Identification of novel miRNAs as diagnostic and prognostic biomarkers for prostate cancer using an *in silico* approach

UNIVERSITY *of the*
WESTERN CAPE

Nasr O M Eshibona

A thesis submitted in fulfilment of the requirements for the degree of Magister Scientiae, at the Department of Biotechnology, Faculty of Science

University of the Western Cape

Supervisor: Dr Ashley Pretorius

Co-supervisor: Dr Musa Gabere

December 2017

DECLARATION OF AUTHORSHIP

I declare that "Identification of novel miRNAs as diagnostic and prognostic for prostate cancer detection using *in silico* approach" is my own work, that it has not been submitted for any degree or examination in any other university and that all the sources I have used or quoted have been indicated and acknowledged by complete reference**.**

Signed:  **Date:** Jan of 2018

ACKNOWLEDGMENTS

First and foremost, I praise Allah, the Almighty for providing me this opportunity and granting me the capability to proceed successfully. You have given me the power to believe in myself and pursue my dreams and blessing in completing my studies. I could never have done this without the faith I have in you, the Almighty.

I would like to thank the valuable assistance received from my supervisor, **Dr Ashley Pretorius**, and co-supervisor Dr **Musa Gabere** for their supervision, knowledge; guidance, encouragement expertise, understanding, and patience. I acknowledge them for proofreading my thesis.

I would like to thank my family for their motivation, support, and encouragement.

I would like to thank Miss Chipampe Patricia Lombe and Dr. Marius Tincho for all the guidance. And to my friend Mr Riziki Martin

My sincerest gratitude goes to the DST/ Mintek Nanotechnology Innovation Centre (NIC) for the funding and financial aid. Grateful acknowledgment.

DEDICATION

To my family

ABSTRACT

Cancer is known as uncontrollable cell growth which results in the formation of tumours in the areas that are affected by the cancer. There are two types of tumours: benign and malignant. This study focus is on prostate cancer (PCa) as one of the most common cancers in men around the world. A previous study has reported that there were 27,132 new cases of cancer in South Africa in 2010. Out of those, 4652 were prostate cancer cases, which make it a considerable issue. The prostate is a gland that forms part of the male reproductive system. Prostate cancer is more apparent in men over the age of 65 years however it can be present in men of a lower age. However it is rare in men under 45 years of age. Prostate cancer start as a small group of cancer cells that can grow into a mature tumour. In the advanced stages, the tumour cells can spread to other tissue by metastases and can lead to death. Current diagnostic tools include Digital Rectal Examination (DRE), the Prostate-Specific Antigen test (PSA) ultra sound, and biopsy.

These methods are not completely accurate and it is insensitive to diagnose all PCa cases. Subsequently, the need for a more accurate and sensitive method for diagnosis of the early stage of the disease is needed. Biomarkers have recently been identified as a viable option for early detection of diseases, for example, biological indicators such as DNA, RNA, protein and microRNAs (miRNAs).

MicroRNAs are small regulatory RNA molecules that modulate the activity of specific messenger RNA (mRNA) targets and play important roles in a wide range of

physiologic and pathologic processes. Their main function is to regulate gene expression in a variety of manners, including translational repression, mRNA cleavage and deadenylation. MiRNAs are increasingly becoming recognized as powerful biomarkers for human diseases. The information potentially held by miRNAs, combined with the fact that they are stable in serum and plasma, has led to a rapidly growing interest in using miRNAs in blood or urine as diagnostic and prognostic biomarkers.

In this study, the main aim was to identify novel miRNAs that can potentially detect PCa at its earliest stage as well as serve as prognostic biomarkers.

Mature miRNAs that have been experimentally implicated in PCa were extracted from the EXIQON database with each miRNA's reference paper showing experimental evidence for their association with PCa. All mature miRNA sequences deposited into the miRBase database, were also collected as well. Duplications were removed using the CD-HIT programme which generated two datasets, first dataset included 150 miRNA sequences that are validated for their association to PCa and were used as the query dataset. The second dataset was generated from all miRNAs that were collected from miRBase which included 2477 miRNAs sequences which were used as a reference dataset.

Both datasets were used in the CD-HIT and BLAST algorithms for a similarity search between them and the result obtained from CD-HT showed 12 clusters which included 16 miRNAs. From BLAST, 258 similarities were detected with only the

common result between the two algorithms selected for farther analysis. Text mining was performed on the 14 common miRNAs using three databases namely PhenomiR, Mircancer, and PubMed. A list of 5 five miRNAs were identified as potential biomarker candidates as none of these miRNAs showed any known link to PCa.

In silico methods were used for target gene identification of these five miRNAs to validate the association of these miRNAs to PCa in the absence of experimental data for these miRNAs. Twelve genes were identified using TiGER, DAVID, and STRING databases. Pathway, prognostic, and expression analysis were performed on these 12 genes and 4 out of 12 showed involvement in known PCa pathways, 9 out of 12 showed good prognostic value for differentiating high and low risk PCa patients whilst 8 out of the 12 genes showed differential expression between PCa tissue and normal tissue as well when compared with other cancer tissues.

By performing *in silico* approaches in this study, a list of 5 miRNAs and their 12 target genes were identified, which potentially could be used as diagnostic and prognostic biomarkers for the management of PCa.

KEYWORDS: Prostate cancer, diagnostic biomarker, prognostic biomarker, expression profiling**,** *In silico***,** MiRNAs**,** Bioinformatics.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Abbreviation	Full Name
%	Percentage
>	Greater than
3'	3 prime end
5'	5 prime end
ADT	Androgen Deprivation Therapy
AFZ	Anterior fibromuscular zone
Ago	Argonaute
AMACR	Alpha-Methylacyl-CoA Racemes
AR	Androgen Receptor
BP	Base Pair
BRCA1	Breast cancer 1
BRCA2	Breast cancer 2
CA 15-3	Cancer antigen 15-3
CRM	Cis-Regulatory Module
CT	Computed Tomography
CTNNB1	Catenin Beta 1
CYP24A1	Cytochrome P450 Family 24 Subfamily A Member1
CZ	Central Zone
DAVID	Database for Annotation, Visualization and Integrated Discovery
Db1	Database 1
Db2	Database 2
DGCR8	DiGeorge Syndrome Critical Region in gene 8
DNA	Deoxyribonucleic acid
DRE	Digital Rectal Examination
ER	Estrogen Receptor

EST	Expressed Sequence Tag
EXP-5	Exportin-5
exRNAs	Extracellular RNAs
FISH	Fluorescence *in situ* hybridization
GEO	Gene Expression Omnibus
HSP90AA1	Heat Shock Protein 90 Alpha Family Class A Member 1
Hr-HPV	High-risk Human Papillomavirus
IGF1	Insulin Like Growth Factor 1
IGFBP3	Insulin Like Growth Factor Binding Protein 1
IGFBP3	Insulin Like Growth Factor Binding Protein 3
IHC	Immunohistochemistry
KEGG	Kyoto Encyclopaedia of Genes and Genomes
KLK2	Kallikrein Related Peptidase 2
KLK3	Kallikrein Related Peptidase 3
KLK4	Kallikrein Related Peptidase 4
MicroRNA	Micro Ribonucleic acid
MirDIP	microRNA Data Integration Portal
MiRmine	Human miRNA Expression Database
MRI	Magnetic Resonance Imaging
mRNA	Messenger ribonucleic acid
MSP	Maximal Segment Pair
MYH11	Myosin Heavy chain 11
NA	Nucleic Acid
NCBI	National Centre for Biotechnology Information
NCI	National Cancer Institute
NCOA2	Nuclear Receptor Coactivator 2
NIH	National Institutes of Health
NKX3.1	Homeobox Protein
NGS	Next Generation Sequence.

NLM	National Library of Medicine
OMIM	Online Mendelian Inheritance in Man
PC3	Prostate Cancer Cell line
PCa	Prostate Cancer
PCR	Polymerase Chain Reaction
PEC	Prostate Epithelial Cells
PR	Progesterone Receptor
Pre-miRNA	Precursor miRNA
Pri-miRNA	Primary miRNA
PSA	Prostate-Specific Antigen
PTEN	Phosphatase and Tensin homolog
PTGS2	Prostaglandin-Endoperoxide Synthase 2
PZ	Peripheral Zone
qRT-PCR	Quantitative real-time polymerase chain reaction
RISC	RNA-induced silencing complex
RLC	RISC loading complex
RNA	Ribonucleic acid
RNase	Ribonuclease
siRNA	Small interfering RNA
SKOV3	Ovarian carcinoma cell line
SRD5A2	Steroid 5 Alpha-Reductase 2
ssRNA	Single-strand RNA
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TF	Transcription factor
TFs	Transcription factors
TiGER	Tissue-specific Gene Expression and Regulation

TIMP-1	Tissue Inhibitor of Metalloproteinase-1
TMPRSS2	Transmembrane Protease, Serine 2
TSG	Tumour Suppressor Gene
TZ	Transition Zone
W	Word size
API	Application Programming Interface
ULK1	Unc-51 Like Autophagy Activating Kinase 1

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Cancer

Cancer is a disease that is known to affect a variety of tissues in the body such as bone marrow, brain, blood, prostate and skin. Cancer is known to be one of the most common diseases (DeSantis *et al.*, 2016) with 1,685,210 new cancer cases reported every year resulting in 595,690 deaths, making cancer a major public health issue worldwide. Cancer is a state of uncontrolled cell growth in which cells divide continuously resulting in tumour formation. Moreover, cancer cells do not undergo programmed cell death, or apoptosis because of the resultant mutations within these cells. Furthermore, cancer cells can distort the immune system, which usually removes the dead cells and any abnormal cells from the body, whilst some cancer cells have the ability to hide from the immune surveillance (Mendes *et al.*, 2016). However, not all the changes in tissues can be considered cancerous. When tissue changes, it can become cancerous if not treated (Mendes *et al.*, 2016). A tumour is classified depending on its ability to spread from the tissue of origin to adjacent tissues namely:

(I) *A benign tumour* is not cancerous and rarely life-threatening; it appears to grow slowly and does not have the ability to spread to other tissue. Usually, it is quite similar to healthy cells. When it grows larger it becomes uncomfortable or presses on other organs.

(II) *A malignant tumour* is a type of a tumour where the cells become aggressive and can spread (metastasize) to adjacent tissues. Cancer is also classified according to the type of cell where it originates from:

i) *Carcinomas*; which is classified as cancer starting in the skin or in the tissue covering internal organs which includes adenocarcinoma, basal cell carcinoma and squamous cell carcinoma.

ii) *Sarcomas*; are classified as cancer that starts in the supportive tissues, such as the cartilage, bone, muscle, fat, and blood vessels.

iii) *Leukemias*; *which* is classified as cancer that is found in blood-forming tissue such as the bone marrow and produces abnormal blood cells.

iv) *Lymphomas and spinal cord cancer*; are classified as cancer, which affects the nervous system.

1.2 Genetics of Cancer

The most common cause of cancer is genetic change whether these are acquired mutations which are more common or germline mutations which are not common in cancer (Rizzolo *et al.*, 2011; Wang *et al.*, 2017). There are different types of genes involved in initiating cancer such as oncogenes and tumour suppressor gene (TSG).

1.2.1 Oncogenes

Proto-oncogenes are grope of genes that cause of normal cell become cancerous after they mutated (Chial *et al.*, 2008). Proto-oncogenes' code for a protein whose function is mostly to stimulate cell division, prevent cell differentiation or to

regulate programmed cell death (Casey, Baylot and Felsher, 2017). When proto-oncogenes are affected by mutations, increasing the production of these proteins, unregulated cell division occurs, causing slowing of the cell differentiation rate and an increased inhibition of cell death. These changes to the cell cause them to become cancerous (American Cancer Society, 2014). Hyperactive oncogenes impart metabolic and growth-promoting advantages to malignant cells, affecting individuals differently. There is a growing concern in understanding the role of oncogenes and their activity in cancer to distinguish between the harmful and dormant tumour subtypes (Dasgupta, Vishwanatha and Srinidhi, 2012).

1.2.2 Tumour suppressor genes (TSGs)

Tumour suppressor genes regulate cell division by encoding proteins that are involved in cell proliferation and repair of DNA damage (Hinds and Weinberg, 1994). Mutagenesis leads to their de-activation which impair the repair of DNA damage resulting in uncontrolled cell growth, which signals the start of tumour progression (Sherr, 2004). TSGs either functions in promoting apoptosis or in the regulation of the cell cycle (Abreu Velez and Howard, 2015). There are several TSGs linked to cancer such as Phosphatase and Tensin Homolog (*PTEN*) which is a tumour suppressor gene associated with to Prostate Cancer (PCa). *PTEN* is involved in pathway regulation of cell growth and division, which is usually uncontrolled in cancer (de la Rosa *et al.*, 2017). *BRCA1* and *BRCA2* genes produce tumour suppressor proteins which aid in the repair of damaged DNA, mutations in these genes are linked to breast cancer (Silver and Livingston, 2012).

1.3 Anatomy of the prostate

The prostate is a gland that is part of the male reproductive system and located between the bladder and bottom of the pelvic cavity. The size and shape of the prostate are like a walnut, and muscle tissue. The prostate gland consists of four regions: The peripheral zone (PZ) and the central zone (CZ) together comprise <95 % of the prostate mass in the prostate of a healthy man. The rest of the prostate gland comprises a peri-urethral transition zone (TZ) and the anterior fibromuscular zone (AFZ) also known as the stroma and the peri-urethral glandular zone (Figure 1.1). The glandular tissue produces about 20 % of seminal fluid and contributes to the viscosity of the semen (Aumüller *et al.*, 1990).

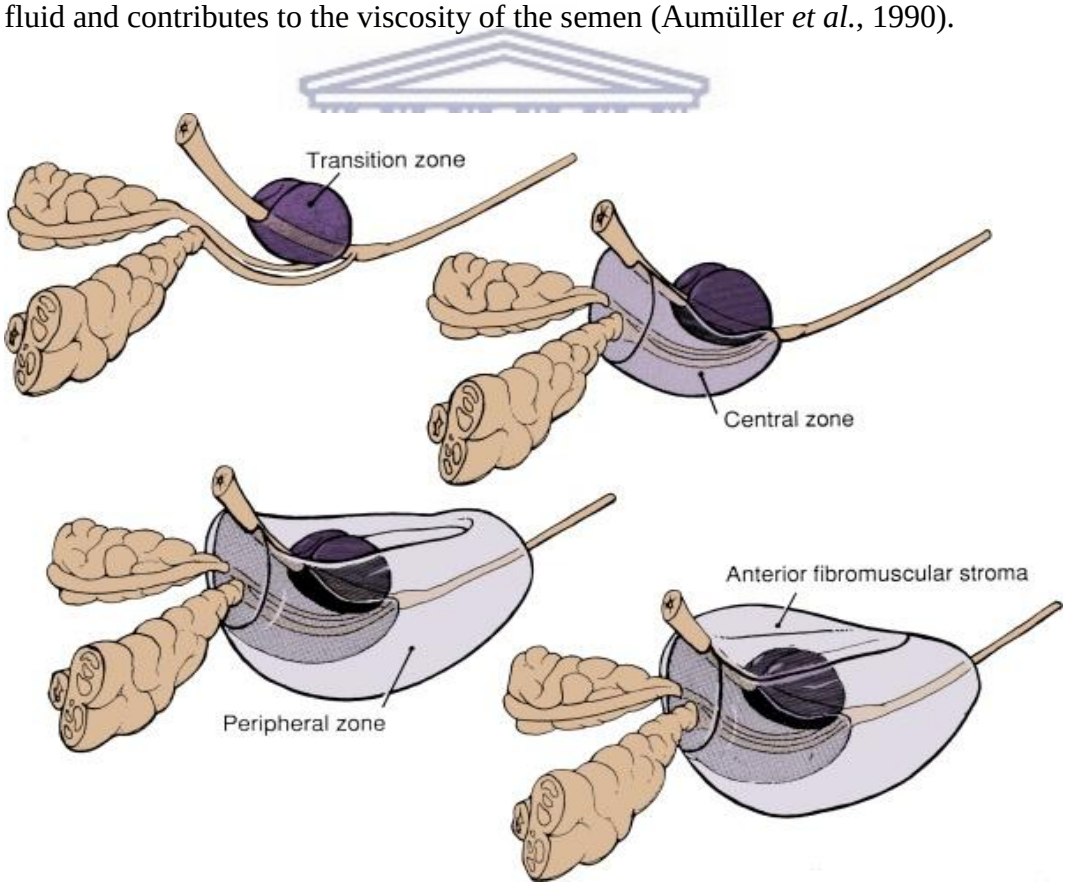


Figure 1.1: Number of zones in the prostate gland (adapted from the National Cancer Institute 2017).

PCa is classified as an adenocarcinoma, or glandular cancer, that starts when normal semen-secreting prostate glandular cells mutate into cancer cells. The region of the prostate gland where an adenocarcinoma is most common is the PZ. Initially, small clumps of cancer cells remain confined to an otherwise healthy prostate gland, PCa often metastasizes to lymph nodes, bone, and other distant sites (Kyprianou, 1994).

1.4 Incidences of Prostate Cancer

In men, PCa is the most commonly occurring cancer and the second leading cause of death after lung cancer (Siegel, Naishadham and Jemal, 2012). However, recent statistics mentioned PCa to be the second most common cancer after skin cancers in men (American Cancer Society, 2016). In addition, one in seven men are diagnosed with PCa according to the American Cancer Society, 2016), whereas an estimated 241,740 men were diagnosed with PCa and about 28,170 died of PCa in 2012 (Siegel, Naishadham and Jemal, 2012). Likewise, 161,360 new cases and 26,730 deaths were observed in the year 2016 (American Cancer Society, 2016). African-American males have the highest incidence of PCa in the world (Siegel, Naishadham and Jemal, 2012). In South Africa in 2013, 6778 new cases of PCa was recorded and distributed among the four races as follows; Blacks 2708, Asians 169, Coloureds 792 and Whites 3114 cases (Ed and Cur, 2016).

1.5 Stages of prostate cancer

There are different methods to classify the various stages of PCa, which all depend on how far the cancer has spread at the time of diagnosis. These stages are

of importance to assist doctors in prescribing a specific treatment for the patient (Tewari, 2013). The stages of PCa are shown in Figure 1.2. In Stage I, the cancer is found in the prostate and usually cannot be detected by digital rectal examination (DRE). At this stage, the cancer is difficult to be detected by the current diagnostic methods. In Stage II, the tumour has grown bigger and is still confined to the prostate. In Stage III, the tumour has broken through the prostate capsule and affects the adjacent tissue such as the seminal vesicles. In Stage IV, metastasis is taking place where a tumour is spreading to other tissues or organs, usually the lymph nodes, liver, and lungs (Chen *et al.*, 2017).

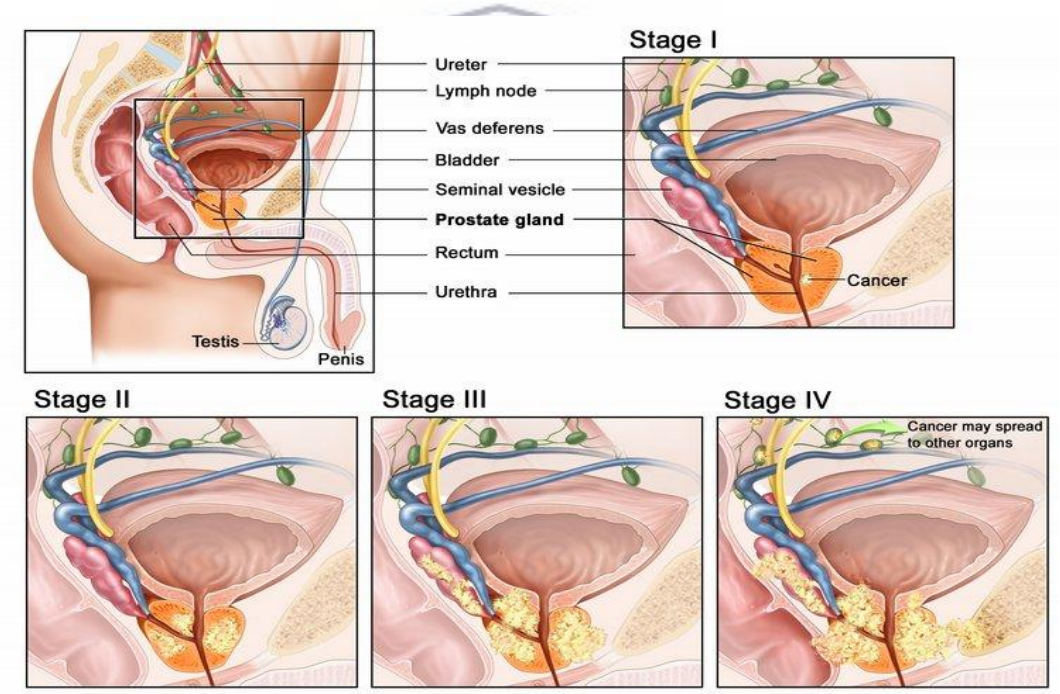


Figure 1.2: Classification of PCa using the Stage Number Method (adapted from the National Cancer Institute, 2016).

1.6 Risk factors for prostate cancer

1.6.1 Age

In a previous study, it was reported that the incidences of PCa increase with age. Therefore, PCa is rare in men younger than 40 years of age. About six out of ten cases of PCa are found in men older than 65 years (Cancer Research UK, 2013). In Another study, the incidence of PCa in men older than 65 years was reported to have dropped from 53 % in 1990 to 27.8 % in 1996 and remained stable thereafter, whilst the number of diagnosed patients younger than 65 increased from 18.6 % in 1991 to 40.7 % in 2000 (Stangelberger, Waldert and Djavan, 2008). Older men tend to be diagnosed with an advanced-stage of the disease and are more likely to die from PCa than younger men (Bian and Hoffman, 2013).

1.6.2 Family history

Family history is significantly linked with PCa risk in epidemiologic study. Moreover, a history of other diseases and certain inherited genetic conditions increase the chances to be at risk for PCa (Helfand, 2016).

1.6.3 Occupation

Manufacturing workers that are exposed to high dose chemical substances are at risk for developing PCa, some occupations have also been observed to increase the risk of PCa such as metalworking, fire-fighters, rubber production, administrative and managerial occupations (Sauvé, Lavoué and Parent, 2016).

1.7 Diagnostic methods for prostate cancer

1.7.1 Prostate-specific antigen

The prostate-specific antigen (PSA) test is one of the diagnostic tools used to detect PCa. PSA is secreted by prostatic epithelial cells and increased concentrations are associated with prostatic disease including cancer. Most doctors consider a PSA level of 4.0 ng/ml and lower as normal, whilst for levels higher than 4.0 ng/ml, a prostate biopsy is recommended to determine the presence of PCa (Walsh, 2006). Measuring PSA levels is more accurate in detecting PCa than Digital Rectal Examination (DRE) or ultrasound, however, combining PSA and DRE provided better results along with an additional ultrasound in those patients with abnormal PSA and DRE results (Tadayon *et al.*, 2016). PSA is also used for screening men with PCa and is used as a marker for identifying recurring disease following to treatment. However, PSA-based screening is also associated with overdiagnosis and overtreatment (Loeb *et al.*, 2014). The limited specificity of PSA as a cancer biomarker is attributed to the fact that the PSA is synthesized by all prostate epithelial cells, whether normal, hyperplastic, or cancerous. Elevated serum PSA levels may reflect the presence of cancer or may be caused by benign prostatic hyperplasia (BPH), infection, and chronic inflammation (Faruq and Vecchione, 2015).

1.7.2 Digital Rectal Examination (DRE)

The DRE is a test used by urologists to detect evident changes in the prostate gland; however, this can only be done on tumours that are relatively large. Another limitation is that most cancers occurring in that part of the anatomy are

not accessible by DRE. Digital rectal examinations are not sensitive enough to detect the small-volume of tumours that are most amenable to cure (Walsh *et al.*, 2014).

1.7.3 Prostate biopsy

Prostate needle biopsy is used to diagnose PCa and to detect the size of a tumour. Not all prostate biopsies are effective, as a false-negative result is sometimes obtained, and the procedure has to be repeated if the PSA level is very high. This makes prostate biopsy insensitive as a diagnostic tool for PCa as well as being invasive (Loeb *et al.*, 2013).

1.7.4 Computed tomography (CT)

Computed tomography (CT) is not recommended for new diagnostic testing for PCa, particularly whilst the cancer is still confined to the prostate. However, once the cancer has spread or after treatment has commenced, this test can be used to see if cancer cells are still present (Picchio, Giovannini and Messa, 2011).

1.7.5 Magnetic Resonance Imaging (MRI)

Magnetic resonance imaging (MRI) can also detect PCa. MRI scans show detailed images of soft tissues in the body. MRI scans use radio waves and strong magnets as a substitute for x-rays. Gadolinium may be injected before the scan for better tissue picture details. MRI scans provides a clear picture of PCa and show if the cancer has spread outside the prostate into the seminal vesicles and other adjacent tissue. This can be very important in determining a patient's treatment options. An

MRI scan is not recommended for newly diagnosed PCa that are likely to be confined to the prostate (Dickinson *et al.*, 2011).

1.8 Treatment of prostate cancer

Treatment depends on the age, stage and other medical conditions (Trewartha and Carter, 2013). There are several options to treat PCa, such as surgery, radiation therapy, and hormonal therapy. However, all the current therapy options for men with PCa are not effective enough and induce several side effects. Furthermore, after surgical ablation of progressive cancer, metastasised tumour cells continue to progress and this is a side effect associated with surgery. The radioactive rays and chemotherapeutic treatment also affects normal cells such as bone marrow function, and can lead to bone necrosis and lung fibrosis (Ghribi *et al.*, 2015). Also, androgen deprivation therapy (ADT) is used for PCa treatment combined with radiation therapy, which is more beneficial for patients with metastatic stage cancer. Androgen hormones are essential for PCa cells to grow and reducing the level of these hormones, inhibit PCa growth (Perlmutter and Lepor, 2007).

1.9 Cancer biomarkers

A biomarker, which is a biological molecule, has been established as a medical diagnostic tool. Biomarkers from body fluids or tissue are used to diagnose a disease or discriminate between a normal and an abnormal condition. These molecules are affected by the presence of disease or a change within biological processes. Moreover, they could be an indicator of the body's response to treatment. Biomarkers can increase the quality of health early in the diagnosis and monitoring of the disease and also during treatment (Madu and Lu, 2010).

1.10 Classification of biomarkers

Biomarkers are classified based on the sequence of events, from the onset of exposure to disease, detection as well as playing an essential role in providing some insight into the prognosis, disease progression and response to treatment (Manne, Srivastava and Srivastava, 2005). There are a variety of biomarkers available, which is utilised in experimental research and clinical settings. Biomarkers are grouped into different categories such as early detection, diagnostics, prognostics and predictive biomarkers, and can be DNA, RNA and protein-based (R Mayeux, 2004). Therefore, biomarkers are invaluable tools for cancer detection, diagnosis, a patient's prognosis and treatment selection. Biomarkers can also function in localising a tumour and determine its stage, subtype and response to therapy. Identification of such signatures in surrounding cells or at more distal and easily sampled sites of the body viz., cells in the mouth (instead of the lung) or urine (instead of the urinary tract) can also influence the management of cancer (Bhatt *et al.*, 2010). A significant focus of clinical research, as evident in many publications, has been the discovery and validation of biomarkers with the primary aim of facilitating disease diagnosis (Bhatt *et al.*, 2010).

1.10.1. Diagnostic biomarkers

Diagnostic biomarkers are used to screen or detect specific types of cancer. These biomarkers may be assessed from tumour specimens. However, most diagnostic biomarker research is performed on serum or plasma. The ideal source material for biomarker analysis is the blood, because of easy sampling methods and being

relatively cheap (Sanjay *et al.*, 2015). The goal of diagnostic biomarkers would be to apply them in a disease-screening setting or in patients with suspected cancer, who are at higher risk of developing the disease (Kulasingam and Diamandis, 2008). Simple diagnostic tests with one or two biomarkers or more complex tests can be developed through assaying novel or putative biomarkers. At the molecular level, such tests can provide faster and more accurate information-rich diagnostics for various diseases (Costa-Pinheiro *et al.*, 2015).

Biomarker-based tests can confirm the clinical diagnosis and may provide other information concerning prognosis and best treatment options. These diagnostic tests may enable identification of the disease or susceptibility to its early stages, before being diagnosed by other means. Opportunities for the prevention of disease progression or better disease management which lead to improved patient outcomes and a reduction in the direct and indirect costs of a disease are provided (Costa-Pinheiro *et al.*, 2015). Improved diagnostics to detect cancer at an early stage, when it is curable with current methods would provide the most significant benefit to cancer patients. For most cancers, 5-10 years' survival often approaches 90 % for cancers detected at an early stage, while it may drop to 10 % or less for cancers detected at a later stage (Hazelton and Luebeck, 2011). Thus screening to detect cancer early saves lives is well established. For instance, Pap smears significantly reduce mortality through the early detection of pre-neoplastic cervical cancer lesions. Moreover, the test is employed widely despite its significant inconvenience, unpleasantness, cost and requirement for clinical expertise (Aebersold *et al.*, 2005). In some countries, high-risk human

papillomavirus (hr-HPV) are implemented as the primary cervical cancer screening method (Litjens *et al.*, 2013).

1.10.2. Prognostic biomarkers

The role of prognostic biomarkers is to acquire knowledge about the natural history of the disease in terms of metastatic potential, the likelihood of tumour progression and the probability of patient survival independently of treatment. Prognostic biomarkers are relevant to cancer research for various reasons, such as when it is used to determine if the expression of a protein, gene or another molecule is directly correlated with an aggressive phenotype (Oldenhuis *et al.*, 2008). This may provide important information about the biology of the disease and which pathways are activated when the phenotype is more aggressive. Furthermore, from a clinical aspect, availability of validated strong and independent prognostic tumour markers may be used to stratify patients in randomised clinical trials aiming at evaluating the effect of diverse drugs (Kulasingam and Diamandis, 2008). Many prognostic biomarkers are linked to different cancers such as the Tissue Inhibitor of Metalloproteinase-1 (*TIMP-1*) expression, which is related to the aggressive forms of multiple myelomas (Terpos *et al.*, 2010). Expression of the progesterone receptor (*PR*) is associated with a better survival rate in breast cancer patients (Harris *et al.*, 2007).

1.10.3. Predictive biomarkers

The function of predictive biomarkers is to determine in which subset of patients, with a particular disease, drug treatment will be more effective. A common clinical observation is that the same treatment may induce excellent responses in

some patients, while in others it will not have any effect. Ideally, after testing for a panel of the predictive biomarkers for different treatment regimens, one could administer the most effective drug only to those patients, who based on the results of the biomarker analysis, will benefit most likely from that specific treatment (Nalejska, Maczyńska and Lewandowska, 2014). Biomarkers can also be used to predict toxicity and be particularly valuable for choosing between drugs with the same activity, but with different toxicity profiles (Kulasingam and Diamandis, 2008). In terms of the drug response, there are various predictive biomarkers in clinical use such as mutations within *BRCA1* and *BRCA2* genes. These mutations are used as predictive biomarkers to evaluating a patient with platinum-sensitive ovarian cancer and to identify patients likely to respond to Poly (ADP-ribose) polymerase (*PARP*) inhibitors (Ledermann *et al.*, 2012). Traditionally, biomarker included proteins, DNA and RNA. However more recently there has been an increasing focus on miRNAs as potential cancer biomarker.

1.11 MiRNAs and Cancer

MicroRNAs are small non-coding RNAs, which regulates the expression of a gene that is used by the cells to produce protein. MiRNAs have important roles in different cellular processes such as the development, proliferation and apoptosis (Davis and Hata, 2009). MiRNA could be a potential biomarker for diseases, which include PCa. In previous studies, irregular expression of miRNAs in PCa tissue and cell lines has been shown (Porkka *et al.*, 2007). For instance, after examining the expression level of miR-145 in human PCa tissue samples, a significant difference between an adjacent normal cell and a cancerous region was

observed (Zaman *et al.*, 2010). Overexpression of miRNA-145 in PC3 cells was also shown to lead to cell cycle constraints and increased apoptosis (Zaman *et al.*, 2010).

1.12 Biogenesis of miRNAs

A diversity of proteins are involved in the processing of miRNAs, with primary stages occurring inside the nucleus. The primary miRNA(pri-miRNA) is cleaved by a Pol II enzyme resulting in the formation of a precursor miRNA (pre-miRNA) by the release of a 60–70 nt (nucleotide) hairpin structure (Lee *et al.*, 2002; Zeng and Cullen, 2003; Wahid *et al.*, 2010). In figure 1.3 the processing step, this cleavage is catalysed by Drosha, which requires cofactors such as Pasha in *Drosophila melanogaster* or *Caenorhabditis elegans* or the human DiGeorge syndrome critical region gene 8 (DGCR8) (Lee *et al.*, 2003; Han *et al.*, 2004; Wahid *et al.*, 2010). Drosha conjugates with either Pasha or DGCR8 forming a large microprocessor complex (Gregory *et al.*, 2004; Han *et al.*, 2004). *In vivo* studies showed that the DGCR8 gene is significant in developmental processes and is well-conserved (Wahid *et al.*, 2010). The animal pri-miRNAs usually include 33 base pairs (bp) in the region of the stem-loop as well as a terminal loop, and a single-strand RNA (ssRNA) flanking fragment. The DGCR8 gene initiates the cutting by process interaction with the ssRNA section, to direct Drosha in completing the cleavage of pri-miRNA. The slicing of RNA duplexes occurs around 11 bp away from the ssRNA-stem loop junction, processing the pri-miRNA to the pre-miRNA with an overhanging 2 nt 3′ and 5′-phosphate group (Lee *et al.*, 2003; Zeng and Cullen, 2005; Han *et al.*, 2006).

The processed pre-miRNAs are further transported through nuclear pore complexes into the cytoplasm for maturation into miRNAs (Nakielny and Dreyfuss, 1999), which is mediated by RanGTP-dependent nuclear transport receptor exportin-5 (EXP5) (Bohnsack, Czaplinski and Gorlich, 2004; Lund *et al.*, 2004). In other studies evidence that pre-miRNA transport is initiated by the EXP5-based recognition of >14-bp double-stranded RNA (ds-RNA) stem-loop followed by a joint binding to both the pre-miRNA and the GTP-bound cofactor Ran inside the nucleus. The EXP5 bounded pre-miRNA is exported from the nucleus, with associated GTP hydrolysis resulting in the release of the pre-miRNA as depicted in figure 1.3 (Gwizdek *et al.*, 2003; Lund *et al.*, 2004; Zeng and Cullen, 2004).

The nuclear cleavage process by Drosha expresses one end of the mature miRNA. The pre-miRNA is discharged into the cytoplasm using EXP5 and is consequently processed by an endonuclease cytoplasmic RNase III enzyme Dicer to generate a mature miRNA (Grishok *et al.*, 2001; Knight, 2001). Dicer is an enzyme that specifically measures 22 nt from the pre-existing terminus of the pre-miRNA and performs the cleavage of a miRNA strand. Dicer is a well-conserved enzyme which is present in nearly all eukaryotic organisms. Several organisms contain multiple types of Dicers; as in *D. melanogaster* which comprises Dicer-1 and Dicer-2, each having distinctive roles. Dicer-1 is involved in the maturation of miRNA, while Dicer-2 is essential for siRNA maturation (Lee *et al.*, 2004). Dicer interacts with other proteins such as the Argonaute (Ago) family proteins in a variety of organisms, RNAi deficient-4 in *C. elegans,* and X mental retardation 1 in *D. melanogaster* (Carmell *et al.*, 2002; Tabara *et al.*, 2002).

In recent studies, it was observed that *D. melanogaster* Dicer-1 involves interaction partner Loquacious (LOQS; also known as R3D1), which comprises three dsRNA binding motifs for pre-miRNA processing (Förstemann *et al.*, 2005; Saito *et al.*, 2005). The human Dicer is coupled with two closely homologous proteins, protein kinase, interferon-inducible double-stranded RNA-dependent activator (PRKRA) and trans-activation response RNA-binding protein (TRBP) (Lee *et al.*, 2006; Chendrimada *et al.*, 2010). The Dicer-associated proteins contribute to the formation of the RNA-induced silencing complex, with little processing activity required (Lee *et al.*, 2006; Chendrimada *et al.*, 2010).

The precise roles of these respective proteins have yet to be resolved. According to the current model, after the generation of an approximately 22 nt miRNA duplex by Dicer cleavage, the miRNA duplex is incorporated into an Ago family protein complex generating an effector complex. Mostly, one strand of the miRNA (passenger strand or miRNA) is degraded, whereas the other strand remains bound to Ago as mature miRNA (guide strand or miRNA). In a few cases, miRNA is loaded into the RNA-induced silencing complex (RISC) and therefore, remains functional. The thermodynamic stability of the two ends of the duplex may determine which strand is to be selected as shown in a few studies (Khvorova, Reynolds and Jayasena, 2003; Wahid *et al.*, 2010).

Dicer, in conjunction with other interacting proteins (TRBP and/or PACT in human and LOQS in fly) and Ago family proteins, contributes to RISC assembly by forming a RISC-loading complex (RLC) (Tomari, 2004; Maniataki and Mourelatos, 2005). The exact mechanism regarding the role of RLC in RNA

loading to Ago is not known. However, evidence suggests that after the processed miRNA duplexes are released from Dicer, the stable end of the miRNA duplex binds to interact with proteins in the RLC, and the unstable end associates with the Ago proteins (Liu, 2003; Tomari, 2004). The endo-nucleolytic enzyme activity of the Ago protein has been demonstrated to be responsible for the removal of the miRNA passenger strands (Diederichs and Haber, 2007).

Most of the miRNAs contain mismatches in the middle, and some Ago proteins lack "slicer" activity, making the passenger strand of the miRNA resistant to cleavage. Evidence suggests that an RNA Helicase (yet to be identified) mediates the unwinding and removal of the unselected strand of the miRNA duplex. After loading, the miRNA guides the RISC to its target mRNA, which is silenced through degradation or translation repression (Jones-Rhoades, Bartel and Bartel, 2006; Wahid *et al.*, 2010).

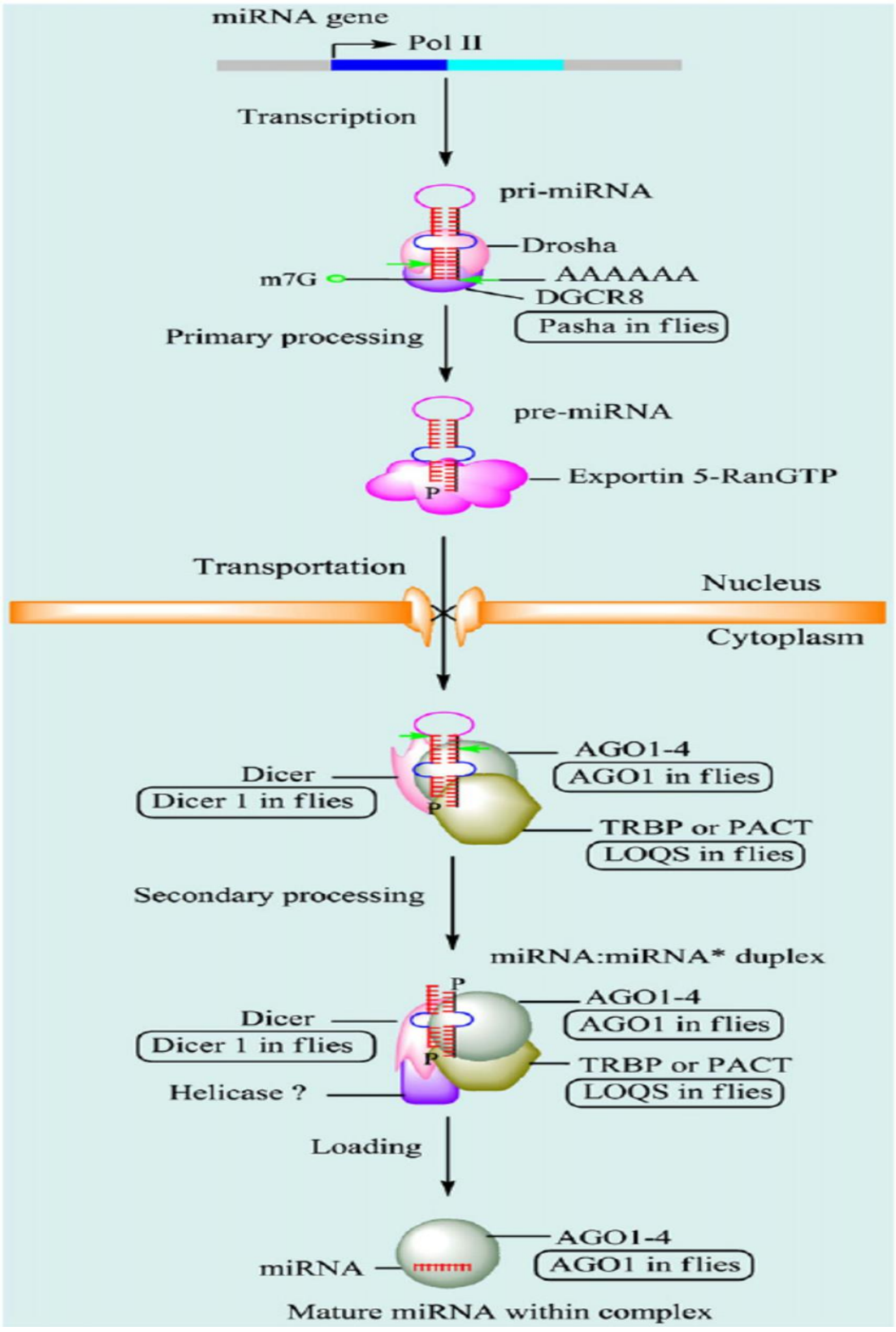


Figure 1. 3: The animal miRNA synthesis pathway and fly factors are in the squares adapted from (Wahid *et al.*, 2010).

1.13 MiRNAs as regulators of gene expression

MicroRNAs regulate the expression of target genes and usually have their specific genes, sometimes they cohabitate with the introns or untranslated regions of genes that encode proteins (Barrett, Fletcher and Wilton, 2012). The regulation of gene expression by miRNAs, which has been proposed to be combinatorial, has different miRNAs acting together on a gene to regulate its precise level of expression during its development and in response to stimuli (Tomari and Zamore, 2005). Estimates of the average number of unique mRNAs that are targets for repression by a typical miRNA vary, depending on the estimation method (Thomson, Bracken and Goodall, 2011). MiRNA regulated involvement has been shown in malignant B-cell development, with chronic lymphocytic leukaemia being the first disease associated with deregulation of miRNAs as tumour suppressor genes (Musilova and Mraz, 2015).

1.14 MiRNAs as oncomiRs and Tumour suppressors

MiRNAs associated with cancer are known as oncomiRs, which play a role in the translation of mRNA to produce proteins that are involved in the regulation of the cell cycle, apoptosis, and necrosis. Certain oncomiRs have been linked to specific cancer such as mir-17-92, which may function as an oncogene and promote cancer development by negatively regulating tumour suppressor genes and/or genes that control cell differentiation or apoptosis (Zhang *et al.*, 2007). Studies have shown that some miRNAs regulate cell proliferation and apoptotic processes that are important in cancer formation (Hammond, 2006). By using multiple molecular techniques, which included Northern blot analysis, real-time PCR,

miRNA microarray measuring up- or down-expression of specific miRNAs, it was found that several miRNAs were directly involved in human cancers, including lung, breast, brain, liver, colon cancer and leukaemia (Melo and Esteller, 2011). Some miRNAs may function as oncogenes or tumour suppressors (Rufino-Palomares *et al.*, 2013).

1.15 Methodology for discovery of biomarkers

The survival rate of a disease relies on its early detection with the most challenging part in the early diagnosis of cancer being that most cases are diagnosed after cancer has almost fully developed (Chari *et al.*, 2015). At present, many methods are proposed to facilitate early detection using body fluids which are easily obtainable. Thus, the treatment can be more effective at the early stage of the disease (Chari *et al.*, 2015). More attention is being paid to identify biomarkers in the body fluids such as genomic biomarkers, which identifies changes in chromosomes, genes, and proteins (Novelli *et al.*, 2008). The new technological advances in proteomics have shown capable initiatives in identifying novel biomarkers of many diseases to investigate and generate a large amounts of data (Ahram and Petricoin, 2008). Molecular profiling supported by the completion of the Human Genome Project and the development of biotechnological tools has revolutionised the exploration of disease biomarkers (Ahram and Petricoin, 2008). There are several diagnostic biomarkers for different types of cancers currently in clinical use which are based on different techniques such as FISH for intense Vysis®Urovysion biomarker for bladder cancer, cologuard® biomarker for colorectal cancer using PCR and

BRACAnalysis CDx® biomarker for ovarian cancer using the PCR technique (Goossens *et al.*, 2015).

The use of Computational Biology (Bioinformatics) methods showed significant results in biomarker discovery, using machine learning techniques such as clustering and support vector machine (SVM) (Yousef *et al.*, 2014).

1.16 Bioinformatics

Bioinformatics is an interdisciplinary field which involves conceptualisation of biology regarding molecules and applying information techniques such as applied mathematics, computer science, and statistics to understand and organise the large-scale data generated from experimental techniques (Luscombe, Greenbaum and Gerstein, 2001). Bioinformatics further involves sub-categories which include the analysis of genomic, proteomic and phylogenetic data, facilitating medical research through enhancing the sophistication of diagnosis and drug design (Luscombe, Greenbaum and Gerstein, 2001). Furthermore, an array of bioinformatics tools have been developed for the analysis of available gene datasets to identify biomarkers for the diagnosis and prognosis of diseases (Phan *et al.*, 2009).

1.17 *In silico* detection of novel biomarkers

In recent decades, bioinformatics has shown high accuracy in biological data analysis for the molecular biology field. Data analysis can be performed on different biological data such as protein, DNA, and RNA to evaluate their genetic changes, their expression in the human body and their involvement in diseases

such as cancer, cardiovascular disease, and osteoporosis etc (Zicari, 2014) *In silico* methods offer a less labourious, time consuming and cost-effective strategy over conventional methods of biomarker discovery. Through the use of *in silico* methods, a priority list of biomarkers can be generated to be validated experimentally. A previous study has successfully shown the use of *in silico* methods to identify 48 tissue specific proteins for colon, lung, prostate, and pancreas as potential cancer biomarkers. However, 14 out of these 48 proteins were well established in other studies as cancer or benign disease serum biomarkers (Prassas *et al.*, 2012). As well *in silico* approach used for another disease, a combination of *in silico* prediction and microarray analyses, successfully identified miR-4487 and miR-595 targeting *ULK1* with experimental verification of this targeting, this could provide the basis to explore *ULK1* and its target miRNAs as potential biomarkers in the future as a Parkinson's disease therapy (Chen *et al.*, 2015).

1.18 Aim and objective

The general aim of this study is to identify novel microRNA(s) as biomarkers for early detection of PCa

I) Identify novel miRNAs as diagnostic and prognostic biomarkers for PCa using an *in silico* approach;

II) Identify the target genes for the identified miRNAs and demonstrate their involvement in PCa using *in silico* approaches;

III) Determine the prognostic value of the identified miRNAs as well as their target genes using several *in silico* methods; and

IV) Perform an expression analysis of the identified miRNAs and their target genes in prostate tissue versus other normal and cancerous tissues using *in silico* analysis.

Chapter 2

Identification of microRNAs as diagnostic and prognostic biomarkers for prostate cancer using an *in silico* approach

2.1 Introduction

Prostate cancer (PCa) is one of the most common diseases in the world and a leading cause of death as a consequence of the shortcomings of current diagnostic methods since most PCa cases are diagnosed at the advanced stages (Sharma, Zapatero-Rodríguez and O'Kennedy, 2017). Currently, physicians rely on the PSA test which is not completely accurate in determining the presence of PCa or other prostatic diseases, which then require additional diagnostic procedures and becomes highly costly for the patient. Current research is focusing on the use of molecular signatures for diagnostic purposes. Based on their change in expression, miRNAs has been suggested as potential biomarkers for use in diagnostics for different types of cancers (Fredsøe *et al.*, 2017).

MiRNAs functions in many different systems through regulation of their target genes (Bartel, 2004). The change in the expression of miRNAs affects the regulation of genes which are translated into proteins involved in cell proliferation and other cancer-related pathways (Johnson *et al.*, 2007). MiRNAs have been known to be involved in a variety of human cancers such as the over-expression of miR-22, which is a primary growth-associated miRNA that leads to inhibition of growth in the ovarian carcinoma cell line (SKOV3) (Lenkala *et al.*, 2014). However, proteomic and microarray studies revealed that miRNA mediated changes in protein expression usually results in altered mRNA expression, signifying that mRNA degradation is the major component of the repression

associated with mammalian miRNAs (Baek *et al.*, 2008; Selbach *et al.*, 2008). In most studies the involvement of miRNAs in the negative regulation of gene expression by base-pairing to the 3' UTR as well as up-regulating their target mRNA translation have been suggested (Vasudevan, Tong and Steitz, 2007). Literature studies suggest that miRNAs can serve as diagnostic biomarkers for a variety of cancers and is easily accessible from a variety of body fluids (Larrea *et al.*, 2016). Therefore, there is a need to investigate the hidden functionalities of miRNAs as well as their target genes to make them suitable candidates for cancer diagnostics.

The study of miRNA functional similarity is a significant parameter used to perform disease miRNA prioritisation, miRNA functional prediction, miRNA synergism identification as well as miRNA clustering (Teng *et al.*, 2013). Sequence analysis is often used to infer functional and structural features of DNA, RNA as well as peptides (Durbin *et al.*, 1998). A variety of tools are available to perform sequence analysis to understand the function of a miRNA. One of which is to perform similarity studies between experimentally validated miRNA sequences and the total number of deposited miRNAs in various databases, using a variety of *in silico* discovery tools which can lead to the identification of miRNAs not previously associated with a particular process or function.

2.2 *In silico* techniques for biomarker identification

In silico technologies aim to facilitate advances in medical science such as biomarker and drug discovery (Terstappen and Reggiani, 2001; Yu, 2009). *In silico* refers to analysis or experimentation carried out in a computer environment,

rather than *in vitro* or *in vivo* given the increasing amount of genetic and molecular data available. Computational models are used as virtual laboratories to test hypotheses rapidly and to prioritise empirical research. These methods have been employed in various studies for the rapid identification of diagnostic, prognostic and predictive biomarkers, such as genes, proteins and miRNAs, for different types of cancer in past decades (Lin and Wooley, 2005).

2.3 Data and text mining databases

A variety of databases were used for data mining and text mining purposes in this study to obtain the required information, with each database employed fully explained in the subsequent sections.

2.3.1 Exiqon

Exiqon houses various databases which include miRsearch 3.0, an online search tool freely available at https://www.exiqon.com/mirsearch. This tool quickly finds and displays specific miRNAs relevant to specific research with detailed information about each miRNAs. With the help of a miRNA name the following variables can be searched: a) miRBase history b) regulation of genes c) potentially co-transcribed microRNAs, d) disease in which the miRNA has been shown to be regulated and e) tissue/samples in which the miRNA has been found. Furthermore, validated targets as well as disease and tissue/sample information, is supported by references with a link to PubMed (Exiqon, n.d).

2.3.2 PubMed

PubMed database available at https://www.ncbi.nlm.nih.gov/pubmed is a free search engine frequently used by academic researchers for different purposes first released in 1996. The United States National Library of Medicine (NLM) at the National Institutes of Health (NIH) maintains the database as part of the Entrez system of information retrieval (Reyes-Aldasoro, 2017). For this study the PubMed database was used as a text mining tool.

2.3.3 MiRcancer

Text mining in MiRcancer available at http://mircancer.ecu.edu/ is based on 75 constructed rules, which represent the common sentence structures typically used to state microRNA expression in cancers. The microRNA-cancer association database, miRCancer, is regularly updated by running a text mining algorithm against PubMed. All miRNA-cancer associations are confirmed manually after automatic extraction. miRCancer currently documents 878 relationships between 236 microRNAs and 79 human cancers through the processing of >26 000 published articles (Xie *et al.*, 2013).

2.3.4 MiRBase

The primary miRNA sequence repository miRBase http://www.mirbase.org/ first released in 2002 contains 218 miRNA loci from five species the latest miRBase release (v20, June 2013) contains 24 521 miRNA loci from 206 species, processed to produce 30 424 mature miRNA products. The rate of deposition of novel microRNAs and the number of researchers involved in their discovery continue to increase, driven largely by small RNA deep sequencing experiments. MiRbase

also provides the role of nomenclature assignment for the novel miRNAs identified by the different research groups (Griffiths-Jones *et al.*, 2006).

2.3.5 PhenomiR

In recent years, published in the scientific literature deregulation of miRNA expression in diseases and other biological processes have been investigated PhenomiR (http://mips.helmholtz-muenchen.de/phenomir/) provides a repository that offers all the scattered information about miRNA expression in a structured and uniform format. This allows users to perform individual queries for specific miRNAs and diseases as well as to use the complete dataset for large-scale statistical analyses. All information in PhenomiR is extracted from published experiments and has been manually curated. The literature reference for each database entry is annotated as a PubMed identifier and is hyperlinked to the PubMed database. Every individual entry of the database refers to an instance of a publication describing a specific disease or bioprocess. Presently, PhenomiR documents data from 296 articles that describe 542 studies. This dataset contains 11,029 data points, each representing one deregulated miRNA in an experiment. A design principle of PhenomiR is to use well-established ontologies and resources (Ruepp *et al.*, 2010).

2.4 Local similarity tools for novel miRNA identification

The following sections will focus on different similarity tools for identification of new miRNAs for PCa based on their similarity to experimentally validated miRNAs.

2.4.1 CD-HIT

CD-HIT performs a very fast clustering of a large dataset of biological sequences which enhances the performance of sequence analysis as well as reduce the redundancy of sequences (Fu *et al.*, 2012). In response to next-generation sequencing based accumulation of biological sequence data, a new CD-HIT program was developed which was accelerated with a novel core parallelisation approach as well as statistical techniques for efficient clustering of the respective datasets (Fu *et al.*, 2012). The algorithm behind CD-HIT is short-word filtering, which can determine that the similarity between two sequences is below a certain value without performing an actual sequence alignment. This algorithm is not limited to protein sequence clustering; it can also be used for different analyses that involve a large number of sequence comparisons (Li and Godzik, 2006).

2.4.2 Local BLAST

BLAST is a heuristics method for fast sequence comparison, which precisely evaluates alignments that improve the measurement of local similarity along with the maximal segment pair (MSP) score which refers to a segment pair whose alignment score (without spaces) cannot be improved by extending it or shortening it. The performance of BLAST, as well as the alignments it generates, are analysed using the mathematical outputs of the stochastic properties of MSP scores. BLAST robustly perform biological sequence database searching, gene identification searches, motif searches, as well as analyse multiple similarity regions in a long biological sequence (Altschul *et al.*, 1990). The two most considered parameters in BLAST similarity searches are: a) E-value which can be

any positive number, with the default being 10, and can be adjusted up to 100 for a desirable limit of output, and b) Word size (W), which should be an integer number; the default for proteins 3 could be changed to 2 to increase sensitivity. For nucleotide sequences, a word size of 11 , is allowed, with the lowest word size being 7 (Koonin and Galperin, 2003).

2.5 Aim and objectives

The aim of this chapter is to identify miRNAs as diagnostic and prognostic biomarkers for PCa using a combination of data and text mining databases as well as sequence similarity tools to:

I. Generate a list of mature miRNAs that are involved in PCa using the Exiqon database;
II. Collect all the mature miRNA sequences using miRBase;
III. Search for miRNA similarity between the miRNA list generated from Exiqon and that generated by miRBase using the CD-HIT and BLAST programs; and
IV. Finalise a list of potential miRNAs for further validation that can serve as diagnostic and prognostic biomarkers for PCa.

2.6 Methodology

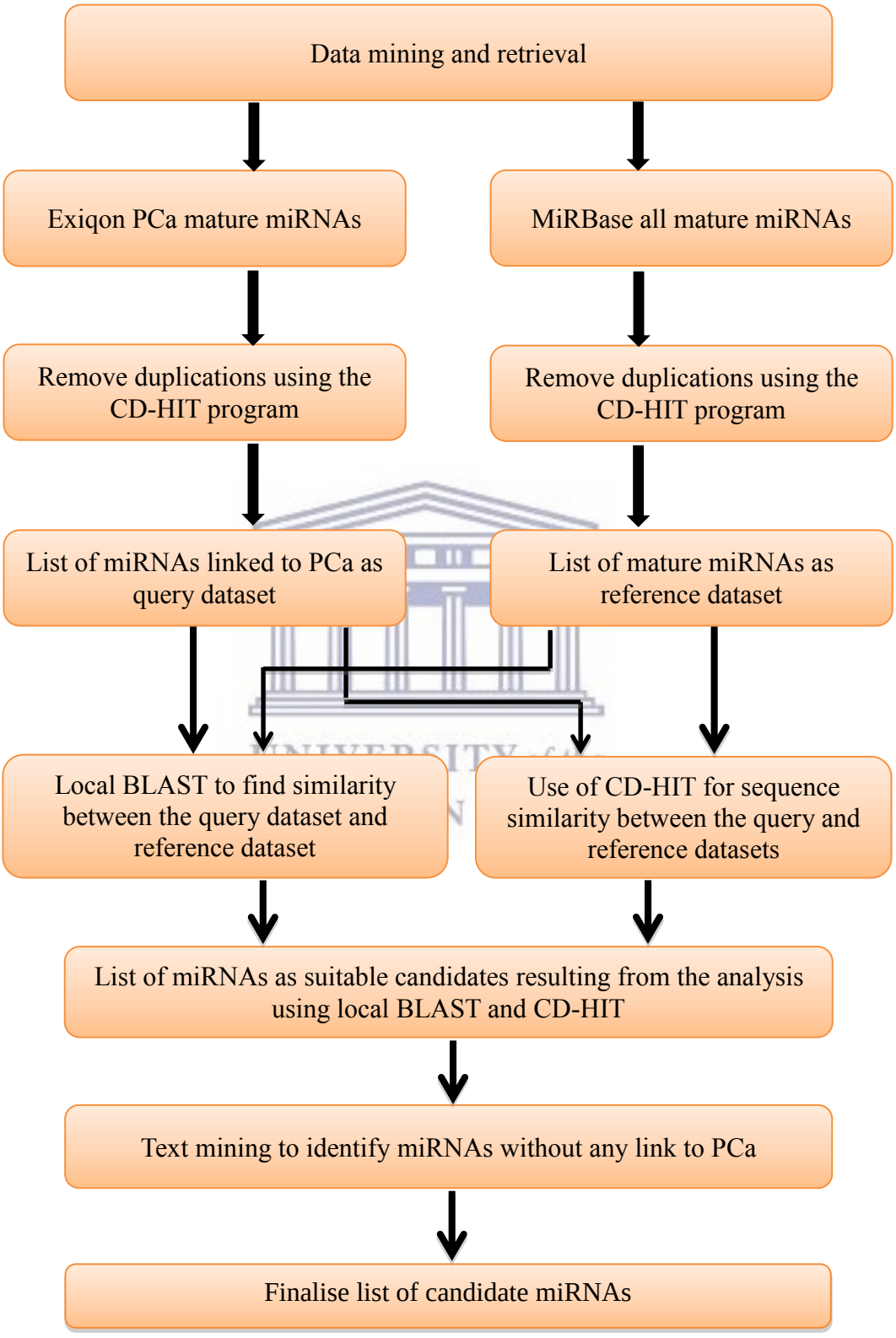


Figure 2.1: Outline of the *in silico* methodology used for miRNAs identification.

2.6.1 Data extraction

2.6.1.1 Data extraction from Exiqon

Exiqon is one of the databases that provide information regarding experimentally validated miRNA as well as their involvement in disease such as cancer. In this study, Exiqon was used to collect all miRNA linked to PCa. The Exiqon database was accessed at https://www.exiqon.com/mirsearch and the keywords "prostate cancer" was used as input in the query option. The search option was then selected. The list of output miRNAs was saved for further analysis.

2.6.1.2 Sequence extraction using MiRbase

MiRBase was used to download all mature *Homo sapiens* miRNA sequences that are available to date using the link (http://mirbase.org/cgi-bin/mirna_summary.pl?org=hsa). All mature miRNAs sequences were saved in text file fasta format for further analysis. The advantage of miRbase is that it provides the sequence precursor as well as the mature sequences. For this study, only mature miRNAs were used as this part of the miRNA performs its function(s).

2.6.1.3 Duplication removal

The CD-HIT program has an option that allows the user to remove duplications from a list of sequences, known as CD-HIT-454 and the re-engineered option CD-HIT-EST. The duplicates in sequence are identified using criteria such as "(1) beginning at the same position, (2) having different lengths, with the short sequence fully aligned with the longer sequence, (3) having a 4 % mismatch

(insertions, deletions, and substitutions) and (4) having only 1 base per insertion or deletion where criteria three (3) and four (4) can be adjusted by the user. Mismatches are allowed in order to tolerate sequencing errors. The CD-HIT program was used to remove duplications from the two files. The first file containing the miRNA sequences involved in PCa used as the query dataset and the second file containing all mature miRNA sequences used as the reference dataset in this study. This program was used by running the basic command line below:

Cd-hit-est -i prostatemiRNA.fa -o resultprostatemiRNA1 -c 0.95 -n 8. (1)

Where *–i* is input filename in fasta format, *-o* is output filename, - *c* is sequence identity threshold, and *-n* is word length.

2.6.2. Identification of potential novel miRNAs for Prostate Cancer

2.6.2.1 Blast for similarity searches

The BLAST program can be used to generate a database of the information of interest. This option was used to first build a database for all the *Homo sapiens* mature miRNAs extracted from MiRBase after removal of duplication to use as the reference set in this study using the command line below:

Makeblastdb –in resultallmiRNA.fasta -Parse_seqids –dbtypenucl……..(2)

Where – *in* is the input file, *-Parse_seqids* is sequence id parsing and –*dbtypenucl* is specifing the input, with the molecule type for this work specifed as nucleotide.

After creating the database containing all the mature miRNAs, the validated miRNAs for PCa was scanned against the database to identify the number of mismatches between the query dataset (validated miRNAs for PCa) and the reference dataset sequences of all known miRNAs). The number of mismatches allowed was three sequences with less than three mismatches having more than 90 % similarity. The command line used to run the program was:

Blastn –db allmiRNA.fasta - evalue 1e-3 - word_size 7 –query resultprostatemiRNA1.fa –out result with 1e-3.out …… (3)

There are four inputs specified with one output generated: *-db* which is the database generated by previous command line (2), - *evalue* specified as 1e-3, - *word_size* is the similarity region, *-query* is the query file, and *–out* is the the result file name.

2.6.2.2 CD-HIT

CD-HIT-EST-2D works on nucleotide sequences. One of the characteristics of CD-HIT-EST-2D is to compare two nucleotide datasets (db1, db2). Db1 contained the non-redundant mature miRNA sequences, which are associated with PCa, which were used as the query dataset and db2 contained all mature miRNAs, which were used as the reference dataset. The input was in FASTA format datasets (db1, db2) and the output was two files: (i) a FASTA file of sequences in

db2 that is dissimilar to db1 and (ii) a text file that lists similar sequences between db1 and db2. The text file showed the similarity in clusters generated and percentage of similarity between sequences. This was done by running the following command line:

Cd-hit-est-2d –I resultprostatemiRNA.fa -I 2 resultallmiRNA.fa - O similar – c 0.88 –n 7……..... ...(4)

Where - *I* is Db1 which includes the validated PCa miRNAs, - *I 2* is Db2 which include all miRNAs, -*O* is the output, -*c* sequence identity threshold and –*n* 7 word-length.

2.7 Results

2.7.1 Data extraction from Exiqon

Exiqon is one of the databases that provide information regarding experimentally validated miRNAs as well as their involvement in diseases such as cancer. In this study, Exiqon was used to collect all miRNA that are currently linked to PCa. There is a variety of the databases that provide this type of information. The advantage of using Exiqon is that provides information only for the mature miRNAs. Consequently, 156 mature miRNAs linked to PCa were collected from the Exiqon database as shown in table A.1 Appendix A.

2.7.2 Sequence extraction using MiRBase

All miRNA sequences were downloaded from miRBase. Two lists of miRNA sequences were generated; first, a list of miRNAs involved in PCa to compare to

the list generated by Exiqon, and secondly, a list of 2588 unique mature miRNA sequences as shown in table 2.1.

Table 2. 1: Number of sequences extracted from Exiqon and miRBase

List of miRNAs linked to PCa	All mature miRNA deposit in miRBase
156 sequences	2588 sequences

2.7.2 Duplication removal

Duplications were removed using CD-HIT as described in section 2.6.1.3. The aim of removing duplications was to avoid repetition of sequences in the same file which will affect the result of the similarity search by generating redundancy in the results file. After removing duplications two files were generated, (1) a query dataset which included 150 miRNAs involved in PCa and (2) a reference dataset including 2477 miRNAs as shown in table 2.2.

Table 2.2: The number of unique miRNA sequence before and after removal of duplicates using the CD-HIT program

Database	miRNAs related to PCa from Exiqon		All miRNAs from miRBase	
	before removing duplication	After removing duplication	before removing duplication	After removing duplication
Number of miRNAs	156	150	2588	2477

2.7.3 Identification of potential novel miRNAs for Prostate Cancer

2.7.3.1 BLAST for similarity searches

Sequence similarity is an important aspect in sequence alignments. In molecular biology proteins and DNA can be similar with respect to their function(s), their structure or their primary sequence of amino or nucleic acids (Ewens and Grant, 2005). BLAST is widely used for similarity searches of homology among different sequences, to evaluate the function of a sequence based on its similarity to a sequence of known function, homologous sequences have similar structures, frequently they have a similar function as well (Dalal and Atri, 2014).

The result obtained from BLAST was based on the parameters or settings in the command line as described in section 2.6.2.1. The results were displayed as a text file with the information of similarity such as (a) mismatching, (b) the start and end point of similarity in the length of the sequence, and (c) the percentage of similarity indicated. After scanning, the number of sequence similarity between the query dataset and reference dataset, there were 258 similar sequences which included the 150 PCa miRNAs in both files. The number of PCa implicated miRNAs was further removed from the 258 similar sequences leaving a 108 sequences.

2.7.3.2 CD-HIT

The clustering of miRNAs has been proposed for relating the function of miRNAs, since miRNAs in gene clusters or families may have functional relationships via regulating or co-ordinately regulating biological processes (Guo *et al.*, 2014). Studies showed the first evidence of a cluster of miRNAs that could

regulate functionally related proteins in cancer (Dews *et al.*, 2006; Xu and Wong, 2008).

MiRNA clusters in this study obtained from CD-HIT were 12 clusters including 16 miRNAs. The clusters were created based on the calculation of their sequence similarity with a specific threshold. Only two clusters showed reverse complementary with minus sight next to percentage of similarity indicated which leaves only 14 unique miRNAs.

The common results from BLAST and CD-HIT programs showed that 16 miRNAs had 90 % similarity and above and were considered as potential PCa miRNAs for the diagnostic and prognostic purposes, since their similarity was with miRNAs linked to PCa and carried forward for further analysis.

Table 2. 3: The numbers of miRNA clusters with their percentage of similarity as calculated by the CD-HIT and BLAST programs. (The highlighted entry indicates the PCa linked miRNA and the entry designated by the alphanumeric letters A, B and C its homologous sequence).

No.	MicroRNAs clusters	Percentage of similarity
1.	23nt,>hsa-miR-1 23nt,>hsa-miR-1A	 at +/95.65%
2.	23nt,>hsa-miR-2 23nt,>hsa-miR-2A	 at +/95.65%
3.	23nt,>hsa-miR-3 23nt,>hsa-miR-3A	 at +/91.30%
4.	23nt,>hsa-miR-4 23nt,>hsa-miR-4A	 at +/91.30%
5.	23nt,>hsa-miR-5 23nt,>hsa-miR-5A	 at +/91.30%
6.	22nt,>hsa-let-6 22nt,>hsa-let-6A 22nt,>hsa-let-6B	 at +/90.91% at +/90.91%
7.	22nt,>hsa-miR-7 22nt,>hsa-miR-7A	 at -/95.45%
8.	22nt,>hsa-miR-8 22nt,>hsa-miR-8A	 at +/95.45%
9.	22nt,>hsa-miR-9 22nt,>hsa-miR-9A 21nt,>hsa-miR-9B	 at +/90.91% at +/90.48%
10.	22nt,>hsa-miR-10 22nt,>hsa-miR-10A 18nt,>hsa-miR-10B 22nt,>hsa-miR-10C	 at +/90.91% at +/94.44% at +/90.91%
11.	22nt,>hsa-miR-11 21nt,>hsa-miR-11A	 at +/90.48%
12.	22nt,>hsa-miR-12 22nt,>hsa-miR-12A	 at -/95.45%

2.8 Text mining

Text mining was performed using different publicly available databases such as PhenomiR, Mircancer, and PubMed. PhenomiR is a database providing information about the regulation and expression of miRNA in different diseases, by using the name of a miRNAs or disease type in the query option. All the

potentially novel miRNAs were used as keywords to search if they were related to PCa. Six out of the 14 miRNAs were associated with PCa whether down or up-regulated and were excluded from the candidate miRNA list. Mircancer provides information about miRNAs and their association with a variety of cancers. MirCancer was used to confirm the implication of the candidate miRNAs in PCa, by using the miRNA name as input in the query option. Only three miRNAs were associated with PCa with reference articles provided. PubMed was used to access all the available articles to confirm the involvement of the miRNAs.

From the candidate list, 5 putative miRNAs was shorted listed that showed no experimental link to PCa (see table 2.4). These miRNAs were carried forward for further analysis in the subsequent chapter to confirm their use in diagnostics and prognostics of PCa.

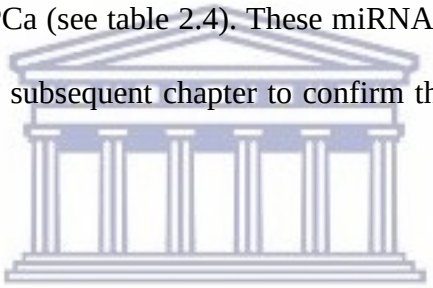


Table 2. 4: List of experimentally validated miRNAs and candidate miRNAs based on text mining

	Candidate MiRNA	PhenomiR	MiRcancer	Pubmed article ID
1.	Hsa-miR-1A	Linked to PCa	Not linked to PCa	17616669
2.	Hsa-miR-2A	Linked to PCa	Linked to PCa	25065599
3.	Hsa-miR-3A	Not linked to PCa	Not linked to PCa	Candidate
4.	Hsa-miR-4A	Not Appear	Linked	26813459
5.	Hsa-let-6A	Linked to PCa	Not Appear	17616669
6.	Hsa-let-6B	Linked to PCa	Not Appear	16192569
7.	Hsa-miR-7A	Linked to PCa	Not Appear	16192569
8.	Hsa-miR-8A	Linked to PCa	Not Appear	16192569
9.	Hsa-miR-9A	Not appear	Not appear	candidate
10.	Hsa-miR-10A	Not linked	Not linked	candidate
11.	Hsa-miR-10B	Not linked	Not linked	20473869
12.	Hsa-miR-11A	Not linked	Not linked	candidate
13.	Hsa-miR-11B	Not linked	Not linked	candidate
14.	Hsa-miR-12A	Linked to PCa	Linked to PCa	16192569

2.9 Discussion

It has been shown that miRNAs are involved in carcinogenesis through the regulation of gene expression (Cinpolat *et al.*, 2017). The recognition of differentially expressed miRNAs in a tumour, as well as normal tissues, may help to establish the pathogenic role of miRNAs in cancer (Di Lisio *et al.*, 2012). A previous study has demonstrated the presence of miRNA differential expression profiles in cancerous tissues as compared to normal tissues (Schetter *et al.*, 2008). These cancer types include lung cancer, leukemia, brain cancer, prostate, and breast cancer (De Bock *et al.*, 2010). The value of miRNA profiles in tumour diagnostics is well-established. For example, 16 miRNAs have been shown to be strongly down-regulated in breast cancer (Aqeilan, Calin and Croce, 2010).

In this study, potentially novel miRNA biomarkers for PCa were identified using similarity search methods between PCa experimentally validated miRNAs and all miRNAs deposited in the miRBase database.

A list of the experimentally validated miRNAs was generated from the Exiqon database containing 156 miRNAs which by given name of the miRNA and the reference articles showed their involvement in PCa. All the mature miRNA sequences were collected using the miRBase database by using the name of individual miRNAs as a keyword. After the search, the mature and precursor sequence was obtained, and only the mature miRNAs were considered for the study. The reference dataset was collected from miRBase and included 2588 miRNAs sequences all deposited for *Homo sapiens*. This dataset was used as a reference dataset to scan the experimentally validated dataset against to identify potentially new miRNAs for PCa. Overlapping (redundancy) data could lead to an

undesirable result file. Therefore, it was essential to remove duplications from both datasets (the reference as well as query dataset). The CD-HIT program was used to remove duplication from the datasets with the resulting file for the PCa experimentally validated miRNAs containing 150 unique miRNA sequences and the reference file containing 2477 unique miRNA sequences.

Local BLAST was used to increase the robustness of the search for similarity. Two steps in BLAST based searches were employed. First, a database was built by using all the mature miRNAs which included 2477 sequences. Second, the PCa experimentally validated miRNAs was scanned against this database with the results file containing 258 miRNAs indicating the percentage of similarity between the sequences. This file also included the 150 miRNAs that showed a link to PCa with the resulting file containing 108 miRNAs after removal of the 150 miRNAs.

CD-HIT was used to find the similarity between the miRNAs from the query and reference datasets (db1 and db2). The results obtained included 12 clusters and contained 16 miRNAs. The cluster contains the miRNAs that are linked to PCa as well as its sequence homolog not linked to PCa with the understanding that these two miRNAs are clustered based on their sequence similarity and can potentially have a similar function(s) and/or participate in similar bioprocesses.

To minimize the amount of results in this study, two criteria were used: (i) common results from both programs, (ii) only miRNAs with equal to and more than 90 % similarity were considered as potential miRNAs for diagnostic and prognostic purposes of PCa.

In conclusion, 16 miRNAs were clustered from CD-HIT as well as BLAST. Two miRNAs were excluded from this study as they were reverse complement sequences and 14 miRNAs were text mined in three different databases to ensure they are not linked to PCa. The five miRNAs that showed no linked to PCa, however, sharing sequence similarity to miRNAs implicated in PCa were selected for further analysis as explained in the subsequent chapters.

Chapter 3

Identification of target genes in prostate cancer for candidate miRNAs using an *in silico* approach

3.1 Introduction

Prostate cancer is a major health problem worldwide, with one of the main causes being epigenetic (Sharma, Kelly and Jones, 2009). Epigenetic mechanisms appear to be responsible for a significant proportion of the alterations in cancer (Riley and Anderson, 2011). A gene is the basic functional unit of DNA which expresses to form the primary structure of proteins. In neoplasia, there is a profound alteration in the patterns of DNA methylation as well as in the structural composition of chromatin. In studies based on the organisational behaviour of chromatin involved in the modulation of gene transcription, the significance of the epigenetic mechanisms present in the initiation and progression of human cancers has been further highlighted (Jones and Baylin, 2002; Wille, 2017). These epigenetic changes include aberrant promoter hypermethylation along with inappropriate gene silencing, which virtually influences each step in tumour progression (Jones and Baylin, 2002).

The current literature available on tumour-profiling studies facilitates the comprehensive comparison of the expression patterns of genes associated with the cell-cycle of different tumour types. These studies have shown that the 'proliferation signature' of the genes whose expression directly correlate with cell-cycle status, tumour mitotic grade as well as their doubling times, were

observed to be the most conserved patterns of expression in cancerous cells (Whitfield *et al.*, 2006). However, the exact number of genes comprising this 'proliferation signature' often varies, but usually, these are the genes that are involved in the vital process of cell proliferation. The understanding of this 'proliferation signature' can be used to improve the understanding of the cell cycle involvement in cancer pathogenesis and can be utilised as a biomarker for cancer diagnosis as well as prognosis (Whitfield *et al.*, 2006).

In addition, genes are regulated by miRNAs which are involved in cellular processes such as differentiation, apoptosis and morphogenesis. miRNAs perform regulation in negative reciprocal feedback loops, in association with protein co-factors that govern signal transduction triggered cell fate decisions (Carthew, 2006). miRNAs are small, non-coding, endogenous RNA molecules that play an important role in a variety of normal and diseased biological processes through post-transcriptional regulation. They can bind to target mRNA transcripts of protein-coding genes and negatively control their translation or cause mRNA degradation (Zheng *et al.*, 2013).

In recent decades, the sophistication of bioinformatics algorithms increased the accuracy regarding the identification of the genes regulated by miRNAs (ReyesHerrera and Ficarra, 2012). The *in silico* methods can efficiently identify the functionalities of genes which are significant in the processes of diagnosis and the prognosis of therapies against a diverse number of cancers (ReyesHerrera and Ficarra, 2012). A variety of bioinformatics databases such as MirDIP, TiGER, DAVID, and STRING were designed for the prediction of miRNA target genes, their functions and interactions, as well as their expression across a variety of

tissues, be it normal or diseased (Huang *et al.*, 2007; Liu *et al.*, 2008a; Shirdel *et al.*, 2011). In this chapter, an array of online databases was used to identify and functionally annotate the target genes for the five miRNAs as potential biomarkers for PCa, identified in Chapter 2.

3.2 Use of available databases for target gene identification

Several online databases, which facilitate data retrieving and analysis, were used to identify the miRNA target genes as well as their involvement in PCa. The details on each database are explained in the subsequent sections.

3.2.1 Use of mirDIP microRNAs Data Integration Portal

MirDIP (http://ophid.utoronto.ca/mirDIP/) is a publically available repository, which integrates the latest information regarding miRNA target predictions, collected from 11 distinct source prediction databases (Stark *et al.*, 2003; Betel *et al.*, 2008). MirDIP combines the expression data from three miRNA databases, namely, MiRanda, TargetScan and picTar, providing a more efficient way than other available methods for users to control the prediction data (Nam *et al.*, 2008). The flexibility in target gene prediction allows the user to choose the appropriate combination of the 11 databases suitable to their query. Furthermore, mirDIP also integrates NAViGaTOR (Brown *et al.*, 2009). Which offers scalable network analysis as well as a visualisation system for viewing the novel miRNA: target predictions (Shirdel *et al.*, 2011). MirDIP predictions in this study were based on eleven criteria as listed below.

(1) The search is primarily based on seed sequence matches that allows base-pair mismatches as well as G: U wobbles, which is significant for miRNA binding. This prediction protocol identifies a high degree of complementarity between the miRNA's 5′ end as well as the 3′ end of the mRNA target sequence, with specific attention to the seed region which is 2–8 bases long from the 5′ end of the miRNA (Wuchty *et al.*, 1999).

(2) Other criteria included conservation of the binding sequence of miRNA to the mRNA target, with a more reliable prediction being made if there is a high level of conservation to the target site. This step was considered as the 'Filtering Step' since the bias in the results will be reduced by using the predictions with as well as without conservation measures (Miranda *et al.*, 2006; Kertesz *et al.*, 2007; Shirdel *et al.*, 2011).

(3) The Free Energy (ΔG) of miRNA: mRNA duplex formation was also considered and calculated using the Vienna Folding package which estimates the energy essential for the formation of the miRNA: mRNA duplex from an entirely dissociated state (McCaskill, 1990).

(4) Another energy-based parameter, i.e, site accessibility is measured in ΔΔG and compares the energy required for the unfolding of a folded 3′ UTR to allow the accessibility of a miRNA to its target site as well as refolding into the miRNA: mRNA duplex (Shirdel *et al.*, 2011).

(5) In order to evaluate the dose-dependent effects of microRNAs on target gene expression, the contribution of its multiple binding sites in the 3′ UTR of a particular gene was also considered. Co-operativity with these binding sites

results in synergistic gene repression (Doench and Sharp, 2004). Previous studies have shown that the distance of novel inter-binding sites is present between 8–40 bp (Sætrom *et al.*, 2007).

(6) Local ALU content is the segments of duplicate DNA interspersed within the human genome, although it is generated by retro-transposons it is named after the enzyme *Alu1*, which is responsible for its cleavage (Batzer and Deininger, 2002). It is used in TargetScan to measure context score and enrichment of A or U base-pairs in the 30 nucleotides up or downstream of the microRNA binding site in the 3′ UTR. Furthermore, it tends to be favourably associated with repression of the target gene's expression (Lewis, Burge and Bartel, 2005; Shirdel *et al.*, 2011).

(7) Local mRNA sequence. The consideration of sequence nearby the microRNA binding site on the 3′ UTR is sometimes considered. Algorithms may examine sequence content for specific nucleotides or examine local sequence effect on site accessibility (Lewis, Burge and Bartel, 2005; Grimson *et al.*, 2007).

(8) Ribosomal shadow is also considered in miRNA target prediction and is taken into consideration by the TargetScan database. It has been hypothesised that repression by miRNAs is caused by a ribosomal shadow effect (Grimson *et al.*, 2007).

(9) The use of miRanda, the first miRNA alignment algorithm, which is similar to the Smith-Waterman algorithm for sequence alignment and uses the same rules of thumb previously established for sequence alignment (Enright *et al.*, 2003; John *et al.*, 2004).

(10) Positional effect. Positional effect rewards miRNA target sites that fall within the first or final quartile of the 3′ UTR after the stop codon (+15 base-pairs), near to the poly (AAAA) tail. This effect is more noticeable in long UTRs (Grimson *et al.*, 2007).

(11) 3′ Pairing. Aside from strong seed region pairing, miRanda also requires nucleotide binding between the microRNA and the target mRNA between bases 12–17 of the 3′ end of the microRNA for miRNA target prediction (Grimson *et al.*, 2007).

3.2.2 Tissue-specific Gene Expression and Regulation

TiGER (Tissue-specific Gene Expression and Regulation) obtainable at http://bioinfo.wilmer.jhu.edu/tiger/ is a database that contains tissue-specific gene expression profiles or expressed sequence tag (EST) data, cis-regulatory module (CRM) data, and tissue-specific transcription factor interaction data in 30 human tissues for each gene contained within the database (Liu *et al.*, 2008b). At present, the database contains expression profiles for 19,526 UniGene genes, combinatorial regulations for 7,341 transcription factor pairs and 6,232 putative CRMs for 2,130 RefSeq genes (Liu *et al.*, 2008).

The gene expression pattern for each UniGene is calculated based on the NCBI EST database. TiGER has identified and catalogued 7261 tissue-specific genes for 30 human tissues based on their expression enrichment and statistical significance. Thus, on average, each tissue expresses approximately 290 tissue-specific genes (Yu *et al.*, 2007).

In addition to EST data, TiGER also identifies transcription factors (TFs) based on patterns of co-occurrence of pairs of DNA binding sites (Yu *et al.*, 2007). TiGER predicts 9060 tissue-specific TF interactions, around 300 for each tissue (Yu *et al.*, 2006). The database uses known interactions as positive controls attributable to the scarcity of tissue-specific interactions to evaluate these results.

Cis-regulatory modules (CRMs) are a stretch of DNA usually about 100-1000 base pairs in length where several transcription factors can bind (Istrail and Davidson, 2005). They are the central cis-elements that control gene expression (Istrail and Davidson, 2005). TiGER calculates the interaction strength between two transcription factor (TF) binding sites and then derives an empirical 'potential energy' for each TF-binding site. This results in energy profiles for the promoter sequences of tissue-specific genes. An energy level less than -1 indicates the existence of a TF module (Liu *et al.*, 2008b).

This development of computational methods for tissue-specific combinational gene regulation, based on transcription factor binding sites, CRMs and ESTs enables the platform to perform a large-scale analysis of tissue-specific gene regulation in human tissues (Liu *et al.*, 2008).

3.2.3 Intersection analysis by Venn diagrams

Venn diagrams tool accessible at http://bioinformatics.psb.ugent.be/webtools/Venn/ is used to calculate the intersection between lists of elements or data. The tool generates a textual output indicating which elements are at each intersection or are unique to a certain list. The user has a choice between symmetric and non-symmetric Venn diagrams.

Currently, the user can calculate the intersections of a maximum of 30 lists (Draw Venn diagram, n.d).

3.2.4 Functional characterisation of genes via DAVID

DAVID (Database for Annotation, Visualization and Integrated Discovery), is a database that has several features and is publicly available at (https://david.ncifcrf.gov/). DAVID is a high-throughput annotation tool that systematically maps a large number of interesting genes to a list of associated Gene Ontology terms and then statistically highlights genes that are over-enriched for those terms (Ashburner *et al*., 2000). This increases the possibility that the researcher will identify the involvement of the gene of interest in different disease states and another biological process (Huang *et al.*, 2007). The annotation tool in DAVID provides several gene annotation options including; LocusLink, Unigene, GenBank, Gene Symbol, RefSeq Gene Name, OMIM, Affymetrix description, Summary and Gene Ontology. Each of these tools can be used for various reasons or functions, In this study, the Gene disease associated tool will be of interest for the identification of the genes that are involved in our disease of interest (Huang *et al.*, 2007).

3.2.5 Gene/protein interaction analysis via STRING

The STRING (https://string-db.org/) database (Search Tool for the Retrieval of Interacting Genes/Proteins) aims to provide comprehensive, yet a quality controlled collection of protein-protein associations for several organisms. The associations are derived from high throughput experimental data, from the mining of databases and literature, and from predictions based on genomic context

analysis with version 10.0 covering 2031 organisms (Szklarczyk *et al.*, 2015). The basic interaction unit in STRING is the functional association, i.e, a specific and productive functional association between two proteins, likely contributing to a common biological purpose. Interactions are derived from multiple sources, (i) known experimental interactions are imported from primary databases; (ii) automated text-mining is applied to uncover statistical and/or semantic links between proteins, based on Medline abstracts and a large collection of full-text articles; (iii) pathway knowledge is parsed from manually curated databases; and (iv) interactions are predicted *de novo* by a number of algorithms using genomic information (Szklarczyk *et al.*, 2015).

3.2.6 KEGG pathway analysis

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a knowledge base for systematic analysis of gene functions, linking genomic information with higher order functional information access at (https://david.ncifcrf.gov/). The genomic information is stored in the GENES database, which is a collection of gene catalogues for all the complete sequenced genomes and some partial genomes with up-to-date annotation of gene functions. The higher order functional information is stored in the PATHWAY database, which contains graphical representations of cellular processes, such as metabolism, membrane transport, signal transduction and cell cycle. The PATHWAY database is supplemented by a set of ortholog group tables for information about conserved sub-pathways (pathway motifs), which are often encoded by positionally coupled genes on the chromosome and which are especially useful in predicting gene functions. The

third database in KEGG is LIGAND for the information about chemical compounds, enzyme molecules and enzymatic reactions. KEGG provides Java graphics tools for browsing genome maps, comparing two genome maps and manipulating expression maps as well as computational tools for sequence comparison, graph comparison and path computation (Kanehisa and Goto, 2000).

3.3 Aim and objectives

The aim of this chapter is to identify the gene targets of the five miRNAs previously identified and to link these genes to PCa. Thus, linking their regulating miRNAs to PCa indirectly since these short-listed miRNAs showed no link to PCa (see Chapter 2).

I. To identify the target gene for the candidate miRNAs using the mirDIP database;
II. Extraction of all genes expressed in prostate tissue based on expression sequence tag (EST) data using the TiGER database;
III. Functional annotation of the target genes to PCa using the DAVID database;
IV. Perform Pathway analysis of the target genes using KEGG; and
V. Gene interaction network analysis of the target genes via STRING.

3.4 Methodology

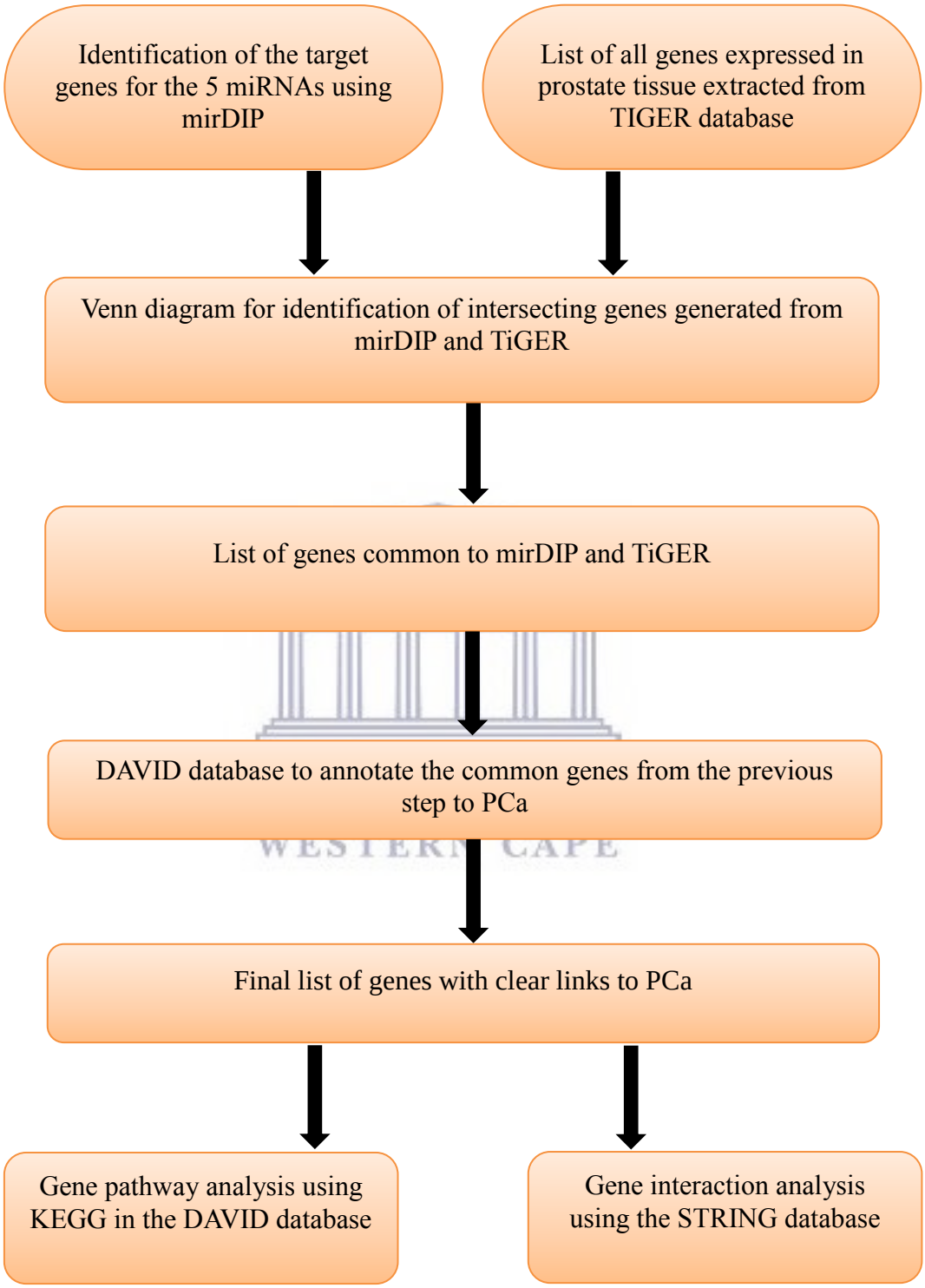


Figure 3.1: Outline of the *in silico* methodology for prioritising of the miRNA target genes

3.4.1 Use of mirDIP microRNAs Data Integration Portal

The mirDIP database was accessed at (http://ophid.utoronto.ca/mirDIP/) on 15/03/2017, for the identification of target genes for each of the miRNA clusters. The previously identified four clusters containing the five miRNAs were submitted separately to the mirDIP database to identify common target genes for each cluster. The database was used with the default settings.

3.4.2 TiGER to extract prostate tissue expressed genes

TiGER was used to extract all the genes that are expressed in prostate tissue by selecting "prostate tissue" as a keyword in the query section. TIGER produce three subsets of data based on three principles, (i) Expressed Sequence Tags (ESTs); (ii) Cis-Regulatory Modules (CRMs); and (iii) Combinatorial Gene Regulation. The expressed sequence tags module was selected for this study. A list of genes expressed in prostate tissue was compiled and saved for further analysis.

3.4.3 Intersection analysis by Venn diagrams

Venn diagrams were used to calculate the intersection between the lists of genes obtained in sections 3.4.1 and 3.4.2 by generating a textual output indicating which genes were in each intersection or were unique to a certain list. This online tool was used to see the intersection between the four lists of target genes generated by mirDIP (representing the four miRNA clusters) with the list of the genes generated by TiGER (prostate specific genes). All the intersections were

submitted at the same time, and the results are shown in figure 3.2. The intersecting genes from each cluster were saved in a separate file.

3.4.4 Involvement of genes in prostate cancer using the DAVID database

DAVID database was accessed at https://david.ncifcrf.gov/ to functionally annotate the miRNA target genes and was used with the following steps, (i) In "Enter Gene List" option, four lists of genes (list of genes from each miRNA cluster) were pasted individually in the query box; (ii) Under the "select identifier" option, official gene symbol was selected; (iii) "List of genes" was selected followed by selection of the "submit" option. On the next page several options are available, the disease option were selected after which GRD_DIESEASE was selected. The chart output places the genes in various cancer processes with special emphasis on PCa for this study. The analysis generated lists of genes involved in PCa and was saved in an Excel file (see table 3.3).

3.5 KEGG Pathway analysis of the genes in prostate cancer

KEGG pathway analyses were carried out in DAVID, Bioinformatics Resources at http://david.abcc.ncifcrf.gov (Huang, Lempicki and Sherman, 2009). The List of genes previously shortlisted was submitted to the DAVID database for pathway analysis using the KEGG functionality within DAVID. Results were given as pathway graphs for the genes that are involved in known PCa processes as shown in figure 3.3.

3.6 Analysis of Gene/protein interaction network

Gene IDs for a list of genes identified by DAVID implicated in PCa were used as input for the generation of gene networks using the STRING database Version 10 (Meiring, 2003; Franceschini *et al*., 2013). First, the genes were used as input to generate an expression network among each other. To produce this expression network, parameters were chosen as follows, (i) a confidence level of 0.7 and (ii) a network depth of 4. Second, an extended network was produced to determine which of the genes showed links to known genes involved in cancer and specifically in PCa using (i) a confidence level of 0.7 and (ii) a network depth of 5. The results are given as a 'stick/line and ball' graphic with the 'balls' (proteins/genes) connected by different colour lines with each colour indicating the type of evidence for the connection between each protein/gene.

3.7 Results

3.7.1 MirDIP miRNAs target gene prediction

The four miRNA clusters were used as input in mirDIP database as explained in section 3.2. Four lists of miRNAs target genes were generated for these miRNA clusters which included five miRNAs. The common target genes resulting from each cluster were filtered. The function(s) of gene targets is regulated by more than one miRNA indicating that these genes are functionally similar and share common features closely related to various cancers such as apoptosis, cell proliferation, transcription regulation and diverse cancer-related pathways (Liu *et al.*, 2016). From the result, a total of 56 339 target genes were identified for the five miRNAs sharing 42 731 genes amongst themselves (see table 3.1)

Table 3.1: The number of target genes in each miRNA cluster and common target genes produced by mirDIP

Cluster	MicroRNA in each cluster	Number of target genes	Number of common target genes
Cluster 1	Hsa-miR-1 Hsa-miR-1A	13 575	12 952
Cluster 2	Hsa-miR-2 Hsa-miR-2A	11 114	10 555
Cluster 3	Hsa-miR-3 Hsa-miR-3A	15 864	9 626
Cluster 4	Hsa-miR-4 Hsa-miR-4A Hsa-miR-4B	15 786	9 598
Total number of target genes		56 339	42 731

3.7.2 TiGER database

A list of genes expressed in prostate tissue was collected from the TiGER database by using "prostate tissue" as a keyword. Based on the EST database a number of genes (201) was collected as displayed in table 3.2, and the data was saved in an excel file. Each of the genes collected from TiGER showed a relatively high expression in prostate tissue. After removal of duplications, 148 unique genes remained.

Table 3.2: Shows the number of genes expressed in prostate tissue as determined by TiGER

Keyword	List of genes from TiGER	List of genes after removing duplications
Prostate tissue	201	148

3.7.3 Venn diagram of intersectiing genes

Venn diagrams were generated by using the online tool representing the intersection between the genes expressed in prostate tissue and the target genes. The intersecting genes for cluster one numbered 66 genes, for cluster two 47 genes, for cluster three 46 genes, and for cluster four 53 genes as shown in figure 3.1.

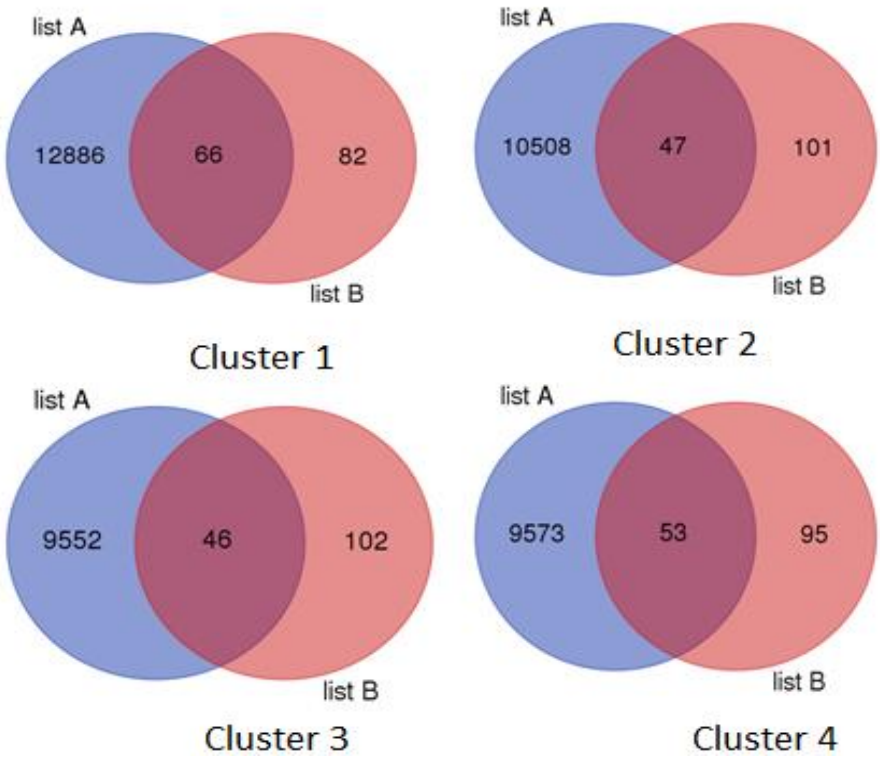


Figure 3.2: Venn diagrams for representing the number of intersecting genes for the four miRNA clusters. List A: target genes generated by mirDIP and List B: genes expressed in prostate tissue generated by TiGER.

3.7.4 DAVID database result involving genes in prostate cancer

As mentioned in section 3.5.4, the DAVID database was used to confirm the link of the target genes in PCa. The list of intersecting genes for each cluster was submitted to DAVID with the number of genes functionally annotated to PCa indicated in table 3.3 several of the genes showed enrichment for a functional

annotation to PCa across the clusters with 12 genes being unique to the four clusters.

Table 3.3: Number of genes from intersection submitted to DAVID

List of genes from intersection	Number of intersecting genes	List of genes annotated to PCa using the DAVID database
Cluster 1	66	10
Cluster 2	47	9
Cluster 3	53	10
Cluster 4	46	8
List of unique genes		12

Table 3.4: The list of gene names for each cluster annotated to PCa

Genes from Cluster 1	Genes from cluster 2	Genes from cluster 3	Genes from cluster 4
NKX3-1	*NKX3-1*	*NKX3-1*	*NKX3-1*
AMACR	*AMACR*	*AMACR*	*AMACR*
AR	*AR*	*AR*	*AR*
CYP24A1	*IGF1*	*CYP24A1*	*CYP24A1*
IGF1	*KLK2*	*IGF1*	*IGF1*
KLK4	*KLK3*	*KLK2*	*MYH11*
MYH11	*KLK4*	*MYH11*	*PTGS2*
PTGS2	*PTGS2*	*PTGS2*	*SRD5A2*
SRD5A2	*TMPRSS2*	*SRD5A2*	
TMPRSS2		*TMPRSS2*	

3.7.5 KEGG pathway analysis through DAVID database

The list of genes that are involved in PCa as determined by DAVID for all clusters was combined and include the following genes, *AMACR, AR, CYP24A1, IGF1, KLK2, KLK3, KLK4, MYH11, NKX3-1, PTGS2, SRD5A2* and *TMPRSS2*. To determine if any of these genes were involved in PCa pathways all of the

genes were submitted to the KEGG functionality within the DAVID database. Five out of the twelve genes were implicated in known PCa pathways using KEGG, viz *NKX3-1 AR*, *IGF1*, *KLK3*, and *SRD5A2* with KEGG ID 05215 as shown in figure 3.3

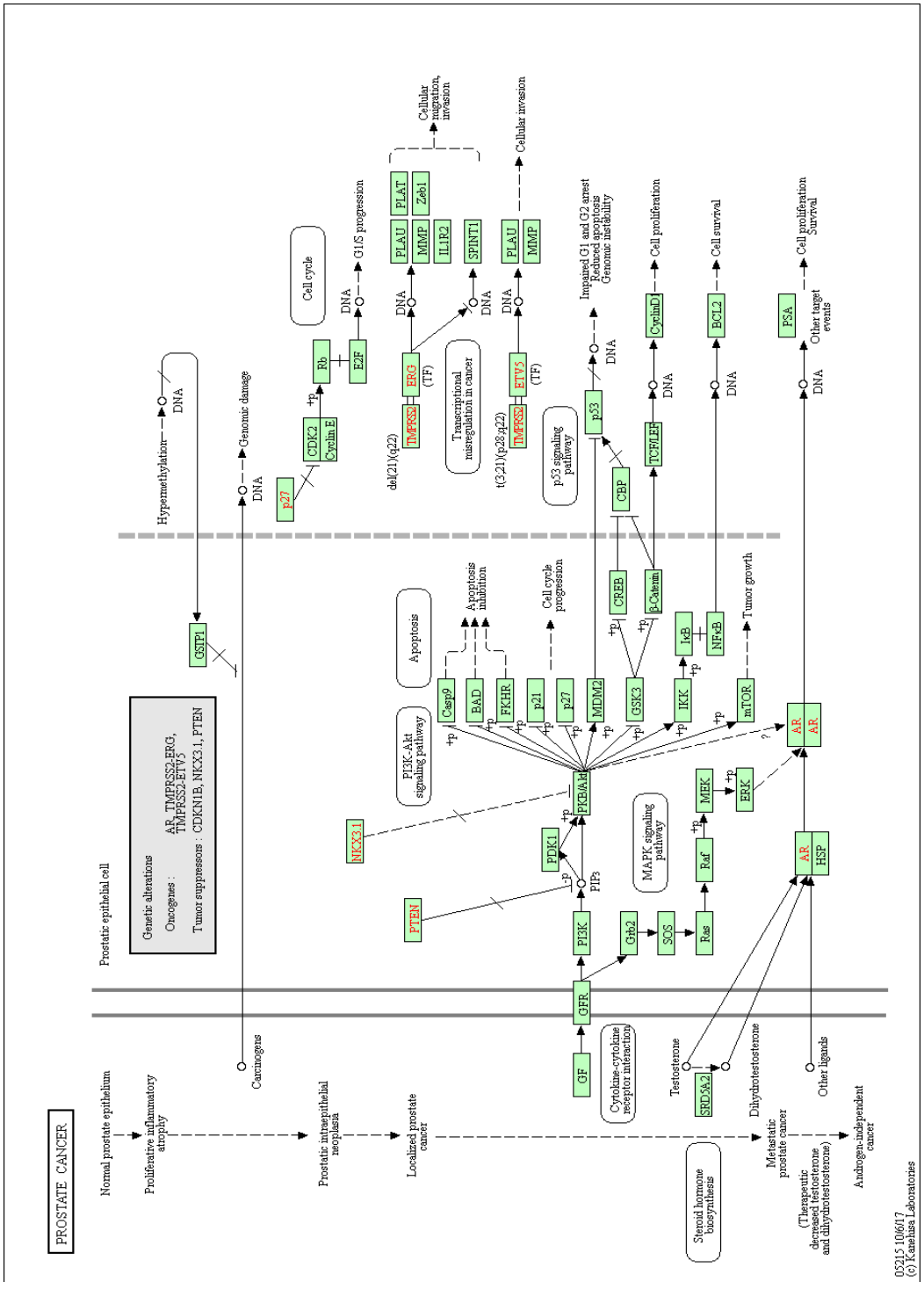


Figure 3.3: Shows the implication of 5 genes (*AR*, *SRD5A2*, *IGF1*, *KLK3*, and *NKX3-1*) out of 12 genes in PCa pathways with the genes highlighted in red. Link for the graph can be obtained at (https://david.ncifcrf.gov/kegg.jsp?path=hsa05215$Prostate%20cancer&termId=550028904&source=kegg)

3.7.6 Gene interaction analysis using STRING database

The twelve genes were submitted to STRING to generate an interaction network amongst themselves as described in section 3.6. The result indicates that there are several lines of evidence for the interaction between 5 out of 12 genes as shown in figure 3.4. From the graphic result, the interactions between *AR* and *TMPRSS2*, *AR* and *KLK3*, *AR* and *NKX3-1* and *KLK2* and *AR* are supported by several lines of evidence which includes experimentally determined, from curated databases and text mining. There was no evidence to link *KLK4*, *CYP24A1* and *PTGS2* to the network. This could be due to the fact that information for these genes was not available at the date of analysis *SRD5A2* did not show up at all in the network, which could also be due to a lack of information on this gene. However, it was placed in known PCa pathways using KEGG.

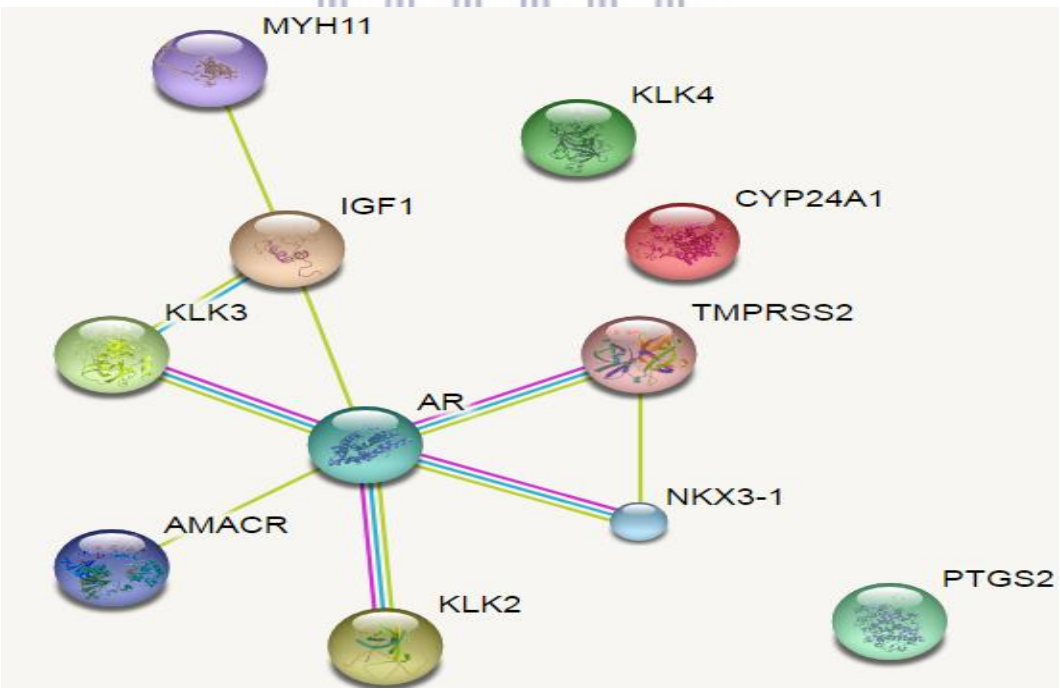


Figure 3.4: Protein Network Visualization generated by STRING. The interactions of the 11 miRNA targeted genes clustered together. The genes are represented by the nodes and the different line colours represent the types of evidence for the interaction. Indicate the evidence and the colours.

The second result from STRING for the miRNA targeted genes are depicted in figure 3.5. From the results, five additional genes, namely *IGFBP3*, *HSP90AA1*, *IGFBP1*, *CTNNB1*, and *NCOA2* showed interaction with the miRNA target genes, with the lines of evidence equally convincing as with the generation of the first interaction network.

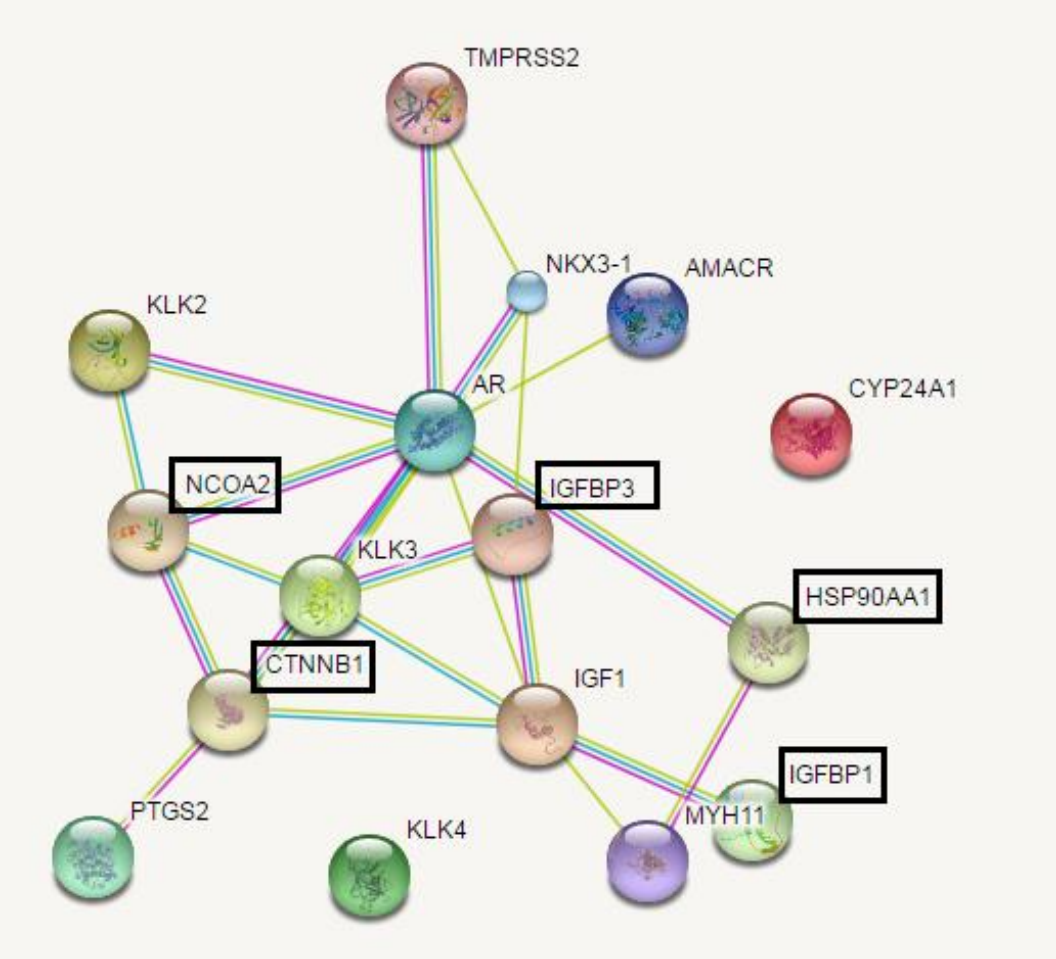


Figure 3.5: Interaction network between the genes and other genes of interest using STRING. The genes in the outlined box were associated with target genes with their evidence.

3.8 Discussion

Genes play a critical roles in different diseases including cancer and they are involved in different cellular processes such as cell proliferation, apoptosis, differentiation, migration and maturation, with all these processes being involved in the performation of tumours (Thompson, 1995; Chen and Lai, 2009) Understanding how genes are expressed and regulated in various tissues under various conditions can help clarify the molecular mechanisms of tissue development and their functions in cancer (Liu *et al*., 2008). The changes in gene expression can be identified or evaluated by both molecular and *in silico* gene expression analysis for diseased and normal states. Gene expression profiling allows for the sub-classification of tumours by providing diagnostic and prognostic information of genes that are differentially expressed in tumours as well as their regulating miRNAs (Thomas *et al*., 2013).

In this chapter, various databases were systemically used to identify the target genes for the five miRNAs identified in the previous chapter. Five potentially novel miRNAs were identified as diagnostic and prognostic biomarkers for PCa using various *in silico* methods.

To understand how these miRNAs are associated with PCa, examination of their target gene regulation was carried out using *in silico* approaches.

The mirDIP database was employed to identify the target genes of the five miRNAs. The prediction of these genes was based on the eleven criteria stipulated in section 3.2.1. From the miRDIP target gene prediction analysis, a total of 56,339 target genes were identified for the five miRNAs divided into four

clusters. After the removal of duplications, the clusters shared 42 731 genes. The number of genes was expected since a single miRNA can regulate the function of thousands of genes. This regulation however is determined by certain stimuli, and physiological conditions, tissue specificity and various other biological factors.

The next step was to download all genes expressed in prostate tissue with the rationale that if a particular gene is expressed in prostate tissue and it is regulated by a miRNA with known links to PCa, the potentially novel PCa miRNAs might function in a similar biological process and regulate the expression of known PCa associated genes. From the TiGER database, a list of 148 genes was collected by using "prostate tissue" as the search keyword. The list from TiGER and the four lists from mirDIP were used to performed intersection analysis using Venn diagrams. The intersections were drawn online between the miRNA target genes from mirDIP for each cluster and the genes expressed in prostate tissue from TiGER. The intersecting genes numbered 66 genes for the first cluster, 47 genes for the second cluster, 46 genes for the third cluster, and 53 genes for the fourth cluster.

All the genes from each cluster highlighted in the intersection analysis were submitted to DAVID. DAVID is a functional annotation database to confirm the involvement of these genes in different types of diseases based on the experimental result. For this study, PCa was selected as the disease category with a few number of intersecting genes from each cluster implicated in PCa as per the results of DAVID.

From cluster one containing 66 genes, 10 genes were implicated in PCa, cluster two 9 out of 47, cluster three 10 out of 46 genes and cluster four 9 out of 53 genes were implicated within PCa. These lists of the genes were extracted, combined and duplications removed resulting in a list of 12 unique genes as targets for the five miRNAs.

To further support these genes as implicated in PCa, pathway analysis was carried out using KEGG within the DAVID database. The list of 12 genes was submitted to DAVID database for KEGG pathway analysis with five out of 12 genes showing involvement in PCa pathways. *AR, TMPRSS2, KLK3, NKX3-1* and *KLK2* were all placed in PCa pathways using KEGG.

The growth and maintenance of the prostate are reliant on androgen acting through the Androgen Receptor (*AR*). *AR* is important in the development and progression of PCa (Heinlein and Chang, 2004). The *AR* plays an essential role in the expression of male phenotype and alteration of *AR* function is associated with PCa (Tan *et al.*, 2015). The wild-type *AR* can induce cellular differentiation or cellular apoptosis in PCa cells (Fu *et al.*, 2003). Somatic missense *AR* gene mutations have been detected in PCa cell lines (Fu *et al.*, 2003).

In a study, it was shown that *NKX3.1* is a prostate-specific homeobox gene which controls differentiation, epithelial cell growth, and stem cell maintenance (Song *et al.*, 2015). Inadequate levels of *PTEN* and *NKX3.1* lead to a reduction in p27 levels and increase proliferation and decrease apoptosis clearly demonstrating their role in tumour formation (Song *et al.*, 2015).

The result of a study demonstrated that the *SRD5A2* rs9282858 polymorphism was remarkably associated with an increased susceptibility to PCa (Fang *et al.*, 2017). *SRD5A2* is primarily expressed within the prostate and testes and plays a major role in the prostate function and disease aetiology. Thus, *SRD5A2* inhibitors such as dutasteride and finasteride have been developed to manage BPH and potentially prevent PCa (Sissung *et al.*, 2014).

TMPRSS2 protein function in prostate carcinogenesis relies on overexpression of ETS transcription factors, such as *ERG* and *ETV1*, through gene fusion (Yu *et al.*, 2010).Also another study suggest that fusion of *TMPRSS2: ERG* may lead to a more aggressive PCa phenotype and possibly account in part to higher grade PCa (Demichelis *et al.*, 2007).

KLK3 also known as prostate specific antigens is already being used as a diagnostic biomarker for PCa, a member of the fifteen-gene family kallikrein-related peptidase (*KLK*), in blood has revolutionized both the detection and management of PCa (Hong, 2014). *KLK*s have more promising use as diagnostic and prognostic value because of the essential similarities between PSA and other *KLK*s (Hong, 2014).

IGF-1 has been shown to stimulate the proliferation of human prostate epithelial cells (PEC) in culture, and is necessary for normal growth and development of the rat and mouse prostate, the link between high circulating serum *IGF-1* levels and the risk of developing advanced PCa later has been established in epidemiological studies (Roberts, 2004). In a previous study, there is significant evidence of associations between total *IGF-1R* and risk of overall and metastatic recurrence,

and between cytoplasmic or internalised *IGF-1R* and biochemical recurrence, following radical radiotherapy but not surgery for PCa (Aleksic *et al.*, 2017).

STRING database confirmed the interaction between the lists of genes which included 12 genes of which only 11 genes showed interaction as shown in figure 3.4. One gene did not appear in the STRING analysis and there might not be enough information published on this gene. Three genes are involved in Biological Processes mainly in prostate gland development namely *AR*, *IGF*, and *NKX3-1* (Holzbeierlein *et al.*, 2004) and six genes are involved in the Molecular Functions namely *KLK2*, *KLK3*, *KLK4*, *TMPRSS2*, *AR* and *NKX3-1*.

AR was central in the gene interaction network, with three lines of evidence connecting it to its interaction partners namely experimentally determined, from curated databases and text mining. The main function of *AR* is that of a DNA-binding transcription factor that regulates gene expression. Some of the genes were linked only by text mining evidence such as *MYH11*, which was linked to *IGF1* and *AMACR* which was linked to *AR*.

Antibodies to *AMACR* were used in an immunohistochemistry experiment to demonstrate the presence of prostate carcinoma since the enzyme is greatly overexpressed in this type of a tumour, with this study performed on prostate tissue (Evans, 2003).

One gene did not appear in the STRING database analysis namely *SRD5A2*. However, epidemiological and biochemical findings showed an association between PCa and a constitutional (germline) missense substitution in *SRD5A2*, which resulted in the replacement of an alanine residue at codon 49 with threonine

(A49T). This substitution is associated with a significantly increased risk of PCa, probably through increased metabolic activation of testosterone to dihydrotestosterone (Makridakis *et al.*, 1999). The result of interaction confirmed that these genes namely *AR, KLK3, KLK2, IGF1* and *NKX3-1* are highly involved in prostate gland development and any change in these genes might result in the performance of a tumour.

Secondly, the extended interaction analysis using STRING five additional genes namely *IGFBP, HSP90AA1, IGFBP3, CTNNB1*, and *NCOA2* showed interactions with the miRNA target genes as shown in figure 3.5.

The central genes in the first interaction network as shown in figure 3.4 *AR* were linked to two more genes namely *NCOA2, HSP90AA* and *IGFBP-3* showing interaction with *IGF1* and *KLK3* by three lines of evidence. *IGFBP-3* protein levels have been reported to decrease during the progression of PCa from benign to metastatic disease. Although production of the protein does not cease completely, *IGFBP-3* is still made at a lower level by PCa cells and secretes into the surrounding environment. However, instead of the full length functional protein, *IGFBP-3* is found to be cleaved. This decreases the affinity of *IGF* binding to *IGFBP-3*, making the growth factors more likely to bind the *IGF1* Receptor and thus promoting cell survival (Brahmkhatri, Prasanna and Atreya, 2015). In a previous study, it was found that PCa patient samples revealed a strong correlation between *NCoA2*-mediated signalling, PCa recurrence and disease progression (Qin *et al.*, 2014). Deprivation of androgen induces *NCoA2*, which mediates activation of PI3K signalling and promotes PCa metastasis and castration-resistant PCa development (Qin *et al.*, 2014). *IGFBP-3* is anti-

angiogenic and a pro-apoptotic protein in PCa. In epidemiologic studies, it was suggested that low *IGFBP-3* is associated with greater risk of aggressive, metastatic PCa (Mehta *et al.*, 2011).

In conclusion, using a combination of *in silico* tools, target genes for the five miRNAs were identified. Through functional annotation, pathway analysis a list of miRNA target genes was created resulting into 12 genes regulated by five miRNAs.

In the subsequent chapters, both the genes and their regulating miRNAs will be further examined as putative biomarkers for PCa management using various *in silico* tools.

Chapter 4

Use of the identified miRNAs and their target genes as potential prognostic biomarkers for PCa using an *in silico* approaches

4.1 Introduction

Prognostic biomarkers provides information on the likely course of cancer in an untreated individual that are objectively measurable (Italiano, 2011). Recently, the discovery of miRNA has provided new insights into cancer research, revealing the role of miRNAs in various biological processes, and evidence shows that their deregulation in many cancers has prognostic and predictive significant value (Pal *et al.*, 2015). For cancer prognosis, miRNA can be complementary to other genomic and proteomic biomarkers (Lan *et al.*, 2015). The possible applications of miRNAs not only in molecular diagnostics but also in molecular prognostics, particularly in cancer, are provided by the discovery of the role of miRNA in numerous pathological processes.

MiRNAs serve as diagnostic, prognostic and predictive biomarkers for different types of cancer by investigating the expression of miRNAs at a different stage of cancer (Kosaka, Iguchi and Ochiya, 2010).

4.2 MiRNA as a prognostic biomarker

A prognostic biomarker aids in monitoring the progression of a disease and provides information on its likely outcome in an untreated individual (Italiano, 2011). Cancer genomics contributed to prognostic biomarker discovery, which

may include molecular signatures such as polymorphisms, genes, mutations, macromolecules or epigenetic regulators (Goswami and Nakshatri, 2013). MiRNAs serve as diagnostic and prognostic biomarkers through the change of their level of expression by comparison between the normal and the cancer sample (Macha *et al.*, 2014). Since the miRNAs are present in the body fluids such as saliva, serum, plasma and blood, their accessibility is facilitated.

MiRNAs are used as biomarkers in various diseases including many cancers (Tricoli and Jacobson, 2007). It has been reported, through measuring the miRNA expression in non-cancerous and cancerous tissue collected from stage I–III squamous cell carcinoma (SCC) patients without any treatment prior to surgery, an association between high expression of miR-31 and poor survival of SCC were made (Cho, 2013). Decreasing survival rates of patients with small-cell lung cancer is associated with high expression of miR-92a-2* (Cho, 2013).

4.3 Predictive biomarker

A predictive biomarker facilitates the monitoring of patients during treatment to measure the response of the patient to the specific treatment. Additional considerations may apply when evaluating the clinical utility of a predictive biomarker in selecting between two therapy options (Burke, 2016). The emergence of innovative technologies in Genomics, Proteomics, Metabolomics and imaging allows researchers to facilitate inclusive analysis of cancer cells. These approaches have already established its power to differentiate cancer cells from normal cells and to identify specific genetic elements involved in cancer (Nair, Sandhu and Sharma, 2014).

4.4 Using the available prognostic and expression databases.

Several databases are available online to perform prognostic and expression analysis for miRNAs and their target genes. In this study some of these databases were utilised and are explained in the subsequent sections.

4.4.1 MiRNA prognostic databases

4.4.1.1 PROGmiRV2 database

PROGmiR is a tool hosted within the PROGmiRV2 database, with the link http://www.compbio.iupui.edu/progmir. This tool can perform prognostic property analysis for a miRNA in various types of cancer using existing data from Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). This tool can be used to study overall survival implications for approximately 1050 human miRNAs in 16 major cancer types (Goswami and Nakshatri, 2012).

4.4.1.2 SurvMicro database

The SurvMicro database is available at http://bioinformatica.mty.itesm.mx/SurvMicro, is a curated and regularly updated online tool that provides information on miRNA expression and clinical outcome offering survival analysis and risk assessment in cancers (Aguirre-Gamboa and Trevino, 2014). Collection of data is performed using keywords related to cancer for searching clinical outcome, different miRNA platforms and 430 samples. The searches are performed in GEO, GEO Meta DB, Array Express, and TCGA (Aguirre-Gamboa and Trevino, 2014).

Gene target prognostic databases

4.4.1.3 SurvExpress database

SurvExpress is a database available at http:// bioinformatica.mty.itesm.mx/SurvExpress. It provides information about the validation of multi-gene biomarkers for clinical outcomes which is important to understand the monitoring of cancer prognosis. An important source of information for virtual validation is the high number of available cancer datasets (Aguirre-Gamboa *et al.*, 2013). The datasets in SurvExpress are collected from Gene Expression Omnibus (GEO) and the Cancer Genome Atlas (TCGA) (Aguirre-Gamboa *et al.*, 2013). At the input page, a list of genes identifier and a specific dataset can be chosen as specified by the user. SurvExpress uses two methods to generate risk groups. The first method (default) generates the risk groups splitting the ordered prognostic indices (higher values for higher risk) by the number of risk groups leaving an equal number of samples in each group (Aguirre-Gamboa *et al.*, 2013). Secondly, to produce risk groups using an optimisation algorithm from the ordered prognostic index (PI).

The outcome results of the database include common measures and plots such as box plots of gene expression across gene groups together with the *p*-value of the corresponding difference, risk group optimisation plot including a visual association of available clinical information to risk groups was used to assess the performance of the survival data (Aguirre-Gamboa *et al.*, 2013). The database includes 20,000 cancer samples dispersed over 140 datasets covering more than 20 tissues or types of cancer (Aguirre-Gamboa *et al.*, 2013).

4.4.1.4 PrognoScan

PrognoScan available at http://www.compbio.iupui.edu/proggene/ identifies prognostic biomarkers for several cancer types. This tool was designed to generate Kaplan-Meier plots for the genes that the user inputs (Goswami and Nakshatri, 2013). The tool was generated by collecting publicly available data from repositories such as Gene Expression Omnibus (GEO), EBI Array Express and recently developed 'The Cancer Genome Atlas' (TCGA), PrognoScan has 64 datasets for 18 different types of cancer, with this tool being the most comprehensive prognostic biomarker identification tool to date (Goswami and Nakshatri, 2013). The data outputs are provided as different graphical results such as expression plots, expression histograms, *P*-value plots and Kaplan-Meier plots as well as an annotation table (Mizuno *et al*., 2009).

4.4.2 MiRNA and gene target expression databases

4.4.2.1 Database of differentially expressed miRNAs in human cancers (dbDEMC version 2.0)

DbDEMC version 2.0 is a database that can be accessed at http://www.picb.ac.cn/dbDEMC/index.html which is a significantly expanded version. The latest version of dbDEMC 2.0 contains expression data on 2224 miRNAs across 36 cancer types through the processing of 200+ expression data sets. In this version, the volume of data has increased with the search and browser functions retained and improved as well as new features being added. This version of the database simplifies the identification of cancer associated miRNAs and a more in depth investigation of their roles in physiological and pathological

processes of cancer development (Yang *et al.*, 2017). Also, miRNA-seq from TCGA was included in dbDEMC 2.0 which is composed high quality miRNA expression profiles for 22 type of cancers and more than 400 data sets were collected initially (Yang *et al.*, 2017).

The new version of dbDEMC 2.0 includes 209 miRNA expression data sets from 143 peer-reviewed publications and those from TCGA. Presently it holds 49202 miRNA–cancer associations for 2224 differentially expressed miRNAs identified from 436 experiments, and a total of 36 cancer types and 73 cancer subtypes are covered. Also, it includes 86% of the miRNAs identified in the human genome from miRBase Release 21, which is the total number of differentially expressed miRNAs.

Meta-profiling analysis added to dbDEMC 2.0 as new function allows users to perform expression profiling to demonstrate the expression of miRNAs among a specific set of cancer types. A list of miRNAs can be used as input, and one of the six types of experimental designs can be selected as well as the cancer type of interest. The meta-profiling analysis generates a heat-map describing the expression change for the inquired miRNA across multiple cancer types. The up and down-regulated expression status of a miRNA is classified by red and green colours respectively according to a confidence score, which is calculated based on studies of each specific cancer type and experimental evidence supporting the differential expression status (Yang *et al.*, 2017).

4.4.2.2 FIREBROWSE database for gene expression analysis

FIREBROWSE is accessed at http://firebrowse.org underlined by a powerful computational infrastructure, application programming interface (API), graphical tools and online reports. Graphical tools like viewGene to explore expression levels, and iCoMut to explore the comprehensive analysis profile of each TCGA disease studied within a single, interactive figure are novel features of this portal. To date it is one of the most integrative open access cancer datasets with over 80K sample aliquots from 11,000+ cancer patients, spanning 38 unique disease cohorts. Using FIREBROWSE**,** TCGA data and analysis in cBioPortal-expression, mutation, copy number, significance analysis, and other operations can be loaded directly (FIREBROWSE, 2016). All tissues utilized by the database in their abbreviated form are shown in table 4.1

Table 4. 1: List of abbreviation used in graphical result from the FIREBROWSE database

ACC	Adrenocortical carcinoma	SARC	Sarcoma
BLCA	Bladder Urothelial carcinoma	STES	Stomach and esophageal carcinoma
BRCA	Breast invasive carcinoma	PRAD	Pancreatic adenocarcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	UCS	Uterine Carcinosarcoma
CHOL	Cholangiocarcinoma	KIRC	Kidney renal clear cell carcinoma
COAD	Colon adenocarcinoma	KIPAN	Pan-kidney cohort
COADREAD	Colorectal adenocarcinoma	THCA	Thyroid carcinoma
DLBC	Lymphoid neoplasm Diffuse large B-cell Lymphoma	SKCM	Skin cutaneous Melanoma
ESCA	Esophageal carcinoma	LGG	Brain lower grade glioma
LUSC	Lung squamous cell carcinoma	GBMLGG	Glioma
OV	Ovarian serous cystadenocarcinoma	GBM	Glioblastoma multiforme
PRAD	Prostate adenocarcinoma	THYM	Thymoma
MESO	Mesothelioma	PCPG	Pheochromocytoma and Paraganglioma
KICH	Kidney Chromophobe	HNSC	Head and neck squamous cell carcinoma
LIHC	Liver hepatocellular carcinoma	KIRP	Kidney renal papillary cell carcinoma
TGCT	Testicular germ cell tumour	READ	Rectum adenocarcinoma
UCEC	Uterine Corpus endometrial carcinoma	UVM	Uveal Melanoma
STAD	Stomach adenocarcinoma	LAML	Acute Myeloid Leukaemia
LUAD	Lung adenocarcinoma		

4.5 Aim and objectives

The aim of this chapter was to perform prognostic and predictive analysis of the five miRNAs and their target genes using various *in silico* databases as well as expression analysis of these miRNAs and their target genes

I. Perform prognostic analysis for the five miRNAs using the PROGmiRV2 and SurvMicro databases;
II. Perform prognostic analysis for the miRNAs target genes using the SurvExpress and PognoScan databases;
III. Perform *in silico* expression analysis for the miRNAs using dbDEMC 2.0; and
IV. Perform *in silico* expression analysis for miRNAs target genes using FIREBROWSE database.

4.6 Methods and Materials

Prognostic and expression analysis was performed on the five miRNAs and their 12 target genes using the available online databases namely, PROGmiR, SurvMicro, SurvExpress, PrognoScan, dbDEMC 2.0 and FIREBROWSE.

4.6.1 Databases for miRNAs and target genes prognostic analysis

4.6.1.1 PROGmiRV2

ProgmiR was accessed at http://xvm145.jefferson.edu/progmir/ and was used for the prognostic analysis of the five miRNAs. All five miRNA names were submitted together as input. For the cancer type "prostate adenocarcinoma" was selected, and for survival measure, "death" was selected and a positive control miRNA was also submitted to confirm the performance of the database.

4.6.1.2 SurvMicro database

SurvMiro was accessed at http://bioinformatica.mty.itesm.mx/SurvMicro and the five miRNAs was used as input to examine the prognostic value for these miRNAs. All miRNA names were pasted in the query box with the following steps followed before analysis, (i) "prostate tissue" was selected from the tissue option; (ii) "prostate adenocarcinoma TCGA" database was selected which included 126 samples; and (iii) under the duplication option two setting were selected, (a) "show all" and (b) data "original. The Cox survival analysis test setting was applied for "censored" data: Survival-months were selected and for "stratification" class, GLEASON was selected.

4.6.1.3 SurvExpress database

SurvExpress accessed at http://bioinformatica.mty.itesm.mx:8080/Biomatec/SurvivaX.jsp was used to evaluate the prognostic value(s) of the 12 miRNA target genes in PCa patient and the following steps were followed: (i) a list of gene identifiers were used as input into the "query" option; (ii) SurvExpress incorporates eight databases and TCGA was selected, as it houses the largest number of samples equalling to 497, (iii) under the duplication option two settings were selected, (a) "show all" and (b) data "original. Cox Survival Analysis was applied with censored data set to survival-days.

4.6.1.4 PrognoScan

PrognoScan was accessed at http://www.abren.net/PrognoScan/ to perform the prognostic analysis for the miRNA target genes by submitting each gene individually into the "query" option. From the list of different cancers provided, "prostate cancer" was selected. The results for this database were given as a Kaplan Meier plot.

4.6.2 Databases miRNAs and target genes expression analysis

4.6.2.1 DbDEMC 2.0 miRNA expression analysis

The DbDEMC 2.0 database was accessed at http://www.picb.ac.cn/dbDEMC/ and was used to evaluate the expression of the five miRNAs in blood for PCa. The list of five miRNAs was pasted within the query box indicated as "miRNAs list" on the home page and for the experimental design section "blood" was selected

and for the cancer type “prostate cancer” was selected to generate a heat map for the expression profiling of the miRNAs.

4.6.2.2 FIREBROWSE database

The FIREBROWSE was accessed at http://firebrowse.org/. First “prostate adenocarcinoma” (PRAD) was selected in the cohort section. Secondly, the gene names were submitted individually into the query box option, with view expression profile selected afterwards and the expression sorted from high to low expression.

4.7 Results

4.7.1 Databases for miRNA prognostic analysis

PROGmiRV2 and SurvMicro databases

All five miRNAs were submitted to PROGmiR and SurvMicro databases and no results were obtained from these databases. Considering that these miRNAs showed no association with PCa as observed in Chapter 2, the results observed here strengthens the notion that these miRNAs are potentially novel for PCa. One miRNAs was used as positive control and the graphical result shown in figure 4.1.

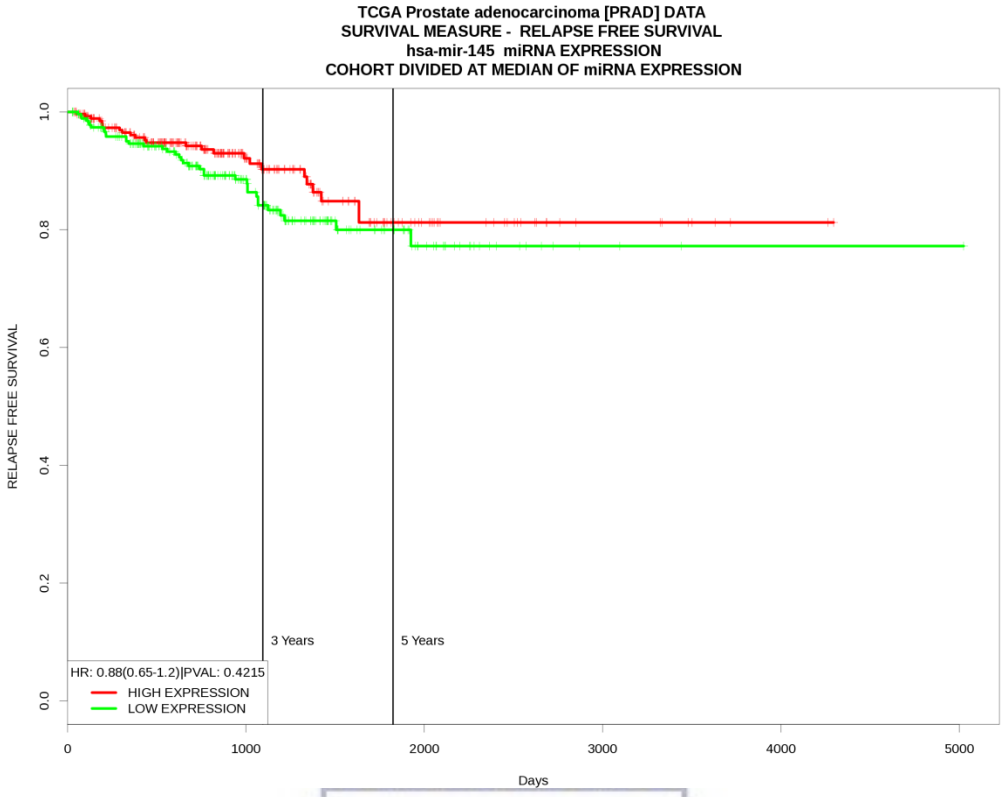


Figure 4.1: Kaplan Meier plot for positive control miRNA generated by the PROGmiRV2 database.

4.7.2 SurvExpress database

The list of miRNA target genes was submitted to SurvExpress for prognostic analysis and the results were displayed for the genes and their associated risk groups based on their individual expression levels. Results for all the genes were obtained as shown in figure 4.2 Results output were represented as box plots indicating the expression levels of genes and risk groups. From the box plots, it was assessed whether gene expression levels differed between the high-risk (red) and low-risk groups (green) as well as the level of significance of expression (p-value < 0.05) using a t- test. The results indicate genes *IGF1*, *PTGS2*, *SRD5A2*, *CYP24A1*, *AMACR*, *AR*, *NKX3-1*, *MYH11*, and *KLK4* showing a distinct difference between the low and the high-risk groups. The difference in expression

between risks groups for these genes can be useful for prognostic outcomes for PCa, i.e. whether a patient's survival is affected by the expression of these genes.

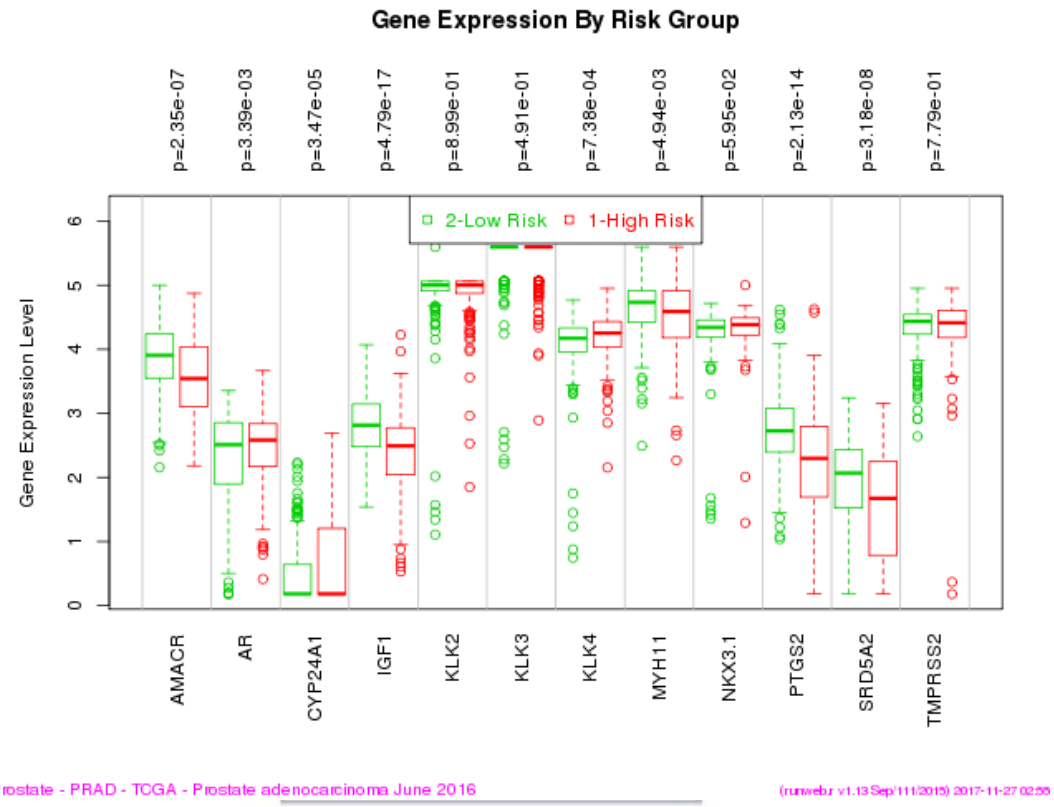


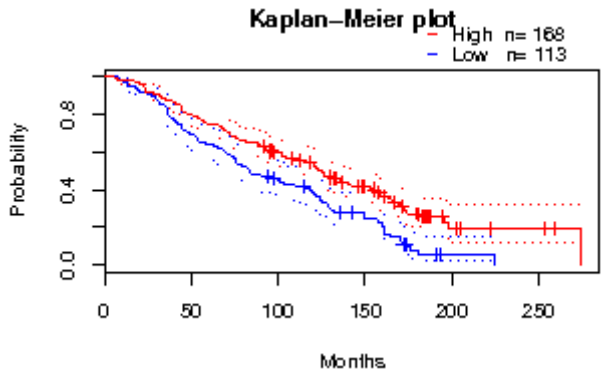
Figure 4.2: Gene expression analysis of the microRNA target genes using SurvExpress. The result outputs are given as box plots showing the expression of the gene in low and high-risk groups.

4.7.3 PrognoScan Results

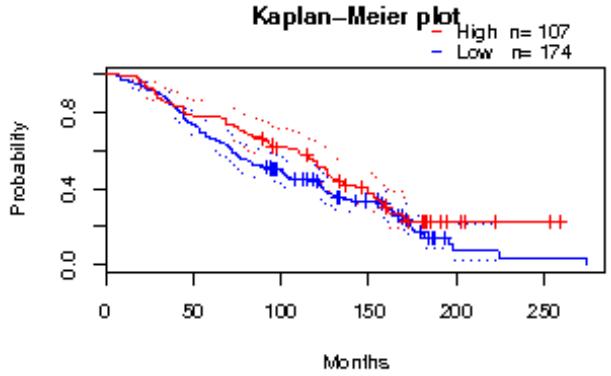
Twelve genes were submitted to PrognoScan for the prognostic analysis as explained in subsection 4.6.1.4. Eleven out of the 12 genes showed link to PCa with one gene not showing any link to PCa in PrognoScan. Three genes (*IGF-1*, *KLK3*, and *KLK4*) showed significant Cox *p-values* 0.003233, 0.001101 and 0.040085, respectively as shown in figure 4.3 for the Kaplan Meier plot results for these genes as well as within the annotation tables (see **Appendix B table 4. A, B,** and **C**). Thus, it implies that these three genes either individually or in

combination has good dicriminatory value between high-risk and low-risk PCa groups, making them potentially useful biomarkers in the management of PCa.

IGF1



KLK3



KLK4

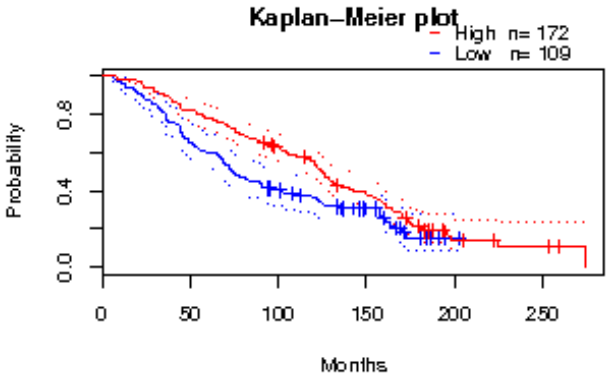


Figure 4.3: Kaplan Meier plots for *IGF1*, *KLK3*, and *KLK4*, showing their levels of expression for the high risk groups in red and for the low risk groups in blue use the PrognoScan database.

4.7.4 Expression analysis of potential MiRNAs

All five miRNAs were submitted to dbDEMC 2.0 for their expression analysis in blood extracted from PCa patient with no results shown from this database which imply that these miRNAs are potentially novel for their association with PCa.

4.7.5 Result from FIREBROWES database

The list of miRNA target genes identified in chapter 3 was submitted to FIREBROWES for gene expression analysis in 37 cancers. In Figure 4.4, the expression of *AR* in all cancer types is shown. It is notable that the *AR* gene expression is higher in the normal prostate tissue compared to prostate cancer tissue but in general is expressed in normal prostate and prostate tumour tissue compare to the among 37 analysed (BRCA, OV, KIRC, LIHC, KIPAN, KIRP, UCEC, GBM, GBMLGG, LGG, SARC, PAAD, LAML, KICH, CHOL, LUAD, THCA, UCS, STAD, BLCA, TGCT, STES, ACC, ESCA, LUSC, MESO, HNSC, PCPG, CESC, READ, COADREAD, COAD, SKCM, UVM, THYM, DLBC).

In Figure 4.5, illustrate the expression of *IGF1* in the 37 cancer types and was highly differentially expressed in ovarian serous cystadenocarcinoma (OV) followed by PRAD. All the other cancer types showed lower expression when compared to PRAD. When comparing between the normal tissue and prostate tumour tissue, expression was lower in the tumour tissue as indicated by the red colour.

In Figure 4.6, *KLK3* is notably over-expressed in both normal prostate tissue and prostate tumour tissue followed by kidney chromophobe (KICH) when compared

to the other cancer tissues. It is expressed higher in prostate tumour tissue than within normal prostate tissue.

In Figure 4.7, the highest expression for *KLK4* is observed in normal and tumour prostate tissue compared to the 37 normal and tumour tissues analysed. *KLK4* was highly expressed in tumour tissue when compared to normal prostate tissue.

In Figure 4.8, *KLK2* was highly expressed in prostate tissue compared to the other 37 normal prostate tissues with its highest expression in prostate tumour tissue.

AMACR showed the highest expression in the kidney renal papillary cell carcinoma (KIRP) tissue followed by its second highest expression in normal prostate tissue and prostate tumour tissue, as shown in Figure 4.9.

CYP24A1 shows high expression in the following 9 normal and tumour tissues (LUAD, HNSC, KIRP, CESC, LUSC, KIPAN, UCEC, UCS) followed by prostate normal and prostate tumour tissue, as shown in figure 4.10.

MYH11 showed high expression in prostate tumour and normal tissue when compared to all other normal and tumour tissue as shown in figure 4.11 followed by stomach adenocarcinoma (STAD) tissue.

NKX3-1 show high expression in normal prostate tissue and prostate tumour tissue when compared to the 37 tissues analysed, as shown in Figure 4.12 followed by

adrenocortical carcinoma (ACC). *NKX3-1* was over-expressed in prostate tumour tissue when compared to normal prostate tissue.

PTGS2 is over-expressed in LUAD, BLCA, LUCS, LAML, HNSC, and PRAD tissues, as shown in Figure 4.13, with *PTGS2* showing a higher expression in normal prostate tissue when compared to prostate tumour tissue.

SRD5A2 showed high expression in prostate tumour tissue compared to the other 37 tumour tissues with its expression down-regulated in prostate tumour tissue when compared to normal prostate tissue, as shown in Figure 4.14.

TMPRSS2 shows a high expression in prostate tumour tissue when compared to the other 37 tumour tissues with a higher expression in prostate tumour tissue compared to normal prostate tissue as shown in Figure 4.15.

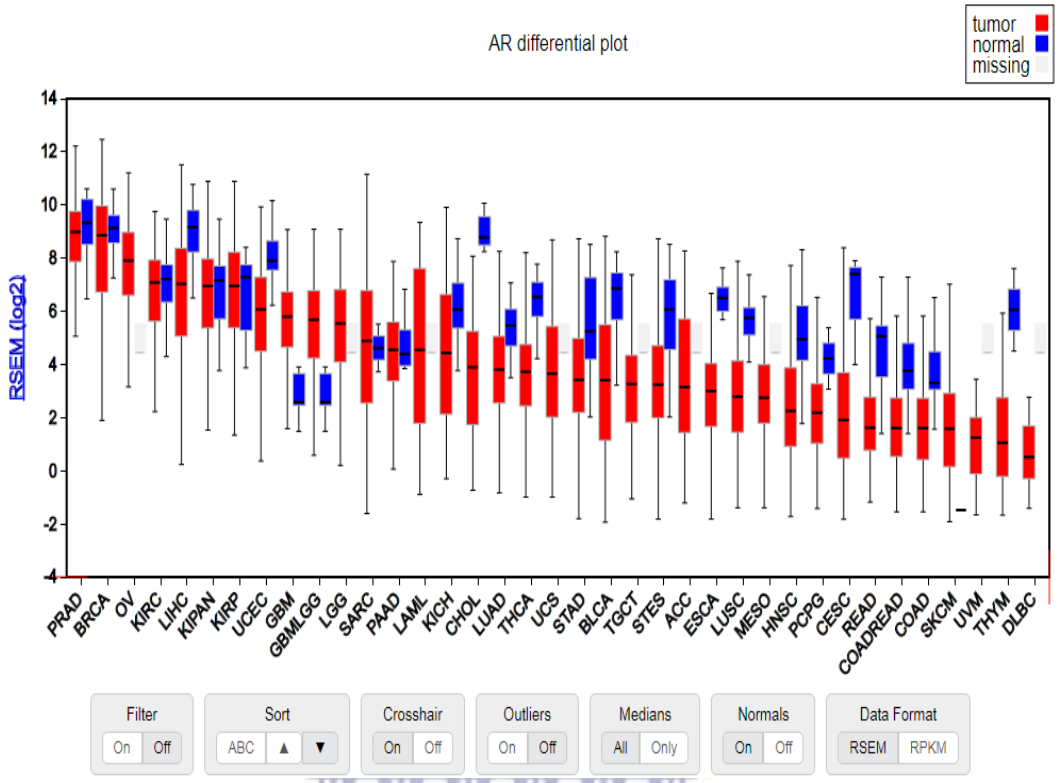


Figure 4.4: Show the expression of the *AR* gene in 37 cancers types with expression sorted from high to low.

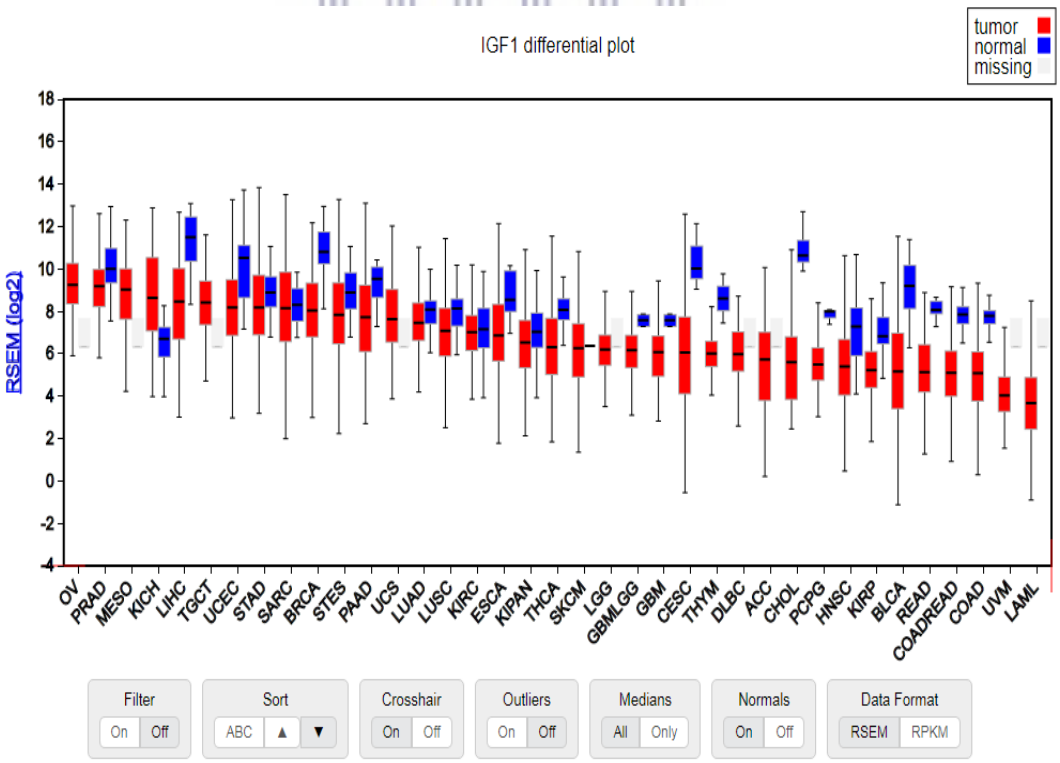


Figure 4.5: Show the expression of the *IGF1* gene in 37 cancers types with expression sorted from high to low

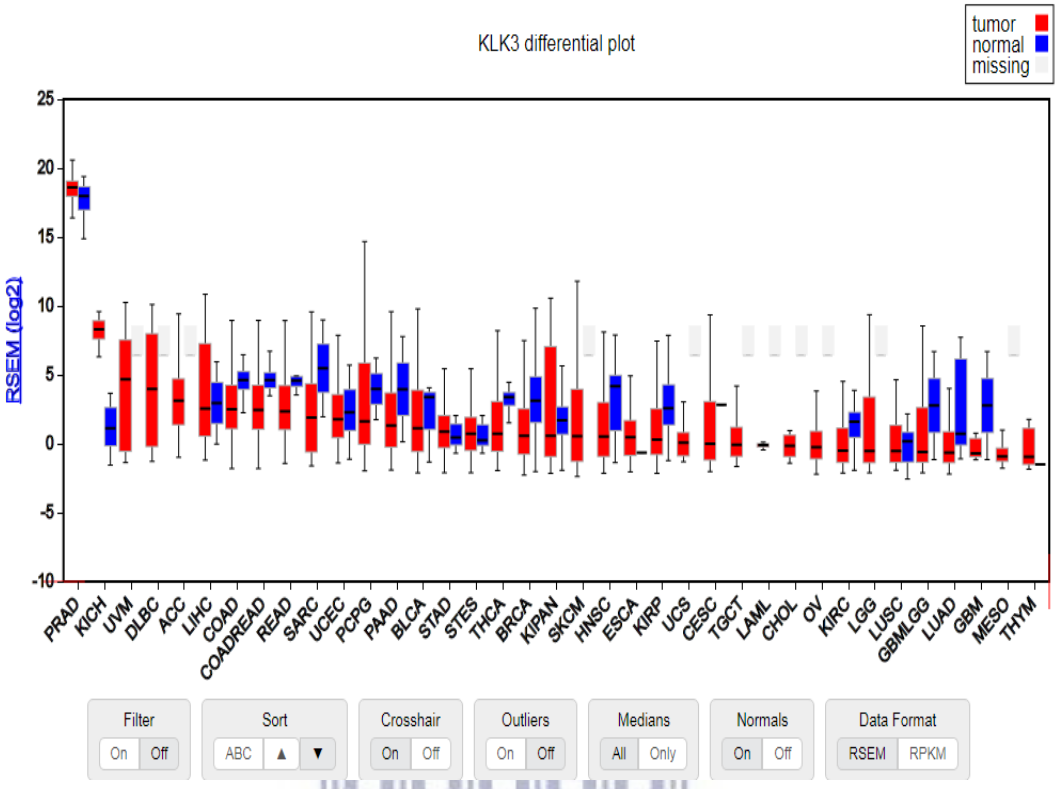


Figure 4.6: Show the expression of *KLK3* in 37 different cancer types sorted from high to low expression.

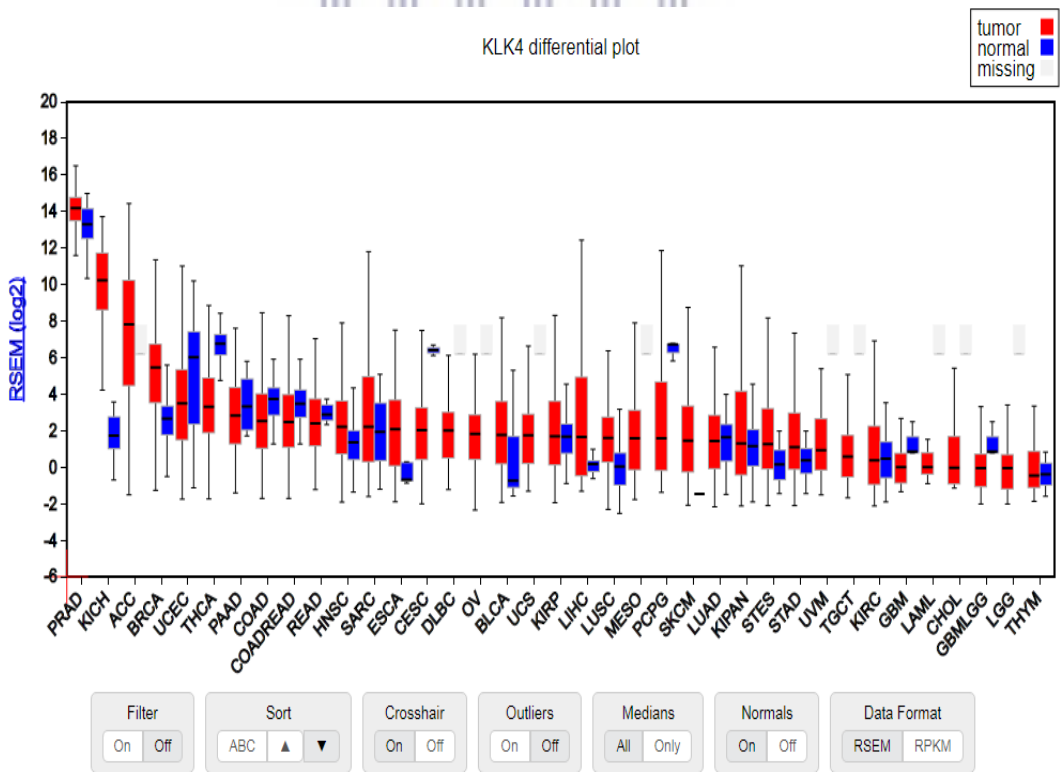


Figure 4.7: Show the expression of *KLK4* in 37 different cancer types sorted from high to low expression.

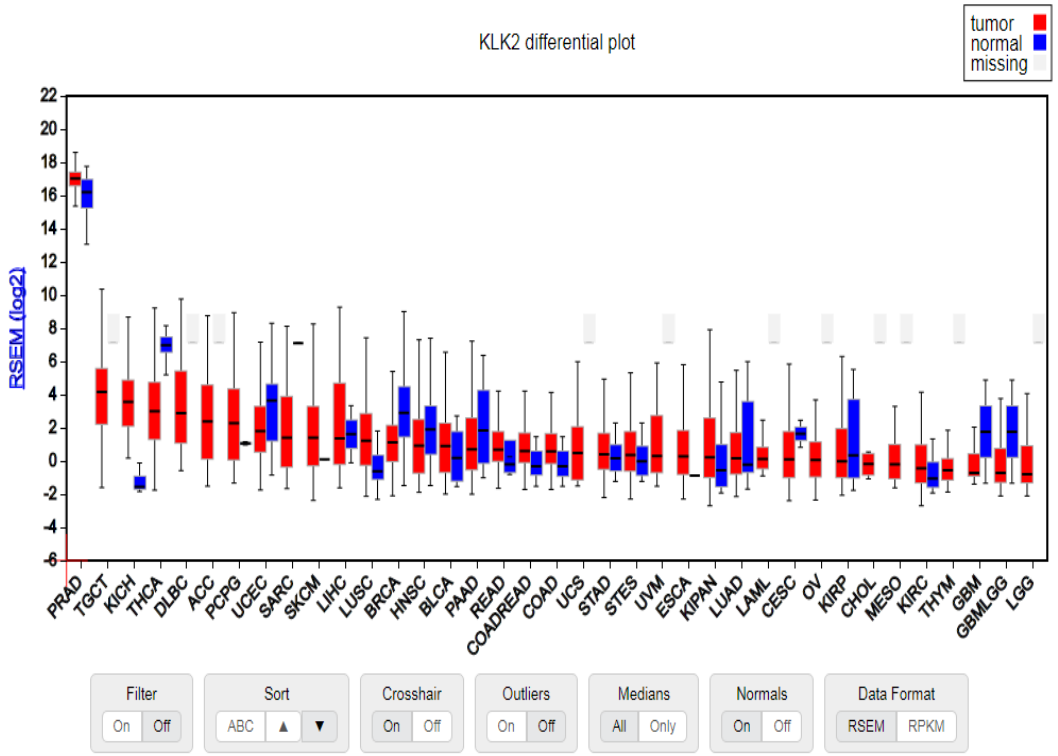


Figure 4.8: Show the expression of *KLK2* in 37different cancer types sorted from high to low expression

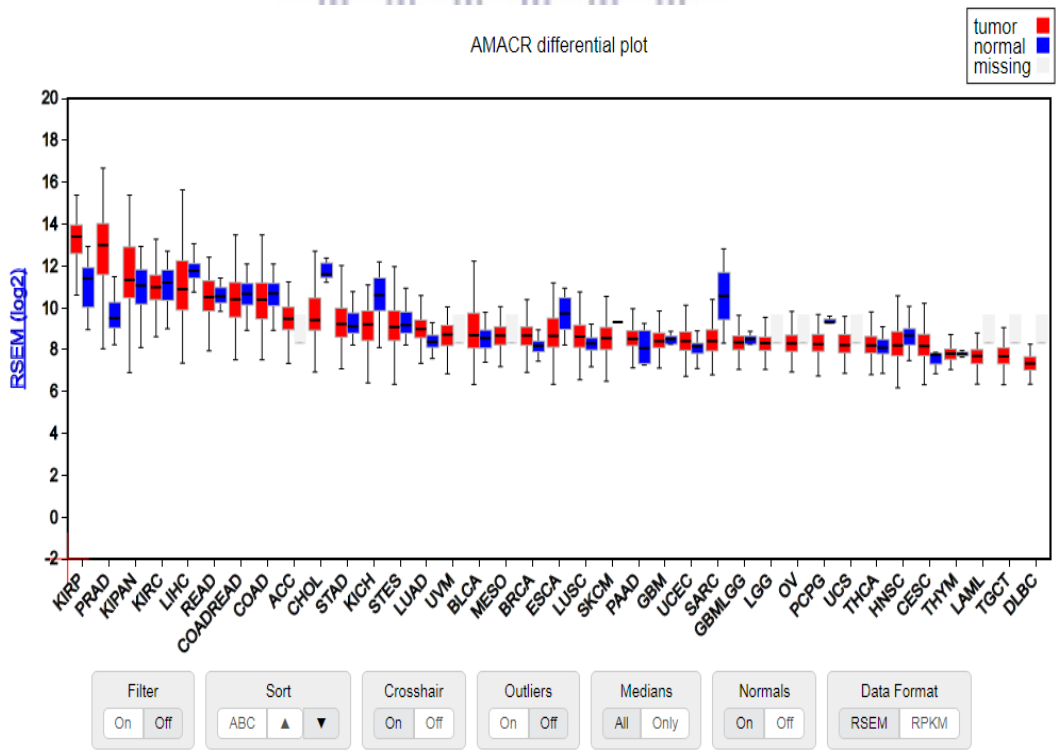


Figure 4.9: Show the expression of *AMACR* in 37 different tumour and normal tissues sorted from high to low expression

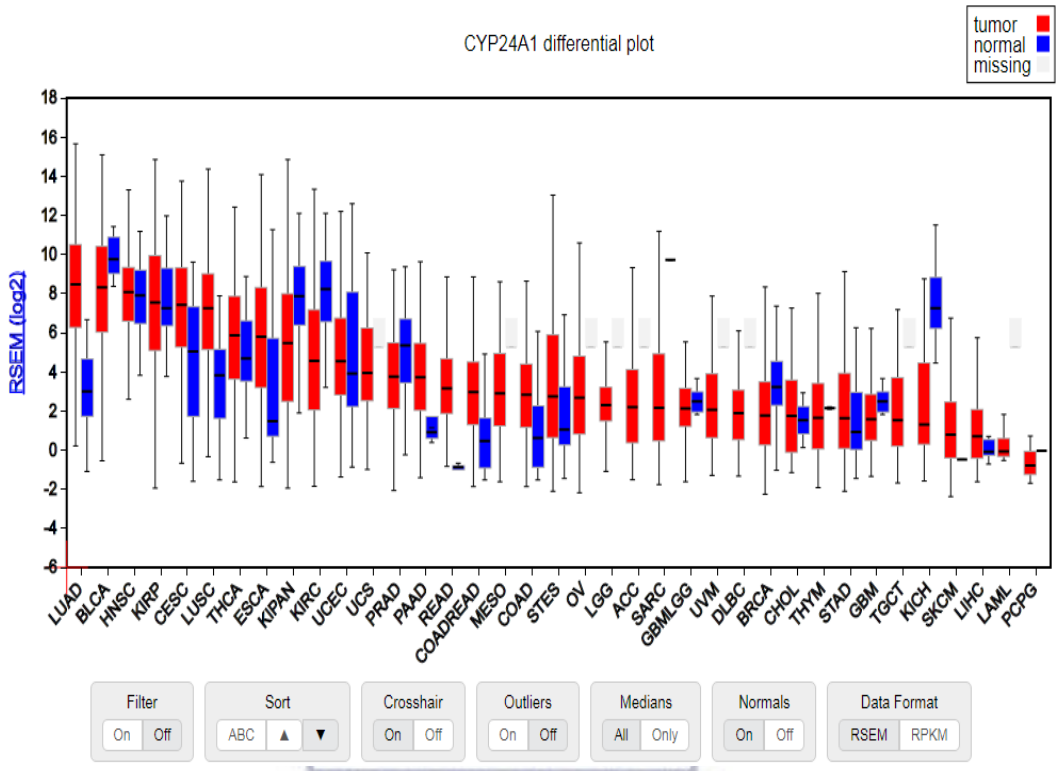


Figure 4.10: Show the expression of *CYP24A1* in 37 different tumour and normal tissues sorted from high to low expression

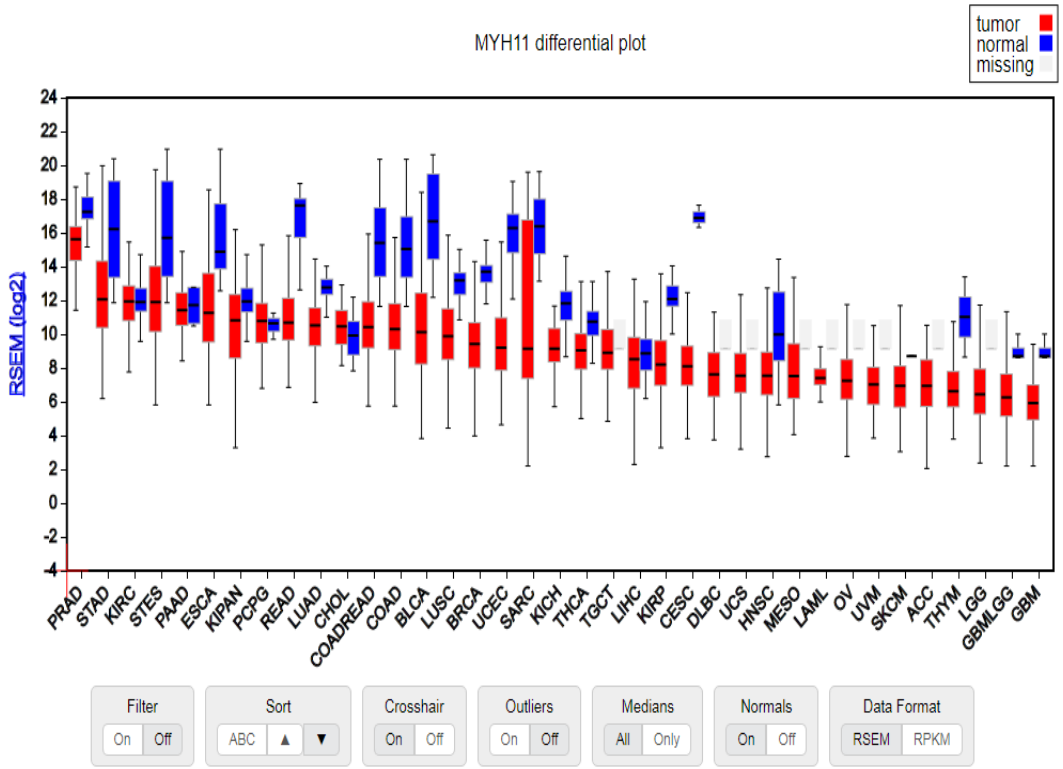


Figure 4.11: Show the expression of *MYH11* in 37 different tumour and normal tissues sorted from high to low expression

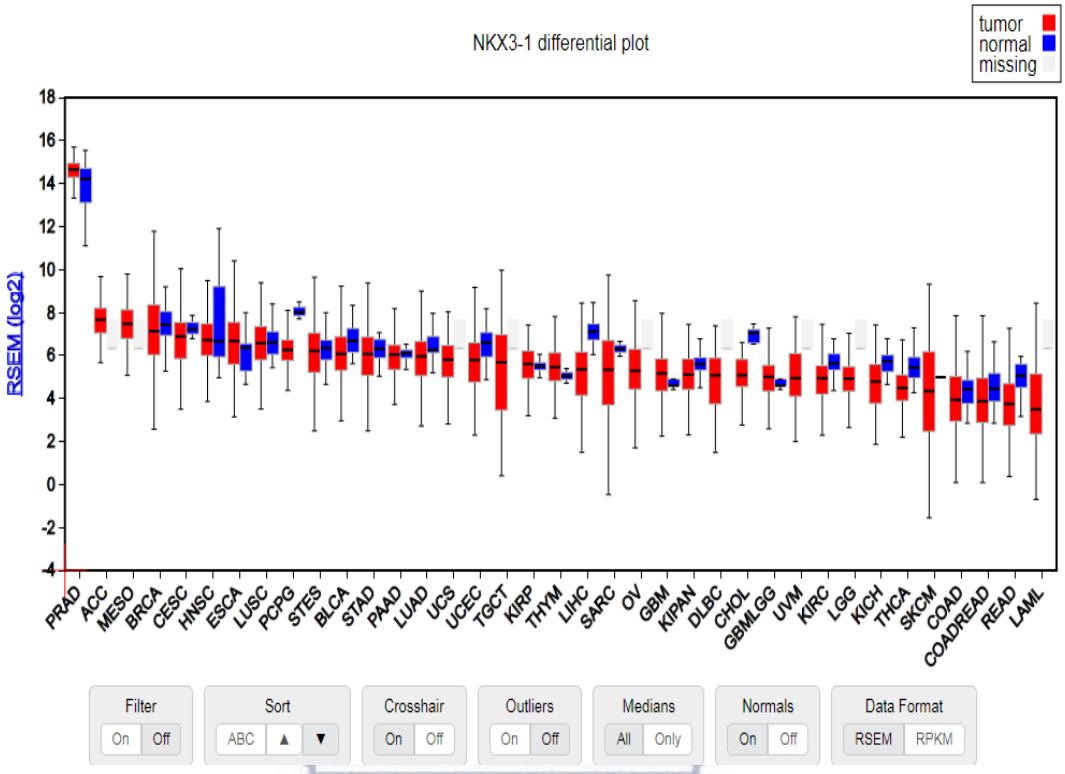


Figure 4.12: Show the expression of *NKX3-1* in 37 different tumour and normal tissues sorted from high to low expression

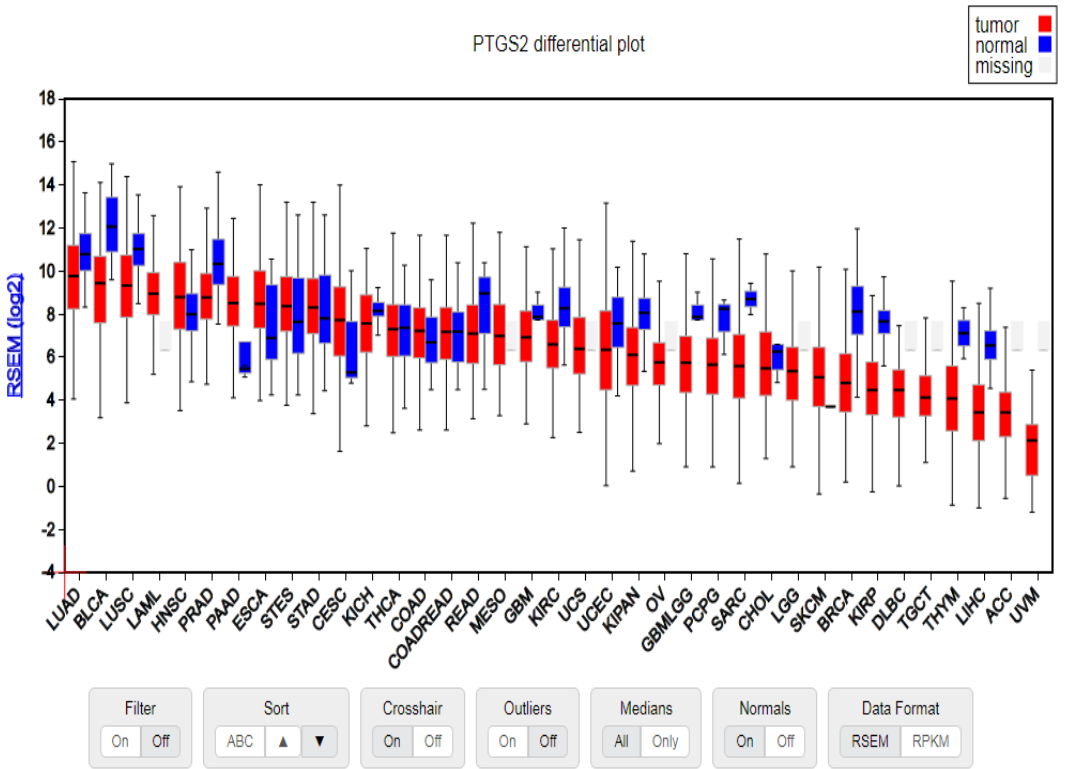


Figure 4.13: Show the expression of *PTGS2* in 37 different tumour and normal tissues sorted from high to low expression

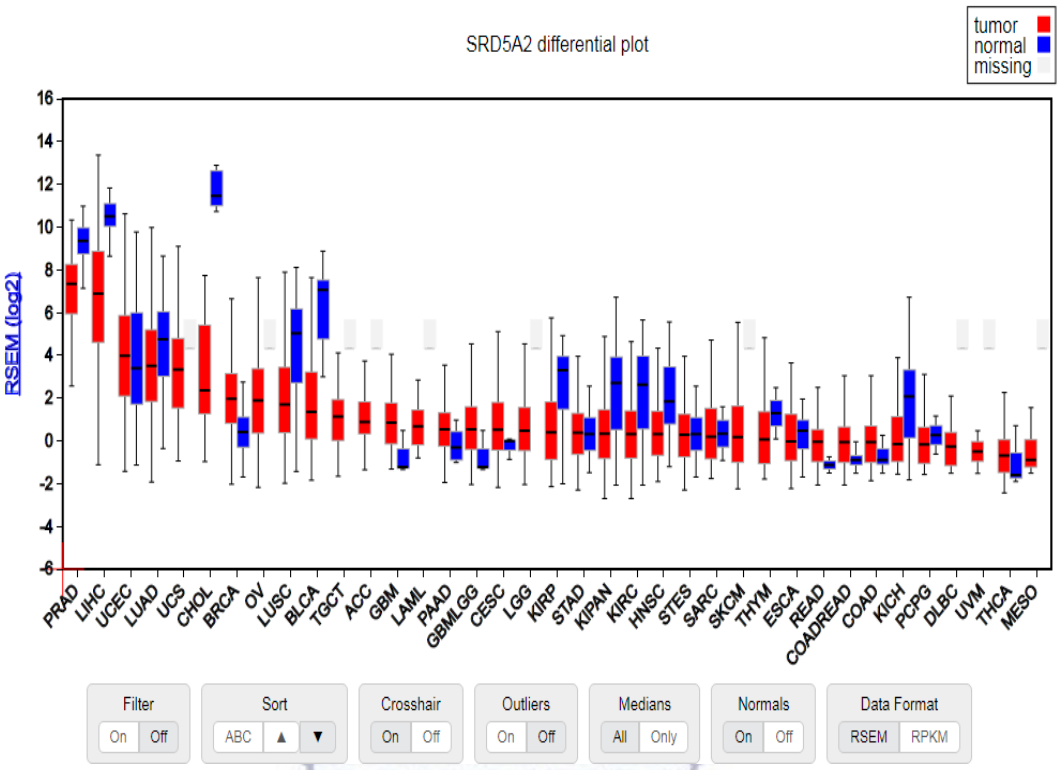


Figure 4.14: Show the expression of *SRD5A2* in 37 different tumour and normal tissues sorted from high to low expression

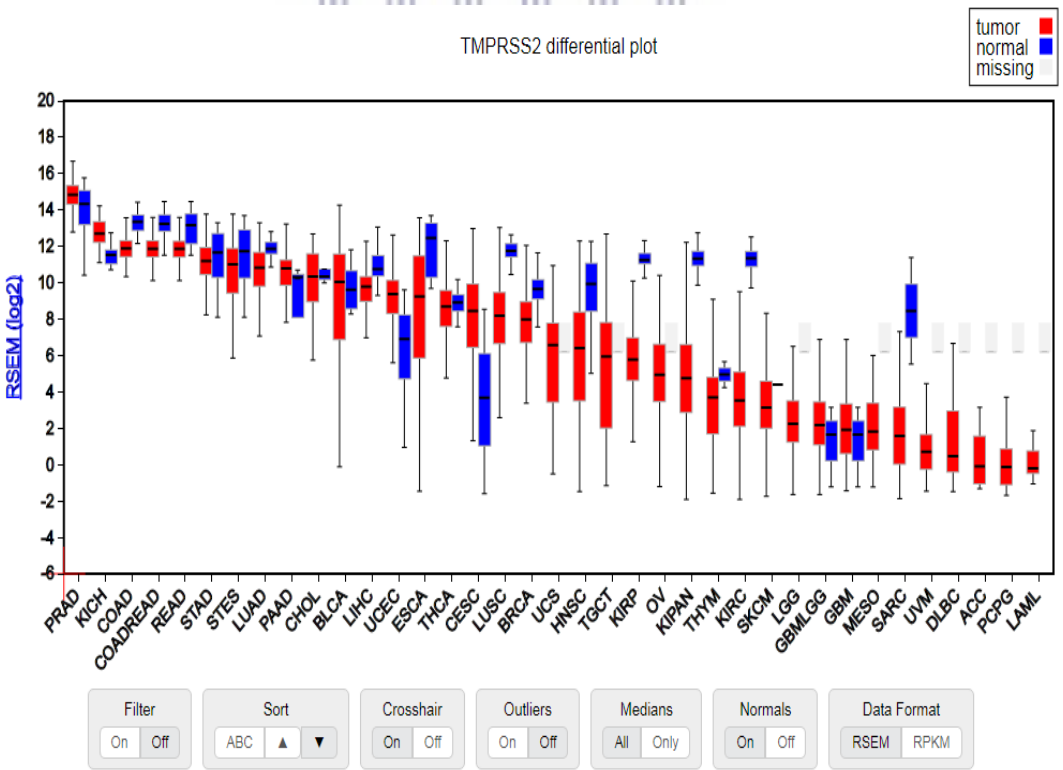


Figure 4. 15: Show the expression of *TMPRSS2* in 37 different tumour and normal tissues sorted from high to low expression

4.8 Discussion

Prostate cancer is one of the leading causes of death and the number of new case are increasing rapidly (Pakzad *et al.*, 2015). Cancer biomarkers can provide information with regard to the cancer stage of diagnosis, progression as well as a predictive indicator pre- and post-treatment (Richard Mayeux, 2004). MicroRNAs can be used as a prognostic biomarker for different types of cancer by investigating the expression of the microRNA at a different stage of cancer, and calculating cancer patient survival rates based on these expression changes (Yang *et al.*, 2016). Prognostic biomarkers aid in detecting molecular alterations in expression assisting in development of new therapeutic strategies based on the risk of the progression of cancer (Mancarella *et al.*, 2017a).

In this work, five miRNAs were identified in chapter 2 and their target genes identified in chapter 3. In particular, the miRNAs, as well as their target genes are evaluated for their prognostic value in PCa using an *in silico* approach.

Prognostic and expression analysis of the miRNAs

First, the five miRNAs were submitted to two prognostic databases which are (PROGmiR and MicroSurv) and one expression database dbDEMC version 2.0. No results were obtained for these miRNAs. These miRNAs showed no links to PCa as evident from the literature mining in chapter 2. Thus, the result obtained here lends more credence to these miRNAs being potentially novel for PCa.

Prognostic and expression analysis of the miRNA target genes

Secondly, a list of the miRNA target genes was submitted to the SurvExpress database to examine the prognostic value for these genes. The results dobtained is indicated by the graph in figure 4.2 which shows the expression of these genes for the high and low-risk PCa groups with their respective p-values indicated. Nine genes namely *IGF1*, *PTGS2*, *SRD5A2*, *CYP24A1*, *AMACR*, *AR*, *NKX3-1*, *MYH11*, and *KLK4* is able to differentiate the high-risk group from the low-risk group for PCa however, not statistically significantly so for all of these nine genes.

Epidemiological studies have reported that high serum *IGF-1* concentration and decreased circulating *IGFBP-3* are correlated with an increased risk of developing PCa which can be interpreted as *IGF-1* having good prognostic value for PCa (Chan *et al.*, 1998; Mancarella *et al.*, 2017a).

Previous study showed that the Insulin-like Growth Factor (*IGF)* system is influenced by *TMPRSS2* as *ERG* (a transcription factor), directly binds the *IGF-1R* gene promoter, thus affecting its expression in PCa. In another study, it was shown that patients with PCa that do not harbour the *TMPRSS2-ERG* rearrangement and who express low levels of *IGF-1R,* represent a subgroup of primary PCa tumours with poor outcome (Meisel-Sharon *et al.*, 2016; Mancarella *et al.*, 2017b).

Another study has identified *PTGS2* rs4648302 as a prognostic predictor for PCa. This variant might alter miRNA binding and thus influences *PTGS2* expression and PCa progression. Clinical studies have suggested that selective *PTGS2*

inhibitors could reduce cancer risk. If validated, this knowledge could ultimately lead to targeted therapies for a subset of high-risk patients harbouring this inherited *PTGS2* variant (Reddy *et al.*, 2000; Smith *et al.*, 2006; Lee *et al.*, 2016).

SRD5A 1 and *2*, are the two enzymes that catalyse the transformation of testosterone to dihydrotestosterone (Audet-Walsh *et al.*, 2017). Interrogation of *SRD5A* mRNA expression in three publicly available datasets confirmed that *SRD5A1* is increased in primary and metastatic PCa compared with non-tumoural prostate tissues, whereas *SRD5A2* is decreased. Activation of the androgen receptor (*AR*), a major oncogenic driver of PCa, induced the expression of *SRD5A1* from twofold to fourfold in three androgen-responsive PCa cell lines (Audet-Walsh *et al.*, 2017). Previously it has been suggested that increase in DNA methylation of *SRD5A2* and *CYP11A1* detectable in plasma have prognostic values for non-invasive PCa by monitoring biochemical recurrence after radical prostatectomy (Horning *et al.*, 2015).

CYP24A1 mRNA was higher in malignant human prostate tissues compared to benign lesions. High *CYP24A1* protein levels were seen in poorly differentiated and highly advanced stages of PCa and correlated with a parallel increase in the tumour proliferation rate (Tannour-Louet *et al.*, 2014). According to a study, *AR* signalling has reported as essential for the growth and maintenance of the prostate as well as the initiation and progression of PCa (Blessing *et al.*, 2017).

It has been reported that *NKX3.1* works as a prostatic tumour suppressor gene. Studies have showed positive results for the staining of the *NKX3.1* protein in

most of primary prostatic adenocarcinomas. This protein has been found down-regulated in many high-grade PCa and completely lost in most of metastatic PCa (Gurel *et al.*, 2010). In another study, the sensitivity and specificity of *NKX3.1* in high-grade prostatic adenocarcinoma when stained with a novel antibody when compared with high-grade urothelial carcinoma (Gurel *et al.*, 2010). It has also been reported that *NKX3.1* correlates with PCa progression (Bowen *et al.*, 2000).

The expression of *KLK4* show association with an increased PCa risk, and its activity favours tumour progression through increasing cell motility and growth. Notably, increasing *KLK4* expression in prostate glandular cells promotes tumour development, positioning the enzyme to play a role in early remodelling of the tumour microenvironment, a process essential for tumour growth (Kryza *et al.*, 2017).

The results from PrognoScan aligned with SurvExpress for two genes namely *IGF1* and *KLK4* which in both databases showed highly significant prognostic value. Also, the results showed good prognostic value for *KLK3* as shown in figure 4.2 in the Kaplan Meier plots.

Expression analysis using FIREBROWES showed the the expression of the miRNA target genes by comparing the expression of each gene across 37 tumour tissues see figures 4.4 to 4.15.

Eight out of 12 genes namely *AR*, *CYP24A1*, *KLK2*, *KLK3*, *KLK4*, *MYH11*, *NKX3.1*, and *TMPRSS2* was found to be differentially expressed in pancreatic adenocarcinoma (PRAD) with a high specificity for PRAD when compared to

other cancerous tissues. Moreover, four out of these eight genes *NKX3.1*, *KLK4*, *CYP24A1*, and *AR* showed good prognostic value using the SurvExpress database.

Taken together no results were obtained for the five miRNAs when their prognostic value was investigated using the PROGmiRV2 and SurvMicro databases. Also, in the expression analysis using dbDEMC 2.0 no expression results were obtained for the miRNAs, the notion is thus strengthened that these are potentially novel for PCa. The prognostic analysis of the 12 miRNA targeted genes showed good prognostic value for five out of the 12 genes with clear implications of these genes in PCa as observed from the literature. In addition, 8 of the target genes shows differential expression in PCa with four of these genes showing good prognostic value to differentiate high risk and low risk PCa groups. Thus, the miRNAs that regulate these genes can indirectly be linked to PCa in the absence of molecular validation data for these miRNAs.

Chapter 5

Conclusion and future work

5.1 Introduction

Prostate cancer is the most common cancer in man and can be treated if diagnosed in its early stages. Only in South Africa the incidences of PCa in 1995 were reported as 2 504 cases and after 10 years, in 2005, 4 631 new cases were reported. Over a 20-year time period from 1986 to 2006, 63 886 PCa cases were reported to the National Cancer Registry (NCR) (Babb *et al.*, 2014). The highest numbers of incidences were reported in the white population group within South Africa (Babb *et al.*, 2014). Relatively small numbers were noticed within the coloured and Asian/Indian population groups. Both the black and white population groups had fairly consistent numbers being reported throughout the years, in particular, for the last 5 years of reporting (Babb *et al.*, 2014). Current diagnostic tools for PCa include prostate specific antigen (PSA), Digital Rectal Examination (DRE), Magnetic Resonance Imaging (MRI) Computed tomography (CT) and Prostate biopsy which all have limitations such as lack of sensitivity and specificity, some of these methods carry high costs, report false-negative and false-positive results and are also invasive (Heidenreich *et al.*, 2011; Alberts *et al.*, 2017). The need for new biomarkers to facilitate the processes of diagnosis of PCa has become an important step for PCa management with miRNAs emerging as worthwhile molecules to investigate for this purpose.

MiRNAs have been found in the body fluids such as serum, plasma, urine, and saliva, which is an easily accessible alternative for invasive biopsy procedures (Lan *et al.*, 2015).

The first discovery of small non-coding RNA (microRNAs) was in 1993 (Almeida, Reis and Calin, 2011). They function in regulating mRNA producing proteins that are involved in the biological processes, such as apoptosis, cell proliferation, thus implicating miRNAs within these processes as well.

MicroRNAs show potential use as diagnostic, prognostic and therapeutic resolutions for different cancers including PCa. Therefore, the main aim of this study was to identify novel microRNAs as biomarkers for the early stage detection of PCa using several *in silico* methods. The advantage of using an *in silico* approaches is that it is rapid, cost-effective, and less labour intensive.

5.2 Chapter 2

Identification of novel miRNAs involved in PCa

The aim in this chapter was to identify novel miRNAs involved in PCa for diagnostic and prognostic purposes. This was achieved by using the CD-HIT program and BLAST. Two lists of miRNAs were generated one from Exiqon which contained 156 miRNAs experimentally validated to be involved in PCa and a second list, which contained 2588 mature miRNA sequences. Using the CD-HIT program and BLAST as described in sections, 16 miRNAs were identified with a potentially novel association with PCa. This list was refined to only five miRNAs through literature mining, which showed no association to PCa

5.3 Chapter 3

Identification of the miRNA target genes

The aim in this chapter was to identify the target genes of these potentially novel miRNAs for PCa. Furthermore, we confirmed the implication of these genes in PCa through various *in silico* analyses. By doing so, the miRNAs can indirectly be linked to PCa in the absence of molecular data for these molecules.

A list of target genes was identified using the mirDIP database as well as a list of all genes implicated to be expressed in prostate tissue was generated using the TiGER database. The intersection between these two lists was determined using Venn diagrams resulting in 212 genes targeted by the five miRNAs. Through functional annotation using the DAVID database as described in section, this number was reduced to 12 unique genes *KLK3*, *KLK2*, *KLK4*, *AR*, *CYP24A1*, *AMACR*, *NKX3.1*, *TMPRSS2*, *SRD5A1*, *PTGS2*, *MYH11*, and *IGF1*. Using a combination of pathway analysis utilizing the KEGG pathway tool as well as protein-protein interaction analysis using STRING, these genes were implicated further in cancer pathways and specifically in PCa onset and progression pathways.

5.4 Chapter 4

Prognostic and expression analysis of the miRNA target genes

Both the miRNAs and their target genes were analysed for their prognostic value for PCa using several *in silico* tools. In addition, the expression of the miRNAs and their target genes were analysed in various tumour tissues as well as normal tissues using several *in silico* databases. For the miRNAs none of the prognostic or the expression databases showed any results to implicate these miRNAs in PCa. This result furthers the notion that these miRNAs are novel for PCa, since the databases utilised are based on experimental evidence to implicate existing miRNA in different disease processes including cancer.

For the prognostic evaluation of the miRNA target genes, 11 out of the 12 genes showed a fair to very good prognostic value for PCa based on the expression of these genes within high- and low-risk groups for PCa survival. Through expression analysis 8 out of the 12 genes showed differential expression between prostate tumour tissue and normal prostate tissue, with the expression for some of these genes specific for prostate tumour tissue.

Table 5.1: Show genes that are involved in PCa pathways, genes having prognostic value and genes that are expressed in prostate tumour tissue.

Pathway analysis	Prognostic analysis	Expression analysis
AR	*AR*	*AR*
NKX3-1	*NKX3-1*	*NKX3-1*
KLK3	*KLK4*	*KLK3*
IGF1	*IGF1*	*KLK2*
	AMACR	*KLK4*
	CYP24A1	*MYH11*
	MYH11	*CYP24A1*
	PTGS2	*TMPRSS2*
	SRD5A2	

5.5 Conclusion

In this thesis five potentially novel miRNAs were identified that could be used for diagnostic as well as prognostic purposes for the management of PCa. These five miRNAs regulates 12 genes of which 11 was showed to have marginal to strong prognostic value for PCa with eight genes showing differential expression for prostate tissue compared to normal prostate tissue as well as other cancerous tissue. Thus, following molecular validation, both the miRNAs as well as their target genes can be utilised as biomarkers for PCa diagnosis as well as prognosis.

5.6 Future work

Follow-up work includes the expression profiling of the miRNAs in cell lines as well as clinical samples, using RT-qPCR to confirm their specificity and accuracy for PCa diagnosis. Additionally, to validate the genes that are regulated by these miRNAs using a luciferase assay which will determine the effect of these miRNAs on the functioning of their target genes. Other molecular diagnostic techniques such as fluorescent *in situ* hybridization (FISH), which is a cytogenetic technique to detect the location of specific gene or specific nucleic acid sequence on chromosomes, can be employed. In addition, the copy number of genes and expression profiles of microRNAs can be assessed to differentiation between normal and cancer PCa tissue using Next Generation Sequencing (NGS). Deep sequencing technologies have the advantage to detect novel sequences, that could be used to determine the function of microRNAs in PCa since the advantage of using these techniques is they have high accuracy and specificity for miRNA detection the normal versus cancer tissue.

Reference

Abreu Velez, A. M. and Howard, M. S. (2015) 'Tumor-suppressor genes, cell cycle regulatory checkpoints, and the skin', *North American Journal of Medical Sciences*, pp. 176–188. doi: 10.4103/1947-2714.157476.

Aebersold, R., Anderson, L., Caprioli, R., Druker, B., Hartwell, L. and Smith, R. (2005) 'Perspective: A program to improve protein biomarker discovery for cancer', *Journal of Proteome Research*, 4(4), pp. 1104–1109. doi: 10.1021/pr050027n.

Aguirre-Gamboa, R., Gomez-Rueda, H., Martínez-Ledesma, E., Martínez-Torteya, A., Chacolla-Huaringa, R., Rodriguez-Barrientos, A., Tamez-Peña, J. G. and Treviño, V. (2013) 'SurvExpress: An Online Biomarker Validation Tool and Database for Cancer Gene Expression Data Using Survival Analysis', *PLoS ONE*, 8(9). doi: 10.1371/journal.pone.0074250.

Aguirre-Gamboa, R. and Trevino, V. (2014) 'SurvMicro: Assessment of miRNA-based prognostic signatures for cancer clinical outcomes by multivariate survival analysis', *Bioinformatics*, 30(11), pp. 1630–1632. doi: 10.1093/bioinformatics/btu087.

Ahram, M. and Petricoin, E. F. (2008) 'Proteomics discovery of disease biomarkers', *Biomarker Insights*, pp. 325–333.

Alberts, A. R., Schoots, I. G., Bokhorst, L. P., Drost, F. J. H., van Leenders, G. J., Krestin, G. P., Dwarkasing, R. S., Barentsz, J. O., Schröder, F. H., Bangma, C. H. and Roobol, M. J. (2017) 'Characteristics of Prostate Cancer Found at Fifth Screening in the European Randomized Study of Screening for Prostate Cancer

Rotterdam: Can We Selectively Detect High-grade Prostate Cancer with Upfront Multivariable Risk Stratification and Magnetic Reson', *European Urology*, pp. 1–8. doi: 10.1016/j.eururo.2017.06.019.

Aleksic, T., Verrill, C., Bryant, R. J., Han, C., Worrall, A. R., Brureau, L., Larré, S., Higgins, G. S., Fazal, F., Sabbagh, A., Haider, S., Buffa, F. M., Cole, D. and Macaulay, V. M. (2017) 'IGF-1R associates with adverse outcomes after radical radiotherapy for prostate cancer', *British Journal of Cancer*, 117(11), pp. 1600–1606. doi: 10.1038/bjc.2017.337.

Almeida, M. I., Reis, R. M. and Calin, G. A. (2011) 'MicroRNA history: Discovery, recent applications, and next frontiers', *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, pp. 1–8. doi: 10.1016/j.mrfmmm.2011.03.009.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) 'Basic local alignment search tool.', *Journal of Molecular Biology*, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

America Cancer Society (2014) 'What are genes ? Gene mutations', *National Society of Genetic Counselors (NSGC)*, p. 10.

American Cancer Society (2016) 'Cancer Facts & Figures 2016', *Cancer Facts & Figures 2016*, pp. 1–9. doi: 10.1097/01.NNR.0000289503.22414.79.

Aqeilan, R. I., Calin, G. A. and Croce, C. M. (2010) 'miR-15a and miR-16-1 in cancer: discovery, function and future perspectives', *Cell Death and Differentiation*, 17(2), pp. 215–220. doi: 10.1038/cdd.2009.69.

Audet-Walsh, É., Yee, T., Tam, I. S. and Giguère, V. (2017) 'Inverse regulation of DHT synthesis enzymes 5α-reductase types 1 and 2 by the androgen receptor in prostate cancer', *Endocrinology*, 158(4), pp. 1015–1021. doi: 10.1210/en.2016-1926.

Aumüller, G., Seitz, J., Lilja, H., Abrahamsson, P. -A, Von Kammer, H. Der and Scheit, K. -H (1990) 'Species- and organ-specificity of secretory proteins derived from human prostate and seminal vesicles', *The Prostate*, 17(1), pp. 31–40. doi: 10.1002/pros.2990170105.

Babb, C., Urban, M., Kielkowski, D. and Kellett, P. (2014) 'Prostate cancer in South Africa: pathology based national cancer registry data (1986-2006) and mortality rates (1997-2009).', *Prostate cancer*, 2014, p. 419801. doi: 10.1155/2014/419801.

Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P. and Bartel, D. P. (2008) 'The impact of microRNAs on protein output.', *Nature*, 455(7209), pp. 64–71. doi: 10.1038/nature07242.

Barrett, L. W., Fletcher, S. and Wilton, S. D. (2012) 'Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements', *Cellular and Molecular Life Sciences*, pp. 3613–3634. doi: 10.1007/s00018-012-0990-9.

Bartel, D. P. (2004) 'MicroRNAs: Genomics, Biogenesis, Mechanism, and Function', *Cell*, pp. 281–297. doi: 10.1016/S0092-8674(04)00045-5.

Batzer, M. A. and Deininger, P. L. (2002) 'alu repeats and human genomic

diversity', *Nature Reviews Genetics*, 3(5), pp. 370–379. doi: 10.1038/nrg798.

Betel, D., Wilson, M., Gabow, A., Marks, D. S. and Sander, C. (2008) 'The microRNA.org resource: Targets and expression', *Nucleic Acids Research*, 36(SUPPL. 1). doi: 10.1093/nar/gkm995.

Bhatt, A. N., Mathur, R., Farooque, A., Verma, A. and Dwarakanath, B. S. (2010) 'Cancer biomarkers - current perspectives', *The Indian Journal of Medical Research*, 132(August), pp. 129–149.

Bian, S. X. and Hoffman, K. E. (2013) 'Management of prostate cancer in elderly men', *Seminars in Radiation Oncology*, pp. 198–205. doi: 10.1016/j.semradonc.2013.01.007.

Blessing, A. M., Rajapakshe, K., Reddy Bollu, L., Shi, Y., White, M. A., Pham, A. H., Lin, C., Jonsson, P., Cortes, C. J., Cheung, E., La Spada, A. R., Bast, R. C., Merchant, F. A., Coarfa, C. and Frigo, D. E. (2017) 'Transcriptional regulation of core autophagy and lysosomal genes by the androgen receptor promotes prostate cancer progression', *Autophagy*, 13(3), pp. 506–521. doi: 10.1080/15548627.2016.1268300.

De Bock, M., De Seny, D., Meuwis, M. A., Chapelle, J. P., Louis, E., Malaise, M., Merville, M. P. and Fillet, M. (2010) 'Challenges for biomarker discovery in body fluids using SELDI-TOF-MS', *Journal of Biomedicine and Biotechnology*, 2010. doi: 10.1155/2010/906082.

Bohnsack, M. T., Czaplinski, K. and Gorlich, D. (2004) 'Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs.',

RNA (New York, N.Y.), 10(2), pp. 185–91. doi: 10.1261/rna.5167604.

Bowen, C., Bubendorf, L., Voeller, H. J., Slack, R., Willi, N., Sauter, G., Gasser, T. C., Koivisto, P., Lack, E. E., Kononen, J., Kallioniemi, O. P. and Gelmann, E. P. (2000) 'Loss of NKX3.1 expression in human prostate cancers correlates with tumor progression', *Cancer Research*, 60(21), pp. 6111–6115. doi: 10.1158/0008-5472.can-06-0963.

Brahmkhatri, V. P., Prasanna, C. and Atreya, H. S. (2015) 'Insulin-like growth factor system in cancer: Novel targeted therapies', *BioMed Research International*. doi: 10.1155/2015/538019.

Brown, K. R., Otasek, D., Ali, M., McGuffin, M. J., Xie, W., Devani, B., van Toch, I. L. and Jurisica, I. (2009) 'NAViGaTOR: Network analysis, visualization and graphing Toronto', *Bioinformatics*, 25(24), pp. 3327–3329. doi: 10.1093/bioinformatics/btp595.

Burke, H. B. (2016) 'Predicting Clinical Outcomes Using Molecular Biomarkers.', *Biomarkers in cancer*, 8, pp. 89–99. doi: 10.4137/BIC.S33380.

Carmell, M. A., Xuan, Z., Zhang, M. Q. and Hannon, G. J. (2002) 'The Argonaute family: Tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis', *Genes and Development*, 16(21), pp. 2733–2742. doi: 10.1101/gad.1026102.

Carthew, R. W. (2006) 'Gene regulation by microRNAs', *Current Opinion in Genetics and Development*, pp. 203–208. doi: 10.1016/j.gde.2006.02.012.

Casey, S. C., Baylot, V. and Felsher, D. W. (2017) 'MYC: Master Regulator of

Immune Privilege', *Trends in Immunology*, pp. 298–305. doi: 10.1016/j.it.2017.01.002.

Chan, J. M., Stampfer, M. J., Giovannucci, E., Gann, P. H., Ma, J., Wilkinson, P., Hennekens, C. H. and Pollak, M. (1998) 'Plasma insulin-like growth factor-I and prostate cancer risk: a prospective study', *Science*, 279(5350), pp. 563–566. doi: 10.1126/science.279.5350.563.

Chari, S. T., Kelly, K., Hollingsworth, M. A., Thayer, S. P., Ahlquist, D. A., Andersen, D. K., Batra, S. K., Brentnall, T. A., Canto, M., Cleeter, D. F., Firpo, M. A., Gambhir, S. S., Go, V. L. W., Hines, O. J., Kenner, B. J., Klimstra, D. S., Lerch, M. M., Levy, M. J., Maitra, A., Mulvihill, S. J., Petersen, G. M., Rhim, A. D., Simeone, D. M., Srivastava, S., Tanaka, M., Vinik, A. I. and Wong, D. (2015) 'Early detection of sporadic pancreatic cancer: summative review.', *Pancreas*, 44(5), pp. 693–712. doi: 10.1097/MPA.0000000000000368.

Chen, C.-M., Lin, C.-L., Chiou, H.-L., Hsieh, S.-C., Lin, C.-L., Cheng, C.-W., Hung, C.-H., Tsai, J.-P. and Hsieh, Y.-H. (2017) 'Loss of endothelial cell-specific molecule 1 promotes the tumorigenicity and metastasis of prostate cancer cells through regulation of the TIMP-1/MMP-9 expression.', *Oncotarget*, 8(8), pp. 13886–13897. doi: 10.18632/oncotarget.14684.

Chen, G. G. and Lai, P. B. S. (2009) *Apoptosis in carcinogenesis and chemotherapy: Apoptosis in cancer, Apoptosis in Carcinogenesis and Chemotherapy: Apoptosis in Cancer*. doi: 10.1007/978-1-4020-9597-9.

Chen, Y., Wang, S., Zhang, L., Xie, T., Song, S., Huang, J., Zhang, Y., Ouyang, L. and Liu, B. (2015) 'Identification of ULK1 as a novel biomarker involved in

MIR-4487 and MIR-595 regulation in neuroblastoma SH-SY5Y cell autophagy', *Scientific Reports*, 5. doi: 10.1038/srep11035.

Chendrimada, T. P., Gregory, R. I., Kumaraswamy, E., Cooch, N., Nishikura, K. and Shiekhattar, R. (2010) 'NIH Public Access', 436(7051), pp. 740–744. doi: 10.1038/nature03868.TRBP.

Chial, B. H., Write, P. D., Right, S. and Education, N. (2008) *Proto-oncogenes to Oncogenes to Cancer, Nature Education*. doi: 10.1038/315758a0.

Cho, W. C. S. (2013) 'Great potential of miRNAs as predictive and prognostic markers for cancer', *Zhongguo fei ai za zhi = Chinese journal of lung cancer*. doi: 10.3779/j.issn.1009-3419.2013.01.12.

Cinpolat, O., Unal, Z. N., Ismi, O., Gorur, A. and Unal, M. (2017) 'Comparison of microRNA profiles between benign and malignant salivary gland tumors in tissue, blood and saliva samples: a prospective, case-control study', *Brazilian Journal of Otorhinolaryngology*, 83(3), pp. 276–284. doi: 10.1016/j.bjorl.2016.03.013.

Costa-Pinheiro, P., Montezuma, D., Henrique, R. and Jerónimo, C. (2015) 'Diagnostic and prognostic epigenetic biomarkers in cancer', *Epigenomics*, 7(6), pp. 1003–1015. doi: 10.2217/epi.15.56.

Dalal, A. and Atri, A. (2014) 'An Introduction to Sequence and Series', *International Journal of Research*, 1(10), pp. 1286–1292. doi: 10.1002/0471250953.bi0301s42.An.

Dasgupta, S., Vishwanatha, J. and Srinidhi, S. (2012) 'Oncogenic activation in prostate cancer progression and metastasis: Molecular insights and future

challenges', *Journal of Carcinogenesis*, 11(1), p. 4. doi: 10.4103/1477-3163.93001.

Davis, B. N. and Hata, A. (2009) 'Cell Communication and Signaling Regulation of MicroRNA Biogenesis : A miRiad of mechanisms', 22, pp. 1–22. doi: 10.1186/1478-811X-7-18.

Demichelis, F., Fall, K., Perner, S., Andrén, O., Schmidt, F., Setlur, S. R., Hoshida, Y., Mosquera, J. M., Pawitan, Y., Lee, C., Adami, H. O., Mucci, L. A., Kantoff, P. W., Andersson, S. O., Chinnaiyan, A. M., Johansson, J. E. and Rubin, M. A. (2007) 'TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort', *Oncogene*, 26(31), pp. 4596–4599. doi: 10.1038/sj.onc.1210237.

DeSantis, C., Rebecca Siegel, M., Ahmedin Jemal, M., Alcaraz, K., Bertaut, T., Fedewa, S., Gansler, T., Goding Sauer, A., McMahon, C., Miller, K., Piercy, A., Richard, C., Simpson, S. and Wagner, D. (2016) 'Cancer Facts & Figures'. doi: 10.1097/01.NNR.0000289503.22414.79.

Dews, M., Homayouni, A., Yu, D., Murphy, D., Sevignani, C., Wentzel, E., Furth, E. E., Lee, W. M., Enders, G. H., Mendell, J. T. and Thomas-Tikhonenko, A. (2006) 'Augmentation of tumor angiogenesis by a Myc-activated microRNA cluster', *Nature Genetics*, 38(9), pp. 1060–1065. doi: 10.1038/ng1855.

Dickinson, L., Ahmed, H. U., Allen, C., Barentsz, J. O., Carey, B., Futterer, J. J., Heijmink, S. W., Hoskin, P. J., Kirkham, A., Padhani, A. R., Persad, R., Puech, P., Punwani, S., Sohaib, A. S., Tombal, B., Villers, A., Van Der Meulen, J. and Emberton, M. (2011) 'Magnetic resonance imaging for the detection, localisation,

and characterisation of prostate cancer: Recommendations from a European consensus meeting', *European Urology*, 59(4), pp. 477–494. doi: 10.1016/j.eururo.2010.12.009.

Diederichs, S. and Haber, D. A. (2007) 'Dual Role for Argonautes in MicroRNA Processing and Posttranscriptional Regulation of MicroRNA Expression', *Cell*, 131(6), pp. 1097–1108. doi: 10.1016/j.cell.2007.10.032.

Doench, J. G. and Sharp, P. A. (2004) 'Specificity of microRNA target selection in translational repression.', *Genes & development*, 18(5), pp. 504–11. doi: 10.1101/gad.1184404.

Durbin, R., Eddy, S., Krogh, a and Mitchison, G. (1998) 'Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids', *Analysis*, p. 356. doi: 10.1017/CBO9780511790492.

Ed, D. N. and Cur, B. A. (2016) 'Cancer Association of South Africa (CANSA) Fact Sheet on the Role of Prostate Specific Antigen (PSA) Screening on Prostate Cancer Diagnosis and Treatment', (February), pp. 1–8.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. and Marks, D. S. (2003) 'MicroRNA targets in Drosophila.', *Genome biology*, 5(1), p. R1. doi: 10.1186/gb-2003-5-1-r1.

Evans, A. J. (2003) 'Alpha-methylacyl CoA racemase (P504S): overview and potential uses in diagnostic pathology as applied to prostate needle biopsies.', *Journal of Clinical Pathology*, 56(12), pp. 892–897. doi: 10.1136/jcp.56.12.892.

Ewens, W. J. and Grant, G. R. (2005) 'The Analysis of Multiple DNA or Protein

Sequences', *Statistical Methods in Bioinformatics - an Introduction*, (I), pp. 181–217. doi: 10.1007/978-1-4757-3247-4_6.

Fang, C., Guo, Z.-Q., Chen, X.-Y., Liu, T.-Z., Zeng, X.-T. and Wang, X.-H. (2017) 'Relationship between *SRD5A2* rs9282858 polymorphism and the susceptibility of prostate cancer', *Medicine (United States)*, 96(19). doi: 10.1097/MD.0000000000006791.

Faruq, O. and Vecchione, A. (2015) 'microRNA: Diagnostic Perspective', *Frontiers in Medicine*, 2. doi: 10.3389/fmed.2015.00051.

Förstemann, K., Tomari, Y., Du, T., Vagin, V. V., Denli, A. M., Bratu, D. P., Klattenhoff, C., Theurkauf, W. E. and Zamore, P. D. (2005) 'Normal microRNA maturation and germ-line stem cell maintenance requires loquacious, a double-stranded RNA-binding domain protein', *PLoS Biology*, 3(7), pp. 1187–1201. doi: 10.1371/journal.pbio.0030236.

Freds??e, J., Rasmussen, A. K. I., Thomsen, A. R., Mouritzen, P., H??yer, S., Borre, M., ??rntoft, T. F. and S??rensen, K. D. (2017) 'Diagnostic and Prognostic MicroRNA Biomarkers for Prostate Cancer in Cell-free Urine', *European Urology Focus*. doi: 10.1016/j.euf.2017.02.018.

Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) 'CD-HIT: Accelerated for clustering the next-generation sequencing data', *Bioinformatics*, 28(23), pp. 3150–3152. doi: 10.1093/bioinformatics/bts565.

Fu, M., Rao, M., Wang, C., Sakamaki, T., Wang, J., Di Vizio, D., Zhang, X., Albanese, C., Balk, S., Chang, C., Fan, S., Rosen, E., Palvimo, J. J., Jänne, O. a,

Muratoglu, S., Avantaggiati, M. L. and Pestell, R. G. (2003) 'Acetylation of androgen receptor enhances coactivator binding and promotes prostate cancer cell growth', *Molecular and cellular biology*, 23(23), pp. 8563–75. doi: 10.1128/MCB.23.23.8563.

Ghribi, L., Waffo-Téguo, P., Cluzet, S., Marchal, A., Marques, J., Mérillon, J. M. and Ben Jannet, H. (2015) 'Isolation and structure elucidation of bioactive compounds from the roots of the Tunisian Ononis angustissima L', *Bioorganic and Medicinal Chemistry Letters*. Elsevier Ltd, 25(18), pp. 3825–3830. doi: 10.1016/j.bmcl.2015.07.076.

Goossens, N., Nakagawa, S., Sun, X. and Hoshida, Y. (2015) 'Cancer biomarker discovery and validation', *Translational cancer research*, 4(3), pp. 256–269. doi: 10.3978/j.issn.2218-676X.2015.06.04.

Goswami, C. P. and Nakshatri, H. (2012) 'PROGmiR: a tool for identifying prognostic miRNA biomarkers in multiple cancers using publicly available data', *Journal of Clinical Bioinformatics*, 2(1), p. 23. doi: 10.1186/2043-9113-2-23.

Goswami, C. P. and Nakshatri, H. (2013) 'PROGgene: gene expression based survival analysis web application for multiple cancers', *Journal of clinical bioinformatics*, 3(1), p. 22. doi: 10.1186/2043-9113-3-22.

Gregory, R. I., Yan, K.-P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N. and Shiekhattar, R. (2004) 'The Microprocessor complex mediates the genesis of microRNAs', *Nature*, 432(7014), pp. 235–240. doi: 10.1038/nature03120.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. and Enright, A. J.

(2006) ‘miRBase: microRNA sequences, targets and gene nomenclature’, *Nucleic Acids Research*, 34(suppl_1), pp. D140–D144. doi: 10.1093/nar/gkj112.

Grimson, A., Farh, K. K. H., Johnston, W. K., Garrett-Engele, P., Lim, L. P. and Bartel, D. P. (2007) ‘MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing’, *Molecular Cell*, 27(1), pp. 91–105. doi: 10.1016/j.molcel.2007.06.017.

Grishok, A., Pasquinelli, A. E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D. L., Fire, A., Ruvkun, G. and Mello, C. C. (2001) ‘Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing’, *Cell*, 106(1), pp. 23–34. doi: 10.1016/S0092-8674(01)00431-7.

Guo, L., Yang, S., Zhao, Y., Zhang, H., Wu, Q. and Chen, F. (2014) ‘Global analysis of miRNA gene clusters and gene families reveals dynamic and coordinated expression’, *BioMed Research International*, 2014. doi: 10.1155/2014/782490.

Gurel, B., Ali, T. Z., Montgomery, E. A., Begum, S., Hicks, J., Goggins, M., Eberhart, C. G., Clark, D. P., Bieberich, C. J., Epstein, J. I. and De Marzo, A. M. (2010) ‘NKX3.1 as a Marker of Prostatic Origin in Metastatic Tumors’, *The American Journal of Surgical Pathology*, 34(8), pp. 1097–1105. doi: 10.1097/PAS.0b013e3181e6cbf3.

Gwizdek, C., Ossareh-Nazari, B., Brownawell, A. M., Doglio, A., Bertrand, E., Macara, I. G. and Dargemont, C. (2003) ‘Exportin-5 mediates nuclear export of minihelix-containing RNAs’, *Journal of Biological Chemistry*, 278(8), pp. 5505–

5508. doi: 10.1074/jbc.C200668200.

Hammond, S. M. (2006) 'RNAi, microRNAs, and human disease', in *Cancer Chemotherapy and Pharmacology*. doi: 10.1007/s00280-006-0318-2.

Han, J., Lee, Y., Yeom, K. H., Nam, J. W., Heo, I., Rhee, J. K., Sohn, S. Y., Cho, Y., Zhang, B. T. and Kim, V. N. (2006) 'Molecular Basis for the Recognition of Primary microRNAs by the Drosha-DGCR8 Complex', *Cell*, 125(5), pp. 887–901. doi: 10.1016/j.cell.2006.03.043.

Han, J., Lee, Y., Yeom, K., Kim, Y., Jin, H. and Kim, V. N. (2004) 'The Drosha – DGCR8 complex in primary microRNA processing', *Genes & Development*, pp. 3016–3027. doi: 10.1101/gad.1262504.mic.

Harris, L., Fritsche, H., Mennel, R., Norton, L., Ravdin, P., Taube, S., Somerfield, M. R., Hayes, D. F., Bast, R. C. and American Society of Clinical Oncology (2007) 'American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer', *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 25(33), pp. 5287–312. doi: 10.1200/JCO.2007.14.2364.

Hazelton, W. D. and Luebeck, E. G. (2011) 'Biomarker-Based Early Cancer Detection: Is It Achievable?', *Science Translational Medicine*, 3(109), p. 109fs9-109fs9. doi: 10.1126/scitranslmed.3003272.

Heidenreich, A., Bellmunt, J., Bolla, M., Joniau, S., Mason, M., Matveev, V., Mottet, N., Schmid, H. P., Van Der Kwast, T., Wiegel, T. and Zattoni, F. (2011) 'EAU guidelines on prostate cancer. Part 1: Screening, diagnosis, and treatment of

clinically localised disease', *European Urology*, pp. 61–71. doi: 10.1016/j.eururo.2010.10.039.

Heinlein, C. A. and Chang, C. (2004) 'Androgen receptor in prostate cancer', *Endocrine Reviews*, pp. 276–308. doi: 10.1210/er.2002-0032.

Helfand, B. T. (2016) 'A comparison of genetic risk score with family history for estimating prostate cancer risk', *Asian journal of andrology*, 18(4), pp. 515–9. doi: 10.4103/1008-682X.177122.

Hinds, P. W. and Weinberg, R. A. (1994) 'Tumor suppressor genes', *Curr.Opin.Genet.Devel.*, 4, pp. 135–141.

Holzbeierlein, J., Lal, P., LaTulippe, E., Smith, A., Satagopan, J., Zhang, L., Ryan, C., Smith, S., Scher, H., Scardino, P., Reuter, V. and Gerald, W. L. (2004) 'Gene expression analysis of human prostate carcinoma during hormonal therapy identifies androgen-responsive genes and mechanisms of therapy resistance', *The American Journal of Pathology*, 164(1), pp. 217–27. doi: 10.1016/S0002-9440(10)63112-4.

Hong, S. K. (2014) 'Kallikreins as Biomarkers for Prostate Cancer', *BioMed Research International*. doi: 10.1155/2014/526341.

Horning, A. M., Awe, J. A., Wang, C. M., Liu, J., Lai, Z., Wang, V. Y., Jadhav, R. R., Louie, A. D., Lin, C. L., Kroczak, T., Chen, Y., Jin, V. X., Abboud-Werner, S. L., Leach, R. J., Hernandez, J., Thompson, I. M., Saranchuk, J., Drachenberg, D., Chen, C. L., Mai, S. and Huang, T. H. M. (2015) 'DNA methylation screening of primary prostate tumors identifies *SRD5A2* and CYP11A1 as candidate

markers for assessing risk of biochemical recurrence', *Prostate*, 75(15), pp. 1790–1801. doi: 10.1002/pros.23052.

Huang, D. W., Lempicki, R. a and Sherman, B. T. (2009) 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.', *Nature Protocols*, 4(1), pp. 44–57. doi: 10.1038/nprot.2008.211.

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., Lane, H. C. and Lempicki, R. A. (2007) 'DAVID Bioinformatics Resources: Expanded annotation database and novel algorithms to better extract biology from large gene lists', *Nucleic Acids Research*, 35(SUPPL.2). doi: 10.1093/nar/gkm415.

Istrail, S. and Davidson, E. H. (2005) 'Logic functions of the genomic cis-regulatory code', *Proceedings of the National Academy of Sciences of the United States of America*, 102(14), pp. 4954–4959. doi: 10.1073/pnas.0409624102.

Italiano, A. (2011) 'Prognostic or predictive? It's time to get back to definitions!', *Journal of Clinical Oncology*, 29(35), p. 4718. doi: 10.1200/JCO.2011.38.3729.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C. and Marks, D. S. (2004) 'Human MicroRNA Targets', *PLoS Biology*, 2(11), p. e363. doi: 10.1371/journal.pbio.0020363.

Johnson, C. D., Esquela-Kerscher, A., Stefani, G., Byrom, M., Kelnar, K., Ovcharenko, D., Wilson, M., Wang, X., Shelton, J., Shingara, J., Chin, L., Brown, D. and Slack, F. J. (2007) 'The let-7 microRNA represses cell proliferation pathways in human cells', *Cancer Research*, 67(16), pp. 7713–7722. doi:

10.1158/0008-5472.CAN-07-1083.

Jones-Rhoades, M. W., Bartel, D. P. and Bartel, B. (2006) 'MicroRNAs and their regulatory roles in plants', *Annual Review of Plant Biology*, 57(1), pp. 19–53. doi: 10.1146/annurev.arplant.57.032905.105218.

Jones, P. a and Baylin, S. B. (2002) 'The fundamental role of epigenetic events in cancer.', *Nature Reviews Genetics*, 3(6), pp. 415–28. doi: 10.1038/nrg816.

Kanehisa, M. and Goto, S. (2000) 'Kyoto Encyclopedia of Genes and Genomes', *Nucleic Acids Research*, 28, pp. 27–30. doi: 10.1093/nar/28.1.27.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. and Segal, E. (2007) 'The role of site accessibility in microRNA target recognition', *Nature Genetics*, 39(10), pp. 1278–1284. doi: 10.1038/ng2135.

Khvorova, A., Reynolds, A. and Jayasena, S. D. (2003) 'Functional siRNAs and miRNAs exhibit strand bias', *Cell*, 115(2), pp. 209–216. doi: 10.1016/S0092-8674(03)00801-8.

Knight, S. W. (2001) 'A Role for the RNase III Enzyme DCR-1 in RNA Interference and Germ Line Development in Caenorhabditis elegans', *Science*, 293(5538), pp. 2269–2271. doi: 10.1126/science.1062039.

Koonin, E. V and Galperin, M. Y. (2003) *Sequence - Evolution - Function*, *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21089240.

Kosaka, N., Iguchi, H. and Ochiya, T. (2010) 'Circulating microRNA in body

fluid: A new potential biomarker for cancer diagnosis and prognosis', *Cancer Science*, 101(10), pp. 2087–2092. doi: 10.1111/j.1349-7006.2010.01650.x.

Kryza, T., Silva, L. M., Bock, N., Fuhrman-Luck, R. A., Stephens, C. R., Gao, J., Samaratunga, H., Lawrence, M. G., Hooper, J. D., Dong, Y., Risbridger, G. P. and Clements, J. A. (2017) 'Kallikrein-related peptidase 4 induces cancer-associated fibroblast features in prostate-derived stromal cells', *Molecular Oncology*, 11(10), pp. 1307–1329. doi: 10.1002/1878-0261.12075.

Kulasingam, V. and Diamandis, E. P. (2008) 'Strategies for discovering novel cancer biomarkers through utilization of emerging technologies', *Nature Clinical Practice Oncology*, 5(10), pp. 588–599. doi: 10.1038/ncponc1187.

Kyprianou, N. (1994) 'Apoptosis : therapeutic significance in the treatment of androgen-dependent and androgen-independent prostate cancer', *World Journal of Urology*, 12(6), pp. 299–303.

de la Rosa, J., Weber, J., Friedrich, M. J., Li, Y., Rad, L., Ponstingl, H., Liang, Q., de Quirós, S. B., Noorani, I., Metzakopian, E., Strong, A., Li, M. A., Astudillo, A., Fernández-García, M. T., Fernández-García, M. S., Hoffman, G. J., Fuente, R., Vassiliou, G. S., Rad, R., López-Otín, C., Bradley, A. and Cadiñanos, J. (2017) 'A single-copy Sleeping Beauty transposon mutagenesis screen identifies new PTEN-cooperating tumor suppressor genes', *Nature Genetics*, 49(5), pp. 730–741. doi: 10.1038/ng.3817.

Lan, H., Lu, H., Wang, X. and Jin, H. (2015) 'MicroRNAs as potential biomarkers in cancer: Opportunities and challenges', *BioMed Research International*. doi: 10.1155/2015/125094.

Larrea, E., Sole, C., Manterola, L., Goicoechea, I., Armesto, M., Arestin, M., Caffarel, M. M., Araujo, A. M., Araiz, M., Fernandez-Mercado, M. and Lawrie, C. H. (2016) 'New concepts in cancer biomarkers: Circulating miRNAs in liquid biopsies', *International Journal of Molecular Sciences*. doi: 10.3390/ijms17050627.

Ledermann, J., Harter, P., Gourley, C., Friedlander, M., Vergote, I., Rustin, G., Scott, C., Meier, W., Shapira-Frommer, R., Safra, T., Matei, D., Macpherson, E., Watkins, C., Carmichael, J. and Matulonis, U. (2012) 'Olaparib Maintenance Therapy in Platinum-Sensitive Relapsed Ovarian Cancer', *New England Journal of Medicine*, 366(15), pp. 1382–1392. doi: 10.1056/NEJMoa1105535.

Lee, C.-H., Pao, J.-B., Lu, T.-L., Lee, H.-Z., Lee, Y.-C., Liu, C.-C., Huang, C.-Y., Lin, V. C., Yu, C.-C., Yin, H.-L., Huang, S.-P. and Bao, B.-Y. (2016) 'Prognostic value of prostaglandin-endoperoxide synthase 2 polymorphisms in prostate cancer recurrence after radical prostatectomy', *International Journal of Medical Sciences*, 13(9). doi: 10.7150/ijms.16259.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S. and Kim, V. N. (2003) 'The nuclear RNase III Drosha initiates microRNA processing.', *Nature*, 425(6956), pp. 415–419. doi: 10.1038/nature01957.

Lee, Y., Hur, I., Park, S.-Y., Kim, Y.-K., Suh, M. R. and Kim, V. N. (2006) 'The role of PACT in the RNA silencing pathway', *The EMBO Journal*, 25(3), pp. 522–532. doi: 10.1038/sj.emboj.7600942.

Lee, Y., Jeon, K., Lee, J., Kim, S. and Kim, V. N. (2002) 'MicroRNA maturation :

stepwise processing and subcellular localization', 21(17), pp. 4663–4670.

Lee, Y. S., Nakahara, K., Pham, J. W., Kim, K., He, Z., Sontheimer, E. J. and Carthew, R. W. (2004) 'Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways', *Cell*, 117(1), pp. 69–81. doi: 10.1016/S0092-8674(04)00261-2.

Lenkala, D., LaCroix, B., Gamazon, E. R., Geeleher, P., Im, H. K. and Huang, R. S. (2014) 'The impact of microRNA expression on cellular proliferation', *Human Genetics*, 133(7), pp. 931–938. doi: 10.1007/s00439-014-1434-4.

Lewis, B. P., Burge, C. B. and Bartel, D. P. (2005) 'Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets', *Cell*, 120(1), pp. 15–20. doi: 10.1016/j.cell.2004.12.035.

Li, W. and Godzik, A. (2006) 'Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences', *Bioinformatics*, 22(13), pp. 1658–1659. doi: 10.1093/bioinformatics/btl158.

Lin, H. and Wooley, J. (2005) *Computational Modeling and Simulation as Enablers for Biological Discovery, Catalyzing inquiry at the interface of computing and biology*. Available at: http://www.nap.edu/catalog/11480.htm%5Cnhttp://books.google.com/books?hl=en&lr=&id=sQOcAgAAQBAJ&oi=fnd&pg=PT21&dq=Catalyzing+Inquiry+at+the+Interface+of+Computing+and+Biology&ots=j_4AHE4ATj&sig=1XVELVxU1nbcRmp3oEz0LRTnm3E.

Di Lisio, L., Martinez, N., Montes-Moreno, S., Piris-Villaespesa, M., Sanchez-

Beato, M. and Piris, M. A. (2012) 'The role of miRNAs in the pathogenesis and diagnosis of B-cell lymphomas', *Blood*, pp. 1782–1790. doi: 10.1182/blood-2012-05-402784.

Litjens, R. J., Hopman, A. H., van de Vijver, K. K., Ramaekers, F. C., Kruitwagen, R. F. and Kruse, A.-J. (2013) 'Molecular biomarkers in cervical cancer diagnosis: a critical appraisal', *Expert Opinion on Medical Diagnostics*, 7(4), pp. 365–377. doi: 10.1517/17530059.2013.808621.

Liu, Q. (2003) 'R2D2, a Bridge Between the Initiation and Effector Steps of the Drosophila RNAi Pathway', *Science*, 301(5641), pp. 1921–1925. doi: 10.1126/science.1088710.

Liu, X., Yu, X., Zack, D. J., Zhu, H. and Qian, J. (2008a) 'TiGER: a database for tissue-specific gene expression and regulation', *BMC Bioinformatics*, 9(1), p. 271. doi: 10.1186/1471-2105-9-271.

Liu, X., Yu, X., Zack, D. J., Zhu, H. and Qian, J. (2008b) 'TiGER: a database for tissue-specific gene expression and regulation.', *BMC bioinformatics*, 9(1), p. 271. doi: 10.1186/1471-2105-9-271.

Liu, Z., Guo, Y., Pu, X. and Li, M. (2016) 'Dissecting the regulation rules of cancer-related miRNAs based on network analysis', *Scientific Reports*, 6(October), p. 34172. doi: 10.1038/srep34172.

Loeb, S., Bjurlin, M. A., Nicholson, J., Tammela, T. L., Penson, D. F., Carter, H. B., Carroll, P. and Etzioni, R. (2014) 'Overdiagnosis and overtreatment of prostate cancer', *European Urology*, pp. 1046–1055. doi:

10.1016/j.eururo.2013.12.062.

Loeb, S., Vellekoop, A., Ahmed, H. U., Catto, J., Emberton, M., Nam, R., Rosario, D. J., Scattoni, V. and Lotan, Y. (2013) 'Systematic review of complications of prostate biopsy', *European Urology*. European Association of Urology, 64(6), pp. 876–892. doi: 10.1016/j.eururo.2013.05.049.

Lund, E., Güttinger, S., Calado, A., Dahlberg, J. E. and Kutay, U. (2004) 'Nuclear export of microRNA precursors.', *Science*, 303(5654), pp. 95–98. doi: 10.1126/science.1090599.

Luscombe, N. M., Greenbaum, D. and Gerstein, M. (2001) 'What is bioinformatics? A Proposed Definition and Overview of the Field', *Methods of information in medicine*, 40(4) , pp. 346–358.

Macha, M., Seshacharyulu, P., Krishn, S., Pai, P., Rachagani, S., Jain, M. and Batra, S. (2014) 'MicroRNAs (miRNAs) as Biomarker(s) for Prognosis and Diagnosis of Gastrointestinal (GI) Cancers', *Current Pharmaceutical Design*, 20(33), pp. 5287–5297. doi: 10.1016/j.biotechadv.2011.08.021.Secreted.

Madu, C. O. and Lu, Y. (2010) 'Novel diagnostic biomarkers for prostate cancer', *Journal of Cancer*, 1(1), pp. 150–177. doi: 10.1016/j.suronc.2009.02.006.

Makridakis, N. M., Ross, R. K., Pike, M. C., Crocitto, L. E., Kolonel, L. N., Pearce, C. L., Henderson, B. E. and Reichardt, J. K. V (1999) 'Association of mis-sense substitution in *SRD5A2* gene with prostate cancer in African-American and Hispanic men in Los Angeles, USA', *Lancet*, 354(9183), pp. 975–978. doi: 10.1016/S0140-6736(98)11282-5.

Mancarella, C., Casanova-Salas, I., Calatrava, A., García-Flores, M., Garofalo, C., Grilli, A., Rubio-Briones, J., Scotlandi, K. and López-Guerrero, J. A. (2017a) 'Insulin-like growth factor 1 receptor affects the survival of primary prostate cancer patients depending on TMPRSS2-ERG status', *BMC Cancer*, 17(1), p. 367. doi: 10.1186/s12885-017-3356-8.

Mancarella, C., Casanova-Salas, I., Calatrava, A., García-Flores, M., Garofalo, C., Grilli, A., Rubio-Briones, J., Scotlandi, K. and López-Guerrero, J. A. (2017b) 'Insulin-like growth factor 1 receptor affects the survival of primary prostate cancer patients depending on TMPRSS2-ERG status', *BMC Cancer*, 17(1). doi: 10.1186/s12885-017-3356-8.

Maniataki, E. and Mourelatos, Z. (2005) 'A human, ATP-independent, RISC assembly machine fueled by pre-miRNA', *Genes and Development*, 19(24), pp. 2979–2990. doi: 10.1101/gad.1384005.

Manne, U., Srivastava, R.-G. and Srivastava, S. (2005) 'Keynote review: Recent advances in biomarkers for cancer diagnosis and treatment', *Drug Discovery Today*, 10(14), pp. 965–976. doi: 10.1016/S1359-6446(05)03487-2.

Mayeux, R. (2004) 'Biomarkers: Potential uses and limitations', *J. Am. Doc. Exp. NeuroThera.*, pp. 182–188.

Mayeux, R. (2004) 'Biomarkers: potential uses and limitations.', *NeuroRx : The Journal of the American Society for Experimental NeuroTherapeutics*, 1(2), pp. 182–8. doi: 10.1602/neurorx.1.2.182.

McCaskill, J. S. (1990) 'The equilibrium partition function and base pair binding

probabilities for RNA secondary structure', *Biopolymers*, 29(6–7), pp. 1105–1119. doi: 10.1002/bip.360290621.

Mehta, H. H., Gao, Q., Galet, C., Paharkova, V., Wan, J., Said, J., Sohn, J. J., Lawson, G., Cohen, P., Cobb, L. J. and Lee, K. W. (2011) 'IGFBP-3 is a metastasis suppression gene in prostate cancer', *Cancer Research*, 71(15), pp. 5154–5163. doi: 10.1158/0008-5472.CAN-10-4513.

Meisel-Sharon, S., Pozniak, Y., Geiger, T. and Werner, H. (2016) '*TMPRSS2-ERG* fusion protein regulates insulin-like growth factor-1 receptor (*IGF1R*) gene expression in prostate cancer: involvement of transcription factor Sp1', *Oncotarget*, 7(32), pp. 51375–51392. doi: 10.18632/oncotarget.9837.

Melo, S. A. and Esteller, M. (2011) 'Dysregulation of microRNAs in cancer: Playing with fire', *FEBS Letters*, pp. 2087–2099. doi: 10.1016/j.febslet.2010.08.009.

Mendes, F., Domingues, C., Rodrigues-Santos, P., Abrantes, A. M., Gonçalves, A. C., Estrela, J., Encarnação, J., Pires, A. S., Laranjo, M., Alves, V., Teixo, R., Sarmento, A. B., Botelho, M. F. and Rosa, M. S. (2016) 'The role of immune system exhaustion on cancer cell escape and anti-tumor immune induction after irradiation', *Biochimica et Biophysica Acta - Reviews on Cancer*. Elsevier B.V., 1865(2), pp. 168–175. doi: 10.1016/j.bbcan.2016.02.002.

Miranda, K. C., Huynh, T., Tay, Y., Ang, Y. S., Tam, W. L., Thomson, A. M., Lim, B. and Rigoutsos, I. (2006) 'A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes', *Cell*, 126(6), pp. 1203–1217. doi: 10.1016/j.cell.2006.07.031.

Mizuno, H., Kitada, K., Nakai, K. and Sarai, A. (2009) 'PrognoScan: a new database for meta-analysis of the prognostic value of genes', *BMC Medical Genomics*, 2(1), p. 18. doi: 10.1186/1755-8794-2-18.

Musilova, K. and Mraz, M. (2015) 'MicroRNAs in B-cell lymphomas: how a complex biology gets more complex', *Leukemia*, 29(5), pp. 1004–1017. doi: 10.1038/leu.2014.351.

Nair, M., Sandhu, S. S. and Sharma, A. K. (2014) 'Prognostic and Predictive Biomarkers in Cancer', *Curr Cancer Drug Targets*, 14(5), pp. 477–504. doi: 10.2174/1568009614666140506111118.

Nakielny, S. and Dreyfuss, G. (1999) 'Transport of Proteins and RNAs in and out of the Nucleus', *Cell*, 99(7), pp. 677–690. doi: 10.1016/S0092-8674(00)81666-9.

Nalejska, E., Mączyńska, E. and Lewandowska, M. A. (2014) 'Prognostic and predictive biomarkers: Tools in personalized oncology', *Molecular Diagnosis and Therapy*, pp. 273–284. doi: 10.1007/s40291-013-0077-9.

Nam, S., Kim, B., Shin, S. and Lee, S. (2008) 'miRGator: An integrated system for functional annotation of microRNAs', *Nucleic Acids Research*, 36(SUPPL. 1). doi: 10.1093/nar/gkm829.

Novelli, G., Ciccacci, C., Borgiani, P., Amati, M. P. and Abadie, E. (2008) 'Genetic tests and genomic biomarkers: Regulation, qualification and validation', *Clinical Cases in Mineral and Bone Metabolism*, pp. 149–154.

Oldenhuis, C. N. A. M., Oosting, S. F., Gietema, J. A. and de Vries, E. G. E. (2008) 'Prognostic versus predictive value of biomarkers in oncology', *European*

Journal of Cancer, 44(7), pp. 946–953. doi: 10.1016/j.ejca.2008.03.006.

Pakzad, R., Mohammadian-Hafshejani, A., Ghoncheh, M., Pakzad, I. and Salehiniya, H. (2015) 'The incidence and mortality of prostate cancer and its relationship with development in Asia', *Prostate International*, 3(4), pp. 135–140. doi: 10.1016/j.prnil.2015.09.001.

Pal, M. K., Jaiswar, S. P., Dwivedi, V. N., Tripathi, A. K., Dwivedi, A. and Sankhwar, P. (2015) 'MicroRNA: a new and promising potential biomarker for diagnosis and prognosis of ovarian cancer.', *Cancer biology & medicine*, 12(4), pp. 328–41. doi: 10.7497/j.issn.2095-3941.2015.0024.

Perlmutter, M. A. and Lepor, H. (2007) 'Androgen deprivation therapy in the treatment of advanced prostate cancer.', *Reviews in urology*, 9 Suppl 1(Suppl 1), pp. S3-8. doi: 10.1016/B978-0-12-800077-9.00059-1.

Phan, J. H., Moffitt, R. A., Stokes, T. H., Liu, J., Young, A. N., Nie, S. and Wang, M. D. (2009) 'Convergence of biomarkers, bioinformatics and nanotechnology for individualized cancer treatment', *Trends in Biotechnology*, 27(6), pp. 350–358. doi: 10.1016/j.tibtech.2009.02.010.

Picchio, M., Giovannini, E. and Messa, C. (2011) 'The role of PET/computed tomography scan in the management of prostate cancer.', *Current opinion in urology*, 21, pp. 230–236. doi: 10.1097/MOU.0b013e328344e556.

Porkka, K. P., Pfeiffer, M. J., Waltering, K. K., Vessella, R. L., Tammela, T. L. J. and Visakorpi, T. (2007) 'MicroRNA expression profiling in prostate cancer', *Cancer Research*, 67(13), pp. 6130–6135. doi: 10.1158/0008-5472.CAN-07-0533.

Prassas, I., Chrystoja, C. C., Makawita, S. and Diamandis, E. P. (2012) 'Bioinformatic identification of proteins with tissue-specific expression for biomarker discovery', *BMC Medicine*, 10. doi: 10.1186/1741-7015-10-39.

Qin, J., Lee, H. J., Wu, S. P., Lin, S. C., Lanz, R. B., Creighton, C. J., DeMayo, F. J., Tsai, S. Y. and Tsai, M. J. (2014) 'Androgen deprivation-induced NCoA2 promotes metastatic and castration-resistant prostate cancer', *Journal of Clinical Investigation*, 124(11), pp. 5013–5026. doi: 10.1172/JCI76412.

Reddy, B. S., Hirose, Y., Lubet, R., Steele, V., Kelloff, G., Paulson, S., Seibert, K. and Rao, C. V. (2000) 'Chemoprevention of colon cancer by specific cyclooxygenase-2 inhibitor, celecoxib, administered during different stages of carcinogenesis', *Cancer Research*, 60(2), pp. 293–297.

Reyes-Aldasoro, C. C. (2017) 'The proportion of cancer-related entries in PubMed has increased considerably; Is cancer truly "the Emperor of All Maladies"?', *PLoS ONE*, 12(3). doi: 10.1371/journal.pone.0173671.

ReyesHerrera, P. H. and Ficarra, E. (2012) 'One Decade of Development and Evolution of MicroRNA Target Prediction Algorithms', *Genomics, Proteomics and Bioinformatics*, 10(5), pp. 254–263. doi: 10.1016/j.gpb.2012.10.001.

Riley, L. B. and Anderson, D. W. (2011) 'Cancer epigenetics', in *Handbook of Epigenetics*, pp. 521–534. doi: 10.1016/B978-0-12-375709-8.00032-0.

Rizzolo, P., Silvestri, V., Falchetti, M. and Ottini, L. (2011) 'Inherited and acquired alterations in development of breast cancer', *Application of Clinical Genetics*, 4, pp. 145–158. doi: 10.2147/TACG.S13226.

Roberts, C. T. (2004) 'IGF-1 and prostate cancer.', *Novartis Foundation symposium*, 262, pp. 193-9-204, 265–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15562830.

Ruepp, A., Kowarsch, A., Schmidl, D., Buggenthin, F., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C. and Theis, F. J. (2010) 'PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes', *Genome biology*, 11(1), p. R6. doi: 10.1186/gb-2010-11-1-r6.

Rufino-Palomares, E. E., Reyes-Zurita, F. J., Lupiáñez, J. A. and Medina, P. P. (2013) 'MicroRNAs as Oncogenes and Tumor Suppressors', in *MicroRNAs in Medicine*, pp. 223–243. doi: 10.1002/9781118300312.ch14.

Sætrom, P., Heale, B. S. E., Snøve, O., Aagaard, L., Alluin, J. and Rossi, J. J. (2007) 'Distance constraints between microRNA target sites dictate efficacy and cooperativity', *Nucleic Acids Research*, 35(7), pp. 2333–2342. doi: 10.1093/nar/gkm133.

Saito, K., Ishizuka, A., Siomi, H. and Siomi, M. C. (2005) 'Processing of pre-microRNAs by the Dicer-1-Loquacious complex in drosophila cells', *PLoS Biology*, 3(7), pp. 1202–1212. doi: 10.1371/journal.pbio.0030235.

Sanjay, S. T., Fu, G., Dou, M., Xu, F., Liu, R., Qi, H. and Li, X. (2015) 'Biomarker detection for disease diagnosis using cost-effective microfluidic platforms', *The Analyst*, 140(21), pp. 7062–7081. doi: 10.1039/C5AN00780A.

Sauvé, J.-F., Lavoué, J. and Parent, M.-É. (2016) 'Occupation, industry, and the risk of prostate cancer: a case-control study in Montréal, Canada.', *Environmental*

Health , 15(1), p. 100. doi: 10.1186/s12940-016-0185-1.

Schetter, A., Leung, S., Sohn, J., Zanetti, K., Bowman, E., Yanaihara, N., Yuen, S., Chan, T., Kwong, D., Au, G., Liu, C., Calin, G., Croce, C. and Harris, C. (2008) 'MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma', *Jama*, 299(4), pp. 425–436. doi: 10.1001/jama.299.4.425.

Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R. and Rajewsky, N. (2008) 'Widespread changes in protein synthesis induced by microRNAs', *Nature.*, 455(7209), pp. 58–63. doi: 10.1038/nature07228.

Sharma, S., Kelly, T. K. and Jones, P. A. (2009) 'Epigenetics in cancer', *Carcinogenesis*, pp. 27–36. doi: 10.1093/carcin/bgp220.

Sharma, S., Zapatero-Rodríguez, J. and O'Kennedy, R. (2017) 'Prostate cancer diagnostics: Clinical challenges and the ongoing need for disruptive and effective diagnostic tools', *Biotechnology Advances*, pp. 135–149. doi: 10.1016/j.biotechadv.2016.11.009.

Sherr, C. J. (2004) 'Principles of Tumor Suppression', *Cell*, pp. 235–246. doi: 10.1016/S0092-8674(03)01075-4.

Shirdel, E. A., Xie, W., Mak, T. W. and Jurisica, I. (2011) 'NAViGaTing the micronome - using multiple microRNA prediction databases to identify signalling pathway-associated microRNAs', *PLoS ONE*, 6(2). doi: 10.1371/journal.pone.0017429.

Siegel, R., Naishadham, D. and Jemal, A. (2012) 'Cancer Statistics , 2012', *CA*

Cancer J Clin, 62, pp. 10–29. doi: 10.3322/caac.20138.Available.

Silver, D. P. and Livingston, D. M. (2012) 'Mechanisms of BRCA1 tumor suppression', *Cancer Discovery*, pp. 679–684. doi: 10.1158/2159-8290.CD-12-0221.

Sissung, T. M., Price, D. K., Del Re, M., Ley, A. M., Giovannetti, E., Figg, W. D. and Danesi, R. (2014) 'Genetic variation: effect on prostate cancer', *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1846(2), pp. 446–456. doi: 10.1016/j.bbcan.2014.08.007.

Smith, M. R., Manola, J., Kaufman, D. S., Oh, W. K., Bubley, G. J. and Kantoff, P. W. (2006) 'Celecoxib versus placebo for men with prostate cancer and a rising serum prostate-specific antigen after radical prostatectomy and/or radiation therapy', *Journal of Clinical Oncology*, 24(18), pp. 2723–2728. doi: 10.1200/JCO.2005.03.7804.

Song, L.-N., Silva, J., Koller, A., Rosenthal, A., Chen, E. I. and Gelmann, E. P. (2015) 'The Tumor Suppressor NKX3.1 Is Targeted for Degradation by DYRK1B Kinase.', *Molecular Cancer Research* , 13(5), pp. 913–922. doi: 10.1158/1541-7786.MCR-14-0680.

Stangelberger, A., Waldert, M. and Djavan, B. (2008) 'Prostate cancer in elderly men', *Reviews in Urology*, 10(2), pp. 111–9. doi: 10.1016/j.cger.2015.07.004.

Stark, A., Brennecke, J., Russell, R. B. and Cohen, S. M. (2003) 'Identification of Drosophila microRNA targets', *PLoS Biology*, 1(3). doi: 10.1371/journal.pbio.0000060.

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. and Von Mering, C. (2015) 'STRING v10: Protein-protein interaction networks, integrated over the tree of life', *Nucleic Acids Research*, 43(D1), pp. D447–D452. doi: 10.1093/nar/gku1003.

Tabara, H., Yigit, E., Siomi, H. and Mello, C. C. (2002) 'The double-stranded RNA binding protein RDE-4 interacts in vivo with RDE-1, DCR-1 and a conserved DExH-box helicase to direct RNA interference in C. elegans', *Cell*, 109, pp. 861–871.

Tadayon, F., Arezegar, H. R., Khorrami, M. H., Hashemi Juzdani, R., Shahdoost, A. A. and Mellat, M. (2016) 'Evaluation of prostatic cancer prevalence in patients with prostatic-specific antigen between 4 and 10 and normal digital rectal examination', *Advanced Biomedical Research*, 5(112). doi: 10.4103/2277-9175.184298.

Tan, M. E., Li, J., Xu, H. E., Melcher, K. and Yong, E. L. (2015) 'Androgen receptor: Structure, role in prostate cancer and drug discovery', *Acta Pharmacologica Sinica*, pp. 3–23. doi: 10.1038/aps.2014.18.

Tannour-Louet, M., Lewis, S. K., Louet, J. F., Stewart, J., Addai, J. B., Sahin, A., Vangapandu, H. V., Lewis, A. L., Dittmar, K., Pautler, R. G., Zhang, L., Smith, R. G. and Lamb, D. J. (2014) 'Increased expression of *CYP24A1* correlates with advanced stages of prostate cancer and can cause resistance to vitamin D3-based therapies', *FASEB Journal*, 28(1), pp. 364–372. doi: 10.1096/fj.13-236109.

Teng, Z., Guo, M., Liu, X., Dai, Q., Wang, C. and Xuan, P. (2013) 'Measuring

gene functional similarity based on group-wise comparison of GO terms', *Bioinformatics*, 29(11), pp. 1424–1432. doi: 10.1093/bioinformatics/btt160.

Terpos, E., Dimopoulos, M. A., Shrivastava, V., Leitzel, K., Christoulas, D., Migkou, M., Gavriatopoulou, M., Anargyrou, K., Hamer, P., Kastritis, E., Carney, W. and Lipton, A. (2010) 'High levels of serum TIMP-1 correlate with advanced disease and predict for poor survival in patients with multiple myeloma treated with novel agents', *Leukemia Research*, 34(3), pp. 399–402. doi: 10.1016/j.leukres.2009.08.035.

Terstappen, G. C. and Reggiani, A. (2001) '*In silico* research in drug discovery', *Trends in Pharmacological Sciences*, pp. 23–26. doi: 10.1016/S0165-6147(00)01584-4.

Tewari, A. (2013) *Prostate Cancer: A comprehensive perspective*, *Prostate Cancer: A Comprehensive Perspective*. doi: 10.1007/978-1-4471-2864-9.

Thomas, M., Poignée-Heger, M., Weisser, M., Wessner, S. and Belousov, A. (2013) 'An optimized workflow for improved gene expression profiling for formalin-fixed, paraffin-embedded tumor samples.', *Journal of Clinical Bioinformatics*, 3(1), p. 10. doi: 10.1186/2043-9113-3-10.

Thompson, C. B. (1995) 'Apoptosis in the pathogenesis and treatment of disease.', *Science (New York, N.Y.)*, 267(5203), pp. 1456–62. doi: 10.1126/science.7878464.

Thomson, D. W., Bracken, C. P. and Goodall, G. J. (2011) 'Experimental strategies for microRNA target identification', *Nucleic Acids Research*, 39(16),

pp. 6845–6853. doi: 10.1093/nar/gkr330.

Tomari, Y. (2004) 'A Protein Sensor for siRNA Asymmetry', *Science*, 306(5700), pp. 1377–1380. doi: 10.1126/science.1102755.

Tomari, Y. and Zamore, P. D. (2005) 'MicroRNA biogenesis: Drosha can't cut it without a partner', *Current Biology*. doi: 10.1016/j.cub.2004.12.057.

Trewartha, D. and Carter, K. (2013) 'Advances in prostate cancer treatment', *Nature reviews. Drug discovery*, 23(9), pp. 8–12. doi: 10.1038/nrd4068.

Tricoli, J. V. and Jacobson, J. W. (2007) 'MicroRNA: Potential for cancer detection, diagnosis, and prognosis', in *Cancer Research*, pp. 4553–4555. doi: 10.1158/0008-5472.CAN-07-0563.

UK, C. R. (2013) 'Prostate Cancer Incidence Statistics', *Cancer Research UK*, p. 1. Available at: http://www.cancerresearchuk.org/cancer-info/cancerstats/types/prostate/incidence/uk-prostate-cancer-incidence-statistics.

Vasudevan, S., Tong, Y. and Steitz, J. A. (2007) 'Switching from Repression to Activation: MicroRNAs Can Up-Regulate Translation', *Science*, 318(5858), pp. 1931–1934. doi: 10.1126/science.1149460.

Wahid, F., Shehzad, A., Khan, T. and Young, Y. (2010) 'Biochimica et Biophysica Acta MicroRNAs : Synthesis , mechanism , function , and recent clinical trials', *BBA - Molecular Cell Research*. Elsevier B.V., 1803(11), pp. 1231–1243. doi: 10.1016/j.bbamcr.2010.06.013.

Walsh, A. L., Considine, S. W., Thomas, A. Z., Lynch, T. H. and Manecksha, R.

P. (2014) ‘Digital rectal examination in primary care is important for early detection of prostate cancer: a retrospective cohort analysis study.’, *The British journal of general practice : The Journal of the Royal College of General Practitioners*, 64(629), pp. e783-7. doi: 10.3399/bjgp14X682861.

Walsh, P. C. (2006) ‘Operating characteristics of prostate-specific antigen in men with an initial PSA level of 3.0 Ng/Ml or lower: Commentary’, *Journal of Urology*, pp. 562–563. doi: 10.1016/S0022-5347(05)00385-X.

Wang, Z., Wilson, C. L., Easton, J., Hedges, D., Liu, Q., Wu, G., Rusch, M., Edmonson, M., Levy, S., Lanctot, J. Q., Caron, E., Shelton, K., Currie, K., Lear, M., Mulder, H. L., Yergeau, D., Rosencrance, C., Vadodaria, B., Sapkota, Y., Brooke, R. J., Moon, W., Rampersaud, E., Ma, X., Wang, S., Chang, T.-C., Rice, S., Thrasher, A., Patel, A., Pepper, C., Zhou, X., Chen, X., Chen, W., Jones, A., Boone, B., Srivastava, D. K., Kesserwan, C. A., Nichols, K. E., Downing, J. R., Hudson, M. M., Yasui, Y., Robison, L. L. and Zhang, J. (2017) ‘Abstract 3001: Germline mutations in cancer predisposition genes and risk for subsequent neoplasms among long-term survivors of childhood cancer in the St. Jude Lifetime Cohort’, *Cancer Research*, 77(13 Supplement), p. 3001. doi: 10.1158/1538-7445.AM2017-3001.

Whitfield, M. L., George, L. K., Grant, G. D. and Perou, C. M. (2006) ‘Common markers of proliferation’, *Nature Reviews Cancer*, 6(2), pp. 99–106. doi: 10.1038/nrc1802.

Wille, J. J. (2017) ‘Role of Epigenetic Regulation of Growth Arrest States in Normal and Cancer Cells’, *Journal of Clinical Epigenetics*, 3(1), pp. 3–4. doi:

10.21767/2472-1158.100047.

Wuchty, S., Fontana, W., Hofacker, I. L. and Schuster, P. (1999) 'Complete suboptimal folding of RNA and the stability of secondary structures', *Biopolymers*, pp. 145–165. doi: 10.1002/(SICI)1097-0282(199902)49:2<145::AID-BIP4>3.0.CO;2-G.

Xie, B., Ding, Q., Han, H. and Wu, D. (2013) 'MiRCancer: A microRNA-cancer association database constructed by text mining on literature', *Bioinformatics*, 29(5), pp. 638–644. doi: 10.1093/bioinformatics/btt014.

Xu, J. and Wong, C. (2008) 'A computational screen for mouse signaling pathways targeted by microRNA clusters', *RNA*, 14(7), pp. 1276–1283. doi: 10.1261/rna.997708.

Yang, J., Ma, D., Fesler, A., Zhai, H., Leamniramit, A., Li, W., Wu, S. and Ju, J. (2016) 'Expression analysis of microRNA as prognostic biomarkers in colorectal cancer.', *Oncotarget*, 8(32), pp. 52403–52412. doi: 10.18632/oncotarget.14175.

Yang, Z., Wu, L., Wang, A., Tang, W., Zhao, Y., Zhao, H. and Teschendorff, A. E. (2017) 'DbDEMC 2.0: Updated database of differentially expressed miRNAs in human cancers', *Nucleic Acids Research*, 45(D1), pp. D812–D818. doi: 10.1093/nar/gkw1079.

Yousef, M., Najami, N., Abedallah, L. and Khalifa, W. (2014) 'Computational Approaches for Biomarker Discovery', *Journal of Intelligent Learning Systems and Applications*, 6(4), pp. 153–161. doi: 10.4236/jilsa.2014.64012.

Yu, B. (2009) 'Role of *in silico* tools in gene discovery', *Molecular*

Biotechnology, pp. 296–306. doi: 10.1007/s12033-008-9134-8.

Yu, J., Yu, J., Mani, R. S., Cao, Q., Brenner, C. J., Cao, X., Wang, X., Wu, L., Li, J., Hu, M., Gong, Y., Cheng, H., Laxman, B., Vellaichamy, A., Shankar, S., Li, Y., Dhanasekaran, S. M., Morey, R., Barrette, T., Lonigro, R. J., Tomlins, S. A., Varambally, S., Qin, Z. S. and Chinnaiyan, A. M. (2010) 'An Integrated Network of Androgen Receptor, Polycomb, and TMPRSS2-ERG Gene Fusions in Prostate Cancer Progression', *Cancer Cell*, 17(5), pp. 443–454. doi: 10.1016/j.ccr.2010.03.018.

Yu, X., Lin, J., Zack, D. J. and Qian, J. (2006) 'Computational analysis of tissue-specific combinatorial gene regulation: Predicting interaction between transcription factors in human tissues', *Nucleic Acids Research*, 34(17), pp. 4925–4936. doi: 10.1093/nar/gkl595.

Yu, X., Lin, J., Zack, D. J. and Qian, J. (2007) 'Identification of tissue-specific cis-regulatory modules based on interactions between transcription factors', *BMC Bioinformatics*, 8(1), p. 437. doi: 10.1186/1471-2105-8-437.

Zaman, M. S., Chen, Y., Deng, G., Shahryari, V., Suh, S. O., Saini, S., Majid, S., Liu, J., Khatri, G. and Tanaka, Y. (2010) 'The functional significance of microRNA-145 in prostate cancer', *British Journal of Cancer*. Nature Publishing Group, 103(2), pp. 256–264. doi: 10.1038/sj.bjc.6605742.

Zeng, Y. A. N. and Cullen, B. R. (2003) 'Sequence requirements for micro RNA processing and function in human cells', pp. 112–123. doi: 10.1261/rna.2780503.known.

Zeng, Y. and Cullen, B. R. (2004) 'Structural requirements for pre-microRNA binding and nuclear export by Exportin 5', *Nucleic Acids Research*, 32(16), pp. 4776–4785. doi: 10.1093/nar/gkh824.

Zeng, Y. and Cullen, B. R. (2005) 'Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences', *Journal of Biological Chemistry*, 280(30), pp. 27595–27603. doi: 10.1074/jbc.M504714200.

Zhang, B., Pan, X., Cobb, G. P. and Anderson, T. A. (2007) 'microRNAs as oncogenes and tumor suppressors', *Developmental Biology*, 302(1), pp. 1–12. doi: 10.1016/j.ydbio.2006.08.028.

Zheng, H., Fu, R., Wang, J. T., Liu, Q., Chen, H. and Jiang, S. W. (2013) 'Advances in the techniques for the prediction of microRNA targets', *International Journal of Molecular Sciences*, 14(4), pp. 8179–8187. doi: 10.3390/ijms14048179.

Zicari, R. V (2014) 'Big Data: Challenges and Opportunities', *Big data computing*, pp. 103–128. doi: 10.1007/978-3-319-13206-8_1.

Appendix A

Supplementary table for chapter 2

Table A.1: list of validated miRNAs in prostate cancer with their reference

MiRNAs name	Accession number	References
hsa-let-7a-3p	MIMAT0004481	Wang M. et al., 2011
hsa-let-7a-5p	MIMAT0000062	Wang M. et al., 2011
hsa-let-7b-5p	MIMAT0000063	Liu C. et al., 2012
hsa-let-7c-5p	MIMAT0000064	Chen Z. H. et al., 2012
hsa-let-7d-5p	MIMAT0000065	Ramberg H. et al., 2011
hsa-let-7e-5p	MIMAT0000066	Chen Z. H. et al., 2012
hsa-let-7g-5p	MIMAT0000414	Prueitt R. L. et al., 2008
hsa-miR-1	MIMAT0000416	Liu Y. N. et al., 2013
hsa-miR-100-5p	MIMAT0000098	Leite K. R. et al., 2011
hsa-miR-101-3p	MIMAT0000099	Hao Y. et al., 2011
hsa-miR-101-5p	MIMAT0004513	Hao Y. et al., 2011
hsa-miR-106a-5p	MIMAT0000103	Liu C. et al., 2012
hsa-miR-106b-5p	MIMAT0000680	Hudson R. S. et al., 2012
hsa-miR-107	MIMAT0000104	Bryant R. J. et al., 2012
hsa-miR-10a-5p	MIMAT0000253	Prueitt R. L. et al., 2008
hsa-miR-10b-5p	MIMAT0000254	Prueitt R. L. et al., 2008
hsa-miR-122-5p	MIMAT0000421	Zhang Z. et al., 2012
hsa-miR-1236-3p	MIMAT0005591	Bryant R. J. et al., 2012
hsa-miR-1256	MIMAT0005907	Li Y. et al., 2012
hsa-miR-125b-1-3p	MIMAT0004592	Feng N. et al., 2011
hsa-miR-125b-2-3p	MIMAT0004603	Feng N. et al., 2011
hsa-miR-125b-5p	MIMAT0000423	Feng N. et al., 2011

hsa-miR-126-3p	MIMAT0000445	Watahiki A. et al., 2011
hsa-miR-127-3p	MIMAT0000446	Cahill S. et al., 2007
hsa-miR-127-5p	MIMAT0004604	Cahill S. et al., 2007
hsa-miR-1285-3p	MIMAT0005876	Chen Z. H. et al., 2012
hsa-miR-1296-5p	MIMAT0005794	Majid S. et al., 2010
hsa-miR-130a-3p	MIMAT0000425	Boll K. et al., 2013
hsa-miR-130b-3p	MIMAT0000691	Bryant R. J. et al., 2012
hsa-miR-132-3p	MIMAT0000426	Formosa A. et al., 2013
hsa-miR-133a-3p	MIMAT0000427	Kojima S. et al., 2012
hsa-miR-133b	MIMAT0000770	Mo W. et al., 2013
hsa-miR-135b-5p	MIMAT0000758	Tong A. W. et al., 2009
hsa-miR-141-3p	MIMAT0000432	Nguyen H. C. et al., 2013
hsa-miR-143-3p	MIMAT0000435	Fan X. et al., 2013
hsa-miR-145-5p	MIMAT0000437	Huang S. et al., 2012
hsa-miR-146a-5p	MIMAT0000449	Xu B. et al., 2012
hsa-miR-146b-3p	MIMAT0004766	Volinia S. et al., 2006
hsa-miR-146b-5p	MIMAT0002809	Volinia S. et al., 2006
hsa-miR-148a-3p	MIMAT0000243	Fujita Y. et al., 2010
hsa-miR-149-5p	MIMAT0000450	Schaefer A. et al., 2009
hsa-miR-151a-5p	MIMAT0004697	Chiyomaru T. et al., 2012
hsa-miR-152-3p	MIMAT0000438	Zhu C. et al., 2013
hsa-miR-15a-5p	MIMAT0000068	Leite K. R. et al., 2011
hsa-miR-15b-5p	MIMAT0000417	Musumeci M. et al., 2011
hsa-miR-16-1-3p	MIMAT0004489	Leite K. R. et al., 2011
hsa-miR-16-2-3p	MIMAT0004518	Leite K. R. et al., 2011
hsa-miR-16-5p	MIMAT0000069	Leite K. R. et al., 2011
hsa-miR-17-3p	MIMAT0000071	Bryant R. J. et al., 2012

hsa-miR-17-5p	MIMAT0000070	Xu Y. et al., 2011
hsa-miR-181a-2-3p	MIMAT0004558	Bryant R. J. et al., 2012
hsa-miR-181b-3p	MIMAT0022692	Volinia S. et al., 2006
hsa-miR-181b-5p	MIMAT0000257	Schaefer A. et al., 2009
hsa-miR-182-3p	MIMAT0000260	Schaefer A. et al., 2009
hsa-miR-182-5p	MIMAT0000259	Hirata H. et al., 2013
hsa-miR-183-5p	MIMAT0000261	Mihelich B. L. et al., 2011
hsa-miR-184	MIMAT0000454	Schaefer A. et al., 2009
hsa-miR-191-5p	MIMAT0000440	Leite K. R. et al., 2011
hsa-miR-193a-3p	MIMAT0000459	Mohan K. V. et al., 2012
hsa-miR-193a-5p	MIMAT0004614	Mohan K. V. et al., 2012
hsa-miR-193b-3p	MIMAT0002819	Xie C. et al., 2012
hsa-miR-194-3p	MIMAT0004671	Tong A. W. et al., 2009
hsa-miR-194-5p	MIMAT0000460	Tong A. W. et al., 2009
hsa-miR-195-5p	MIMAT0000461	Mahn R. et al., 2011
hsa-miR-196b-5p	MIMAT0001080	Hulf T. et al., 2011
hsa-miR-198	MIMAT0000228	Bryant R. J. et al., 2012
hsa-miR-199a-3p	MIMAT0000232	Leite K. R. et al., 2011
hsa-miR-199a-5p	MIMAT0000231	Leite K. R. et al., 2011
hsa-miR-19a-3p	MIMAT0000073	Mo W. et al., 2013
hsa-miR-200a-3p	MIMAT0000682	Liu Y. N. et al., 2013
hsa-miR-200b-3p	MIMAT0000318	Liu Y. N. et al., 2013
hsa-miR-200c-3p	MIMAT0000617	Liu Y. N. et al., 2013
hsa-miR-203a	MIMAT0000264	Boll K. et al., 2013
hsa-miR-204-5p	MIMAT0000265	Turner D. P. et al., 2011
hsa-miR-205-5p	MIMAT0000266	Boll K. et al., 2013
hsa-miR-20a-3p	MIMAT0004493	Bryant R. J. et al., 2012

hsa-miR-20a-5p	MIMAT0000075	Li X. et al., 2012
hsa-miR-214-3p	MIMAT0000271	Volinia S. et al., 2006
hsa-miR-21-5p	MIMAT0000076	Li T. et al., 2012
hsa-miR-218-1-3p	MIMAT0004565	Leite K. R. et al., 2011
hsa-miR-218-5p	MIMAT0000275	Leite K. R. et al., 2011
hsa-miR-221-3p	MIMAT0000278	Sun T. et al., 2012
hsa-miR-222-3p	MIMAT0000279	Fuse M. et al., 2012
hsa-miR-223-3p	MIMAT0000280	Volinia S. et al., 2006
hsa-miR-22-3p	MIMAT0000077	Szczyrba J. et al., 2012
hsa-miR-224-5p	MIMAT0000281	Prueitt R. L. et al., 2008
hsa-miR-22-5p	MIMAT0004495	Szczyrba J. et al., 2012
hsa-miR-23a-5p	MIMAT0004496	Bryant R. J. et al., 2012
hsa-miR-23b-3p	MIMAT0000418	He H. C. et al., 2012
hsa-miR-24-3p	MIMAT0000080	Szczyrba J. et al., 2012
hsa-miR-25-3p	MIMAT0000081	Leite K. R. et al., 2011
hsa-miR-26a-1-3p	MIMAT0004499	Mahn R. et al., 2011
hsa-miR-26a-2-3p	MIMAT0004681	Mahn R. et al., 2011
hsa-miR-26a-5p	MIMAT0000082	Mahn R. et al., 2011
hsa-miR-27a-3p	MIMAT0000084	Mo W. et al., 2013
hsa-miR-27b-3p	MIMAT0000419	Sun T. et al., 2012
hsa-miR-296-3p	MIMAT0004679	Wei J. J. et al., 2011
hsa-miR-296-5p	MIMAT0000690	Wei J. J. et al., 2011
hsa-miR-29a-3p	MIMAT0000086	Li Y. et al., 2012
hsa-miR-29b-1-5p	MIMAT0004514	Ru P. et al., 2012
hsa-miR-29b-2-5p	MIMAT0004515	Ru P. et al., 2012
hsa-miR-29b-3p	MIMAT0000100	Ru P. et al., 2012
hsa-miR-301a-3p	MIMAT0000688	Bryant R. J. et al., 2012

hsa-miR-30c-1-3p	MIMAT0004674	Volinia S. et al., 2006
hsa-miR-30c-2-3p	MIMAT0004550	Volinia S. et al., 2006
hsa-miR-30c-5p	MIMAT0000244	Chen Z. H. et al., 2012
hsa-miR-31-5p	MIMAT0000089	Fuse M. et al., 2012
hsa-miR-320a	MIMAT0000510	Hsieh I. S. et al., 2013
hsa-miR-32-5p	MIMAT0000090	Jalava S. E. et al., 2012
hsa-miR-326	MIMAT0000756	Bryant R. J. et al., 2012
hsa-miR-330-3p	MIMAT0000751	Siddiqui I. A. et al., 2011
hsa-miR-330-5p	MIMAT0004693	Siddiqui I. A. et al., 2011
hsa-miR-331-3p	MIMAT0000760	Bryant R. J. et al., 2012
hsa-miR-331-5p	MIMAT0004700	Epis M. R. et al., 2012
hsa-miR-34a-5p	MIMAT0000255	Kong D. et al., 2012
hsa-miR-34c-3p	MIMAT0004677	Hagman Z. et al., 2011
hsa-miR-34c-5p	MIMAT0000686	Hagman Z. et al., 2011
hsa-miR-370-3p	MIMAT0000722	Wu Z. et al., 2012
hsa-miR-373-3p	MIMAT0000726	Iczkowski K. A. et al., 2010
hsa-miR-375	MIMAT0000728	Nguyen H. C. et al., 2013
hsa-miR-378a-3p	MIMAT0000732	Nguyen H. C. et al., 2013
hsa-miR-379-5p	MIMAT0000733	Bryant R. J. et al., 2012
hsa-miR-409-3p	MIMAT0001639	Nguyen H. C. et al., 2013
hsa-miR-409-5p	MIMAT0001638	Nguyen H. C. et al., 2013
hsa-miR-432-5p	MIMAT0002814	Bryant R. J. et al., 2012
hsa-miR-449a	MIMAT0001541	Noonan E. J. et al., 2009
hsa-miR-452-5p	MIMAT0001635	Liu C. et al., 2012
hsa-miR-484	MIMAT0002174	Bryant R. J. et al., 2012
hsa-miR-488-3p	MIMAT0004763	Sikand K. et al., 2011
hsa-miR-513a-5p	MIMAT0002877	Bryant R. J. et al., 2012

hsa-miR-519d-3p	MIMAT0002853	Long Q. et al., 2011
hsa-miR-520c-3p	MIMAT0002846	Iczkowski K. A. et al., 2010
hsa-miR-520c-5p	MIMAT0005455	Iczkowski K. A. et al., 2010
hsa-miR-521	MIMAT0002854	Josson S. et al., 2008
hsa-miR-572	MIMAT0003237	Bryant R. J. et al., 2012
hsa-miR-574-3p	MIMAT0003239	Bryant R. J. et al., 2012
hsa-miR-574-5p	MIMAT0004795	Bryant R. J. et al., 2012
hsa-miR-577	MIMAT0003242	Bryant R. J. et al., 2012
hsa-miR-582-3p	MIMAT0004797	Bryant R. J. et al., 2012
hsa-miR-609	MIMAT0003277	Bryant R. J. et al., 2012
hsa-miR-616-3p	MIMAT0004805	Ma S. et al., 2011
hsa-miR-619-3p	MIMAT0003288	Bryant R. J. et al., 2012
hsa-miR-622	MIMAT0003291	Chen Z. H. et al., 2012
hsa-miR-624-5p	MIMAT0003293	Bryant R. J. et al., 2012
hsa-miR-625-3p	MIMAT0004808	Bryant R. J. et al., 2012
hsa-miR-642a-5p	MIMAT0003312	Epis M. R. et al., 2012
hsa-miR-647	MIMAT0003317	Long Q. et al., 2011
hsa-miR-708-5p	MIMAT0004926	Saini S. et al., 2012
hsa-miR-92a-1-5p	MIMAT0004507	Li Y. et al., 2012
hsa-miR-92a-2-5p	MIMAT0004508	Li Y. et al., 2012
hsa-miR-92a-3p	MIMAT0000092	Li Y. et al., 2012
hsa-miR-93-5p	MIMAT0000093	Poliseno L. et al., 2010
hsa-miR-96-5p	MIMAT0000095	Mihelich B. L. et al., 2011
hsa-miR-98-5p	MIMAT0000096	Ting H. J. et al., 2013
hsa-miR-99a-5p	MIMAT0000097	Sun D. et al., 2011
hsa-miR-99b-5p	MIMAT0000689	Sun D. et al., 2011

Appendix B

Supplementary table for chapter 4

Table 4.A: Annotation table for *IGF1* from PrognoScan database

DATA POSTPROCESSING	None
PROBE_NAME	DAP3_5880 [6K DASL]
PROBE_DESCRIPTION	insulin-like growth factor 1 (somatomedin C)
GENE_SYMBOL	*IGF1*
GENE_DESCRIPTION	insulin-like growth factor 1 (somatomedin C)
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	0.40
MINIMUM *P*-VALUE	0.000219
CORRECTED *P*-VALUE	0.006963
$\ln(HR_{high} / HR_{low})$	-0.52
COX *P*-VALUE	0.003233
ln(HR)	-0.26
HR [95% CI]	0.77 [0.64 - 0.92]

Table 4.B: Annotation table for *KLK4* from Prognoscan database

DATA POSTPROCESSING	None
PROBE_NAME	DAP2_0881 [6K DASL]
PROBE_DESCRIPTION	kallikrein-related peptidase 4
GENE_SYMBOL	*KLK4*
GENE_DESCRIPTION	kallikrein-related peptidase 4
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	0.39
MINIMUM *P*-VALUE	0.007654
CORRECTED *P*-VALUE	0.131774
$\ln(HR_{high} / HR_{low})$	-0.38
COX *P*-VALUE	0.040851
ln(HR)	-0.28
HR [95% CI]	0.76 [0.58 - 0.99]

Table 4.C: Annotation table for *KLK3* from PrognoScan database

DATA POSTPROCESSING	None
PROBE_NAME	DAP1_1066 [6K DASL]
PROBE_DESCRIPTION	kallikrein-related peptidase 3
GENE_SYMBOL	*KLK3*
GENE_DESCRIPTION	kallikrein-related peptidase 3
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	0.62
MINIMUM P-VALUE	0.088109
CORRECTED P-VALUE	-
ln(HRhigh / HRlow)	-0.25
COX P-VALUE	0.001101
ln(HR)	-1.15
HR [95% CI]	0.32 [0.16 - 0.63]