblastic leukemia. *Genes Dev,* 6**,** 1608-20.

ISAACS, W. B., CARTER, B. S. & EWING, C. M. 1991. Wild-Type P53 Suppresses Growth of Human Prostate-Cancer Cells Containing Mutant P53 Alleles. *Cancer Research,* 51**,** 4716-4720.

JOHN, L. B. & WARD, A. C. 2011. The Ikaros gene family: transcriptional regulators of hematopoiesis and immunity. *Mol Immunol,* 48**,** 1272-8.

KANEHISA, M. & GOTO, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research,* 28**,** 27-30.

KANEHISA, M., GOTO, S., FURUMICHI, M., TANABE, M. & HIRAKAWA, M. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research,* 38**,** D355-D360.

KIM, J., JIA, L., STALLCUP, M. R. & COETZEE, G. A. 2005. The role of protein kinase A pathway and cAMP responsive element-binding protein in androgen receptor-mediated transcription at the prostate-specific antigen locus. *J Mol Endocrinol,* 34**,** 107-18.

KORKMAZ, K. S., KORKMAZ, C. G., RAGNHILDSTVEIT, E., KIZILDAG, S., PRETLOW, T. G. & SAATCIOGLU, F. 2000. Full-length cDNA sequence and genomic organization of human NKX3A - alternative forms and regulation by both androgens and estrogens. *Gene,* 260**,** 25-36.

KYOTO ENCYCLOPEDIA OF GENES AND GENOMES (KEGG) PATHWAY DATABASE. Available: http://www.genome.jp/kegg/pathway.html.

LEE, J. C., LUSIS, A. J. & PAJUKANTA, P. 2006. Familial combined hyperlipidemia: upstream transcription factor 1 and beyond. *Curr Opin Lipidol,* 17**,** 101-9.

LESLIE, M. C. & BAR-ELI, M. 2005. Regulation of gene expression in melanoma: new approaches for treatment. *J Cell Biochem,* 94**,** 25-38.

LUO, J. & LIU, X. 2012. Polo-like kinase 1, on the rise from cell cycle regulation to prostate cancer development. *Protein Cell,* 3**,** 182-97.

MCKINSEY, T. A., ZHANG, C. L. & OLSON, E. N. 2002. MEF2: a calcium-dependent regulator of cell division, differentiation and death. *Trends Biochem Sci,* 27**,** 40-7.

MITRI, Z., CONSTANTINE, T. & O'REGAN, R. 2012. The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy. *Chemother Res Pract,* 2012**,** 743193.

MORA, L. B., BUETTNER, R., SEIGNE, J., DIAZ, J., AHMAD, N., GARCIA, R., BOWMAN, T., FALCONE, R., FAIRCLOUGH, R., CANTOR, A., MURO-CACHO, C., LIVINGSTON, S., KARRAS, J., POW-SANG, J. & JOVE, R. 2002. Constitutive

activation of Stat3 in human prostate tumors and cell lines: direct inhibition of Stat3 signaling induces apoptosis of prostate cancer cells. *Cancer Res,* 62**,** 6659-66.

MUKHOPADHYAY, T., ROTH, J. A. & MAXWELL, S. A. 1995. Altered expression of the p50 subunit of the NF-kappa B transcription factor complex in non-small cell lung carcinoma. *Oncogene,* 11**,** 999-1003.

NAIR, S. K. & BURLEY, S. K. 2006. Structural aspects of interactions within the Myc/Max/Mad network. *Curr Top Microbiol Immunol,* 302**,** 123-43.

NERLOV, C. 2007. The C/EBP family of transcription factors: a paradigm for interaction between gene expression and proliferation control. *Trends Cell Biol,* 17**,** 318-24.

NI, Z., LOU, W., LEMAN, E. S. & GAO, A. C. 2000. Inhibition of constitutively activated Stat3 signaling pathway suppresses growth of prostate cancer cells. *Cancer Res,* 60**,** 1225-8.

OBINATA, D., TAKAYAMA, K., URANO, T., MURATA, T., KUMAGAI, J., FUJIMURA, T., IKEDA, K., HORIE-INOUE, K., HOMMA, Y., OUCHI, Y., TAKAHASHI, S. & INOUE, S. 2012. Oct1 regulates cell growth of LNCaP cells and is a prognostic factor for prostate cancer. *Int J Cancer,* 130**,** 1021-8.

OUYANG, X., JESSEN, W. J., AL-AHMADIE, H., SERIO, A. M., LIN, Y., SHIH, W. J., REUTER, V. E., SCARDINO, P. T., SHEN, M. M., ARONOW, B. J., VICKERS, A. J., GERALD, W. L. & ABATE-SHEN, C. 2008. Activator protein-1 transcription factors are associated with progression and recurrence of prostate cancer. *Cancer Res,* 68**,** 2132-44.

OZANNE, B. W., SPENCE, H. J., MCGARRY, L. C. & HENNIGAN, R. F. 2007. Transcription factors control invasion: AP-1 the first among equals. *Oncogene,* 26**,** 1-10.

PATEL, D. & CHAUDHARY, J. 2012. Increased expression of bHLH transcription factor E2A (TCF3) in prostate cancer promotes proliferation and confers resistance to doxorubicin induced apoptosis. *Biochem Biophys Res Commun,* 422**,** 146-51.

PERSENGIEV, S. P. & GREEN, M. R. 2003. The role of ATF/CREB family members in cell growth, survival and apoptosis. *Apoptosis,* 8**,** 225-8.

PHILLIPS, T. & HOOPES, L. 2008. Transcription factors and transcriptional control in eukaryotic cells. *Nature Education,* 1**,** 119.

POTTHOFF, M. J. & OLSON, E. N. 2007. MEF2: a central regulator of diverse developmental programs. *Development,* 134**,** 4131-40.

QI, H., FILLION, C., LABRIE, Y., GRENIER, J., FOURNIER, A., BERGER, L., EL-ALFY, M. & LABRIE, C. 2002. AIbZIP, a novel bZIP gene located on chromosome 1q21.3 that is highly expressed in prostate tumors and of which the expression is up-regulated by androgens in LNCaP human prostate cancer cells. *Cancer Res,* 62**,** 721-33.

REBHAN, M., CHALIFA-CASPI, V., PRILUSKY, J. & LANCET, D. 1998. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics,* 14**,** 656-64.

SAFRAN, M., DALAH, I., ALEXANDER, J., ROSEN, N., INY STEIN, T., SHMOISH, M., NATIV, N., BAHIR, I., DONIGER, T., KRUG, H., SIROTA-MADI, A., OLENDER, T., GOLAN, Y., STELZER, G., HAREL, A. & LANCET, D. 2010. GeneCards Version 3: the human gene integrator. *Database (Oxford),* 2010**,** baq020.

SANKPAL, U. T., GOODISON, S., ABDELRAHIM, M. & BASHA, R. 2011. Targeting Sp1 transcription factors in prostate cancer therapy. *Med Chem,* 7**,** 518-25.

SUMITOMO, M., TACHIBANA, M., NAKASHIMA, J., MURAI, M., MIYAJIMA, A., KIMURA, F., HAYAKAWA, M. & NAKAMURA, H. 1999. An essential role for

nuclear factor kappa B in preventing TNF-alpha-induced cell death in prostate cancer cells. *J Urol,* 161**,** 674-9.

SURGET, S., KHOURY, M. P. & BOURDON, J. C. 2014. Uncovering the role of p53 splice variants in human malignancy: a clinical perspective. *Oncotargets and Therapy,* 7**,** 57-67.

SUVA, M. L., RHEINBAY, E., GILLESPIE, S. M., PATEL, A. P., WAKIMOTO, H., RABKIN, S. D., RIGGI, N., CHI, A. S., CAHILL, D. P., NAHED, B. V., CURRY, W. T., MARTUZA, R. L., RIVERA, M. N., ROSSETTI, N., KASIF, S., BEIK, S., KADRI, S., TIROSH, I., WORTMAN, I., SHALEK, A. K., ROZENBLATT-ROSEN, O., REGEV, A., LOUIS, D. N. & BERNSTEIN, B. E. 2014. Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell,* 157**,** 580-94.

TANAKA, T., TANAKA, K., OGAWA, S., KUROKAWA, M., MITANI, K., YAZAKI, Y., SHIBATA, Y. & HIRAI, H. 1997. An acute myeloid leukemia gene, AML1, regulates transcriptional activation and hemopoietic myeloid cell differentiation antagonistically by two alternative spliced forms. *Leukemia,* 11 Suppl 3**,** 299-302.

UZZO, R. G., CRISPEN, P. L., GOLOVINE, K., MAKHOV, P., HORWITZ, E. M. & KOLENKO, V. M. 2006. Diverse effects of zinc on NF-kappaB and AP-1 transcription factors: implications for prostate cancer progression. *Carcinogenesis,* 27**,** 1980-90.

VERNIMMEN, D., BEGON, D., SALVADOR, C., GOFFLOT, S., GROOTECLAES, M. & WINKLER, R. 2003. Identification of HTF (HER2 transcription factor) as an AP-2 (activator protein-2) transcription factor and contribution of the HTF binding site to ERBB2 gene overexpression. *Biochem J,* 370**,** 323-9.

WANG, Z. P., CAI, S. X., LIU, D. B., XU, X. & LIANG, H. P. 2006. Anti-inflammatory effects of a novel peptide designed to bind with NF-kappaB p50 subunit. *Acta Pharmacol Sin,* 27**,** 1474-8.

WATERS, K. M., SONTAG, R. L. & WEBER, T. J. 2013. Hepatic leukemia factor promotes resistance to cell death: implications for therapeutics and chronotherapy. *Toxicol Appl Pharmacol,* 268**,** 141-8.

YIN, H., RADOMSKA, H. S., TENEN, D. G. & GLASS, J. 2006. Down regulation of PSA by C/EBPalpha is associated with loss of AR expression and inhibition of PSA promoter activity in the LNCaP cell line. *BMC Cancer,* 6**,** 158.

YU, Y., WAN, Y. & HUANG, C. 2009. The biological functions of NF-kappaB1 (p50) and its potential as an anti-cancer target. *Curr Cancer Drug Targets,* 9**,** 566-71.

ZHANG, J., WILKINSON, J. E., GONIT, M., KECK, R., SELMAN, S. & RATNAM, M. 2008. Expression and sub-cellular localization of the CCAAT/enhancer binding protein alpha in relation to postnatal development and malignancy of the prostate. *Prostate,* 68**,** 1206-14.

ZHANG, S., XIE, M., REN, G. & YU, B. 2013. CDC5, a DNA binding protein, positively regulates posttranscriptional processing and/or transcription of primary microRNA transcripts. *Proc Natl Acad Sci U S A,* 110**,** 17588-93.

ZHONG, H., CHILES, K., FELDSER, D., LAUGHNER, E., HANRAHAN, C., GEORGESCU, M. M., SIMONS, J. W. & SEMENZA, G. L. 2000. Modulation of hypoxia-inducible factor 1alpha expression by the epidermal growth factor/phosphatidylinositol 3-kinase/PTEN/AKT/FRAP pathway in human prostate cancer cells: implications for tumor angiogenesis and therapeutics. *Cancer Res,* 60**,** 1541-5.

Chapter 4

Prognostic and Predictive validation of Genes using an *in silico* approach

4.1. Introduction

Genome-wide molecular profiles have served as sources for *in silico* discovery and/or validation of predictive/prognostic molecular biomarkers as well as therapeutic targets (Chen *et al*., 2014). Prognostic biomarkers are measurements made at diagnosis that provide information about a patient's prognosis in the absence of treatment or in the presence of standard treatment. Predictive biomarkers are measured at baseline to identify patients who are likely or unlikely to benefit from a specific treatment (Simon, 2009). Prognostic markers help to stratify patients for treatment by identifying patients with different risks of outcome (e.g. recurrence of disease) and are important tools in the management of cancer and many other diseases (Riley *et al*., 2003).

To obtain more accurate predictions, researchers have developed predictive tools (directed at predicting the probability of an outcome without considering the effect of time) and prognostic tools (directed at predicting the probability of an outcome over time) that are based on statistical models. Predictive and prognostic nomograms (devices that make predictions) have been introduced to predict the risk of the outcome of interest for the individual patient (Shariat *et al.,* 2008).

Several tools are available to conduct online survival analysis on genes of interest using publicly available data (Goswami and Nakshatri, 2014).

For a molecular biomarker to be considered as a clinical test, the marker must demonstrate clinical magnitude of benefit, feasibility of clinical implementation and low costs. This incorporation of any clinical frameworks can be technically possible in various web-based

tools of genomic survival analysis by engaging disease domain experts in their development (Chen *et al*., 2014).

Gene expression profiles are routinely applied to identify diagnostic and predictive biomarkers or novel targets for cancer and a few predictive markers identified *in silico* have also been validated for clinical, functional or mechanistic relevance in disease progression (Alinezhad *et al.*, 2016).

Several clinical features of prostate cancer including tumour stage, degree of tumour cell differentiation or Gleason score (GS) and the serum PSA are used in routine clinical practice to separate men into groups at low, intermediate, and high risk for tumour recurrence following local therapy (Singh *et al*., 2002). Majority of patients who undergo prostatectomy have low to intermediate risk clinical features and determining the prognosis for these patients remains difficult. Attempts to explore genetic correlates of tumour behaviour have found alterations in a number of candidate genes associated with prostate cancer progression and no single gene has been shown to have sufficient prognostic utility to warrant clinical implementation (Singh *et al*., 2002).

To further examine the genes that were identified as possible novel biomarkers for prostate cancer by bioinformatics analysis, survival analysis methodologies have been adapted to the analysis of genomics data to link molecular information with clinical outcomes of interest especially prognosis. Genes recognized as being potential biomarkers for prostate cancer were evaluated to assess their relationship to prognosis of this disease.

Validation studies provide conclusion about the expression of the candidate genes by using further bioinformatics tools. Uniting several databases enhances more accurate estimates of biomarkers for cancer diagnosis and prognosis. For the analysis of the candidate biomarkers

identified in Chapter 2, for prognostic value, several databases have been used to validate the six candidate genes in this study.

4.2. SurvExpress®

SurvExpress is a cancer-wide gene expression web-based database with clinical outcomes and a tool that provides survival analysis, risk assessment of cancer datasets and validations of survival biomarkers for cancer outcomes (Aguirre-Gamboa *et al*., 2013). It is a cancer database containing more than 20,000 samples and 130 datasets with censored clinical information covering tumours over 20 tissues. SurvExpress is the largest, most versatile, and quickest free tool available. SurvExpress can be accessed at http://bioinformatica.mty.itesm.mx/SurvExpress using a biomarker gene list as input (Aguirre-Gamboa *et al*., 2013).

4.3. PROGgene

PROGgene is a web accessible tool available at www.compbio.iupui.edu/proggene for researchers to identify potential prognostic biomarkers. It is useful in accelerating biomarker discovery in cancer and quickly providing results that may indicate disease-specific prognostic value of specific biomarkers. It has compiled data from public repositories such as Gene Expression Omnibus (GEO), European Bioinformatics Institute (EBI) Array Express and The Cancer Genome Atlas (Goswami and Nakshatri, 2013). The database can be used to create prognostic (Kaplan-Meier, KM) plots for mRNAs and genes of interest using data in different cancers. A total of 64 datasets from 18 cancer types make this tool the most comprehensive prognostic biomarker identification tool to date and is primarily a hypothesis generation tool, which is meant to provide pursuable gene biomarkers in cancers of choice (Goswami and Nakshatri, 2013).

4.4. PrognoScan

PrognoScan can be used for assessing the biological relationship between gene expression and prognosis. The database is publicly accessible at http://gibk21.bse.kyutech.ac.jp/PrognoScan/index.html. It is used to evaluate potential tumour markers and therapeutic targets and would accelerate cancer research. Cancer microarray datasets with clinical annotation were collected from the public domain including Gene Expression Omnibus (GEO) and individual laboratory web sites (Mizuno *et al*., 2009). The collection includes more than 40 datasets of various cancer types. The probe annotations were retrieved from GEO and ArrayExpress which was mapped to an Entrez Gene ID by querying the accompanied public identifier in UniGene database (Mizuno *et al*., 2009).

PrognoScan employs the minimum p-value approach for grouping patients for survival analysis that finds the optimal cut-point in continuous gene expression measurements (Mizuno *et al*., 2009). Patients are divided into two (high and low) expression groups and the risk differences of the two groups are estimated by log-rank test. The optimal cut-point that gives the most pronounced pvalue (Pmin) is selected and conducts multiple correlated testing, p-value correction is conducted to control the error rate (Mizuno *et al*., 2009).

4.5. Kaplan-Meier plots used as survival curves for prognostic value

Kaplan-Meier plots using the Kaplan-Meier method were used to construct survival curves. The method, called the Kaplan-Meier estimator (also known as the product limit estimator), is based on a mathematical formula using information from those who have died and those who have survived to estimate the proportion of patients survived at any point during a study. The estimator is plotted over time. It computes the probabilities of occurrence of an event at a certain point in time. The resulting curve is called the Kaplan-Meier curve, which is a series of horizontal steps of declining magnitude that, when a large enough sample is taken,

approaches the true survival function for that population (Goel *et al*., 2010, Kaplan and Meier, 1958). The Kaplan-Meier approach estimates the survival curve in the presence of censored observations. Censoring means the total survival time for that subject cannot be accurately determined (Rich *et al.*, 2010).

Most survival analyses in cancer journals use some or all of Kaplan-Meier (KM) plots, log rank tests, and Cox (proportional hazards) regression (Clark *et al*., 2003). The main outcome under assessment is the time to an event of interest being the survival time, the time from complete remission to relapse or progression as equally as to the time from diagnosis to death.

In each of the 3 gene expression databases used, KM plots were used to construct survival curves for the candidate prostate cancer biomarkers identified by this study due to the fact that the Kaplan-Meier method takes into account censored data, giving an estimate of recovery rates, probability of death, measurement of multiple variables to determine correlation of events.

4.6. Aim

The aim of this chapter is to evaluate the six identified candidate genes from Chapter 2, as prognostic and/or predictive biomarkers using various bioinformatics databases.

4.7. Materials and Methods

4.7.1. SurvExpress for the validation of prognostic biomarkers

The list consisting of the 6 identified potential prostate cancer biomarkers were submitted to SurvExpress (http://bioinformatica.mty.itesm.mx/SurvExpress) homepage for validation as possible prognostic markers. The six genes (ERP44, GP1BA, IFNG, SEPT2, TNFRSF13C, and TNFSF4) were entered as a panel as well as individually and "prostate" was selected as

tissue type. The default settings were used for the options "duplicated genes" (Average: All probe sets/records will be averaged per sample) and "data" [Original (Quantile-Normalized)]. The genes were then analysed using four different datasets. For this study, there were eight Prostate datasets available and four datasets were selected with the most number of samples and relevant clinical data; (a) Taylor MSKCC with 140 samples and clinical data of recurrence, Gleason score and stage, (b) Sboner Rubin GSE16560 with 281 samples and clinical data of Gleason score, (c) Kollmeyer-Jenkins Prostate GSE10645-GPL5858 with 596 samples and clinical data of Survival, Age, PSA, Stage, Grade (d) PRAD-TCGA with 497 samples and clinical data of survival.

Search criteria analysis was performed according to the parameters in table 4.1. The rest of the parameters were left as default. The results generated, after the specification of parameters, showed Kaplan-Meier curves, plots and tables. Results were looked at as a panel and individually to identify prognostic or predictive biomarkers in prostate cancer from the list of putative biomarkers identified.

Table 4.1: Search criteria for the four datasets used in SurvExpress

Dataset	**Censored**	**Stratification**
Taylor MSKCC	Recurrence_months	Death
Sboner Rubin GSE16560	Survival_month	Outcome
Kollmeyer-Jenkins GSE10645-GPL5858	Survival_after_RRP	Prostate.cancer.specific.death
PRAD-TCGA	Survival_days	No stratification

4.7.2. Prognostic value of the candidate biomarkers using PROGgene

The six candidate genes separated by commas were entered in the 'Input Genes' space provided and cancer type selected was 'prostate'. Survival measure chosen was 'death' and bifurcate gene expression remained as default 'Median'. Using the median gene expression value as bifurcating point allowed samples to be divided into High and Low gene expression groups. A result in only one dataset was generated: 'GSE16560 – Molecular Sampling of Prostate Cancer: a dilemma for predicting disease progression' with data for only two out of the six genes found: GP1BA and TNFSF4. Kaplan-Meier plots were created as the results output for both genes.

4.7.3. Analysis of the biological relationship between gene expression and prognosis within the PrognoScan database.

The multiple genes were entered in the gene identifier space provided and submitted. This resulted in a "Summary table" for each gene. Headings for each table include dataset, cancer type, subtype, endpoint, cohort, contributor, array type; probe ID, number of patients, optimal cutpoint, p-value (Pmin and Pcor). A statistically significant value of Pcor is given in red font. Each dataset has a link to the public domain where the raw data is archived (Mizuno *et al*., 2009). Clicking the "probe ID" in the list revealed a detailed report of the individual gene which indicated further annotations for the dataset.

The following plots where used to describe the prognostic value of each gene found in the PrognoScan databases (See figures 4.6, 4.7 and 4.8) (i) Expression plot showing patients ordered by the expression values of the given gene; the X-axis represents the accumulative number of patients and the Y-axis represents the expression value. Straight lines (cyan) show the optimal cut points that dichotomize patients into high (red) and low (blue) expression groups. (ii) Expression histogram resulting in the distribution of the expression value is

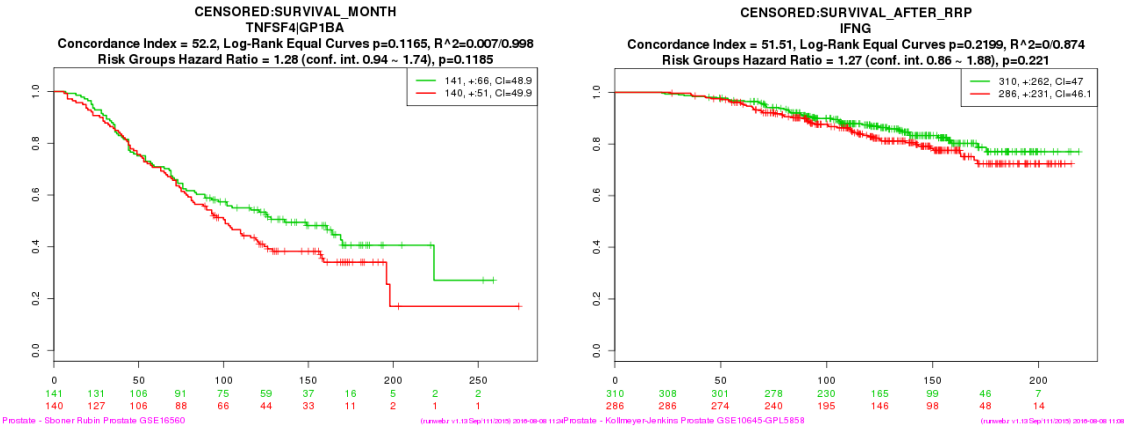
presented where the X-axis represents the number of patients and the Y-axis represents the expression value on the same scale as the expression plot. The line of the optimal cutpoint is also shown (cyan). (iii) P-value plot for each potential cut point of expression measurement; the X-axis represents the accumulative number of patients on the same scale as the expression plot and the Y-axis represents raw P-values on a log scale. The cut point to minimize the P-value is determined and indicated by the cyan line. The grey line indicates the 5% significance level and (iv) Kaplan-Meier plot displays survival curves for high (red) and low (blue) expression groups at the optimal cut point are plotted; the X-axis represents time and the Y-axis represents survival rate. 95% confidence intervals for each group are also indicated by dotted lines (Mizuno *et al*., 2009).

4.8. Results and Discussion

4.8.1. Analysis of the six biomarkers as a panel using SurvExpress

SurvExpress, an online biomarker validation tool and database, was used to explore the patient survival outcome in relation to the expression of the candidate biomarkers for prostate cancer.

(A) (B)



(C) (D)

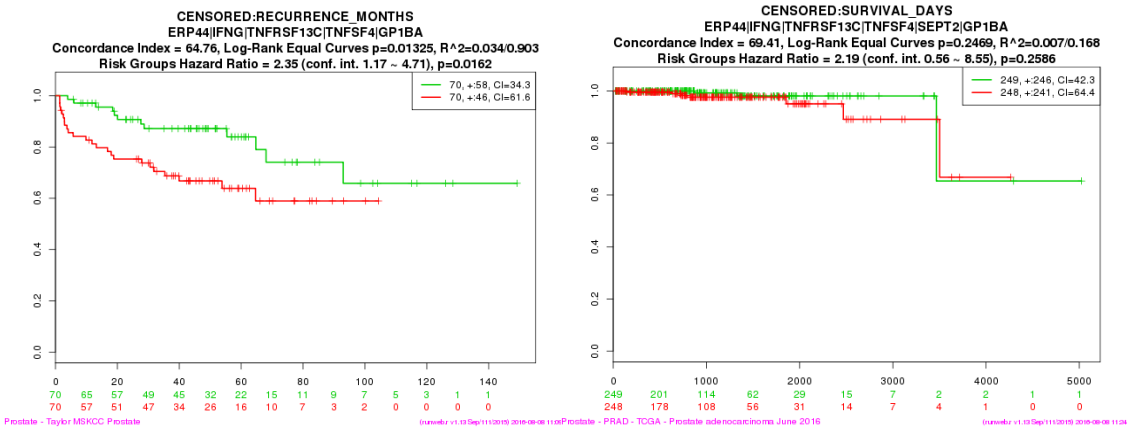


Figure 4.1: Kaplan-Meier curves and performance of the six candidate biomarker panel in four datasets by SurvExpress: (A) Survival by month on Sboner Rubin prostate cancer dataset (B) Survival after Radical retropubic prostatectomy (RRP) on Kollmeyer Jerkins prostate cancer dataset (C) Recurrence on Taylor prostate cancer dataset and (D) Survival by days on Prostate adenocarcinoma (PRAD) - June 2016 dataset (refer to sections 4.2, 4.3, 4.4 and 4.5 for patient data comparison of graphs in each database).

The Sboner Rubin (Fig. 4.1A), Kollmeyer-Jenkins (Fig. 4.1B) and PRAD (Fig.4.1D) datasets demonstrated association of the biomarkers with patient survival. Taylor dataset also indicates clear association with disease recurrence (Fig. 4.1C). Analysis on PRAD data sets indicates that the genes are significantly associated with Prostate adenocarcinoma survival (Fig. 4.1D). The results of Kaplan–Meier (KM) plots for the four authors are summarized in Figure 4.1.

4.8.1.1. Dataset: Taylor MSKCC Prostate

Resulting in one gene not found SEPT2 and the rest of the 5 genes as seen in the Kaplan-Meier survival curve showed highly significant differences between low and high expression of this gene panel in the probability that prostate cancer patients display recurrence of disease.

4.8.1.2. Dataset: Sboner Rubin GSE16560

Results for 4 genes out of the 6 were not found (ERP44, SEPT2, TNFRSF13C, and IFNG) using this tool. As a panel of genes in this dataset, 2 genes: TNFSF4 and GP1BA were found but showed only slight prognostic value to the end of the study where patients highly expressing these genes (red) show poor prognosis compared to patients with low expression of these biomarkers. Biomarkers when combined could significantly predict the prognostic outcome of prostate cancer based on the difference in outcome between the higher and the lower expression of these genes (as shown in figure 4.1).

4.8.1.3. Dataset: Kollmeyer-Jenkins Prostate GSE10645-GPL5858

Results showed 5 genes not found: ERP44, GP1BA, SEPT2, TNFRSF13C, and TNFSF4, with only 1 gene (IFNG) found in this dataset, showing that after radical retropubic prostatectomy patients survives resulting in no prognostic value of this gene.

4.8.1.4. Dataset: PRAD - TCGA - Prostate adenocarcinoma

Although all genes were found in this dataset, the expression of genes resulted in no differential expression in low and high risk patients, which estimated the probability of survival in days with censoring occurring at the start of the study, giving an indication that the genes present poor prognostic value for prostate cancer.

Table 4.2: Result of survival analysis showing the significant genes based on p-values ($p \leq 0.05$) using SurvExpress

	Summary Result
Biomarker	ERP44, GP1BA, IFNG, SEPT2, TNFRSF13C, TNFSF4
Database	Prostate - PRAD - TCGA - Prostate adenocarcinoma June 2016
Genes	6
CI	0.694142
Log.Rank	0.2469253
Hazard.Ratio	2.191466
C.I.Hazard.Ratio	[0.561838339122112 - 8.54787633194716]
pHR	0.2585771
Correlation	0.006903972
Significant.Genes	1
Marginal.Genes	1
Cox.Interesting	GP1BA
DEG	4
DEG.Interesting	GP1BA, IFNG, SEPT2, TNFSF4
Notes	Hazard Ratio was estimated by fitting a CoxPH using risk group as covariate.

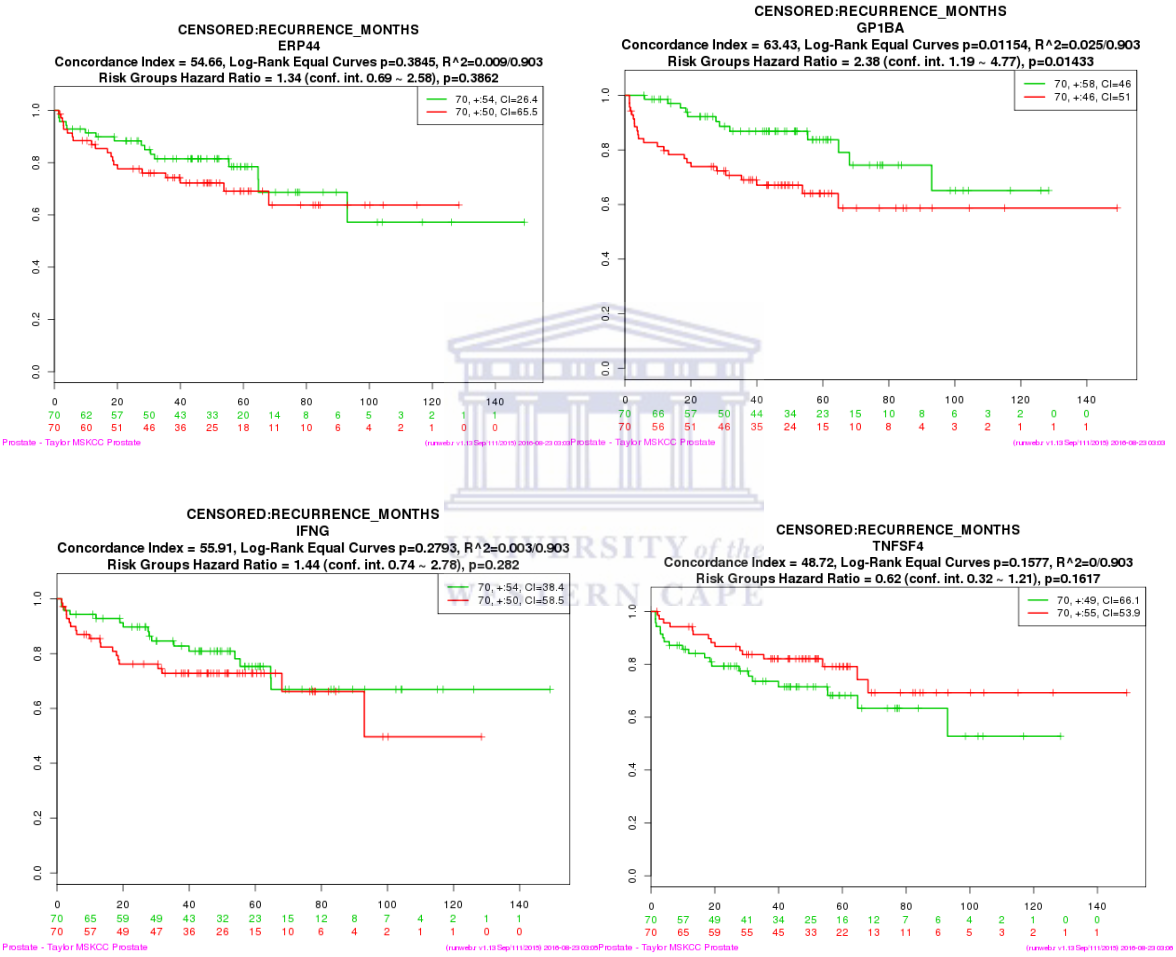
From the results seen, one gene out of the six on the dataset PRAD-TCGA Prostate adenocarcinoma showed promise as a good distinguishable prognostic markers based on the p-value ($p < 0.05$), GP1BA was significantly expressed with a p-value= 0.0350104.

This result shows that in different datasets some genes are significant whereas that same genes are not significant in another dataset and this reason can be explained in a study that showed that the performance of biomarkers may differ in different populations, based on clinical information, probes per gene, the gene expression technology and the conditions used (Aguirre-Gamboa *et al*., 2013).

4.8.2. Analysis of the six biomarkers individually using SurvExpress

As individual genes, the six candidate genes in SurvExpress generated little or no prognostic value, compared to the genes used as a panel; with certain genes showing no data within the dataset (see Figures 4.2, 4.3 and 4.4).

4.8.2.1. Dataset: Taylor MSKCC Prostate



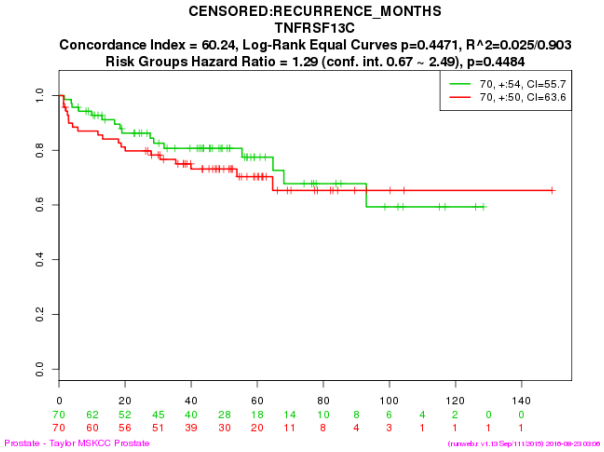


Figure 4.2: Kaplan-Meier curves and performance of the individual candidate biomarkers in dataset Taylor prostate cancer data by SurvExpress

In this dataset, there is a good prognostic value for recurrence of disease in the beginning stages of the study for gene GP1BA based on the significant p-value of 0.01154. ($p \leq 0.05$ is statistically significant). The rest of the genes found in this dataset, based on recurrence of disease, showed no prognostic value.

4.8.2.2. Dataset: Sboner Rubin GSE16560

GP1BA TNFSF4

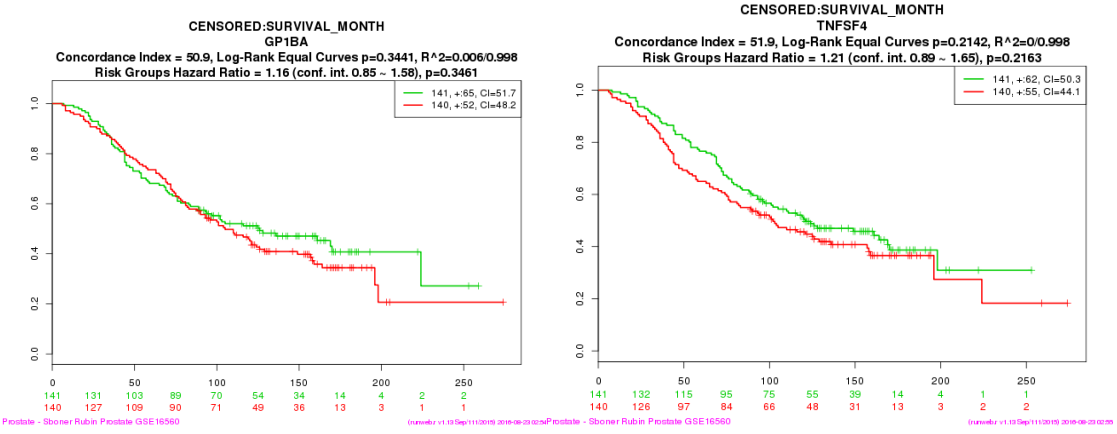


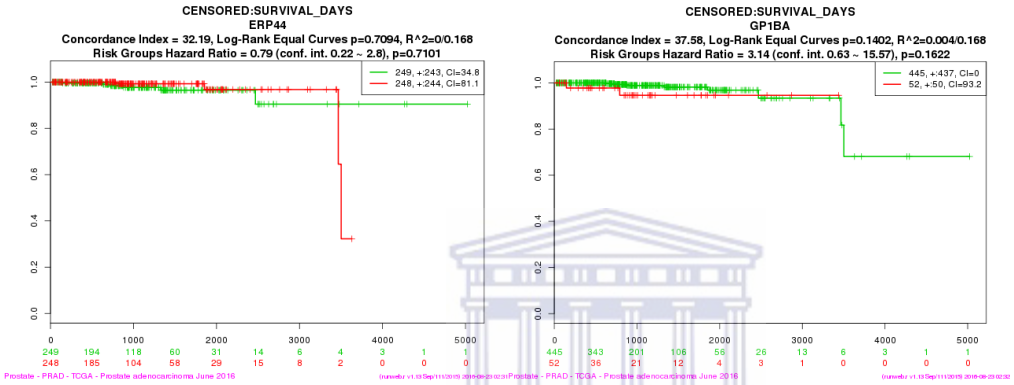
Figure 4.3: Kaplan-Meier curves and performance of the individual candidate biomarkers in dataset Sboner Rubin prostate cancer by SurvExpress

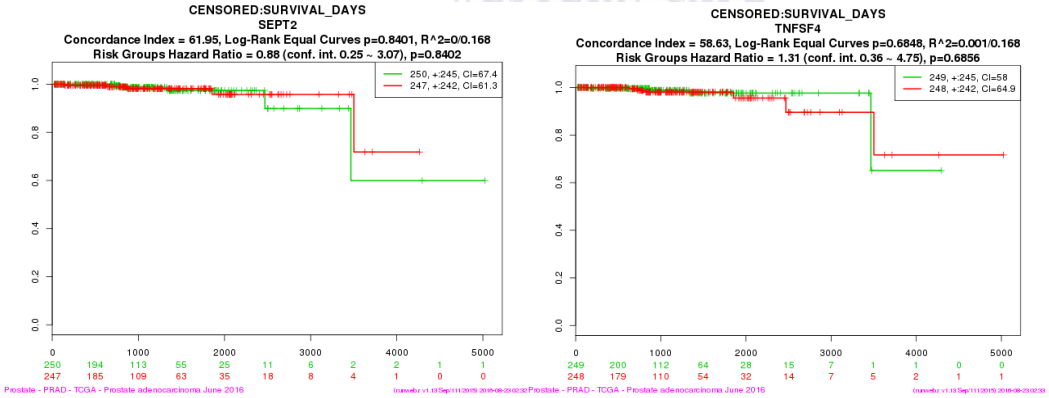
Displaying only the two genes found in this dataset, both genes again resulted in no significant prognostic value, to the end of the study in patients expressing these genes.

4.8.2.3. Dataset: Kollmeyer-Jenkins Prostate GSE10645-GPL5858

For gene, IFNG in this dataset (only 1 gene found) – the output showed the same results as the panel of genes as seen in figure 4.1B.

4.8.2.4. Dataset: PRAD - TCGA - Prostate adenocarcinoma





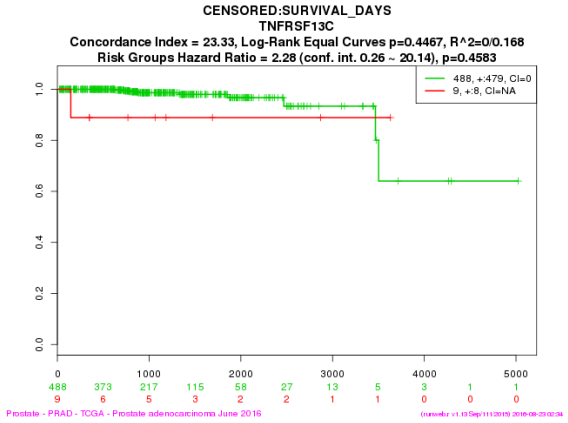


Figure 4.4: Kaplan-Meier curves and performance of the individual candidate biomarkers in dataset PRAD - TCGA - Prostate adenocarcinoma June 2016 by SurvExpress

From the survival curves of this dataset (Figure 4.2), the 5 genes found did not show significant value in predicting the disease outcome in relation to prostate adenocarcinoma.

4.8.3. Determining the prognostic value the of candidate biomarkers within the PROGgene database

To further determine the prognostic value of the individual biomarker candidates in overall survival of prostate cancer, the PROGgene database was used. Using this software the outcome was individual gene results with no gene panels looked at for the six candidate genes. The same dataset GSE16560 as in SurvExpress was used in PROGgene, with only slight variations in the survival curves (comparison of figure 4.3) but the outcome being the same. During this study of dataset GSE16560, all patients with the expression of these genes, whether low or highly expressed, resulted in no prognostic value for prostate cancer.

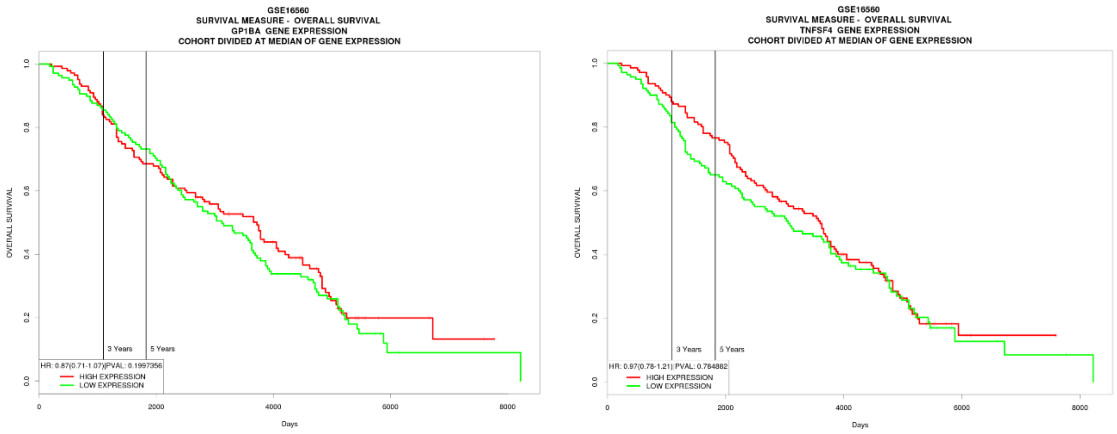


Figure 4.5: Kaplan-Meier plot for genes GP1BA and TNFSF4 in GSE16560 dataset by PROGgene

4.8.4. Expression analysis and prognostic value of candidate genes using PrognoScan

The prognostic value of the six candidate biomarkers was also determined using the PrognoScan database and the results represented with expression plots, expression histograms, P-value plots and Kaplan-Meier plots (survival curves). Again, not all genes were found in this database, with the following three genes: SEPT2, GP1BA (see appendix D) and TNFSF4 (See appendix E) and only one gene, SEPT2 (figure 4.6), showed possible prognostic value. The probability of survival with this gene highly expressed is significant in the early stages of the disease, showing possibility as a good prognostic biomarker at that stage. There is differential expression between the low and high expression during this stage for SEPT2. See table 4.3 displaying a significant p-value for this gene.

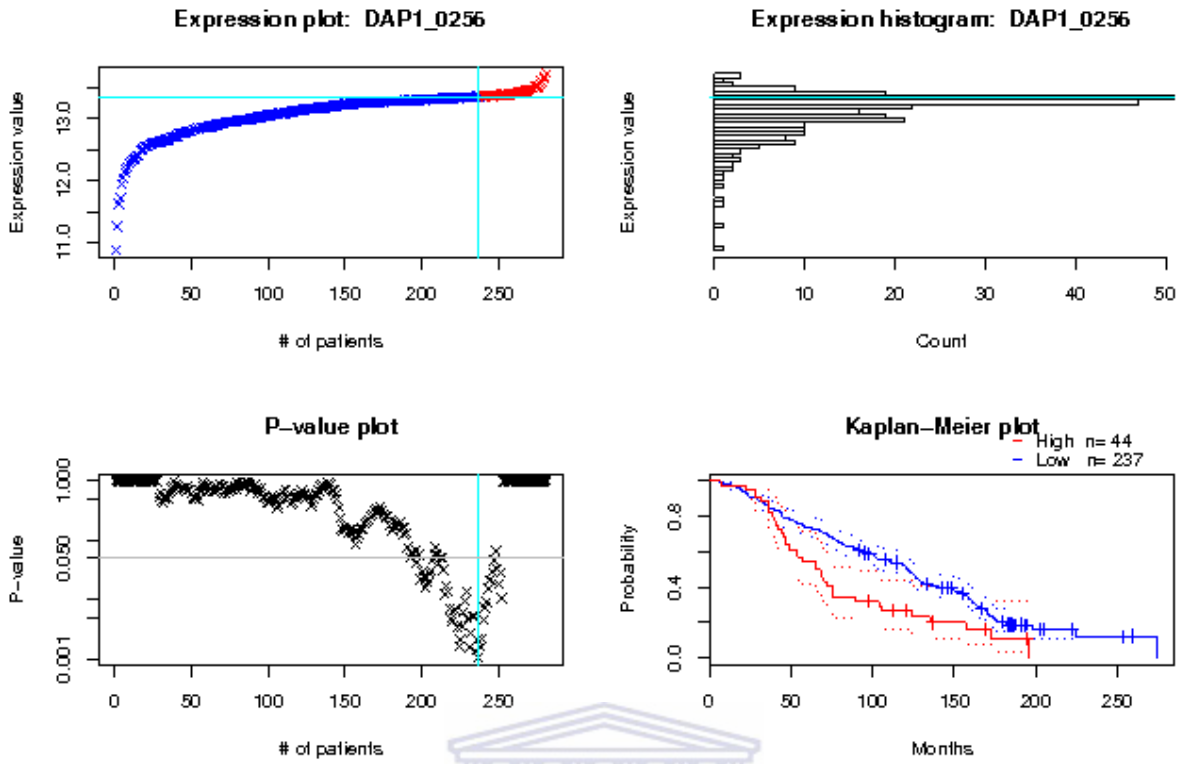


Figure 4.6: Expression plot, Expression histogram, P-value plot and Kaplan-Meier plots for high and low SEPT2 -expressing groups in prostate cancer

Table 4.3: PrognoScan SEPT2 gene result-Statistically significant gene with corrected p-value= 0.026191

DATA POSTPROCESSING	None
PROBE_NAME	DAP1_0256 [6K DASL]
PROBE_DESCRIPTION	septin 2
GENE_SYMBOL	SEPT2
GENE_DESCRIPTION	septin 2
DATASET	GSE16560
CANCER_TYPE	Prostate cancer
SUBTYPE	
N	281
ENDPOINT	Overall Survival
PERIOD	Months
COHORT	Sweden (1977-1999)
ARRAY TYPE	6K DASL
CONTRIBUTOR	Sboner
SAMPLE PREPARATION	DASL
CUTPOINT	0.84
MINIMUM *P*-VALUE	0.001033
CORRECTED *P*-VALUE	0.026191
ln(HR_{high} / HR_{low})	0.59
COX *P*-VALUE	0.254584
ln(HR)	0.22
HR [95% CI]	1.25 [0.85 - 1.83]

4.9. Conclusion

Despite the introduction of PSA screening, the mortality from prostate cancer has remained relatively high. Although the benefits of PSA screening are widely debated, this serum marker remains one of only a few preoperative parameters of prognostic utility (Henshall *et al*., 2003). *In silico* biomarker validation could be a substantially more cost-effective strategy for biomarker development, which typically requires costly and lengthy processes. Survival analysis tools and resources, clinically deployed genome-based biomarkers are still scarce, highlighting the unresolved challenges in biomarker development from genomic studies (Chen *et al*., 2014).

Novel and clinical markers for prostate cancer diagnosis, prognosis, and prediction is essential to the optimal identification and treatment of this disease and to bring potential biomarkers from the laboratory environment into clinical use at the patient bedside requires a comprehensive pursuit and rigorous analysis (Tricoli *et al*., 2004).

The prognostic gene signatures related to patient outcome such as survival time and tumour stage must be genes that are important in tumour development and progression (Li *et al*., 2015).

Larger patient cohorts are needed for prostate cancer, as compared to other cancers such as breast cancer, for data outcomes not to be ambiguous (Sutcliffe*et al*., 2009). This was evident in the comparison of the PROGgene and SurvExpress results of GSE16560 dataset. A limitation is that the cohort was not big enough therefore the genes were difficult to assess for prognosis for prostate cancer patients in general.

Although, the results of the SurvExpress analysis revealed gene expression differences that were not significantly sufficient to be distinguished as strong prognostic biomarkers, one marker, GP1BA did stand out, supporting its prognostic value based on the statistical p-value.

Another gene, from the Prognoscan database SEPT2 shows promise in that has some prognostic value in the early stages of the disease.

This study provides some promising evidence that bioinformatics data mining can be a highly beneficial means to identify novel biomarkers, although combined with clinical biomarker validation by qRT-PCR, using a molecular approach and functional evaluation of candidate genes, it can be considered for a detailed follow-up study on selected candidates in the near future.

4.10. References

ADEOLA, H. A., CALDER, B., SOARES, N. C., KAESTNER, L., BLACKBURN, J. M. & ZERBINI, L. F. 2016. In silico verification and parallel reaction monitoring prevalidation of potential prostate cancer biomarkers. *Future Oncol,* 12**,** 43-57.

AGUIRRE-GAMBOA, R., GOMEZ-RUEDA, H., MARTINEZ-LEDESMA, E., MARTINEZ-TORTEYA, A., CHACOLLA-HUARINGA, R., RODRIGUEZ-BARRIENTOS, A., TAMEZ-PENA, J. G. & TREVINO, V. 2013. SurvExpress: An Online Biomarker Validation Tool and Database for Cancer Gene Expression Data Using Survival Analysis. *Plos One,* 8.

ALINEZHAD, S., VAANANEN, R. M., MATTSSON, J., LI, Y., TALLGREN, T., TONG OCHOA, N., BJARTELL, A., AKERFELT, M., TAIMEN, P., BOSTROM, P. J., PETTERSSON, K. & NEES, M. 2016. Validation of Novel Biomarkers for Prostate Cancer Progression by the Combination of Bioinformatics, Clinical and Functional Studies. *PLoS One,* 11**,** e0155901.

CHEN, X., SUN, X. & HOSHIDA, Y. 2014. Survival analysis tools in genomics research. *Hum Genomics,* 8**,** 21.

CLARK, T. G., BRADBURN, M. J., LOVE, S. B. & ALTMAN, D. G. 2003. Survival analysis part I: Basic concepts and first analyses. *British Journal of Cancer,* 89**,** 232-238.

GLINSKY, G. V., GLINSKII, A. B., STEPHENSON, A. J., HOFFMAN, R. M. & GERALD, W. L. 2004. Gene expression profiling predicts clinical outcome of prostate cancer. *J Clin Invest,* 113**,** 913-23.

GOEL, M. K., KHANNA, P. & KISHORE, J. 2010. Understanding survival analysis: Kaplan-Meier estimate. *Int J Ayurveda Res,* 1**,** 274-8.

GOSWAMI, C. P. & NAKSHATRI, H. 2013. PROGgene: gene expression based survival analysis web application for multiple cancers. *J Clin Bioinforma,* 3**,** 22.

GOSWAMI, C. P. & NAKSHATRI, H. 2014. PROGgeneV2: enhancements on the existing database. *BMC Cancer,* 14**,** 970.

HENSHALL, S. M., AFAR, D. E., HILLER, J., HORVATH, L. G., QUINN, D. I., RASIAH, K. K., GISH, K., WILLHITE, D., KENCH, J. G., GARDINER-GARDEN, M., STRICKER, P. D., SCHER, H. I., GRYGIEL, J. J., AGUS, D. B., MACK, D. H. & SUTHERLAND, R. L. 2003. Survival analysis of genome-wide gene expression profiles of prostate cancers identifies new prognostic targets of disease relapse. *Cancer Res,* 63**,** 4196-203.

KAPLAN, E. L. & MEIER, P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American statistical association,* 53**,** 457-481.

LI, W., YU, J., LIAN, B., SUN, H., LI, J., ZHANG, M., LI, L., LI, Y., LIU, Q. & XIE, L. 2015. Identifying prognostic features by bottom-up approach and correlating to drug repositioning. *PLoS One,* 10**,** e0118672.

MIZUNO, H., KITADA, K., NAKAI, K. & SARAI, A. 2009. PrognoScan: a new database for meta-analysis of the prognostic value of genes. *BMC Med Genomics,* 2**,** 18.

RICH, J. T., NEELY, J. G., PANIELLO, R. C., VOELKER, C. C., NUSSENBAUM, B. & WANG, E. W. 2010. A practical guide to understanding Kaplan-Meier curves. *Otolaryngol Head Neck Surg,* 143**,** 331-6.

PROGGENE. Available: www.compbio.iupui.edu/proggene.

PROGNOSCAN. Available: http://gibk21.bse.kyutech.ac.jp/PrognoScan/index.html

RILEY, R. D., ABRAMS, K. R., SUTTON, A. J., LAMBERT, P. C., JONES, D. R., HENEY, D. & BURCHILL, S. A. 2003. Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. *Br J Cancer,* 88**,** 1191-8.

SARDANA, G., DOWELL, B. & DIAMANDIS, E. P. 2008. Emerging biomarkers for the diagnosis and prognosis of prostate cancer. *Clin Chem,* 54**,** 1951-60.

SBONER, A., DEMICHELIS, F., CALZA, S., PAWITAN, Y., SETLUR, S. R., HOSHIDA, Y., PERNER, S., ADAMI, H. O., FALL, K., MUCCI, L. A., KANTOFF, P. W., STAMPFER, M., ANDERSSON, S. O., VARENHORST, E., JOHANSSON, J. E., GERSTEIN, M. B., GOLUB, T. R., RUBIN, M. A. & ANDREN, O. 2010. Molecular sampling of prostate cancer: a dilemma for predicting disease progression. *BMC Med Genomics,* 3**,** 8.

SHARIAT, S. F., KARAKIEWICZ, P. I., ROEHRBORN, C. G. & KATTAN, M. W. 2008. An updated catalog of prostate cancer predictive tools. *Cancer,* 113**,** 3075-99.

SINGH, D., FEBBO, P. G., ROSS, K., JACKSON, D. G., MANOLA, J., LADD, C., TAMAYO, P., RENSHAW, A. A., D'AMICO, A. V., RICHIE, J. P., LANDER, E. S., LODA, M., KANTOFF, P. W., GOLUB, T. R. & SELLERS, W. R. 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell,* 1**,** 203-9.

SIMON, R. 2009. Advances in clinical trial designs for predictive biomarker discovery and validation. *Current Breast Cancer Reports,* 1**,** 216-221.

SURVEXPRESS. Available: http://bioinformatica.mty.itesm.mx/SurvExpress.

SUTCLIFFE, P., HUMMEL, S., SIMPSON, E., YOUNG, T., REES, A., WILKINSON, A., HAMDY, F., CLARKE, N. & STAFFURTH, J. 2009. Use of classical and novel biomarkers as prognostic risk factors for localised prostate cancer: a systematic review. *Health Technol Assess,* 13**,** iii, xi-xiii 1-219.

TRICOLI, J. V., SCHOENFELDT, M. & CONLEY, B. A. 2004. Detection of prostate cancer and predicting progression: current and future diagnostic markers. *Clin Cancer Res,* 10**,** 3943-53.

Chapter 5

General Discussion and Summary

5.1 General discussion

The measurement of serum PSA is currently the most useful biomarker to aid in earlier detection of prostate cancer and is performed on millions of men worldwide. Urgently needed in the study of disease is the development of biomarkers that can detect curable disease earlier and not detect advanced disease better (Petricoin *et al*., 2004).

miRNAs are important regulators of biologic processes in prostate cancer progression and in recent years, much effort has been invested in improving patient care by substituting procedures such as DRE or prostate biopsy with miRNA analysis in patient serum or plasma. In prostate cancer, diagnosis and follow-up monitoring after therapies are some of the major challenges for its clinical management. Patients undergo repetitive biopsies, being invasive and not decisive, even if coupled with PSA and digital rectal examination (DRE) (Fabris *et al*., 2016). Although several markers have reasonable operating characteristics, no individual marker is ideal and therefore it is possible that a combination of biomarkers may provide better predictive value (Makarov *et al*., 2009).

There are several molecular events that are believed to occur in a large percentage of prostate carcinomas such as down-regulation of the NKX3.1 homeobox gene represents a frequent and critical event in prostate cancer initiation and is likely to involve multiple mechanisms. Studies have suggested a role for MYC overexpression in prostate cancer initiation as nuclear MYC protein is up-regulated in many PIN lesions and the majority of carcinomas (Shen and Abate-Shen, 2010). PTEN was originally identified as a tumour suppressor that is frequently mutated or deleted in prostate cancer. miRNAs have specific roles in regulation of critical target genes, as the cluster miR-106b-25 negatively regulates PTEN expression. Suitable

combinations of markers may be successful in cumulatively predicting outcomes, as enabled by new technologies such as molecular systems pathology (Shen and Abate-Shen, 2010). Throughout prostate development and maturation, genes such as nkx3.1, FOXA1 and AR function to mediate gland formation and cellular differentiation. Activation of developmental genes within differentiated prostate epithelium can promote hyperplastic proliferation and/or progression of prostate cancer. Many of detected mutations still require validation to determine if there is any functional consequence and it represents the current paradigm shift from single molecule research to genome-wide analysis (Schrecengost and Knudsen, 2013).

Despite extensive research efforts, very few biomarkers of prostate cancer have been successfully implemented into clinical practice today. Biomarkers for prostate cancer should be addressed to distinguish BPH from prostate cancer, to detect the aggressive forms from the indolent cases, and to identify metastatic cancer predictors (Tefekli and Tunc, 2013).

5.2. Summary

Chapter 2

In chapter 2, the 13 miRNAs that were identified as potential prostate cancer miRNA targets and their 21 target genes were generated and shown to be related to prostate cancer (Khan, PhD thesis, 2015) were expanded on by identifying their co-expressed genes using the STRING database and a total of 300 additional genes/proteins were identified and extracted. The gene list was then refined and duplicates were eliminated, resulting in a total of 231 genes from the STRING database as possible potential biomarkers for early diagnosis of prostate cancer. Cell surface gene products were identified using the GO database with the 231 genes identified in STRING generating a number of mapped identifications of 26 genes in relation to their expression on the cell surface with a P –value of <0.05. A confirmation of the generated gene list by functional annotation clustering in DAVID database resulted in 20

clusters that were queried using the GO terms: cell surface, regulation of cell death, regulation of cell proliferation, apoptosis and cell cycle process and in turn also resulted in the same 26 genes as GO, which then was used in the TiGER and GeneHub-GEPIS databases known for tissue specific enriched genes and expression analysis. These databases were used to eliminate genes already experimentally linked to prostate cancer.

Literature mining: PubMed and Google Scholar were used, in order to obtain genes of greater relevance to be validated as novel biomarkers for prostate cancer. The final candidate gene list was narrowed down to the following 6 genes: ERP44, GP1BA, IFNG, SEPT2, TNFRSF13C and TNFSF4 for prostate cancer and would be easily detectable in bodily fluids since the GO analysis of these genes is enriched for cell surface expression.

Chapter 3

The 6 genes identified *in silico* (Chapter 2) were associated with TFs and pathways in Chapter 3. The TFs were identified using GeneCards and were annotated to confirm regulatory control in cancer processes and specifically prostate cancer. 44 TFs were found to be associated to the candidate genes and 15 of these TFs showed no experimental information about the genes they regulate and any direct links to prostate cancer. Identification of Pathways in the KEGG database were linked to the TFs that target the promoter regions of the candidate genes involved in prostate cancer and resulted in various TFs with some links to cancer however only 4 (NKX3.1, p53, E2F and CREB), showed association to the Prostate Cancer pathway. CREB, E2F, Nkx3-1 and p53 TFs were discovered to be linked to the genes IFNG, GP1BA, SEPT2 and TNFRSF13C respectively therefore showing that these genes can be assessed as potential biomarkers for prostate cancer.

Chapter 4

In this chapter, the six identified candidate genes from Chapter 2 were evaluated as prognostic and/or predictive biomarkers using SurvExpress as a panel as well as individually. The genes were analysed using four different datasets. One gene out of the six, GP1BA was significantly expressed, on the dataset PRAD-TCGA Prostate adenocarcinoma and showed promise as a good distinguishable prognostic marker based on the p-value ($p < 0.05$). As individual genes, the six candidate genes in SurvExpress generated little or no prognostic value, compared to the genes used as a panel; with certain genes showing no data within the dataset. The prognostic value of the candidate biomarkers using PROGgene database allowed samples to be divided into High and Low gene expression groups resulted in no prognostic value for prostate cancer. Expression analysis and prognostic value of candidate genes using PrognoScan showed one gene, SEPT2 with differential expression significant at early stages of the disease being possibility a good prognostic biomarker.

5.3. Future Work for this study

This study provides some promising evidence that bioinformatics data mining can be a highly beneficial means to identify novel biomarkers, although combined with clinical biomarker validation by qRT-PCR, using a molecular approach and functional evaluation of candidate genes, it can be considered for a detailed follow-up study on selected candidates in the near future to evaluate their roles in disease initiation and progression of prostate cancer using cell lines as well as patient samples.

Large scale protein interactions and miRNA target prediction data were used in this study and future directions for this work could use additional protein interaction networks, different miRNA target prediction algorithms and different expression data sets to reveal more miRNA regulated genes in prostate cancer as diagnostic biomarkers.

The 6 genes of interest identified (Chapter 2) would require validation at a molecular level. Various types of assays can be used in the biomarker method validation process and range from the relatively low technology end such as immunohistochemistry (IHC) to immunoassays to the high technology end including platforms for genomics, proteomics and multiplex ligand-binding assays. A genomics approach such as gene expression analysis microarrays has become the standard technology used for target identification and validation. Immunoassays are routinely used for protein biomarker assessments due to its straightforward clinical application and translation into a potential diagnostic assay. The analytical platforms used are based on nuclear magnetic resonance spectroscopy and the combination of liquid chromatography with mass spectroscopy. Molecular and functional imaging technologies are used to assess cell proliferation and apoptosis, cellular metabolism, angiogenesis and vascular dynamics. The right choice of assay is an important first step to successful biomarker method validation (Chau *et al*., 2008).

5.4. Future perspectives

Future research should focus on validation of already existing biomarkers and the discovery of new markers to identify men with aggressive prostate cancer, PSA is not specific for prostate cancer and serum levels are elevated in common benign diseases (Makarov *et al*., 2009).

The future biomarker studies for prostate cancer should focus on biomarker candidates that address the current gaps in biomarker development, including prognostic and predictive biomarkers. One approach to identifying predictive biomarkers is to focus on genomic disease signatures, which influence the biological characteristics of an individual cancer (Prensner *et al.*, 2012).

Challenges for future studies will be to integrate epidemiological studies with molecular investigations and clinical analyses to gain fundamental insights into how environmental, dietary, and lifestyle influences contribute to the development of prostate cancer, much work remains to be done to enhance the overall rate of prostate cancer survival.

High-throughput technology methods and advances in molecular biology are helping and accelerating the exploration to useful biomarkers. Future studies are needed to reconfirm the features of the existing biomarkers and further discover novel potential ones to better predict the presence of the disease (Qu *et al*., 2014).

With the emergence of novel high throughput omics-based technologies, there is a need for better *in silico* computational and bioinformatics tools to improve clinical inferences drawn from huge databases (Adeola *et al*., 2016). The ideal biomarker should be economical, consistent, non-invasive, easily accessible, and quickly quantifiable. A drawback of the PSA test has resulted in a concerted effort to develop replacement-screening tools for prostate cancer. It will be a challenge to replace PSA entirely due to its minimally invasive nature and low cost but there is a pressing need to complement PSA with biomarkers that can increase the specificity and sensitivity of a screen. A panel of diagnostic and prognostic biomarkers that will work in conjunction with PSA will be ideal (Velonas *et al*., 2013).

Also, prognostic markers are particularly important at the time of initial diagnosis of prostate cancer since it varies widely in patient outcome. No prognostic marker can accurately predict outcome for an individual patient, it provides a probability estimate of outcome for a heterogeneous population of patients. Prognostic markers may be crucial to reduce overtreatment of patients with indolent malignancy and so minimizing the side effects of adjuvant systemic therapies, and to avoid under treatment of patients with aggressive and life-threatening malignancy, which would be recommended in receiving the most appropriate

local and systemic therapy. Tumour biomarkers might be useful for risk assessment, screening for early cancer detection, diagnosis, prognosis, selection and monitoring of anticancer therapy (Mordente *et al*., 2015).

5.5. References

ADEOLA, H. A., CALDER, B., SOARES, N. C., KAESTNER, L., BLACKBURN, J. M. & ZERBINI, L. F. 2016. In silico verification and parallel reaction monitoring prevalidation of potential prostate cancer biomarkers. *Future Oncol,* 12**,** 43-57.

CHAU, C. H., RIXE, O., MCLEOD, H. & FIGG, W. D. 2008. Validation of analytical methods for biomarkers employed in drug development. *Clinical cancer research: an official journal of the American Association for Cancer Research,* 14**,** 5967.

FABRIS, L., CEDER, Y., CHINNAIYAN, A. M., JENSTER, G. W., SORENSEN, K. D., TOMLINS, S., VISAKORPI, T. & CALIN, G. A. 2016. The Potential of MicroRNAs as Prostate Cancer Biomarkers. *Eur Urol,* 70**,** 312-22.

MAKAROV, D. V., LOEB, S., GETZENBERG, R. H. & PARTIN, A. W. 2009. Biomarkers for prostate cancer. *Annu Rev Med,* 60**,** 139-51.

MORDENTE, A., MEUCCI, E., MARTORANA, G. E. & SILVESTRINI, A. 2015. Cancer Biomarkers Discovery and Validation: State of the Art, Problems and Future Perspectives. *Adv Exp Med Biol,* 867**,** 9-26.

PETRICOIN, E. F., ORNSTEIN, D. K. & LIOTTA, L. A. 2004. Clinical proteomics: Applications for prostate cancer biomarker discovery and detection. *Urol Oncol,* 22**,** 322-8.

QU, M., REN, S. C. & SUN, Y. H. 2014. Current early diagnostic biomarkers of prostate cancer. *Asian J Androl,* 16**,** 549-54.

SCHRECENGOST, R. & KNUDSEN, K. E. 2013. Molecular pathogenesis and progression of prostate cancer. *Semin Oncol,* 40**,** 244-58.

SHARMA, P., ZARGAR-SHOSHTARI, K. & POW-SANG, J. M. 2016. Biomarkers for prostate cancer: present challenges and future opportunities. *Future Science OA,* 2.

SHEN, M. M. & ABATE-SHEN, C. 2010. Molecular genetics of prostate cancer: new prospects for old challenges. *Genes Dev,* 24**,** 1967-2000.

TEFEKLI, A. & TUNC, M. 2013. Future prospects in the diagnosis and management of localized prostate cancer. *ScientificWorldJournal,* 2013**,** 347263.

VELONAS, V. M., WOO, H. H., DOS REMEDIOS, C. G. & ASSINDER, S. J. 2013. Current status of biomarkers for prostate cancer. *Int J Mol Sci,* 14**,** 11034-60.

APPENDICES

Appendix A: Genes found within the Prostate cancer Pathway (Extracted from the KEGG Pathway database)

CDKN1B; cyclin dependent kinase inhibitor 1B
CDK2; cyclin dependent kinase 2
CCNE1; cyclin E1
CCNE2; cyclin E2
RB1; RB transcriptional corepressor 1
E2F1; E2F transcription factor 1
E2F2; E2F transcription factor 2
E2F3; E2F transcription factor 3
INS; insulin
PDGFA; platelet derived growth factor subunit A
PDGFB; platelet derived growth factor subunit B
PDGFC; platelet derived growth factor C
PDGFD; platelet derived growth factor D
EGF; epidermal growth factor
TGFA; transforming growth factor alpha
IGF1; insulin like growth factor 1
INSRR; insulin receptor related receptor
PDGFRA; platelet derived growth factor receptor alpha
PDGFRB; platelet derived growth factor receptor beta

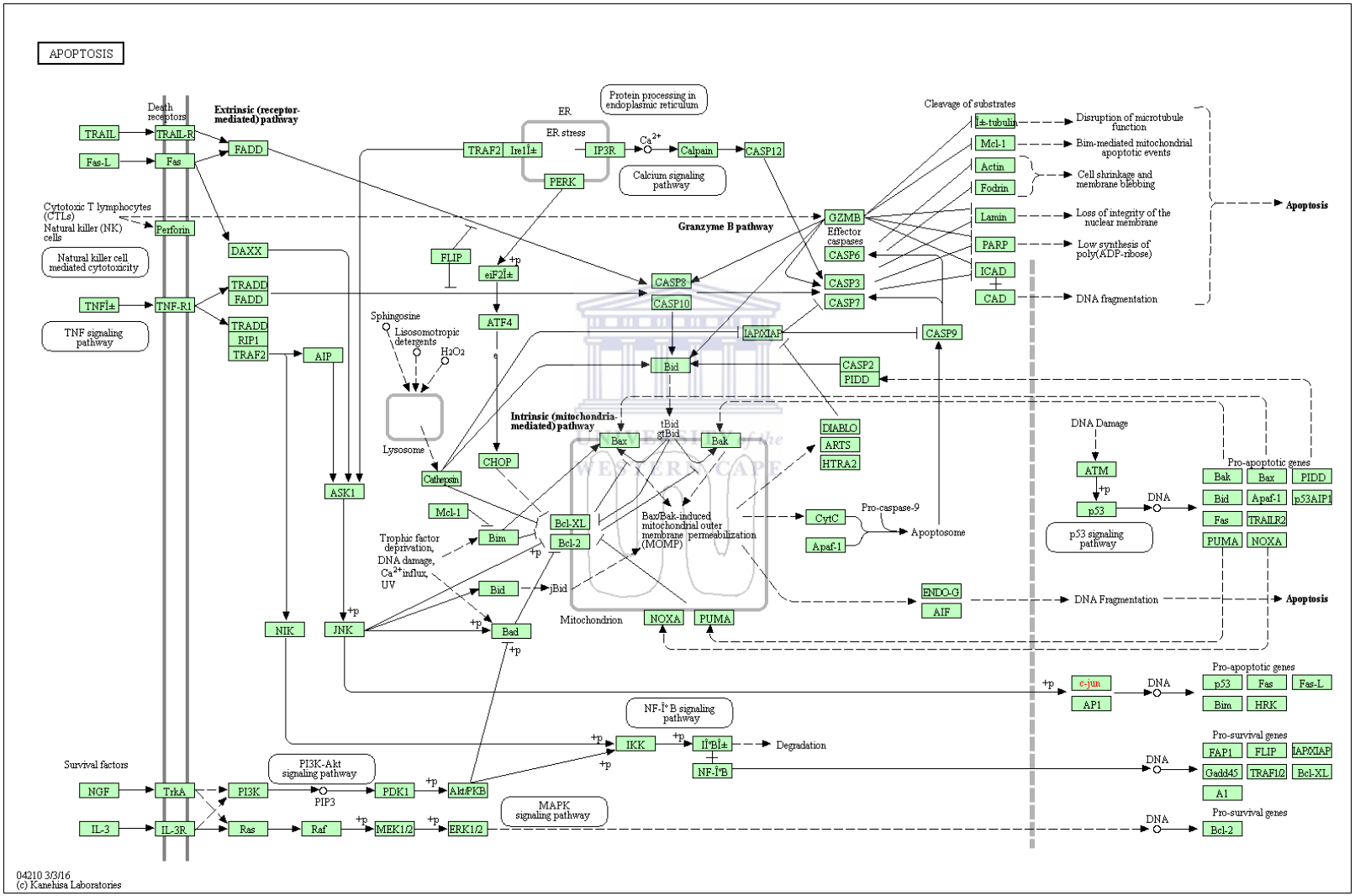
FGFR1; fibroblast growth factor receptor 1
FGFR2; fibroblast growth factor receptor 2
EGFR; epidermal growth factor receptor
ERBB2; erb-b2 receptor tyrosine kinase 2
IGF1R; insulin like growth factor 1 receptor
PIK3CA; phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
PIK3CD; phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit delta
PIK3CB; phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit beta
PIK3CG; phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit gamma
PIK3R1; phosphoinositide-3-kinase regulatory subunit 1
PIK3R5; phosphoinositide-3-kinase regulatory subunit 5
PIK3R2; phosphoinositide-3-kinase regulatory subunit 2
PIK3R3; phosphoinositide-3-kinase regulatory subunit 3
PTEN; phosphatase and tensin homolog
PDPK1; 3-phosphoinositide dependent protein kinase 1
NKX3-1; NK3 homeobox 1
AKT1; AKT serine/threonine kinase 1
AKT2; AKT serine/threonine kinase 2
AKT3; AKT serine/threonine kinase 3
CASP9; caspase 9
BAD; BCL2 associated agonist of cell death
FOXO1; forkhead box O1

CDKN1A; cyclin dependent kinase inhibitor 1A
MDM2; MDM2 proto-oncogene
TP53; tumor protein p53
GSK3B; glycogen synthase kinase 3 beta
CREB1; cAMP responsive element binding protein 1
ATF4; activating transcription factor 4 33
CREB3; cAMP responsive element binding protein 3
CREB3L1; cAMP responsive element binding protein 3 like 1
CREB3L2; cAMP responsive element binding protein 3 like 2
CREB3L3; cAMP responsive element binding protein 3 like 3
CREB3L4; cAMP responsive element binding protein 3 like 4
CREB5; cAMP responsive element binding protein 5
CTNNB1; catenin beta 1
CREBBP; CREB binding protein
EP300; E1A binding protein p300
TCF7; transcription factor 7 (T-cell specific, HMG-box)
TCF7L1; transcription factor 7 like 1
TCF7L2; transcription factor 7 like 2
LEF1; lymphoid enhancer binding factor 1
CCND1; cyclin D1
CHUK; conserved helix-loop-helix ubiquitous kinase
IKBKB; inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase beta

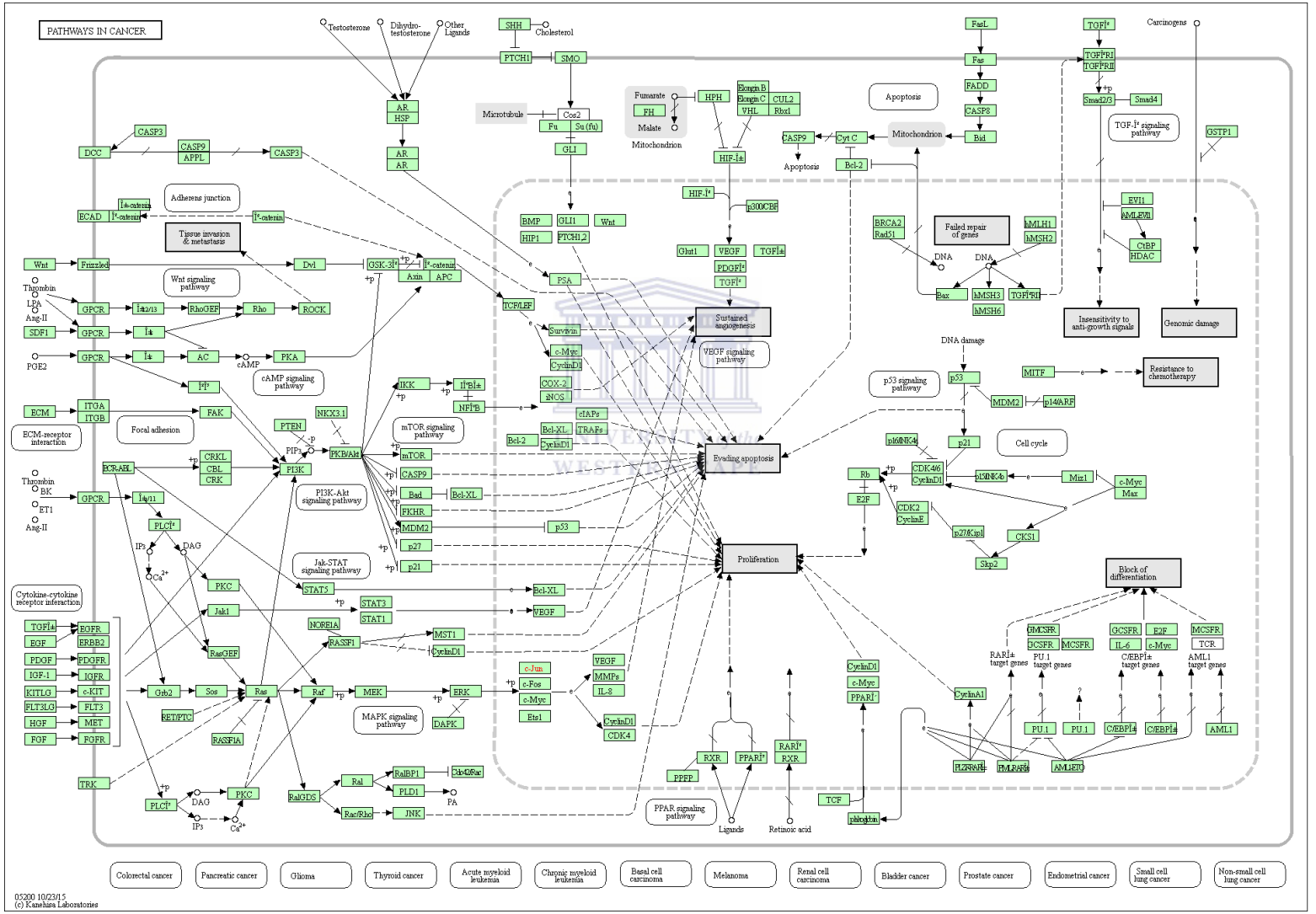
IKBKG; inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma
NFKBIA; NFKB inhibitor alpha
NFKB1; nuclear factor kappa B subunit 1
RELA; RELA proto-oncogene, NF-kB subunit
BCL2; BCL2, apoptosis regulator
MTOR; mechanistic target of rapamycin
GRB2; growth factor receptor bound protein 2
SOS1; SOS Ras/Rac guanine nucleotide exchange factor 1
SOS2; SOS Ras/Rho guanine nucleotide exchange factor 2
HRAS; HRas proto-oncogene, GTPase
KRAS; KRAS proto-oncogene, GTPase
NRAS; neuroblastoma RAS viral oncogene homolog
ARAF; A-Raf proto-oncogene, serine/threonine kinase
BRAF; B-Raf proto-oncogene, serine/threonine kinase
RAF1; Raf-1 proto-oncogene, serine/threonine kinase
MAP2K1; mitogen-activated protein kinase kinase 1
MAP2K2; mitogen-activated protein kinase kinase 2
MAPK1; mitogen-activated protein kinase 1
MAPK3; mitogen-activated protein kinase 3
SRD5A2; steroid 5 alpha-reductase 2
AR; androgen receptor
HSP90AA1; heat shock protein 90 alpha family class A member 1

HSP90AB1; heat shock protein 90 alpha family class B member 1
HSP90B1; heat shock protein 90 beta family member 1
KLK3; kallikrein related peptidase 3
GSTP1; glutathione S-transferase pi 1

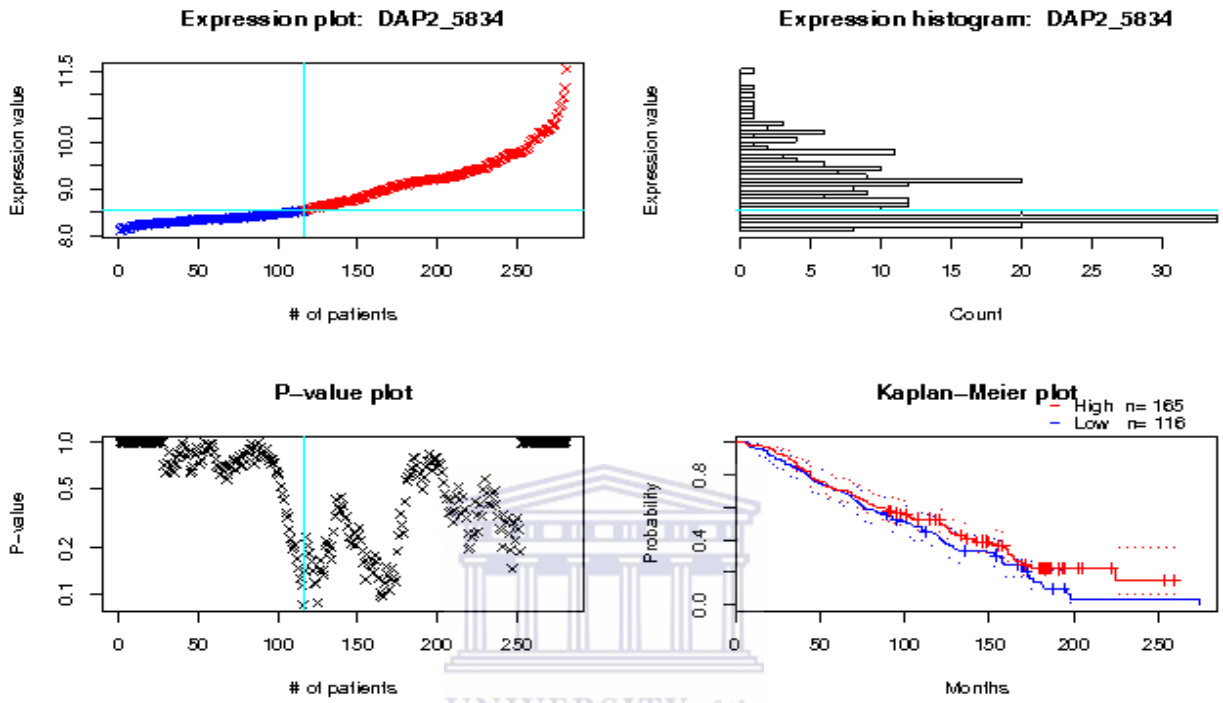
Appendix B: Apoptosis Pathway (Extracted from the KEGG Pathway database)



Appendix C: Pathways in Cancer (Extracted from the KEGG Pathway database)



Appendix D: Expression plot, Expression histogram, P-value plot and Kaplan-Meier plots for high and low GP1BA-expressing groups in prostate cancer. Expression analysis and prognostic value of GP1BA gene using PrognoScan database



Appendix E: Expression plot, Expression histogram, P-value plot and Kaplan-Meier plots for high and low TNFSF4-expressing groups in prostate cancer. Expression analysis and prognostic value of TNFSF4 gene using PrognoScan database