Identification of biomarkers associated with cervical cancer: A combined *in silico* and molecular approach

Thesis presented in the fulfilment of the requirements for the

Magister Scientae

Department of Biotechnology, DST/Mintek Nanotechnology Innovation Centre,
University of the Western Cape, Cape Town, South Africa

Supervisor: Dr. Ashley Pretorius

Co-supervisors: Dr Mervin Meyer

DECLARATION OF AUTHORSHIP

Last name: Ludaka **First name: Namhla**

I declare that the work presented here is, to the best of my knowledge and belief, original and the result of my own investigations, except as acknowledged, and has not been submitted, either in part or whole, for a degree at this or any other University.

Formulations and ideas taken from other sources are cited as such and all work which was the result of joint effort, have been acknowledged by complete references. This work has not been published.

………………………… ....………………

Signature Date

ACKNOWLEDGEMENTS

I would like to thank my father in heaven, the omnipresent God, for being the source of my strength and the strength of my life. Thank you God for everything, for all the answered prayers and always reminding me that I am a more than a conqueror through Christ my Lord. A special thanks to my supervisors Dr. A. Pretorius and Dr. M. Meyer for the unconditional support, guidance and help during my master‘s degree. Specially, I feel the need to acknowledge them for having believed in my capabilities and for making me believe in myself. It has been a privilege working with them. Thank you very much. Thank you to NRF, NIC/MINTEK and MRC for the funding and making this project a reality. In addition, I would like to thank all of my group colleagues (BRG/NIC) for all the teaching, for having supported me whenever I was feeling down, for all the laughs that made me feel good and for all the experience shared during these months. Their friendship was, without any doubt, very important to overcome the difficulties faced. Big thanks to Mrs Potelwa for the love, support, encouragement and for unlocking my potential.

From the bottom of my heart a big thank you to my family, especially my mother and grandmother (Sange Totose and Noluthando Ludaka), from the youngest to the elderly for the prayers, love, support, sacrifices and patience not only during the execution of this master thesis, but also through my entire academic course. They were truly important in all of my victories and they are definitely ‘the only love that doesn‘t make me weak‘. Thank You to my church family (DCF) for the love, laughs and more importantly their unending prayers. I could not end these acknowledgments without a special thanks to my best friend and partner for having appeared in my life when I needed him the most. For all the love, support, care and patience during the good and the bad days and for all the advices and for always listening, I do thank him a lot.

ABSTRACT

Cervical cancer is the leading cause of cancer mortality among black women in South Africa. It is estimated that this disease kills approximately 8 women in South Africa every day. Cervical cancer is caused by the human papillomavirus (HPV) with the most common screening method for cervical cancer being Papanicolaou (Pap) smear, test amongst others. However, less than 20% of South African women go for these tests. There are several reasons why women do not go for these tests but the invasiveness of the test is one of the major causes for the low rate of screening. Lateral flow devices offer medical diagnosis at the point-of-care, allowing for the quick initiation of the appropriate therapeutic response. These tests are more cost-effective for the healthcare delivery industry, and can potentially be used by patients to self-test in the privacy of their homes and allow them to make informed decisions about their health. Therefore, the aim of this study was to use computational methods to identify serum biomarkers for cervical cancer that can be used to develop a point-of-care diagnostic device for cervical cancer.

An *in silico* approach was used to identify genes implicated in the initiation and development of cervical cancer. Several bioinformatics tools were employed to extract a list of genes from publicly available cancer repositories. Multiple gene enrichment analysis tools were employed to analyze the selected candidate genes. Through this pipeline, ~28190 genes were identified from the various databases and were further refined to only 10 genes. The 10 genes were identified as potential cervical cancer biomarkers. A subcellular compartmentalization analysis clustered the proteins encoded by these genes as cell surface, secretory granules and extracellular space/matrix proteins. The selected candidate genes were predicted to be specific for cervical cancer tissue in a cancer tissue specificity meta-analysis study. The expression levels of the candidate genes were compared relative to each other and a graph constructed using gene expression data generated by GeneHub-GEPIS and TiGER databases. Further gene enrichment analysis was performed such as protein-protein interactions, transcription factor analysis, pathway analysis and co-expression analysis, with 9 out of the 10 of the candidate genes showing co-expression.

A gene expression analysis done on cervical cancer cell lines, other cancer cell lines and normal fibroblast cell line revealed differential expression of the candidate genes. Three candidate genes were significantly expressed in cervical cancer, while the seven remaining genes showed over expression in other cancer types. The study serves as basis for future investigations to diagnosis of cervical cancer, as well as for cancers. Thus, they could also serve as potential drug targets for cancer therapeutics and diagnostics.

Key Words: cervical cancer, early diagnosis, biomarkers, bioinformatics, gene enrichment analysis

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

2D-Gel	Two-dimensional Polyacrylamide gel electrophoresis
ADAM 12	ADAM metallopeptidase domains 12
AEs	Adverse Effects
AFP	Alpha Fetoprotein
AIS	Adenocarcinoma *In Situ*
ASIR	Age Standardised Incidence Rate
ASR	Age Standardised Rate
ATP	Adenosine Triphosphate
B/MRP-14	Myeloid-related protein 14
BA	Beta-Actin
CA 19-9	Carbohydrate Antigen 19-9
CA125	Cancer Antigen 125
CANSA	Cancer Association of South Africa
CC	Cervical Cancer
CCDB	Cervical Cancer gene Database
CEA	Carcinoembryonic Antigen
C/EBPα	CCAAAT enhancer binding protein alpha
CGAP	Cancer Genome Anatomy Project
CIN	Cervical Intraepithelial Neoplasia
CRM	*Cis*-Regulatory Module
CT	Computed Tomography

DAVID	Database for Annotation, Visualization and Integrated Discovery
DEPC	Diethylpyrocarbonate
DEU	Digital Expression Unit
DM	Data Mining
DMEM	Minimal Essential Medium
DMSO	Dimethyl Sulphoxide
DNA	Deoxyribonucleic Acid
EBI	European Bioinformatics Institute
ECD	Extracellular Domain
ECM	Extracellular Matrix
EDTA	Ethylene Diamine Tetra-acetic acid
EMEM	Eagle's Minimal Essential Medium
EST	Expressed Sequence Tag
FBS	Fetal Bovine Serum
FDA	Food and Drug Administration
FIGO	International Federation of Gynaecology and Obstetrics
GAPDH	Glyceraldehyde-3-Phosphate Dehydrogenase
GEA	Gene Expression Atlas
GEB	Gene Expression Barcode
GEO	Gene Expression Omnibus
GDB	Genome Database
GMAP	Genomic Mapping and Alignment Program

GO	Gene Ontology
GOI	Genes of Interest
GPCR	G-protein Coupled Receptor
HER2	Human Epidermal Growth Factor Receptor 2
HIV	Human Immunodeficiency Virus
HLA	Human Leukaemia Antigen
HPA	Human Protein Atlas
HPRT1	Hypoxanthine Phosphoribosyltransferase1
HPV	Human Papilloma Virus
HR	High Risk
HSIL	High-grade Squamous Intraepithelial Lesion
HTMS	High Throughput Mass Spectrometry
IAP	Immunosuppressive Acidic Protein
IARC	International Agency for Research on Cancer
IE	Information Extraction
iHOP	Information Hyperlinked Over Proteins Database
intOGen	Integrative Oncogenomics
IR	Information Retrieval
KB	Knowledge Base
KDD	Knowledge Discovery in Databases
KEGG	Kyoto Encyclopedia of Genes and Genomes
KLK 2/3/5/6/14	Kallikrein-Related Peptidase 2/35/6/14

LEEP	Loop Electrosurgical Procedure
LLETZ	Loop Excision of the Transformation Zone
LOE	Level of Evidence
LR	Low Risk
LSIL	Low-grade Squamous Intraepithelial Lesion
MDLC	Multidimensional Liquid Chromatography
MeSH	Medical Subject Headings
MGI	Mouse Genome Informatics
MMPs	Matrix Metalloproteinases
MPSS	Massively Parallel Signature Sequencing
MRI	Magnetic Resonance Imaging
mRNA	Messenger Ribonucleic Acid
NCI	National Cancer Institute
NER	Named Entity Recognition
NGS	Next‑Generation Sequencing
NIH	National Institutes of Health
OS	Ostiums
OMIM	Online Mendelian Inheritance in Man
Pap smear	Papanicolaou smear
PBS	Phosphate Buffered Saline
PCA	Principal Component Analysis
PCNA	Proliferating Cell Nuclear Antigen

PET	Positron Emission Tomography
POC	Point-Of-Care
PPARγ	Perixosome Proliferator-Activated Gamma
PSA	Prostate-Specific Antigen
RB	Retinoblastoma
RB1	Retinoblastoma 1
RBL1	Retinoblastoma-like 1
RBL2	Retinoblastoma-like 2
rDNAse	Recombinant Deoxyribonuclease
REST	Relative Expression Software Tool
RPMI	Roswell Park Memorial Institute Medium
SAGE	Serial Analysis of Gene Expression
SCC	Squamous Cell Carcinoma
SCJ	Squamous Columnar Junction
SDS	Sodium Dodecyl Sulphate
SIL	Squamous Intraepithelial Lesions
SNP	Single Nucleotide Polymorphism
SOM	Self-Organising Maps
STIs	Sexually Transmitted Infections
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
TCEP	Tris (2-carboxyethyl) phosphine
TF	Transcription Factor

TiGER	Tissue-specific Gene Expression and Regulation
TM	Text Mining
Tris	Tris [Hydroxymethyl] aminoethane
TSGs	Tissue-Specific Genes
TZ	Transformation Zone
UBC	Ubiquitin C
UniProt	Universal Protein Knowledgebase
VEGF	Vascular Endothelial Growth Factor
VIA	Visual Inspection with Acetic Acid
VIAM	Visual Inspection with Magnification
VILI	Visual Inspection with Lugol's Iodine
VLP	Virus-Like Particle
WHO	World Health Organisation

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

LIST OF TABLES

Chapter 1

Chapter 2

Chapter 4

CHAPTER 1: Literature Review

1.1. Cancer Overview

Cancer is a very complex disease that develops due to the accumulation of genetic and epigenetic changes that allow for deviation from normal cellular and environmental control (Kumar *et al*., 2009). Cancer is a common term used for a group of more than 100 diseases that affect any part of the body and is also known as neoplasms or tumours. The most prominent characteristics of cancer are the uncontrolled proliferation of cells beyond their normal boundaries, the ability to invade and invade adjoining body tissue and that they can metastasize to other tissue. Cancer morbidity or mortality is a result of metastases (Kumar *et al*., 2009). Cancer is a disease that affects people of all races and age groups and is responsible for about 13% of all deaths. According to the World Health Organisation (WHO) there were around 12 million new cases of cancer in 2008, and 7 million deaths resulting from cancer in the same year. Among the known cancer types, lung cancer is the most common cause of death in both genders, followed by breast cancer in females and prostate cancer in males (Iyoke and Ugwu, 2013). According to estimates by the International Agency for Research on Cancer (IARC), in 2008, 53% of the 12.7 million new cases of cancer and 63% of the 7.6 million cancer deaths occurred in developing countries (Iyoke and Ugwu, 2013).

A portion of the cancer types are specific to a particular sex, such as uterine sarcoma, cervical and ovarian cancer in females and prostate cancer in males. There are more than a 100 distinct types of human cancers that have been described and different tumour subtypes are found within specific organs (Grizzi and Chiriva-Internati, 2006). Cancerous cells are neoplastic and they proliferate in defiance to normal cell control, thus preventing normal cellular function, tissue, and organ formation with the potential of becoming malignant and colonising surrounding tissues (Abbott *et al.*, 2006). Cancers are classified by the type of cell and tissue they originate from. Approximately 90% of cancers are characterised as carcinomas, which develop from epithelial cells with the most common sites being the breast, lung, prostate and colon (Petersen *et al.*, 2003). The high prevalence of carcinomas is due to the fact that epithelial cells are the site of active cell proliferation and are constantly exposed to both physical and chemical carcinogens. Sarcomas are tumours that originate from connective tissue or mesenchymal cells, whereas cancers that arise from hemopoietic cells are lymphomas and leukaemia. All the different types of cancers can be further subdivided based on specific cell type, location in the body and the structure of the tumour (Petersen *et al.*, 2003). The majority of cancers occur as a result of initial somatic hypermutation of a single cell that must undergo a number of mutations in order to become cancerous, a process that is described in Hanahan and Weinberg's hypothesis of 6 cumulative alterations (Hanahan and Weinberg, 2000). Surgery is used to remove benign tumours and they consist of neoplastic cells which remain clustered together in a single mass. When aggressive cancerous cells divide, they penetrate the basement membrane. The result of this local invasion is neoplastic cell adherence to adjacent tissue and through, production of proteolytic enzymes and factors such as matrix metalloproteinases (MMPs); degradation of the extracellular matrix (ECM) is caused in surrounding tissue. As a result cancer cells are able to cross tissues into neighbouring stroma, thus facilitating deeper invasion and metastasis. This process is responsible for tumours becoming malignant as the cells invade the surrounding tissue and cross the basal lamina to enter the blood stream (Laerum, 1997). Tumours that are metastatic are able to survive in circulation and have the ability to travel through blood vessels to other organs and through the lymphatic vessels to lymph nodes. As a consequence, the development of secondary tumours at a site far away from the primary tumour is a hallmark of cancers that are more aggressive and rendering them difficult to eradicate (Liotta and Kohn, 2001).

1.1.1. Pathogenesis of Cancer

There are 6 hallmarks of cancer that define malignancy as hypothesised by Hanahan and Weinberg: (i) sustaining proliferative signalling, (ii) evading growth suppressors, (iii) activating invasion and metastasis, (iv) enabling replicative immortality, (v) inducing angiogenesis and (vi) resisting cell death as shown in figure 1.1 (Hanahan and Weinberg 2011). It is believed that the majority of cancer cell genotypes are a manifestation of these six essential alterations in cell physiology that collectively dictate malignant growth.

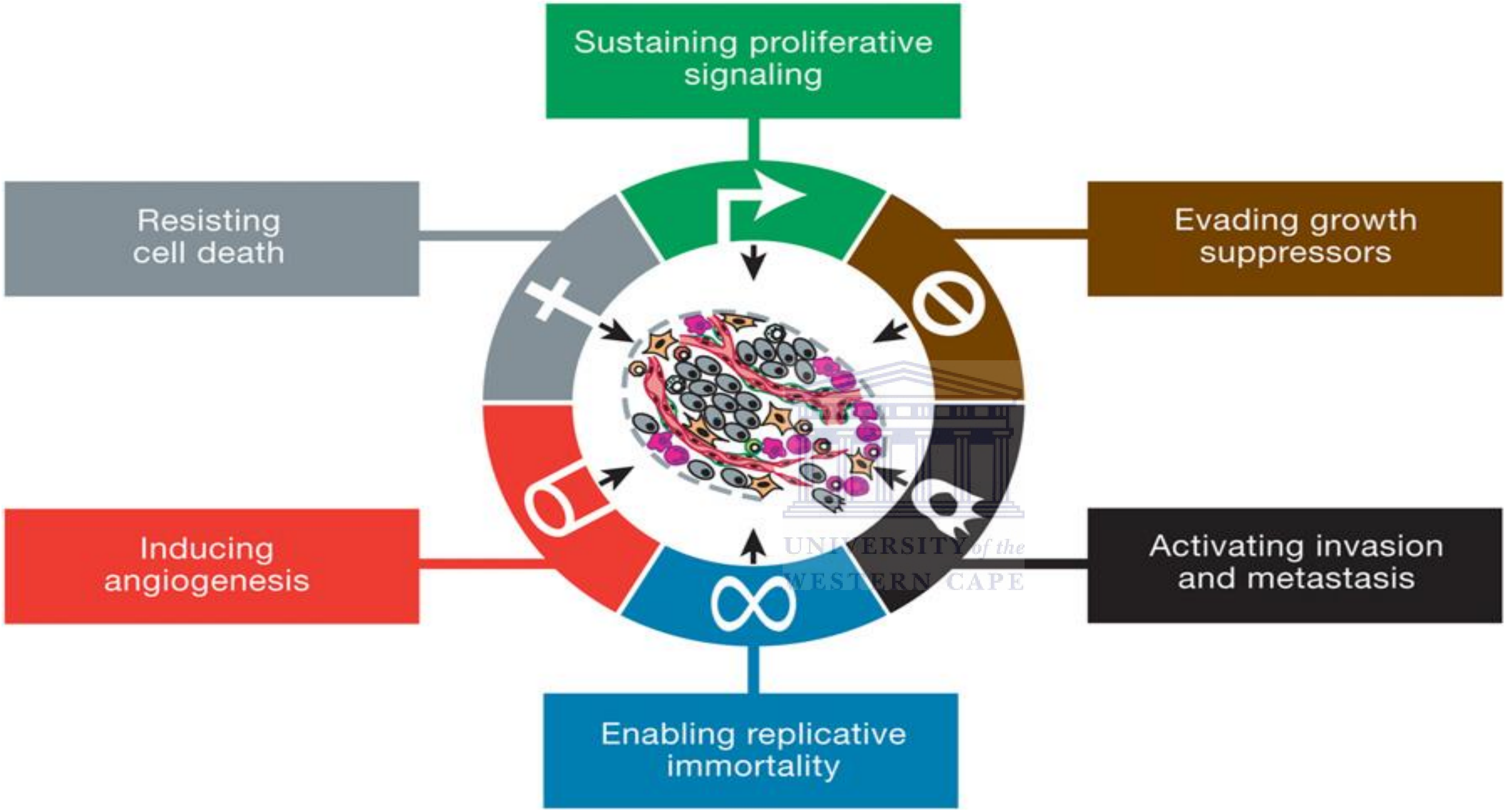

Sustaining proliferative signaling
Resisting cell death
Evading growth suppressors
Inducing angiogenesis
Activating invasion and metastasis
Enabling replicative immortality

Figure 1.1: The six hallmark capabilities of cancer, adapted from Hanahan & Weinberg 2011.

1.1.1.1. Sustaining Proliferative Signalling

The ability of cancer cells to sustain chronic proliferation is the most fundamental trait of cancer cells. Normal tissues carefully control the production and release of growth-promoting signals that instruct entry into and progression through the cell growth-and-division cycle, thereby ensuring a homeostasis of cell number and thus maintenance of normal tissue architecture and function. Cancer cells, by deregulating these signals, become masters of their own destinies (Hanahan and Weinberg 2011). The tumor cells generate many of their own growth signals by altering extracellular growth signals, transcellular transducers of those signals, or intracellular circuits that translate these signals into action (Mees *et al*., 2009). The enabling signals are conveyed in large part by growth factors that bind cell-surface receptors, typically containing intracellular tyrosine kinase domains. The latter proceed to emit signals via branched intracellular signalling pathways that regulate progression through the cell cycle as well as cell growth (that is, increases in cell size); often these signals influence yet other cell-biological properties, such as cell survival and energy metabolism. Tumour cells are therefore less dependent on exogenous growth stimulation (Hanahan and Weinberg 2011).

1.1.1.2. Evading Growth Suppressors

Cancer cells have the ability to evade antiproliferative signals, such as soluble growth inhibitors and immobilized inhibitors, in order to progress (Hanahan and Weinberg 2000). These signals block proliferation by either forcing the cells out of the active proliferative cycle into the G0 phase of the cell cycle, or by inducing the cells to permanently give up their proliferative potential by making them enter into post mitotic states. At the molecular level, most antiproliferative signals are funnelled through retinoblastoma 1 (RB1) and its two relatives, retinoblastoma-like 1 (RBL1) and retinoblastoma-like 2 (RBL2) proteins. Disruption of the RB1 pathway allows proliferation, causing cells insensitive to antigrowth factors that normally operate along this pathway to block progression (Weinberg, 1995).

Cancer cells also circumvent powerful programs that negatively regulate cell proliferation and many of these programs depend on the actions of tumour suppressor genes. The two prototypical tumour suppressors encode the RB (retinoblastoma) and tumor protein p53

(TP53 proteins); which operates as central control nodes within two key complementary cellular regulatory circuits that govern the decisions of cells to proliferate or alternatively, activate senescence and apoptotic programs (Hanahan and Weinberg 2011).

1.1.1.3. Resisting Cell Death

In the face of aberrant and potentially cancerous growth signalling, normal cells activate programmed cell death (Apoptosis). Apoptosis is also activated in response to DNA damage, although other cellular stresses can also be features of cancerous cells. Apoptosis thus represents a crucial mechanism to avoid accumulation of damaged cells and mutations that can culminate in cancer formation (Fulda, 2009). Programmed cell death, is triggered by a variety of physiological signals, and will unfold in a precisely choreographed series of steps, including disruption of cellular membranes, breakdown of cytoplasmic nuclear skeletons, degradation of chromosomes, and fragmentation of the nucleus (Fulda, 2009). Cancer cells acquire the ability to evade this induction of cell death and this ability is crucial for maintaining tumour growth and allowing cancerous cells to form in the first stage of disease development. Part of the ability of tumour cells to evade programmed cell death is derived from survival signals supplied by the stromal compartment (Hanahan and Weinberg 2011). Failure of apoptosis is essential to all types of cancer and the apoptotic program is present in a latent form in almost all cell types in the body (Hanahan and Weinberg 2011).

Cells use a variety of ways to avoid cell death; some of these strategies are less clearly understood than others. One of the most commonly mutated tumour suppressor genes is the p53 gene. 50% of the tumour apoptosis evasive characteristics are a result of p53 protein inactivation. More than half of all types of human cancers have a mutated or missing gene for p53, resulting in a damaged or missing P53 protein. As an alternative to achieving the loss of P53, cancer cells can compromise the activity of P53 by increasing the inhibitors of P53, or silencing the activators of P53. The P13 Kinase-AKT/ PKB pathway is another pathway used by tumour cells to evade programmed cell death. This pathway is concerned with the anti-apoptotic survival signals. This survival signalling circuit is found to be up-regulated either by extracellular factors such as IGF-1and IL-3, or by intracellular signals involving RAS, thus leading to evasion of apoptosis (Wong, 2011).

1.1.1.4. Inducing Angiogenesis

Induced angiogenesis is the process by which cancer cells induce and sustain the growth of new blood vessels. Oxygen and nutrients are crucial for cell function and survival, they are supplied by the vasculature. Cells within aberrant proliferative lesions initially lack this angiogenic ability, reducing their capability to expand. Cancer cells on the other hand overcome this restriction by inducing and sustaining angiogenesis (Hanahan and Folkman 1996). Tumour growth depends on angiogenesis to first provide oxygen and nutrients to proliferating cells and to then provide a physical route for metastasis transport. Tumours, like normal tissues also require sustenance as well as the ability to evacuate metabolic wastes and carbon dioxide. The tumour-associated neovasculature generated by the process of angiogenesis addresses these needs (Hanahan and Weinberg 2011).

Tumour cells make use of the vascular network and penetrate into the blood using it as a transport channel, circulating through the intravascular channel and then proliferate to distant sites. Since nutrients and oxygen are a necessity for cancer cells, angiogenesis becomes an important factor for the progression of cancer. Angiogenesis therefore becomes a multistep process of which angiogenic factors (VEGF, bFGF TGF-alpha etc.) play a huge role. It involves degradation of the basement membrane in tissues; endothelial cells which are activated by angiogenic factors (Nishida *et al*., 2006). Upon further migration, they proliferate and stabilise and angiogenic factors will further continue the angiogenic process forming new blood vessels. Angiogenesis therefore creates a vascular support for cancer cells and absence thereof will result in tumours becoming necrotic or even becoming apoptotic. Angiogenic factors have thus received remarkable attention, with the VEGF family being one of the major angiogenic agents in neoplastic tissues (Nishida *et al*, 2006).

1.1.1.5. Activating Invasion and Metastasis

Invasion of normal host tissue by the tumour and metastasis ultimately leads to death of the cancer patient. Cancer cells acquire the ability to become motile and migrate from the original tumour site and this is the acquisition of the invasive and metastatic phenotype. Changes promoting invasion happens at the cellular level, including changes in the expression of surface markers which allow the cells to adhere to the surrounding tissues.

Metastasis is a particularly complex process, but usually occurs by cancer cells invading blood vessels and hitchhiking through the circulatory system to other sites of the body (Hanahan and Weinberg 2011).

1.1.1.6. Enabling Replicative Immortality

The ability of cancerous cells to avoid controlled growth often involves finding ways of repairing or lengthening the telomeres to prevent them from shortening, allowing indefinite replication (Hanahan and Weinberg 2011). About 10% of cancers develop as result of inherited genetic susceptibility; therefore a 90% risk of cancer development can be attributed to a combination of diet, environment, cultural and lifestyle factors (Bell, 2005). Mitogenic growth factors are required by normal cells to stimulate proliferation; however cancerous cells have developed autonomy from this mode of down-regulation. Cells that are cancerous have the ability to gain self-sufficiency for growth signalling by overexpression or constitutive activation of growth factor receptors, thus synthesising their own growth factors and deregulating downstream signalling targets (Peters *et al*., 2001). Cancerous cells evade negative anti-growth signals which act by forcing cells out of the cell cycle into a resting, quiescent state (G0), or by inducing cells to undergo differentiation into a post-mitotic state. This can also be achieved by abrogation of growth-inhibitory receptor function (Zeimet *et al*., 2000). Cancerous cells have the ability to evade programmed apoptotic cell death by abrogation of pro-apoptotic sensor receptor signalling that monitors the extracellular and intracellular environment. Cells in every tissue survive by being in close proximity to a blood vessel to supply oxygen and nutrients. Cancer cells have developed mechanisms to encourage growth of new blood vessels from pre-existing vessels (angiogenesis) by changing the angiogenesis inducers such as vascular endothelial growth factor (VEGF) and inhibitors such as thrombospondin-1 (Hanahan and Folkman 1996). As previously mentioned cancer can originate in any part of the body and cancers that develop in the female reproductive system are known as women's reproductive cancers. These encompass cancer of the breast, cervix, ovaries, endometrium, vagina and vulva. The most persistently occurring cancers of the reproductive system in women worldwide are breast, cervical and ovarian cancer (Getahun, *et al*., 2013).

1.2. Cervical Cancer

1.2.1. Global Cervical Cancer Burden

Cervical cancer (CC) is the third most prevalent cancer amongst women globally, with approximated 83,195 new cases and 35,673 mortality cases in 2012 (Human Papillomavirus and Related Diseases Report, South Africa, 2014). More than 80% of the global burden occurs in developing countries, where cervical cancer accounts for 13% of all female cancers. However, cervical cancer is not limited to the developing world as it is considered in Europe as a significant public health problem with an incidence rate of 10.6 per 100 000. As a result of better developed prevention programmes, Western Europe has a lower incidence and mortality rate when compared to Central and Eastern Europe (Mareea and Moitse, 2014). In western countries, the incidence and mortality of CC have declined substantially over the past decades, whereas in developing countries there is a slight increase in mortality (Figure 1.2). This is probably due to the lack of screening and the greater impact of infectious cofactors in the latter regions. Age-adjusted incidence rates vary from about 10 per 100 000 per year in many industrialized countries to more than 40 per 100 000 in some developing countries. More than 88% of deaths occur in low-income countries and it is predicted to increase to 91.5% by 2030 (Maree and Wright, 2010).

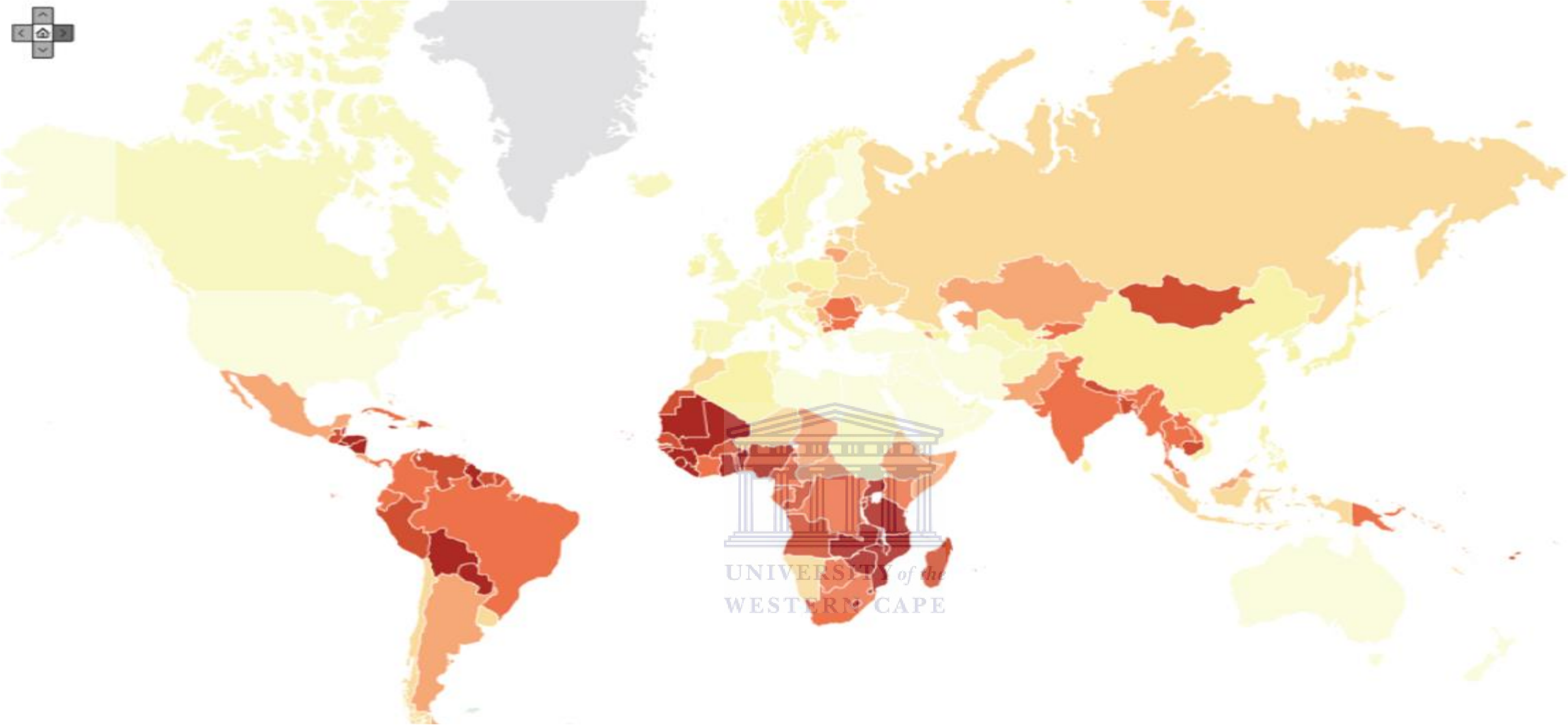


Figure 1.2: Estimated cervical cancer incidence worldwide in 2008. GLOBOCAN 2008, International Agency for Research on Cancer (IARC). The red and dark highlighted areas have the highest incidence rates.

Cervical cancer is the second most frequently diagnosed cancer in women after breast cancer and the leading cause of cancer death in African women (Jemal *et al*., 2012). Eastern and Western Africa are the highest risk regions with Age Standardised Rate (ASR) greater than 30 per 100 000, Southern Africa with ASR 26.08 per 100 000, South-Central Asia (ASR 24.6 per 100 000), and South America and Middle Africa with 23.9 and 23.0 ASRs respectively. Low-risk regions are Western Asia, Northern America and Australia/New Zealand with ASRs less than 6 per 100 000. Cancer of the cervix is most prevalent in women in Eastern Africa, Melanesia and South-Central (Globocan, 2008).Cervical Cancer was responsible for approximately 275 000 deaths in 2008 and 88% of these cases occurred in developing countries. Africa reported 53 000 death cases, 31 700 in the Caribbean and Latin America and Asia with 159 800 deaths (Figure 1.2) (Globocan, 2008).

According to Globocan 2012, cervical cancer is the fourth most common cancer in women, and the seventh overall, with an estimated 528 000 new cases in 2012. There were an estimated 266 000 deaths from cervical cancer worldwide in 2012, accounting for 7.5% of all female cancer deaths. Almost nine out of ten (87%) cervical cancer deaths occur in the less developed regions. According to the IARC, there were 453 531 cases of cervical cancer in developing countries in 2008 representing 89% of global estimates. Also 273 000 deaths occur worldwide every year due to cervical cancer out of which 83% occur in developing countries ((Iyoke and Ugwu, 2013). It is estimated that 80-90% of cervical cancer cases in developing countries occur amongst women of age 35 and older. Cervical cancer develops very slowly from precancerous lesions to advance cancer. On a global scale the incidence of cancer is low in women under the age of 25 years, however, the incidence increases at age 35 to 40 years and reaches the maximum in women of ages 50 and 60 years (Alliance for Cervical Cancer Prevention, 2005). It is estimated that about 83% of all new cases of cervical cancer and 85% of all deaths related to cervical cancer are occurring in developing countries as shown in Figure 1.2 (Anorlu, 2008).

1.2.2. Cervical Cancer Burden in Africa and South Africa

In Africa, there is an absence of accurate information relating to the extent of cancer and this is mostly due to cancer registries being limited. However, this does not reflect the magnitude of the cancer problem on the African continent. The African continent has a population of 267.9 million women from age 15 years and older who are at risk of developing cervical cancer, with an estimation of 80 000 women diagnosed with cervical cancer annually and over 60 000 women perishing from this disease (Denny, 2012). The incidence of cervical cancer in Africa varies substantially by region, with the highest rates in Africa (ASIR>40 per 100 000) all found in Eastern, Southern and Western Africa as shown in figure 1.3 (Denny, 2012). There are many compelling factors that contribute to the high prevalence of cervical cancer in Africa. These include limited human and financial resources, competing health needs, poorly developed healthcare systems, war and civil strife, women being uninformed and disempowered and the nature of cytological-based screening programmes (Mareea and Moitse, 2014).

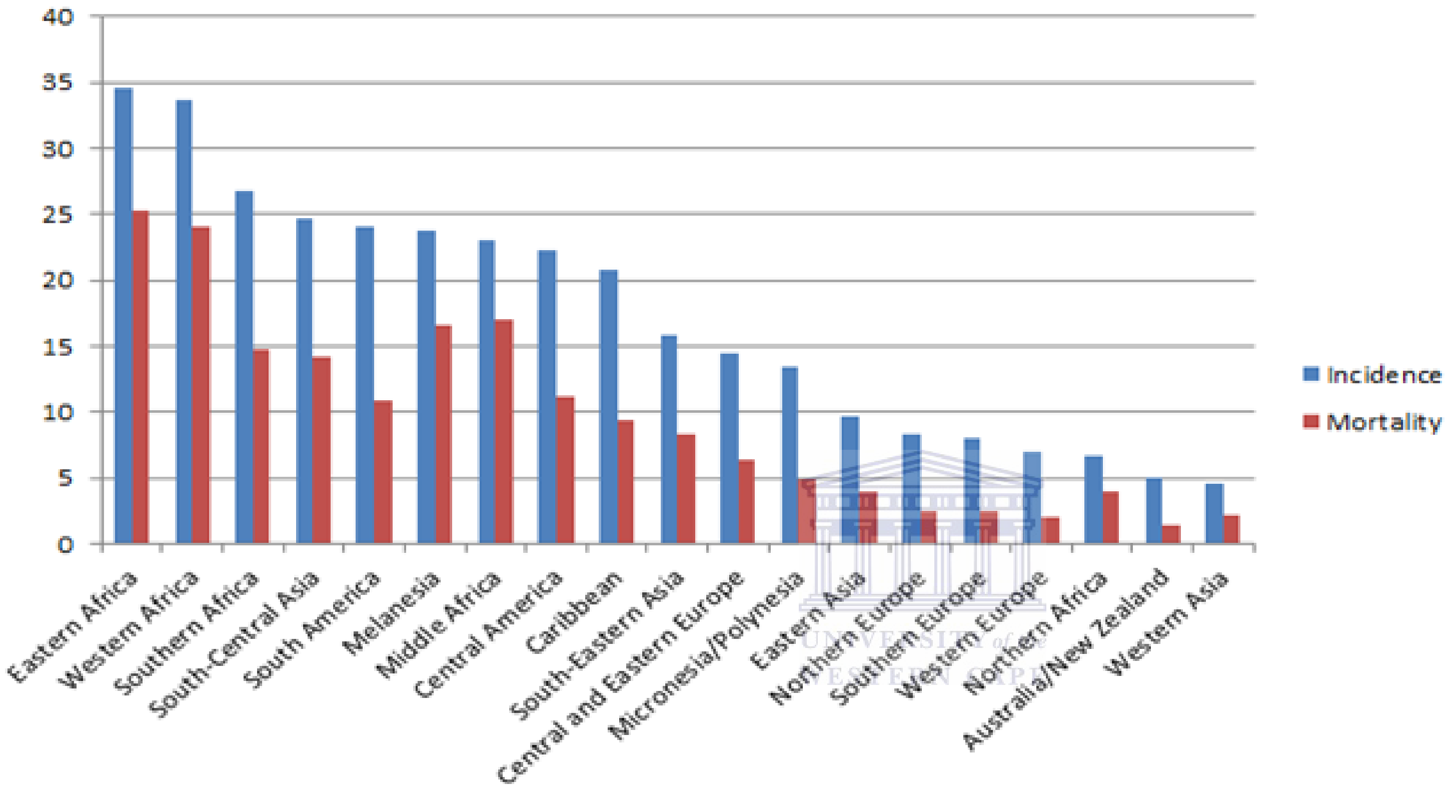


Figure 1.3: Incidents and mortality of cervical cancer in 2008. Source Jemal *et al*., 2011,

http://onlinelibrary.wiley.com/doi/10.3322/caac.20107/pdf

Other factors include lack of access to preventative measures, late diagnosis, treatment and palliation for cancer related disease. The African continent has poor accessibility to cancer therapies (Denny, 2012). In South Africa, cervical cancer is the second most diagnosed cancer in women with, 7735 new cases being diagnosed yearly and is the most common cancer in women aged 15-44 years (Table 1.1 and Figure 1.4) (Human Papillomavirus and Related Diseases Report, 2014). Cervical cancer has the highest incidence rate in South Africa, followed by breast cancer, lung cancer, cancer of the trachea and bronchi (Maree and Wright, 2010). In South African women, cervical cancer is the primary cause of death with approximately 4248 new cancer deaths occurring each year with the highest death rate among black women between the ages of 66-69 years (Figure 1.4). South African women, particularly black women, most often present late with cancer that is already in an advanced stage (Francis *et al*., 2011).

Table 1.1: Incidence of cervical cancer in South Africa (estimations for 2012)

Indicator	South Africa	Southern Africa	World
Annual number of new cancer cases	7,735	8,652	527,624
Crude incidence rate[α]	30.2	29.3	15.1
Age-standardized incidence rate[α]	31.7	31.5	14.0
Cumulative risk (%) at 75 years old	3.1	3.1	1.4

*Adapted from Human Papillomavirus and Related Diseases Report, 2014.

The total age-adjusted incidence rate (ASIR) of cancer in the black population is far lower than the corresponding white population. Black women are most at risk of developing cervical cancer when compared to their white and coloured counterparts. According to the Department of Health of KwaZulu-Natal, cancer of the cervix accounts for 18.5% of gynaecological cancers, with approximately 5000 new cases being reported annually (Walker *et al*., 2002).

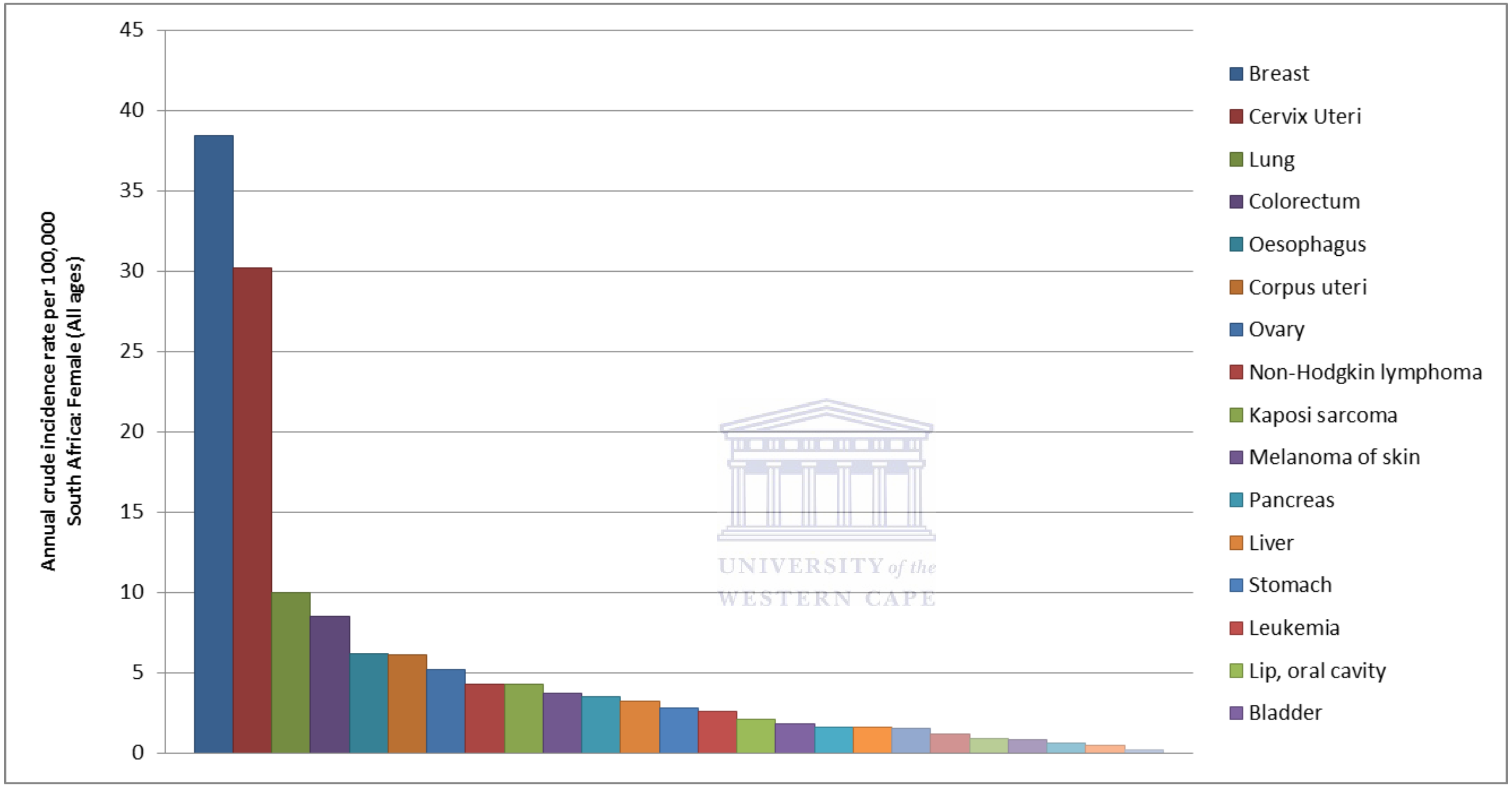


Figure 1.4: Cervical cancer mortality compared to other cancers in women of all ages in South Africa, adapted Human Papillomavirus and Related Diseases Report, 2014.

1.3. Physiology and Anatomy of the Cervix

The cervix is the lower portion of uterus that connects to the upper portion of the uterus. It is cylindrical in shape and connects the vagina and the uterus as depicted in figure 1.5 (Junqueira and Carneiro, 2005). There are two narrow openings present in the cervix namely the internal and external ostiums (os). The location of the internal os is the topmost portion of the cervix and opens into the uterus and the external os is located at the minor portion of the cervix and opens into the vagina (Stevens and Lowe, 2005). The endocervical canal or canal of the cervix, which can change in width and length, is the passageway between the external ostiums and the uterine cavity. The cervix is divided into two parts, the endocervix which is the portion proximal to the uterus and the ectocervix proximal to the vagina. The endocervix has a fusiform shape and is composed of a single layer of mucous-secreting columnar epithelium and the ectocervix has a convex, elliptical surface and is composed of nonkeratinized stratified squamous epithelium (Arends *et al*., 1998). The transformation zone (TZ) or squamo-columnar junction is the portion adjacent to the edge of the endocervix and ectocervix, where the columnar epithelium is converted to squamous epithelium by a process known as metaplasia. The transformation zone is the area where the majority of abnormal changes occur and is susceptible to carcinogens and diseases (Ross and Pawlina, 2006).

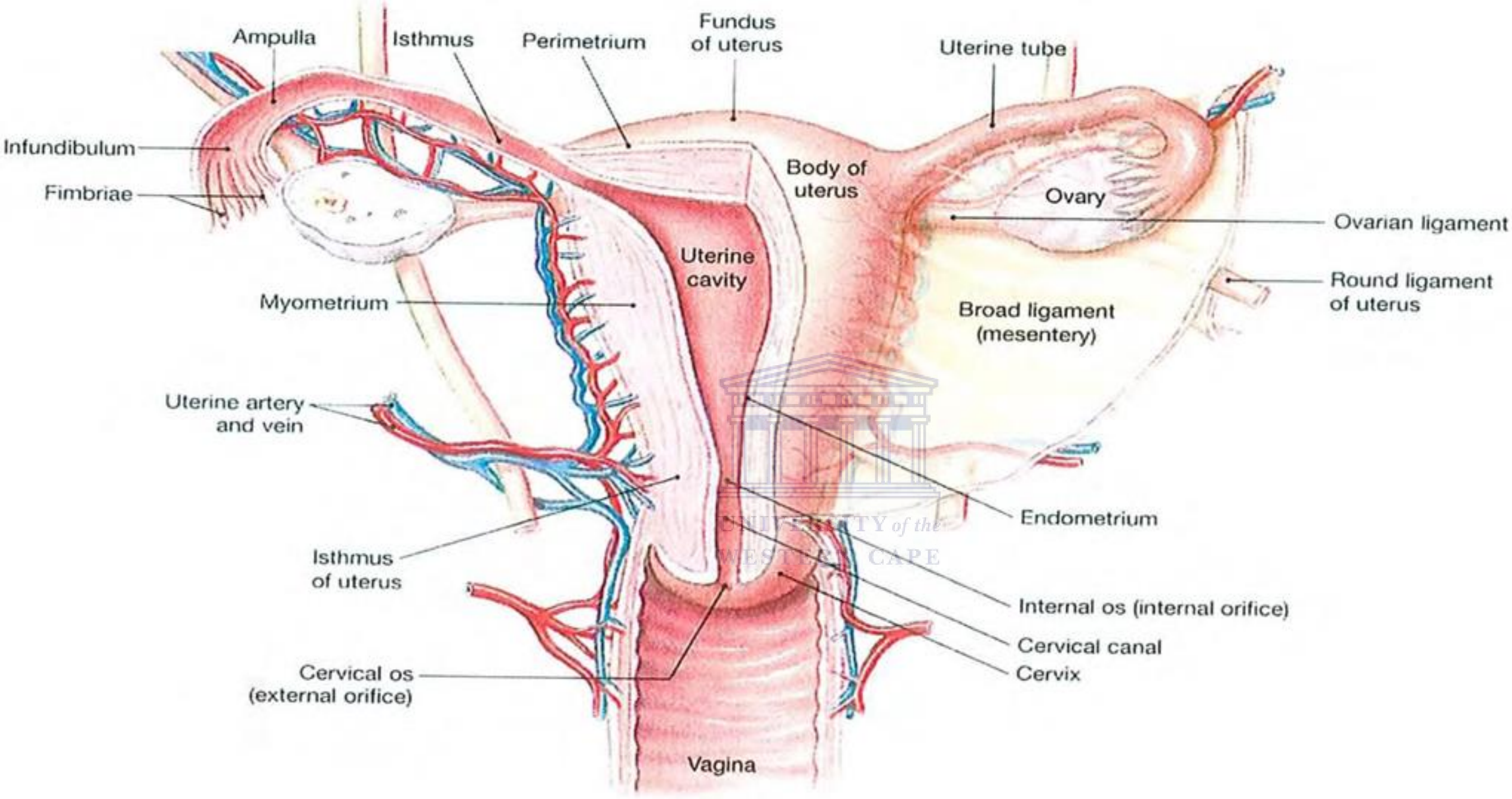


Figure 1.5: Diagram of the uterus indicating the cervix, internal and external ostiums and the cervical canal (Martini and Bartholomew, 2007).

1.4. Risk Factors for Cervical Cancer

The majority of women will be exposed to the human papilloma virus (HPV) virus at some stage in their life, however, only a fraction develop persistence of infection and subsequent life-threatening cervical disease, thus implicating other factors in cervical cancer pathogenesis. Risk factors associated with squamous cervical cancer, besides infection with HPV, include early age of coitus debut, numerous sexual partners, hormonal contraceptives, high parity, smoking and other sexually transmitted infections (STIs) (Dahlstrom *et al*., 2011). The daunting challenge is to determine to what extent these risk factors possess inherent capabilities to induce cervical cancer, or if they advance proxy measures for the risk for present and/or past high-risk (HR) HPV infection. Infection with HPV is the main factor responsible for initiation of cervical cancer through sexual intercourse. Out of the 100 different HPV types, about 40 of these affect the genital areas. The other types infect the skin on other areas of the body such as hands or feet. HPV 6 and 11 are responsible for causing warts which develop in a period of six to eight weeks (Likes and Itano, 2003).

The HPV is very difficult to identify because it’s asymptomatic, hence there is a need for routine cervical check-ups and HPV testing (Godfrey, 2007). There are about 13 high risk HPV strains that cause high-grade cervical cell abnormalities, with the high risk strains detected in 90% of cervical cancer incidences, with 70% of these caused by HPV 16 and 18. Infection by one HPV strain does not necessarily guarantee that a person is not susceptible to infection by a second or more strains. Individuals infected with mucosal HPV, about 5% to 30% get infected with more than one viral type simultaneously (Pink book, 2011). A woman that has multiple sexual partners is at an increased risk of contracting HPV which is dominant in men (Likes and Itano, 2003). Also at risk of HPV infection are women with three or more full term pregnancies and this is due to the different hormonal changes associated with pregnancy. Girls aged 17 years or younger are also at an increased risk when they have their first full term pregnancy and are more at risk of developing cervical cancer later on in life when compared to women who get pregnant at age 25 years and above (American Cancer Society, 2010).

The risk associated with hormonal contraceptives could be explained by their hormonal influence on the cervical mucosa, thus rendering it more susceptible to persistent/progressing infection. According to Appleby *et al.*, 2007, this can also be attributed to a higher risk for HR-HPV due to a concurrent tendency to (1) the infrequent use of condoms since already on contraception and (2) more likely to indulge in sexual activity than those not using contraceptives. Therefore, it is important to adjust for information on these factors as far as possible, and with large scale analyses, an association to contraceptive use still needs to be demonstrated (Appleby *et al.*, 2007). High parity has been consistently found to increase the risk of squamous cell cervical carcinoma among HPV positive women. It leads to direct exposure of the transformation zone in the cervix to HPV infection, thus increasing the risk of cervical cancer (Munoz *et al*., 2002).

1.5. Human Papilloma Virus and Cervical Cancer

Human papilloma virus is the most common sexually transmitted virus. HPVs are double stranded DNA viruses, which are very small, with their genomes encoded by approximately 8000 base pairs. There are nearly 100 types of HPV, with different variations in their genetic and oncogenic potential. Cervical cancer is caused by HPV strains that belong to a few phylogenetically related “High-risk” (HR) species (alpha-5, 6, 7, 9, 11) of the mucosotropic alpha genus. The types found most frequently associated with CC (-16, -18, -31, -33, -35, -45, -52, -58) and four less-common types (-39, -51, -56, -59) are classified in Group 1. The remaining types of HPV in the HR alpha species are classified as “possibly carcinogenic” (Group 2. 2A: -68; 2B: -26, -30, -34, -53, -66,-67, -69, -70, -73, -82, -85, -97). Finally, HPV -6 and -11, which belong to the alpha-10 species, were not classifiable as to their carcinogenicity in humans (Group 3) and were also described as low risk (LR) strains (Abreu *et al.*, 2012). Worldwide, the most common HR-HPV strains are -16 and -18, and approximately 70% of CC is due to these genotypes. LR-HPV strains, principally -6 and -11, are predominantly involved in the development of genital warts (Abreu *et al.*, 2012). HPV 16 accounts for about half of the cervical cancer cases in the United States and Europe (Gómez and Santos, 2007). The HPV 18 strain has been also associated with adenocarcinoma of the cervix, but the connection is less pronounced and is age dependent (Gómez and Santos, 2007).

1.6. Development of Cervical Cancer

More than 80% of the population is infected with HPV at some point in their life. In rare cases (1%), this infection will eventually lead to cervical cancer and simultaneous infections with multiple HPV types are common (Southern Africa Litigation Centre, 2012). The majority of HPV infections, irrespective of the type, are asymptomatic and resolve over a short period of time without treatment, as the woman's immune system will usually suppress or eliminate the HPV infection. The HPV infection persists in a small percentage of women (Southern Africa Litigation Centre, 2012). Cervical cancer develops over time when persistent HPV infection triggers alterations in the cells of the cervix, called precursor lesions or cervical intraepithelial neoplasia (CIN), or recently referred to as squamous intraepithelial lesions (SIL) (Southern Africa Litigation Centre, 2012). When these pre-cancerous lesions are left untreated, they can eventually lead to cancer. The lesions can progress from low grade (CIN 1) to high grade (CIN 2 and CIN 3) as their size, shape and number increases. Following its natural course, progression of the disease is slow and can take as long as 10 to 20 years from the initial infection with HPV to invasive cancer (Southern Africa Litigation Centre, 2012).

There are four steps involved in the development of cervical cancer: (i) HPV transmission, (ii) viral persistence, (iii) progression of a clone of persistently infected cells to precancer and (iv) invasion. Backward steps can also occur such as clearance of HPV infection and regression of precancer to normality as indicated in figure 1.6 (Schiffman *et al*., 2007). There is a high chance for progression to precancerous lesions and ultimately invasive lesions, when HPV acquisition is followed by HPV persistence instead of clearance. HPV 16 and 18 account for about 70% of all Squamous Cell Carcinoma (SCC) and for up to 85% of all adenocarcinomas. HPV has the ability to incorporate into human DNA, with HPV 16, 18 and 45 being predominant HPV types in cervical cancer as they are more likely to integrate into the human genome than other HPV types (Hoste *et al.*, 2013). When these three types of HPV cause cervical cancer, the patients are diagnosed on average 4 to 5 years earlier than those caused by other high-risk types (Hoste *et al*., 2013). HPV 16 and 18 positive Low-grade Squamous Intraepithelial Lesion (LSIL) is more likely to progress to CC than LSIL containing other HPV genotypes. HPV 16 and 18 account for 35% of LSIL but nearly 70% of CC worldwide. HPV 16 is more persistent and more likely to progress to CIN3+ (CIN3, carcinoma *in situ* and invasive CC) than other high risk HPV types (Hoste *et al*., 2013). The onco-proteins E6 and E7 deactivate essential processes associated with tumour genes like p53 and pRb functioning (Botha and Dochez, 2012). This ubiquitous infection does not lead to disease progression in all infected individuals and certain processes make individuals more susceptible to the development of pre-malignant and malignant disease. Since HPV is almost exclusively an epithelial disease, the virus is poorly presented to the adaptive immune system, which is important for induction of long term immunity. Most natural infections of HPV do not cause significant immunoglobulin responses and therefore the immune response after natural infections is not very pronounced (Botha and Dochez, 2012). After the initial exposure to HPV there is an incubation period of between 1 and 8 months after which the first HPV-related lesions might appear. There is active growth of the virus for a period of between 3 and 6 months but usually there are host-immune responses that will in most cases clear the infection by about 9 months (Botha and Dochez, 2012). A large percentage of the population will have sustained clinical remission but a small proportion will develop chronic infection and become HPV-DNA positive on repeated testing. These are the individuals that will be at highest risk for the development of pre-malignant conditions and later invasive cancer (Botha and Dochez, 2012).

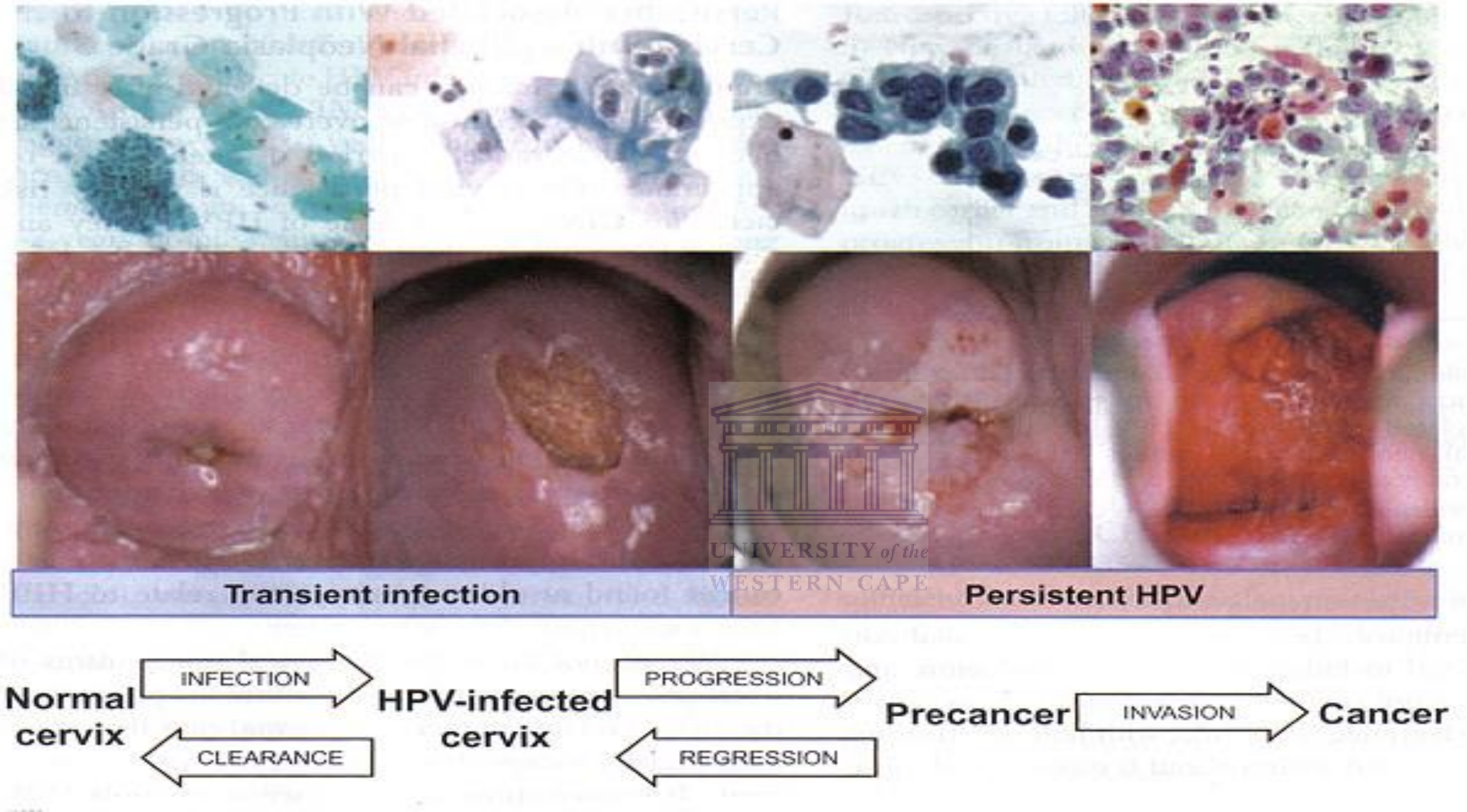


Figure 1.6: Major steps in cervical cancer carcinogenesis. HPV infection followed by clearance by the immune system or progression to precancer and invasion. The top row shows cytology and the bottom row shows colposcopy (Schiffman *et al*., 2007).

1.7. Types of Cervical Cancer

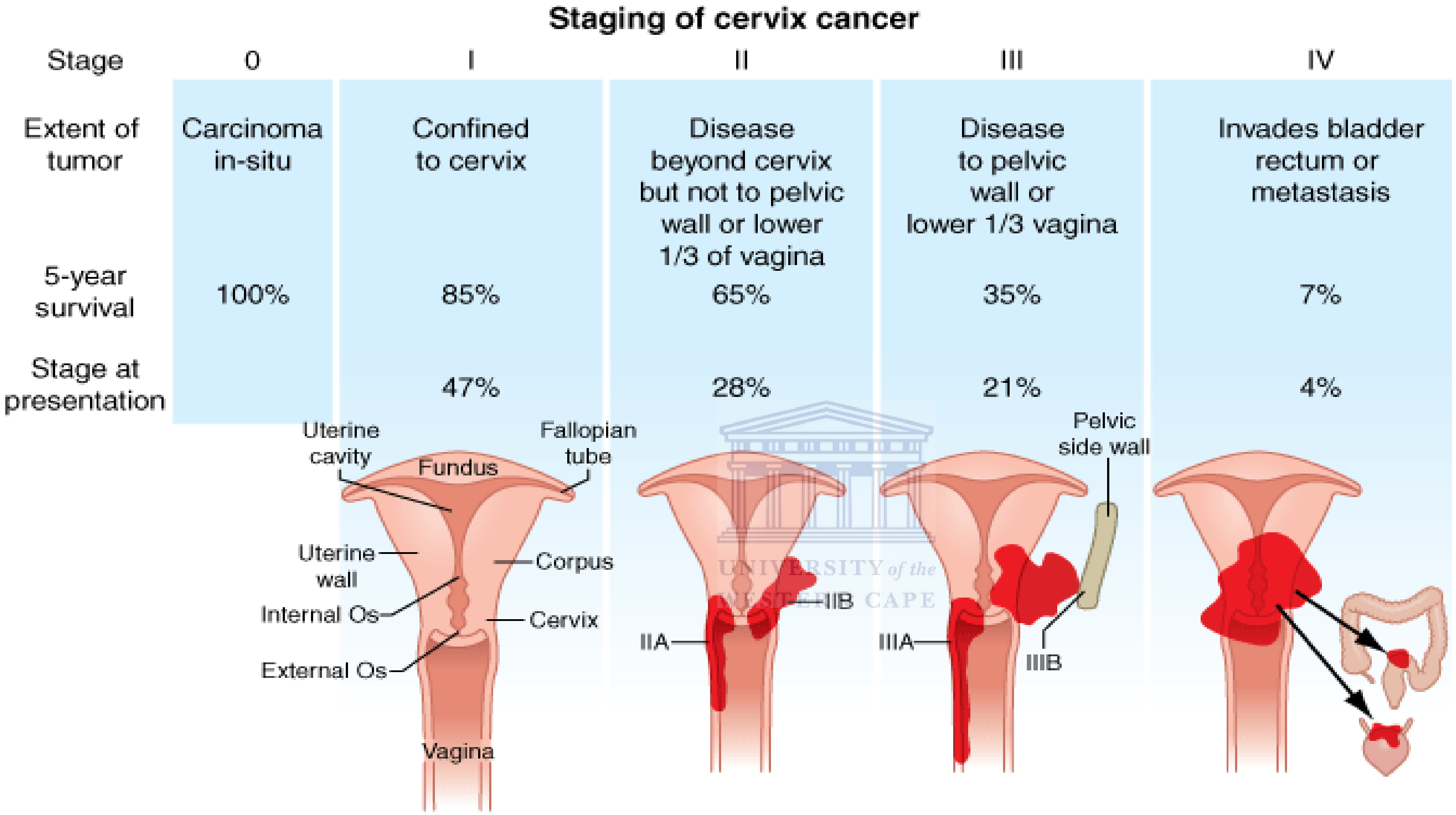
There are two types of cells that line the surface of the cervix: which are the glandular cells and the squamous cells. Glandular cells are found in the middle and upper third of the cervix close to the lining of the uterus and these cells have a column-shape or columnar appearance. Squamous cells are thin, flat cells that line the bottom third of the cervix. The transformation zone or the squamocolumnar junction is the borderline between the glandular cells and squamous cells and this is the area where cervical dysplasia and cancer usually occur (Jefferies, 2008). These cells eventually develop into different types of cancer which include the following:

- **Squamous cell carcinoma**: Cancer that develops from the cells covering the outer surface of the cervix at the top of the vagina (ectocervix).
- **Adenocarcinoma**: is the type of cancer that develops from the glandular cells lining the cervical canal or in the upper portion of the cervix (endocervix).
- **Adenosquamous carcinomas**: these are rare, mixed cancer cell types which contain features of both squamous cells and adenocarcinoma.
- **Small cell carcinoma and cervical sarcoma**: are the other rare cancer types that can develop in the cervix (<1% of all cervical cancers) (Jefferies, 2008).

Squamous cell carcinoma and adenocarcinoma are the most frequent types of cervical cancer and they are responsible for 85-90 % and 10-15 % of all cervical cancer cases respectively. They develop from the distinctive precursor lesions CIN / SIL and adenocarcinoma *in situ* (AIS) respectively (Lax, 2011). Invasive cancer occurs when the abnormal cells invade the deep muscle, fibrous tissue and the organs surrounding the uterus. Staging is the process of finding out how far the cancer has spread and the stage of cervical cancer informs about the size of the tumour, how deeply the tumour has invaded tissues within and around the cervix and whether there is metastasis to lymph nodes or distant organs. Determining the stage of the cancer is critical in determining what type of treatment should be offered (Southern Africa Litigation Centre, 2012).

1.8. Classification and Stages of Cervical Cancer

There are various cervical cancer stages that indicate the extent and site of infection. The International Federation of Gynaecology and Obstetrics (FIGO) system is used in staging of invasive cervical cancer based on clinical criteria like size, penetration depth within the cervix and spreading within and beyond the cervix as shown in figure 1.7. According to the FIGO system, cervical cancer is divided into 4 different stages, with 10 sub-stages from IA to IVB (Koyama *et al*., 2007). In stage I, the cancer cells are strictly limited to the cervix and are divided into stage 1A, 1B1 and 1B2 depending on the depth of penetration. Stage II cancerous cells extend beyond the cervix to the upper two thirds of the vagina (IIA) or the parametrial tissue (IIB) and not to the pelvic side wall. Furthermore, stage III describes a tumour that has spread to the lower third of the vagina (IIIA) or to the pelvic side wall (IIIB) and in stage IV, the cancer is advanced and has invaded the mucosa of the bladder or rectum (IVA) or has metastasised to distant sites outside the pelvis (IVB) (Koyama *et al*., 2007).



Staging of cervix cancer

Stage	0	I	II	III	IV
Extent of tumor	Carcinoma in-situ	Confined to cervix	Disease beyond cervix but not to pelvic wall or lower 1/3 of vagina	Disease to pelvic wall or lower 1/3 vagina	Invades bladder rectum or metastasis
5-year survival	100%	85%	65%	35%	7%
Stage at presentation		47%	28%	21%	4%

Figure 1.7: Illustration of the various stages of cervical cancer according to FIGO, adapted from Fauci *et al*., 2008.

1.9. Current Therapy and Prognosis of Cervical Cancer

Normally, cervical cancer early stages are either treated with surgery, including radical hysterectomy or pelvic lymph node dissection or a combination of chemotherapy and radiation (Rasty *et al*., 2009). Stage IA cervical cancers are treated with surgery with a five-year survival rate exceeding 95%, however, for stage IB or early stage IIA surgery and chemo radiotherapy are the choice of treatment with a five-year survival rate of 80%. The locally advanced stages IIB, III and IVA are treated with chemo radiotherapy, with both brachytherapy and external radiation being given in combination with adjuvant cisplatin-based chemotherapy which has been demonstrated to increase the efficacy of radiotherapy (Klopp and Eifel 2011). The five-year survival rates are 65%, 40% and less than 20% for stage IIB, III and IVA respectively, with stage IVB benefiting from local radiotherapy, which is combined with carboplatin and 5-FU-based chemotherapy and the disease is not curable once it reached this stage (Klopp and Eifel 2011). There are other tumour characteristics that are not included in the FIGO staging system that also influence prognosis of cervical cancer such as tumour volume as determined with high accuracy using imaging techniques in particular Magnetic Resonance Imaging (MRI) (Chiang and Quek, 2003).

The presence of lymph node metastases is another factor that influences survival of cervical cancer patients. The lymph node metastasis incidence correlates with other parameters of poor prognosis such as increasing stage, diameter of the tumour, lymphovascular space involvement and parametrial involvement with another important independent prognostic factor for cervical cancer being the presence of positive lymph nodes. Also of important prognostic significance is the number of positive lymph nodes, site and number of nodal sites (Creasman and Kohler, 2004). An important aspect of the treatment regime is the incidence of serious side effects after therapy. As a result of the anatomical location of the cervix in the pelvis, the lower uterus, bladder and posterior urethra are exposed to radiation during treatment of cervical cancer. A consequence of this is several urinary adverse effects (AEs). The probability of developing grade 1 and 2 AEs following radiotherapy for cervical cancer has been reported to be 28%, increasing by an additional 17.4% in five years (Elliott and Malaeb, 2011). The acute toxic effects caused by treatment last for a short time and may be resolved with medical management. However, the long term toxic effects may cause permanent impairment in the quality of life of the survivors.

Since the recurrence rate is very high and the incidence of side effects is quite frequent, there is still a great need for improved treatment strategies (Klopp and Eifel, 2011).

1.10. Screening and Diagnosis of Cervical Cancer

There are two approaches used for the control of cervical cancer: primary and secondary prevention. Primary prevention involves a risk-reduction approach through behavioural interventions for sexual health care seeking behaviour or through the mass immunization against high risk HPVs. These preventive methods can eliminate the probability of the development of the disease and, in the case of cervical cancer, one would prevent its onset by eliminating the risk of being infected with HPV. This can be achieved either by abstaining from sexual intercourse or through HPV vaccination (Sehgal and Singh, 2009). Secondary prevention includes screening for precancerous lesions and treating them. This stops the progression of the disease once the individual has already been infected. Routine screenings for cervical cancer precursors followed by appropriate treatment is an effective preventative measure in curbing the incidence of cervical cancer (Sehgal and Singh, 2009).

1.10.1. Screening of Cervical Cancer

The objective of screening programmes is to lessen the rate of mortality and morbidity due to cervical cancer and to reduce the number of patients suffering from cervical cancer. Basic screening programmes can lead to down-staging of cervical cancer, which in its own right offers benefits for patients (Botha *et al*., 2010). The lack of efficient high quality precancer screening, treatment resources and poor or lack of infrastructure results in an increased number of deaths in developing countries as a result of cervical cancer (Alliance for Cervical Cancer Prevention, 2009). Cancer of the cervix is preventable and highly curable by screening especially for those women who are asymptomatic for precancerous cervical cancer lesions. Early detection leads to faster and more successful treatment. It has been demonstrated by various studies that women who have been screened at least once in their lifetime between the ages of 30 and 40 have a reduced cervical cancer risk by 25-36% (Cervical Cancer Action, 2007).

It has been recommend by the American Congress of Obstetricians and Gynaecologists that cervical cytology screening begins at age 21 years and thereafter be repeated every 2 years for women aged 21-29 years and every 3 years for women aged 30 years or older who have had three prior normal pap smears. It is recommend that women who are infected with human immunodeficiency virus (HIV), immunosuppressed, and women previously treated for CIN 2, CIN 3 or cancer should undergo frequent screening. Women aged 65–70 years with three prior consecutively normal pap smears, and no abnormal pap smears over a period of 10 years may discontinue screening (Brown and Trimble, 2012).

1.10.2. Current Status of Cervical Screening in South Africa

There are two healthcare systems in South Africa with 80% of the population depending on the public sector, providing cost-free healthcare and 20% utilises private health care as they have medical insurance or can afford to pay for it (SouthAfrica.Info 2012). However, the South African Department of Health, in 1999, developed and adopted a National Cancer Control Policy which included a national programme for cervical cancer screening. The screening programme allows asymptomatic women aged 30 years and older three Pap smears within a ten-year period in their lifetime. The cervical cancer screening programme is implemented at district level at nurse-led primary health care clinics, which serve as an entrance to the public health care service (Mojaki *et al.,* 2010). The rationale behind the starting age was based on the fact that cervical cancer affects women in early to late middle age. The goal of the programme was to screen 70% of women in the targeted age group within 10 years from the initiation of the programme and to decrease cervical cancer incidence by 64% (Smith *et al*., 2003).

However, according to Gakidou, *et a*l., 2008 only 20% of the target population was screened. In some areas of the country the screening programme has been implemented but not throughout the country. The outcome is that presently there is no population-wide screening programme in South Africa. However, in some areas partial screening does take place and in the private sector opportunistic screening is commonly practised (Botha *et al*., 2010). In South African women, cervical cancer remains the second most common cancer and is most common in black women, accountable for 31% of the cancer burden in this group. The reflection of these figures is debatable as the registry is an under-representation. There are far-reaching implications associated with not being screened, as women with micro-invasive cervical cancer may not experience noticeable signs and symptoms and only seek health care when symptoms are evident and the disease is advanced (Maree and Moitse, 2014).

1.10.3. Cervical Cancer Screening Methods

Screening of cervical cancer is a way of preventing the disease from developing and diagnosing it at an early pre-cancerous stage. The three screening modalities are cytology, HPV detection and visual inspection.

1.10.3.1. The Papanicolaou Test

The Papanicolaou (Pap) smear test, also known as exfolative cervicovaginal cytology, is the most common technique used for screening and diagnosis of cervical cancer in its early stages. Women who are 18 years and older and sexually active are encouraged to undergo annual Pap smear tests. Cells are collected from the cervix by inserting a speculum inside the vagina and removing cells using a cotton swab or a small brush (Duraisamy *et al*., 2011). The cells are fixed on a glass slides and are sent to a cytology laboratory and evaluated by a trained cytologist or cytotechnician who determines the cell classification as atypical squamous cells of undetermined significance, low grade squamous intraepithelial lesions and high grade squamous intraepithelial lesions. Should abnormalities be encountered, additional tests will be performed (Duraisamy *et al*., 2011). The value of the Pap smear in screening of cervical cancer cannot be disputed as the method has resulted in a reduction in cervical cancer related mortalities. However, there are several limits to this test such as sensitivity to detect precursors of cervical cancer which is less than 50%, inadequate collection and transfer of cells to the slide, presence of obscuring blood, inhomogeneous distribution of abnormal cells and inflammation or thick areas of overlapping epithelial cells. There is an occurrence of false-negative results associated with the Pap test and it is said that it is unlikely to detect 60% of the general cervical cancer cases (Duraisamy *et al*., 2011).

1.10.3.2. Pelvic Examination

The pelvic examination is also an important technique in cervical cancer detection and is very similar to the Pap smear. A speculum is inserted into the vagina and the doctor examines the vagina and surrounding organs both visually and manually. The doctor inserts gloved hands into the vagina and feels the cervix and surrounding areas with the fingers (Duraisamy *et al*., 2011).

1.10.3.3. HPV DNA Screening

This procedure is targeted at identifying high risk HPV strains 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 59 and 68 which are commonly associated with high grade cervical intraepithelial neoplasia (HSIL) and invasive cancer of the cervix. Various methods such as southern blot hybridation are used with this screening method, however they are very laborious and tedious and are not suitable for clinical usage because they demand use of fresh tissue and not easy to conduct in mass screening programmes (Kerkal and Kulkarni, 2006). However, Hybrid capture 2 assay is a more suitable technique used and is utilised mostly in HPV-DNA screening. Samples that are used for screening are obtained from cell suspensions acquired from liquid based cytology or use of the cytocervical brush (Kerkal and Kulkarni, 2006).

The FDA has approved three types of tests to detect oncogenic HPV DNA: Hybrid Capture 2 test, Cervista_ HPV HR test and Cervista® HPV 16/18. The Hybrid Capture 2 test was approved in 2003 and it detects 13 oncogenic HPV types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, and 68) by making use of full genome probes complementary to HPV DNA, specific antibodies, signal amplification, and chemiluminescent detection. The Cervista_ HPV HR test was approved in 2009 and detects 14 HR HPV types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 66, and 68) using a signal amplification method for detecting specific nucleic acid sequences. This method uses a primary reaction that occurs on the targeted DNA sequence and a secondary reaction that produces a fluorescent signal. There are two limitations associated with these tests. Firstly, neither test can differentiate between single HPV genotype infections and multiple concurrent HPV genotype infections and secondly, they cannot quantitate viral load. The third HPV DNA test, Cervista_ HPV 16/18 was approved in 2009 and it only detects HPV 16 and 18, the genotypes most commonly associated with cancer, using a similar method to the Cervista_ HPV HR assay (Brown and Trimble, 2012).

1.10.4. Visual Approaches to Screening

Cervical cancer screening is limited by social infrastructure of the society being screened and financial resources, relying on methods that are low-cost and require fewer visits to the clinic. Therefore, alternative methods of screening may be implemented quickly and cheaply, such as visual inspection alone or with a magnifying device are presently utilised in low resource settings (Brown and Trimble, 2012). There are three approaches involved: (i) Visual inspection of the cervix with acetic acid (VIA), (ii) visual inspection after application of Lugol's iodine (VILI) and (iii) visual inspection with magnification (VIAM) (Duraisamy *et al*., 2011).

1.10.4.1. Visual Inspection with Acetic Acid

For this procedure the cervix is examined following the application of acetic acid. A bivalve speculum is used to expose the cervix and 4% dilute solution of acetic acid is applied to the cervix. Lesions that stain acetowhite are regarded as positive for VIA and those with dull white plaques and faint borders are considered low grade VIA, whereas those with sharp borders are considered high grade VIA. If no acetowhite lesions are detected, the test is considered negative. Various studied have demonstrated VIA to be reliable, sensitive and cost effective in comparison with Pap smear testing, especially in low resource countries (Duraisamy *et al*., 2011). Visual inspection using acetic acid wash has a sensitivity of 79% (95% CI 73 to 85%) and a specificity of 85% (95% CI 81 to 89%) for the detection of CIN2+ lesions (Brown and Trimble, 2012).

1.10.4.2. Visual Inspection with Lugol's Iodine (VILI)

After the cervix is examined using VIA, it is painted with Lugol's iodine solution and examined with the naked eye. Small high-grade lesions are easier to identify within the larger low-grade area. Abnormal areas of squamous epithelium (CIN or inflammation) does not stain brown. VILI is believed to be more accurate and more reproducible than VIA and to be much better than a Pap smear in identifying CIN (Duraisamy *et al*., 2011). The use of Lugol's iodine solution can increase sensitivity marginally, by 10% compared to other screening methods, and does not change the specificity (Brown and Trimble, 2012).

1.10.4.3. Visual Inspection with Magnification

This is the technique that visualises the cervix under low magnification after application of acetic acid. Using a magnifying device to aid in evaluating the cervix has comparable sensitivity and specificity to VIA only. The sensitivity and specificity of visual detection are dependent on the skill of the provider and vary widely (Brown and Trimble, 2012).

1.10.5. Cervical Cancer Diagnostic Methods

If an abnormal result is obtained following a screening process, often additional tests needs to be performed to determine the extent of the pre-cancer or cancer.

1.10.5.1. Colposcopy

Colposcopy is magnified visual examination of the ectocervix, squamous columnar junction (SCJ) and endocervical canal. It is accompanied by a biopsy of any abnormal-looking tissue and is also similar to a Pap smear (Duraisamy *et al*., 2011). A solution that stains abnormal cells white is applied to the cervix and the doctor views the cells using a high-powered microscope to detect abnormal cancerous cells. Colposcopy is used as a diagnostic test and not a screening test. Other techniques that are used include Cone biopsy, Endocervical Curettage, Loop electrosurgical procedure (LLETZ/LEEP) and imaging (Duraisamy *et al*., 2011).

1.10.5.2. Cone Biopsy

Cone Biopsy is a procedure that is also known as conization, where a cone-shaped piece of tissue is removed from the cervix. The cone base is formed by the exocervix and the apex of the cone is from the endocervical canal with the transformation zone (the area in the cervix where pre-cancers and cancer are likely to develop) is contained within the cone specimen. Cone biopsy is also used as treatment to remove many pre-cancers and some very early cancers completely. If a huge amount of tissue is removed, the result is a higher risk of giving birth prematurely if the individual it to become pregnant (Duraisamy *et al*., 2011).

1.10.5.3. Endocervical Curettage

Endocervical curettage also called endocervical scraping is a procedure where a curette is inserted into the endocervical canal to scrape the inside of the canal to remove some of the tissue which is sent to the laboratory for analysis. Side effects associated with this procedure include light bleeding and abdominal cramping pain (Duraisamy *et al*., 2011).

1.10.5.4. Loop Electrosurgical Procedure

The Loop Electrosurgical Procedure also called loop excision of the transformation zone (LLETZ). This is a procedure where a thin wire loop heated with electrical current is used to remove tissue. Mild cramping during and after the procedure and mild-to-moderate bleeding for several weeks are some of the side effects associated with this method (Duraisamy *et al*., 2011).

1.10.5.5. Imaging

Certain imaging techniques are performed in order to determine if the cancer has spread beyond the cervix. These include Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) scans. The usefulness of a CT scans is to help identify if the cancer has spread to the lymph nodes in the abdomen and pelvis (Duraisamy *et al*., 2011).

1.10.6. HPV Vaccines

The most cost-effective method to prevent cervical cancer against infectious agents is through vaccination. The primary goal of HPV vaccination is to reduce the incidence of cervical cancer and precancerous lesions. The other objective is to decrease the rate of cancers and other benign lesions related with HPV infection (Zarchi *et al*., 2009). There are two vaccines available Gardasil® (Merck, Sharpe and Dohme) and Cervarix® (GlaxoSmithKline). The vaccines are based on the virus-like particle (VLP) technology, where viral genes encoding surface proteins are used to produce empty virus shells, capable of inducing effective immune responses without any infectious or malignant potential. Both vaccines have prophylactic properties and do not clear existing HPV infection or treat HPV-related disease (Souter, 2012). The purpose should be to administer both vaccines before the start of sexual activity and the first exposure to HPV infection as this is when the individual would derive the most benefit from these vaccines (Souter, 2012). Both vaccines are highly immunogenic and nearly all individuals develop an antibody response one month after completing the three-dose series. It is not yet known how long these vaccines protective efficacies last; however, to date it has been maintained for the duration of the observation periods: 9.4 years for Cervarix® and five years for Gardasil® (Souter, 2012).

1.10.6.1. Gardasil®

Gardasil® is a quadrivalent vaccine that contains virus-like particles for HPV subtypes 6, 11, 16 and 18. It is designated for prevention of cervical, vaginal and vulva pre-cancers and cancers as well as anogenital warts that are caused by HPV types 6, 11, 16 and 18 in females aged 9-26 years. The vaccine is administered as a series of three intramuscular injections at 0.2 and 6 months. It is delivered into the deltoid muscle or in the higher anterolateral area of the thigh (Souter, 2012). Internationally, Gardasil® is used up to the age of 26 years in men and 45 years in women. In South Africa, Gardasil® is also registered for use in boys aged 9-17 years for the prevention of anogenital warts (Souter, 2012).

1.10.6.2. Cervarix®

Cervarix® is a bivalent vaccine that contains virus-like particles for HPV subtypes 16 and 18. It is indicated for the prevention of cervical pre-cancers and cancers caused by HPV types 16 and 18 in females from nine years of age. Cervarix® is given as a series of three intramuscular injections at 0, 1 and 6 months, delivered into the deltoid muscle (Souter, 2012). Both Gardasil and Cervarix vaccines are preventive, not curative for HPV infection or HPV-related diseases. Therefore, HPV vaccine is most useful when given to girls and women prior to infection. Vaccination is able to reduce up to 70% of cervical cancer related to HPV infection and even prevents precancerous and cancerous lesions of the genitalia. It must be remembered that the HPV vaccine does not eliminate the need for continued Pap smears as 30% of cervical cancers are caused by HPV types that are not included in these vaccines (Souter, 2012).

1.10.7. Limitations of Current Screening Methods

The major challenge for cervical cytology is the need to detect rare events. A specimen collected by liquid-based cervical cytology contains a minimum of 5 000 normal squamous cells, with most samples containing 50 000 or more normal cervical squamous epithelial cells as well as benign endocervical cells and inflammatory components. On the other hand, HSILs may often be based on the detection of only a very small number of abnormal cells; frequently in the range of 10–100 dysplastic cells per slide (Ling *et al*., 2008). Current methods for cervical cancer screening are not only labour-intensive but are also highly subjective and have a relatively low sensitivity and specificity rate for the detection of some high-grade clinically significant lesions. With the liquid-based Pap test, the sensitivity of cervical screening has increased to about 80% from 65% compared to the conventional Pap smear, resulting in an improvement of the overall clinical, economic, and patient outcomes. However, the specificity of liquid-based Pap test dropped from 95% with conventional Pap smear to about 75% (Ling *et al*., 2008). Unfortunately only a small percentage of cervical cancer patients are diagnosed in the early stages (stage 0 and 1). This has prompted the search for cervical cancer biomarkers, although biomarkers have been identified, these have not yet been successful in the detection of CC.

1.11. Biomarker Applications in Cancer

The discovery of biomarkers in cancer is becoming extremely important and it is very clear that people would benefit tremendously by a greater availability of such effective molecular indicators that can be monitored non-invasively from readily accessible body fluids. A widely accepted and comprehensive medical definition describes a biomarker as a characteristic that can be objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes and pharmacologic responses to a therapeutic intervention (Aebersold *et al.*, 2005). The National Cancer Institute (NCI) defines a biomarker as a biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease. Cancer biomarkers can be messenger ribonucleic acid (mRNA), deoxyribonucleic acid (DNA), proteins, metabolites or processes such as apoptosis, angiogenesis or proliferation (Kulasingam and Diamandis, 2008). The markers are produced either by the tumour itself or by other tissues in response to the presence of cancer or other associated conditions such as inflammation. Such biomarkers can be found in a variety of fluids, tissues and cell lines. Contrary to common belief, biomarkers are not only or necessarily molecules, they are any type of measurable change which may have clinical relevance (Kulasingam and Diamandis, 2008).

For instance biomarkers can be certain proteins present on the tumour or released by the tumour in the blood, which serves as an indication for recurrence of the disease after curative surgical intervention. A biomarker can also be the expression level of mRNA or the presence of a gene mutation targeted by a drug (Kulasingam and Diamandis, 2008). Various characteristics are used to classify biomarkers such as molecular biomarkers or imaging biomarkers, with molecular biomarkers being tumour-associated proteins, mRNA or DNA fragments which are used either individually or in signatures of multiple molecules. Additionally, different functional subgroups of proteins, such as enzymes, glycoproteins, oncofetal antigens and receptors, may serve as useful biomarkers. Furthermore, tumour changes such as genetic mutations, amplifications, translocations and changes in microarray profiles (signatures) may also be utilized as tumour markers (Smith *et al.*, 2003). Imaging biomarkers are anatomical, physiological, biochemical or molecular characteristics which are detectable using certain parameters or features from imaging modalities such as MRI, CT or Positron Emission Tomography (PET) (Smith *et al.*, 2003). There are various advantages in discovery of new biomarkers, such as improved diagnosis and predicting response to both current treatment regimens and to novel molecularly targeted agents (Eifel, 2006). Cancer biomarkers have several applications such as early diagnosis of cancers, improved reproducibility of the histopathological diagnoses, surveillance of individuals at risk and post therapy monitoring (Wentzensen and von Knebel Doeberitz, 2007).

1.12. Classification of Biomarkers

Biomarkers are classified based on the sequence of events from the onset of exposure to disease detection and they play an essential role in providing some insights into prognosis, disease progression and response to treatment (Manne *et al*., 2005). There is a variety of biomarkers available; they are utilized in experimental research and clinical settings. Biomarkers are grouped in to different categories such as early detection, diagnostics, prognostics and predictive biomarkers. Other types are classified based on biomolecules such as DNA, RNA and proteins (Mayeux, 2004). Therefore, biomarkers are invaluable tools for cancer detection, diagnosis, patient prognosis and treatment selection. They can also function in localizing the 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 lung) or urine (instead of urinary tract) can also influence the management of cancer (Bhatt *et al*., 2010). A significant focus of clinical research, as evident by the large number of publications, has been the discovery and validation of biomarkers with the primary aim of facilitating disease diagnosis (Bhatt *et al*., 2010).

1.12.1. Diagnostic Biomarkers

Diagnostic biomarkers are used to assist in making a specific diagnosis and in relation to cancer these biomarkers may be assessed on tumour specimens; however, diagnostic biomarker research is performed on serum or plasma. The ideal source material for biomarker analysis is the blood, because it is easy to sample and relatively cheap. Furthermore, in preclinical research, it is common to encounter molecules of interest, which can be actively released in the bloodstream by tumour cells or leak out as a result of the high cell turnover rates typical of tumours (Kulasingam and Diamandis, 2008). The ultimate goal of diagnostic biomarkers would be to apply them in a disease screening setting or in suspected cancer patients, those that are at higher risk of developing the disease. For example, the presence of Bence–Jones protein in urine remains one of the strongest diagnostic indicators of multiple myeloma (Kulasingam and Diamandis, 2008). Novel or putative biomarkers may be developed into simple diagnostic tests assaying one or two biomarkers or more complex tests, where multiple biomarkers are assayed. Based on improved knowledge at the molecular level, such tests can provide faster, more accurate or information-rich diagnostics of many diseases (Hempel *et al*., 2008).

Biomarker-based tests can confirm clinical diagnosis and may provide other information concerning prognosis and best treatment options. These diagnostic tests may enable identification of disease or disease susceptibility in its early stages, before it can be diagnosed by other means, providing opportunities for prevention of disease progression or better disease management leading to better patient outcomes and a reduction in the direct and indirect costs of disease (Hempel *et al*., 2008). Improved diagnostics to detect cancer at an early stage, when it is curable with current methods would provide the greatest benefit to cancer patients. For the majority of cancers, 5-10 year survival often approaches 90% for cancers detected at an early stage, whilst it may drop to 10% or less for cancers detected at a later stage. It is well established that screening to detect cancer earlier saves lives. For instance, the Pap smear strongly reduces mortality through 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).

1.12.2. Prognostic Biomarkers

The role of prognostic biomarkers is to acquire knowledge about the natural history of the disease in terms of metastatic potential, likelihood of tumour progression and probability of patient survival independently of treatment. Prognostic biomarkers are relevant in cancer research for various reasons such as, when it is used to determine that the expression of a protein, gene or other molecules is directly correlated with an aggressive phenotype. This may give 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 randomized clinical trials aiming at evaluating the effect of diverse drugs (Kulasingam and Diamandis, 2008).

1.12.3. 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. It is indeed a common clinical observation 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 predictive biomarkers for different treatment regimens, one could administer the most effective drug only to those patients who, on the basis of the results from the biomarker analysis, will benefit most likely from that specific treatment. Biomarkers can also be used to predict toxicity and be particularly valuable for choosing between drugs with the same activity, but different toxicity profiles (Kulasingam and Diamandis, 2008).

1.13. Non-invasive Monitoring of Biomarkers

People at risk for development of cancer or with cancer would benefit tremendously by superior methods for determining cancer risk, detecting and localising cancer at its earliest stage, profiling for therapeutic decision making and monitoring response to therapy in real time. For some of these applications, it will not be known whether a tumour exists or, if it does its anatomical site. Thus, there is a need for biomarkers that can be monitored noninvasively in readily available bodily fluids (Aebersold *et al*., 2005). Tumours seep DNA and proteins into circulation and they also induce dramatic alterations of the surrounding stroma (e.g. alterations in basement membranes, angiogenesis and lymphogenesis) and release proteases that digest normal tissues and plasma proteins (Aebersold *et al*., 2005). Therefore, it is rational to expect many biomarkers to be present in blood and other fluids. Indeed, several individual plasma proteins (i.e., prostate-specific antigen (PSA), cancer antigen 125 (CA125), Carcinoembryonic antigen (CEA), and alpha fetoprotein (AFP) antigen) are in clinical use as markers of the presence of a tumour, response to therapy or of tumour recurrence (Aebersold *et al*., 2005). Any proteins that are differentially expressed in cancer tissue when compared to normal tissue, or any proteins that are known to be involved in the cancer process, are good sources of candidate biomarkers for cancer.

1.14. Characteristics of an Ideal Biomarker

An ideal biomarker ought to explain the occurrence of a moderate proportion of the disease in the community and must have several qualities in order to be clinically applicable. Firstly, the biomarker test must be safe and easy to perform, meaning it must be as non-invasive as possible using external body fluids or blood (Fathi *et al*., 2014). The biomarker test should be done at the bedside or as a relatively simple laboratory test using a rapid and reliable standardised platform. Secondly, a biomarker should be highly specific for the disease and preferably be able to identify subtypes and causes of the disease. Thirdly, a biomarker should be sensitive for early detection as possible. Additionally, the sensitivity and specificity of the biomarker should be relatively high, thus reducing false-positive and false-negative values (Fathi *et al*., 2014). An ideal cancer biomarker should be measured easily, reliably and cost-effectively using an assay with high analytical sensitivity and specificity. An ideal biomarker should be present in detectable quantities at early or preclinical stages and the quantitative levels of the cancer biomarker should reflect tumour burden (Hanash *et al*., 2008).

1.15. Cancer Biomarkers in Clinical Use

Cancer biomarkers are present in tumour tissues or serum and they include a wide range of molecules such as DNA, mRNA, transcription factors, secreted proteins and cell surface receptors. Proteins designated as clinical cancer biomarkers are those offered commercially by ARUP or by Mayo Medical Laboratories, also offered for internal use by either NIH or the Fred Hutchinson Cancer Research Centre (Hanash *et al*., 2008). Table 1.2 contains cancer biomarkers that are approved by the Food and Drug Administration (FDA) (Sahab *et al*., 2008), and table 1.3 shows certain types of biomarkers and their applications.

Table 1.2: FDA Approved Biomarkers

Cancer	Biomarker
Prostate	PSA

Breast	CA15.3, Her-2/neu, CA27-29
Ovarian	CA125
Testicular	Human corionic
Thyroid	Thyroglobulin
Pancreas	CA19–9

*****Adapted from Sahab *et al*, 2007

Table 1.3: Different types of Cancer Biomarkers

Cancer Biomarker	Tumour	Application	Typical Sample
AFP	Hepatocellular carcinoma, Hepatoblastoma	Diagnostic and prognostic	Blood
BRCA-1, BRCA-2	Breast cancer	Diagnostic	
CA125	Epithelial ovarian carcinoma, Fallopian tube cancer	Diagnostic and prognostic	Blood
CA 15-3	Breast cancer	Diagnostic and prognostic	Blood
CA 19-9	Pancreatic cancer, Bladder cancer	Diagnostic and prognostic	Blood
CEA	Colorectal cancer	Diagnostic and prognostic	Serum
hCG	Germ cell tumours (ovarian, Testicular)	Diagnostic	Serum
PSA	Prostate cancer	Diagnostic and prognostic	Blood
Thyroglobulin	Papillary and follicular thyroid cancer	Diagnostic and prognostic	Serum

*Adapted from Fathi *et al*., 2014

1.16. Mechanisms of Biomarker Elevation in Biological Fluids

The levels of proteins are physiologically maintained in body fluids and thus, in disease states proteins may become elevated as a result of the disease by various mechanisms. These include and not limited to gene over-expression, angiogenesis, invasion and destruction of tissue architecture and lastly increased protein secretion and shedding (Jarjanazi *et al.*, 2008). Firstly, increased quantities of proteins may be a result of increases in the specific gene or chromosome copy number (gene amplification), epigenetic modifications such as DNA methylation and increased transcriptional activity. The imbalance between gene repressors and activators causes an increase in transcriptional activity (Jarjanazi *et al.*, 2008). Secondly, invasion of tissues by the tumour may allow release of molecules into the interstitial fluids directly, reabsorbed by the lymphatics and subsequently into the blood. In the case of epithelial cancer types, proteins must break through the basement membrane of the invading tumour before entering the circulation (Jarjanazi *et al.*, 2008). Thirdly, approximately 20-25% of proteins are secreted, thus elevated protein levels may occur as a result of aberrant secretion or shedding of membrane-bound proteins containing an extracellular domain (ECD) (Jarjanazi *et al.*, 2008).

Furthermore, single nucleotide polymorphisms may cause alterations in the signal peptide of proteins resulting in atypical secretion patterns. Cancer-associated glycoproteins may be released into circulation due to the change in the polarity of the cancer cells. Also, increased protease expression may lead to increased ECD cleavage of membrane bound proteins, resulting in increased circulating levels of these cleaved products (Jarjanazi *et al.*, 2008). There are five major mechanisms by which molecules can be elevated in biological fluids during initiation and progression of cancer. Such molecules could serve as effective cancer biomarkers. A representation of the different human body fluids that could be used as a source of biomarkers for specific types of cancers is shown in table 1.4. The various mechanisms of elevation are outlined below.

Table 1.4: Human biological fluids: a source for biomarker discovery

Human biological fluid	Cancer type
Plasma	Broad spectrum of diseases
Serum	Broad spectrum of diseases
Cerebrospinal fluid	Brain
Nipple aspirate fluid	Breast
Breast cyst fluid	Breast
Ductal lavage	Breast
Cervicovaginal fluid	Cervical and endometrial
Stool	Colorectal
Pleural effusion	Lung
Bronchoalveolar lavage	Lung
Saliva	Oral
Ascites fluid	Ovarian
Pancreatic juice	Pancreatic
Seminal plasma	Prostate and testicular
Urine	Urological

*Adapted from Jarjanazi *et al*., 2008

1.16.1. Gene Overexpression

The protein encoded by the gene can be expressed in increased quantities as a result of increases in gene or chromosome copy number (i.e. gene amplification) or through increased transcriptional activity. The latter process could be the result of imbalances between gene repressors and gene activators. Epigenetic changes, such as DNA methylation, are also known to affect gene expression. On a larger scale, chromosomal translocations can result in gene regulation by promoters that are sometimes enhanced by steroid hormones; transposons can serve a similar role (Kulasingam and Diamandis, 2008).

1.16.2. Increased Protein Secretion and Shedding

Another way by which molecules can be elevated in biological fluid is aberrant secretion or shedding of membrane-bound proteins with an extracellular domain, given that 20-25% of all proteins are secreted. Alterations in the signal peptide of proteins as a result of single nucleotide polymorphisms can result in atypical secretion patterns (Kulasingam and Diamandis, 2008). The release of cancer-associated glycoproteins into circulation is caused by change in polarity of cancer cells, this result in elevation of molecules in biological fluids. Increased circulating levels could also be caused by increased expression of proteases that cleave the ECD portion of membrane proteins. Alfa-fetoprotein is one of the many proteins secreted into circulation, it is rapidly released from both normal and cancer cells. Human epidermal growth factor receptor 2 (HER2) currently serving as a breast cancer biomarker is a classic example of shedding of membrane proteins into bodily fluids (Kulasingam and Diamandis, 2008).

1.16.3. Angiogenesis, Invasion and Destruction of Tissue Architecture

Invasion of tissue by the tumour might allow direct release of molecules into the interstitial fluid and subsequent delivery by the lymphatics into the blood. For epithelial cancer types, the proteins must break through the basement membrane of the invading tumour before they appear in the blood. For instance PSA is abundantly expressed by prostatic columnar epithelial cell and secreted into the glandular lumen (Kulasingam and Diamandis, 2008).

1.17. Bioinformatics

Bioinformatics is the application of computer technology to the management of biological information. It is the science of storing, extracting, organizing, analysing, interpreting and utilizing information from biological sequences and molecules. Bioinformatics has been driven by advances in sequencing of DNA and mapping techniques. Rapid developments over the past few decades in genomic, other molecular research technologies and developments in information technologies have combined to produce remarkable amounts of information related to molecular biology. The main aim of bioinformatics is to increase the understanding of biological processes (Raza, 2010). One of the numerous ways to apply bioinformatics methods to cancer, relating to signalling, proliferation, communication and specificity of disease metabolisms is through cancer bioinformatics. Another developing science is clinical bioinformatics, merging medical informatics, clinical informatics, bioinformatics, mathematics, omics science and information technology together. Clinical bioinformatics is considered as one of the crucial factors for addressing important clinical challenges in early diagnosis, predictive prognosis and effective therapies in cancer patients. The development of cancer bioinformatic-specific methodologies or the introduction of new and advanced bioinformatics tools is strongly desired to address the specific challenge of cancer (Wu *et al.,* 2012). The application, specificity and integration of methodologies, computational tools, software and databases which can be utilised to explore the molecular mechanisms of cancer and identify and validate novel biomarkers, network biomarkers and individualized medicine in cancer should be seriously considered (Wu *et al.,* 2012).

Cancer bioinformatics is anticipated to play a vital role in identifying and validating biomarkers specific to clinical phenotypes connected to early diagnosis, measurements to monitor the progress of the disease and the response to therapy and predictors for the improvement of a patient’s quality of life. The first and critical step in discovering and developing new diagnostic and therapies for disease is to understand the interaction between bioinformatics and clinical informatics (Wu *et al.,* 2012). Bioinformatics can enable clinicians to answer fundamental questions based on individual patient details including disease characteristics, laboratory results, proteomic, genomic and metabolic information. The progression of biomarker discovery is impossible without bioinformatics, which connects individual discovery processes, including experimental design, study execution and bioanalytic analysis. Bioinformatics has supported translational research, which has provided critical tools for transforming data into medical practise and has prompted biomarker breakthroughs and drug development. The development of cheaper, less invasive tests that will benefit both clinicians and patients can be permitted by the use of clinically validated biomarkers (Suh *et al.,* 2013).

1.17.1. Analysis of Gene Expression

Expression of various genes can be determined by measuring mRNA levels with various techniques such as microarrays, expressed cDNA sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signature sequencing (MPSS), or various applications of multiplexed *in-situ* hybridization and so forth (Raza, 2010).

1.17.2. Analysis of Protein Expression

There are various ways to measure gene expression including mRNA and protein expression; however protein expression is one of the best indicators of actual gene activity since proteins are generally final catalysts of cellular activity. Protein microarrays and high throughput mass spectrometry (HTMS) can provide a preview of the proteins present in a biological sample. Bioinformatics plays a major role in making sense of protein microarray and HT MS data (Raza, 2010).

1.18. Sources of Biomarkers

One of the major concerns in biomarker research is the accessibility of the source of biological matrix. Amongst a wide variety of available body fluids, blood is considered the most promising and other fluids such as urine, amniotic fluid, synovial fluid, saliva, nipple aspirate fluid, and cerebrospinal fluid to mention a few cannot offer a protein profile as representative as that of blood and availability of these samples may be restricted. Blood as a source of biomarkers is easily accessible; its collection is minimally invasive, low risk and cheap (Tambor *et al*., 2010).

1.18.1. Blood

Blood is the most commonly used biological fluid for biomarker analysis in clinical practice. The advantages of using blood, serum and plasma as a source to mine for biomarkers is that it can be obtained through a minimally invasive procedure, it is abundantly available and some constituents of blood reflect diverse pathological states. The prime importance of blood is that it is in contact with virtually all cells of the organism and due to specific secretion, shedding from the surface or non-specific leakage, tissue-related proteins are released into the blood stream. Thus, pathologically affected cells with deregulated proteomes may create a specific barcode by releasing disease-related proteins into circulating blood. Furthermore, the barcode can also be represented by molecules resulting from the organism's response to the disease (Zhang *et al*., 2007).

1.18.2. Proximal Fluids and Tissue

Exploring biological fluids proximal to tumours is an attractive strategy for the identification of tumour-secreted proteins. Various types of fluid and effusion offer access to the proteins from tumour tissue that may be released into extracellular fluids through secretion or through cell and tissue breakdown (Hanash *et al*., 2008). Proximal fluids are an alternative to blood as a source of biomarkers, even though they are not as representative as blood, their expedience increases if the nidus of a disease is in close proximity with the particular body fluid. For instance, urine may be a prospective source of kidney disease biomarkers or cerebrospinal fluid for central nervous system diseases (Quintana *et al*., 2009). According to Jurisicova *et al*, certain proteins originating from the tissue could subsequently appear and be monitored in the bloodstream. The secretion and shedding of tumour proteins into the bloodstream are expected to occur as a result of leaky capillary beds, protease cleavage and high rates of cell death within the tumour mass. However, these samples are often complex incorporating many different types of cells. Often, tumour biopsies may not simply contain tumour tissues but also include blood components as well as normal tissue (Jurisicova *et al*., 2008).

Urine has been used for hundreds of years as a source for biomarker discovery and has become an extremely valuable body fluid for the discovery of biomarkers. Urine is able to provide important diagnostic information about certain biological disorders and conditions. It is a powerful diagnostic fluid utilized in point-of-care devices and biomarkers for various cancers have become evident in urine. For instance urinary protein biomarkers for prostate cancer and breast cancer are calgranulin B/MRP-14 and ADAM 12 respectively. The use of urinary biomarkers to diagnose diseases is a long-standing practice. Proteins that are found to be differentially expressed can be used as biomarkers for the diagnosis and prognosis of the disease and may be used as therapeutic targets for future treatment and management of the disease (Tantipaiboonwong et *al*., 2005).

1.19. Methods for Cancer Biomarker Discovery

The discovery of biomarkers have proven to be one of the most broadly applicable and successful means of translating molecular and genomic data into clinical practise. The comparisons between healthy and diseased tissues have highlighted the importance of tasks such as class discovery (detecting novel subtypes of a disease) and class prediction (determining the subtype of a new sample) (Segata *et al*., 2011). There are various advantages of identifying a biomarker or a panel of biomarkers based on the expectation that it will lead to the development of a sensitive and reliable assay that is easily readable. That capability developed and validated in a platform, leads to the ability to develop an assay that is able to detect biomarkers (i.e. proteins) at extremely low concentrations (Larner, 2008). In order to ensure long-term and widespread success the assay platform needs to be as non-invasive as possible. Following the development of an assay kit, the ultimate goal would be to translate this assay into a user friendly, hand-held point-of-care (POC) device which is able to monitor this panel of biomarkers in body fluids such as blood or urine with minimal invasive procedures (Larner, 2008). Advances in genomics, proteomics, transcriptomics, and metabolomics have generated many candidate biomarkers with the potential for diagnostic and clinical value.

1.19.1. Genomics

Genomics is the study of all nucleotide sequences in the genome of an organism. The widely used genomic technologies in cancer research include single nucleotide polymorphism (SNP) array, next‑generation sequencing (NGS) technologies, such as Roche 454, ABI SOLiD, Illumina Solexa and Helicos. Genomics is applied in cancer biomarker discovery to seek specific biomarkers related to genome alterations caused by cancer, for instance DNA sequence changes, copy number aberrations, chromosomal rearrangements and epigenetic modifications such as DNA methylation. There are various advantages for DNA methylation since it has emerged as highly promising biomarkers offering stability and easy detection using PCR or array-based approaches in blood, sputum and urine thus making it suitable for non-invasive clinical detection (Zhang *et al*., 2011). The human leukaemia antigen (HLA) was found to be associated with renal cell carcinoma subsequent to cytokine therapy. Furthermore, haplotypes are observed in most key genes thus contributing to cancer progression by creating allelic imbalances. Srinivas *et al*, 2001 noted that SNPs can be mostly present at the initiation of many epithelial tumours thus causing loss in heterozygosity in many tumour suppressor genes. All these serve as possible gene markers for cancer. Wang *et al* used integrative transcriptomics and proteomics for identification of novel liver cancer biomarkers.

1.19.2. Transcriptomics

Transcriptomics is a technique that measures the relative amount of all mRNAs in an organism in order to determine the patterns and levels of gene expression. DNA microarray is a power technique used in Transcriptomics. Microarray-based gene expression profiling of human cancers has generated hundreds of novel diagnostic and prognostic biomarkers as well as therapeutic targets (Zhang *et al*., 2011).

1.19.3. Proteomics

Proteomics is used to study proteins expressed in a cell, tissue or organism, including all protein isoforms and posttranscriptional modifcations. There are two proteomic approaches: gel-based and gel-free proteomics. In the gel-based technique, proteins are separated and quantified using two-dimensional polyacrylamide gel electrophoresis (2D-Gel), with mass spectrometry to identify molecules of interest (Zhang *et al*., 2011). In the gel-free approach, “shotgun" proteomics are employed in the combined use of multidimensional liquid chromatography (MDLC) combined with tandem mass spectrometry. The basic strategies include digesting proteins into peptides and sequencing them using tandem mass spectrometry and identifying the generated sequences using automated database searches. In an attempt to identify cancer biomarkers, proteomics has been widely applied to analyse serum, saliva, cerebrospinal fluid and nipple-aspirated fluid (Zhang *et al*., 2011).

Recent proteomics studies in pancreatic cancer have identified proteins differentially regulated in cancer samples and have led to the discovery of several candidate biomarkers. An example of a known biomarker discovered through the use of proteomic technologies is CA19-9. This is the only widely used marker for pancreatic cancer and is frequently elevated in pancreatic cancer but can also be expressed in other malignancies. Moreover, CA19-9 levels can be elevated in such benign conditions as acute and chronic pancreatitis, hepatitis, and biliary obstruction. The sensitivity and specificity of CA19-9 are 80–90%, limiting its value as a screening marker for the general populace (Chen *et al*., 2005).

1.19.4. Secreted Protein Approach

A candidate biomarker should be a secreted protein, because secreted proteins have the highest probability of entering the circulation. Examining tissues or biological fluids near to the tumour site of origin could facilitate identification of candidate molecules for further investigation. This approach is supported by the increasing evidence that tumour growth and progression is dependent on the malignant potential of the tumour cells as well as on the microenvironment surrounding the tumour e.g. stroma, endothelial cells and immune and inflammatory cells (Kulasingam and Diamandis, 2008). Some of the widely used cancer biomarkers such as CEA, CA125 and HER2 are membrane-bound proteins, which are shed into the circulation. However, the identification of secreted proteins in tissues or other biological fluids does not necessarily mean that the proteins will be detectable in the sera of cancer patients. Serum-based diagnostic tests rely on the stability of the protein, its clearance, its association with other serum proteins and the extent of post-translational modifications (Kulasingam and Diamandis, 2008).

1.19.5. Cancer Biomarker Family Approach

The cancer biomarker family approach is based on the premise that if a member of a protein family is already an established biomarker, then other members of that family might also be good cancer biomarkers. For instance, PSA is a member of the human tissue kallikrein family and kallikreins are secreted enzymes with trypsin-like or chymotrypsin-like serine protease activity. This enzyme family consists of 15 genes clustered in tandem on chromosome 19q13.4.63. Currently PSA (KLK3) and KLK2 have important clinical applications as prostate cancer biomarkers. KLK6 has been studied as a novel biomarker for ovarian cancer. Similarly, KLK3, KLK5 and KLK14 have been shown to be increased in the serum of patients with breast cancer, thereby potentially serving as diagnostic markers (Kulasingam and Diamandis, 2008).

1.20. Serum Markers for Cervical Cancer

The Pap smear test has resulted in a remarkable decline in cervical cancer incidence and mortality rates; however it has its imperfections. The Pap smear has an average sensitivity of 51% to detect CIN and average specificity of 98%. In order to improve these qualitative parameters several attempts have been made amongst them the use of liquid-based cytology, repetition of Pap smears every 1-3 years and also the addition of HR-HPV detection. Regardless of these efforts, the number of false-positive and false-negative results is considerably high (Litjens *et al*., 2013). Increasing the rate of screening among groups of women who are at a higher risk of cervical cancer will reduce the incidence and mortality associated with this malignancy. Another approach would be to establish appropriate serum testing for early diagnosis of cervical cancer. Additional disease markers are thus needed to identify women at risk. The most commonly used serum marker for squamous cell cervical carcinoma, which makes up 85–90% of all cervical carcinomas, is the squamous cell carcinoma antigen (SCC). Elevated levels of serum SCC have been detected in 28–85% of cervical squamous cell carcinomas (Ueda *et al*., 2010).

Table 1.5 outlines cervical cancer diagnostic serum markers in clinical use, with the positive rates (elevated serum levels) detected for the indicated serum markers, in cases of squamous cell carcinoma (squamous), adenocarcinoma (adeno), or for all histological types. In some studies elevated levels of SCC were demonstrated to have a predictive value for prognosis. Pre-treatment levels of SCC have been shown to be related to the stage of the disease, size of the tumor, depth of the stromal invasion, the lymph-vascular space involvement, and lymph node metastasis. The SCC marker has been used also in cervical cancer patients for follow-up examination and increased levels were shown to precede the clinical detection of recurrence of the disease (Ueda *et al*., 2010).

Table 1.5: Diagnostic serum markers for cervical cancer in clinical use

Serum markers	Positive rate
SCC	Squamous 28–85%
CYFRA 21-1	Squamous 42–52%
CA 125	Adeno 27–75%
CA 19-9	Adeno 35–42%
CEA	Adeno 26–48%
IAP	43–51%

*Adapted from Ueda *et al*., 2010

The serum tumour marker CYFRA 21-1 1 (serum fragments of cytokeratin 19) is used for squamous cell carcinoma of the uterine cervix. Elevated levels have been detected in 42–52% of patients. Pre-treatment levels of CYFRA 21-1 are related to stage of the disease, size of the tumor, depth of the stromal invasion, the lymph-vascular space involvement, and lymph node metastasis. Elevated levels of CYFRA 21-1 do not have predictive value for prognosis and have been reported to be useful in monitoring response to radiotherapy and chemotherapy. It is used also in follow-up examination of cervical cancer patients (Ueda *et al*., 2010). Raised serum CA 125 levels are associated with the stage of the cervical disease and are of some prognostic significance. Immunosuppressive acidic protein (IAP) marker is elevated in cervical carcinomas. IAP levels are linked to disease stage and lymph node metastasis and are of predictive value for prognosis (Ueda *et al*., 2010). The number of candidate biomarkers for the diagnosis of cervical cancer is overwhelming. However, the majority of these biomarkers have been tested on histological samples only. A lack of sensitivity and specificity has, so far, given most of the tumor markers in current use an unsatisfactory predictive value.

1.21. Problem Identification

Development of cervical cancer involves sequential progression from normal cervical epithelium to preneoplastic cervical intraepithelial neoplasia and finally invasive cervical cancer. It has been shown by increasing evidence that early detection by testing for HR HPV and cervical papilloma smears have declined mortality rate associated with cervical cancer. However, these methods lack the capability to detect directly the development of cervical cancer. Therefore new and less invasive biomarkers are desired for the improvement of detection and prognostic outcome of cervical cancer (Ma *et al*., 2014). Consequently, there is a need for a diagnostic tool that is non-invasive by allowing for the detection of cervical cancer in bodily fluids. It should be sensitive enough to detect cancer in its early as well as pre-invasive stages and it should be consistently accurate across all ethnicities and ages. Additionally, it should be specific for cervical cancer with minimal generation of false positives or false negative results (Kumar and Sarin, 2009). The most significant molecular signatures implicated in cervical cancer are the HPV oncogenes E6 and E7. Unfortunately, they are ambiguous (non-specific) because they have been found to be associated with subset of head and neck cancers. If molecular signatures are used to signal the disease, ideally they should be present when the individual has cervical cancer and absent when the individual is healthy (Mishra and Verma, 2010). Discovering molecules that are solely expressed in the cervix tissue is not always feasible since cancerous cells are renegade normal cells. Henceforth, there is ongoing research for molecules differentially expressed or altered in a manner that discriminates them from normal cells (Tiffin *et al*., 2005). Considering the importance of the issue, the current study was initiated with the aim to identify potential biomarkers that can help in cervical cancer diagnosis, as well as a parallel study to identify biomolecules that can detect HPV infections with high specificity as well as sensitivity.

1.22. References

1. Abbott, R.G., Forrest, S., et *al*. 2006. Simulating the hallmarks of cancer. *Artif Life* 12(4): 617-34.

2. Abreu, A.L.P., Souza, R.P., Gimenes, F., and Consolaro, M.E.L. 2012. A review of methods for detects human Papillomavirus infection. *Virology Journal*, 9:262.

3. 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*, 4:1104-1109.

4. Alliance for Cervical Cancer Prevention, 2005. Preventing cervical cancer worldwide. 2-29.

5. Alliance for cervical cancer prevention, 2009. New evidence on the impact of cervical cancer screening and treatment using HPV DNA tests, visual inspection, or cytology; cervical cancer prevention fact sheet. 2009, 1-3. Consulted 2.5.2013.

6. American Cancer Society. 2010, 1-9. Cervical cancer. Consulted 2.05.2013 http://www.cancer.org/acs/groups/cid/documents/webcontent/003094-pdf.pdf.

7. Anorlu, R.I. 2008. Cervical cancer: the sub-Saharan African perspective. *Reprod Health Matters* 16:41-49.

8. Appleby, P., Beral, V., de Gonzalez, B.A., Colin, D., Franceschi, S., Goodhill, A., *et al*. 2007. Cervical cancer and hormonal contraceptives: collaborative reanalysis of individual data for 16,573 women with cervical cancer and 35,509 women without cervical cancer from 24 epidemiological studies. *Lancet*; 370:1609-21.

9. Arends, M.J., Buckley, C.H., and Wells, M. 1998. Aetiology, pathogenesis, and pathology of cervical neoplasia. *J Clin Pathol*; 51:96-103.

10. Bell, D.A. 2005. Origins and molecular pathology of ovarian cancer. *Mod. Pathol*, 18:19-32.

11. Bhatt, A.N., Mathur, R., Farooque, A., Verma, A., and Dwarakanath, B.S. 2010. Cancer biomarkers - Current perspectives. *Indian J Med Res*, 132:129-149.

12. Botha, H. M., and Dochez, C. 2012. Introducing human papillomavirus vaccines into the health system in South Africa. *Vaccine*, 30S:28-34.

13. Botha, H., Cooreman, B., Dreyer, G., Lindeque, G., Mouton, A., Guidozzi, F., Koller, T., Smith, T., Hoosen, A., Marcus, L., Moodley, M., and Soeters, R. 2010. Cervical cancer and human papillomavirus: South African guidelines for screening and testing. *South Afr J Gynaecol Oncol;* 2(1):23-26.

14. Brown, A.J., and Trimble, C.L. 2012. New technologies for cervical cancer screening. Best Practice and Research. *Clinical Obstetrics and Gynaecology,* 26:233–242.

15. Cervical cancer Action. 2007. New options for cervical cancer screening and Treatment in Low-resource settings. Consulted 9.6.2013 http://www.cervicalcanceraction.org/pubs/CCA_cervical_cancer_screening_treatmen.pdf

16. Chen, R., Pan, S., Brentnall, T.A., and Aebersold, A. 2005. Proteomic Profiling of Pancreatic Cancer for Biomarker Discovery. *Molecular and Cellular Proteomics* 4:523–533.

17. Chiang, S.H., and Quek, S.T. 2003. Carcinoma of the cervix: role of MR imaging. *Ann Acad Med Singapore*, 32:550-556.

18. Creasman, W.T., and Kohler, M.F. 2004. Is lymph vascular space involvement an independent prognostic factor in early cervical cancer? *Gynecol Oncol*, 92:525-529.

19. Dahlstrom, A. L., Andersson, K., Luostarinen, T., Thoresen, S., Ogmundsdottir, H., Tryggvadottir, L., *et al*. 2011. Prospective seroepidemiologic study of human papillomavirus and other risk factors in cervical cancer. *Cancer Epidemiol Biomarkers Prev*; 20:2541-2550.

20. Denny, L. 2012. Cervical cancer prevention: New opportunities for primary and secondary prevention in the 21st century. *International Journal of Gynaecology and Obstetrics*, 119:80–84.

21. Duraisamy, K., Jaganathan, K.S., and Jagathesh, C. B. 2011. Methods of Detecting Cervical Cancer. *Advances in Biological Research*, 5(4):226-232.

22. Eifel, P.J. 2006. Chemoradiotherapy in the treatment of cervical cancer. *Semin Radiat Oncol*, 16:177-185.

23. Elliott, S.P., and Malaeb, B.S. 2011. Long-term urinary adverse effects of pelvic radiotherapy. *World J Urol*, 29:35-41.

24. Faro, A., Giordano, D., and Spampinato, C. 2012. Combining literature text mining with microarray data: advances for system biology modelling, *Briefings in bioinformatics*, 13(1):61-82.

25. Fathi, E., Mesbah-Namin, S.A., and Farahzadi, R. 2014. Biomarkers in Medicine: An Overview. *British Journal of Medicine and Medical Research*, 4(8):1701-1718.

26. Fauci, A.S., Kasper, D.L., Braunwald, E., Hauser, S.L., Longo, D.L., Jameson, J.L., and Loscalzo, J. 2008. Harrison’s Principle of Internal Medicine, 17th Edition.

27. Francis, S.A., Battle-Fisher, M., Liverpool, J., Hipple, L., Mosavel, M., Soji Soogun, S, and Mofammere, N. 2011. A qualitative analysis of South African women’s knowledge, attitudes, and beliefs about HPV and cervical cancer prevention, vaccine awareness and acceptance, and maternal-child communication about sexual health. *Vaccine*, 29:8760-8765.

28. Fulda, S. 2009. Tumor resistance to apoptosis. *Int. J. Cancer*, 124:511-515.

29. Gakidou, E., Nordhagen, S., and Obermeyer, Z. 2008. Coverage of cervical cancer screening in 57 countries: Low average levels and large inequalities, *PLoS Med,* 5(6):0863-0868.

30. Getahum, F., Mazengia, F., Abuhay, M., and Birhanu, Z. 2013. Comprehensive knowledge about cervical cancer is low among women in Northwest Ethiopia. *BMC Cancer*, 13:2.

31. Globocan, 2008. Cancer incidence, mortality and prevalence worldwide. International Agency for Research on Cancer. Http: //globocan.iarc.fr/. (Access June 2013).

32. Godfrey, J. 2007. Towards optimal Health, Diane M.Harper, M.D., M.S., M.P.H. Discusses the HPV vaccine and prevention of cervical cancer. *Journal of women΄s Health*, 16(10)139-1401.

33. Gómez, D.T., and Santos, J.L. 2007. Human papillomavirus infection and cervical cancer: pathogenesis and epidemiology. Communicating Current Research and Educational Topics and Trends in Applied Microbiology A. Méndez-Vilas (Ed.)

34. Grizzi, F., and Chiriva-Internati, M. 2006. Cancer: looking for simplicity and finding complexity. *Cancer Cell Int*, 6:4.

35. Hanahan, D., and Folkman, J.1996. Patterns and emerging mechanisms of the angiogenic switch during tumourigenesis. *Cell*, 86(3):353-64.

36. Hanahan, D., and Weinberg, R. A. 2000. The hallmarks of cancer. *Cell*, 100(1):57-70.

37. Hanahan, D., and Weinberg, R.A. 2011. Hallmarks of Cancer: The Next Generation. *Cell*, 144:646-674.

38. Hanash, S.M., Pitteri, S.J., and Faca, V.M. 2008. Mining the plasma proteome for cancer biomarkers. *Nature*, 452(3), 571-579.

39. Hempel, W., Sziraczky, G., Swalm, L., and Takacs, L. 2008. Biomarkers: Impact on Biomedical Research and Health Care: Case Reports, OECD Science, Technology and Industry Analytical Paper, Directorate for Science, Technology and Industry, OECD, Paris.

40. Hoste, G., Vossaert, K., and Poppe, W.A.J. 2013. The Clinical Role of HPV Testing in Primary and Secondary Cervical Cancer Screening. Hindawi Publishing Corporation Obstetrics and Gynecology International, 1-8.

41. Human Papillomavirus and Related Diseases Report, SOUTH AFRICA. Version posted on www.hpvcentre.net in March 17th, 2014.

42. Iyoke, C.A., and Ugwu, G.O. 2013. Burden of gynaecological cancers in developing countries. *World J Obstet Gynecol*, 2(1):1-7.

43. Jarjanazi, H., Savas, S., Pabalan, N., Dennis, J., and Ozcelik, H. 2008. Biological implications of SNPs in signal peptide domains of human proteins. *Proteins,* 70:394-403.

44. Jefferies, H. 2008. Cervical cancer 1: an overview of screening and diagnosis. N*ursing times*; 104(44):26-7.

45. Jemal, A., Bray, F., Center, M., Ferlay, J., Ward, E., and Forman, D. 2011. Global cancer statistics.CA. *A cancer journal for clinicians*, 61(2):61-90.

46. Jemal, A., Bray, F., Forman, D., O'Brien, M., Ferlay, J., Center, M., and Parkin, M. 2012. Cancer Burden in Africa and Opportunities for Prevention Cancer.

47. Junqueira, L.C., and Carneiro, J. 2005. Chapter 22: The female reproductive system. In Basic Histology. The McGraw-Hill Companies. 435-455.

48. Jurisicova, A., Jurisica, I., and Kislinger, T. 2008. Advances in ovarian cancer proteomics: the quest for biomarkers and improved therapeutic interventions. *Expert Rev Proteomics*, 5: 551-560.

49. Kerkar, R., and Kulkarni, Y. 2006. Screening for cervical cancer: an overview. *Journal of obstetrics and Gynecology of India*, 56(2):115-122.

50. Klopp, A.H., and Eifel, P.J. 2011. Chemo-radiotherapy for cervical cancer in 2010. *Curr Oncol Rep*, 13:77-85.

51. Koyama, T., Tamai, K., and Togashi, K. 2007. Staging of carcinoma of the uterine cervix and endometrium. *Eur. Radiol*, 17:2009-2019.

52. Kulasingam, V., and Diamandis, E.P. 2008. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nature clinical practice oncology*, 5(10):588-599.

53. Kumar, B., Yadav, P.R., Goel, H.C, Rizvi, M.M.A. 2009. Recent Development in Cancer: Therapy by the use of Nanotechnology. *Digest Journal of Nanomaterials and Biostructures*, 4(1):1-12.

54. Kumar, M., and Sarin, S.K. 2009. Biomarkers of Disease in Medicine. *Current Trends in Science*: Platinum Jubilee Special, 403-417.

55. Laerum, O. D. 1997. Local spread of malignant neuroepithelial tumours. Acta Neurochir.

56. Larner, S. 2008. Biomarkers: The future of diagnosis and therapy for traumatic brain injury. Retrieved 2013, from International Brain Injury Association: internationalbrain.org/?q=node/77.

57. Lax, S. 2011. Histopathology of cervical precursor lesions and cancer. *Acta Dermatoven APA*, 20(3).

58. Likes, W., and Itano, J. 2003. Human papillomaviruses and cervical cancer; not just a sexually transmitted diseases. *Clinical journal of oncology Nursing*, USA, 7(3):271-306.

59. Ling, J., Wiederkehr, U., Cabiness, S., Shroyer, K.R., and Robinson, J.P. 2008. Application of Flow Cytometry for Biomarker-Based Cervical Cancer Cells Detection. *Diagnostic Cytopathology*, 36(2):76-84.

60. Liotta, L. A., and Kohn, E. C. 2001. The microenvironment of the tumour-host interface. *Nature*, 411(6835):375-9.

61. Litjens, J.N.T.M.R., Hopman, H.N.A., van de Vijver, K.K., Ramaekers, C.S.F., Kruitwagen, F.P.M.R., and Kruse, A.J. 2013. Molecular biomarkers in cervical cancer diagnosis: a critical appraisal. *Expert Opin. Med. Diagn*. 7(4):365-377.

62. Ma, Q., Wan, G., Wang, S., Yang, W., Zhang, J., and Yao, X. 2014. Serum microRNA-205 as a novel biomarker for cervical cancer patients. *Cancer Cell International*, 14(81):1-7.

63. Manne, U., Srivastava, R.G., Srivastava, S. 2005. Keynote review: Recent advances in biomarkers for cancer diagnosis and treatment.

64. Maree, J.E., and Moitse, K.A. 2014. Exploration of knowledge of cervical cancer and cervical cancer screening amongst HIV positive women. *Curationis*, 37(1), Art. 1209, 7 pages.

65. Maree, J.E.M., and Wright, S. C.D. 2010. How would early detection be possible? An enquiry into cancer related knowledge, understanding and health seeking behaviour of urban black women in Tshwane, South Africa. *European Journal of Oncology Nursing*, 14:190-196.

66. Martini, F.H., Bartholomew, E.F. 2007. The reproductive system, in essential of anatomy and physiology. *Pearson Education Inc*.: San Fransisco; 19: 612-643.

67. Mayeux, R. 2004. Biomarkers: Potential Uses and Limitations, NeuroRx: *The Journal of the American Society for Experimental NeuroTherapeutics.*

68. Mees, C., Nemunaitis, J., and Senzer, N. 2009. Transcription factors: their potential as targets for an individualized therapeutic approach to cancer. *Cancer Gene Ther*, 16:103-112.

69. Mishra, A., and Verma, M. 2010. Cancer Biomarkers: Are We Ready for the Prime Time? *Cancers*, 2:190-208.

70. Mojaki, M., Basu, D., Letskokgohka, M., and Govender, M. 2010. Referral steps in district health system are side-stepped. *South African Medical Journal*, 101(2):109.

71. Munoz, N., Franceschi, S., Bosetti, C., Moreno, V., Herrero, R., Smith, J.S., Shah, K.V., Meijer, C.J., and Bosch, F.X. 2002. Role of parity and human papillomavirus in cervical cancer: the IARC multicentric case-control study. *Lancet,* 359:1093-101.

72. Nishida, N., Yano, H., kamura, T., and Kojiro, M. 2006. Angiogenesis in Cancer. *Vascular Health and Risk Management*, 2(3):213-219.

73. Peters, J., and Loud, J., *et al*. 2001. Cancer genetics fundamentals. *Cancer Nurs* 24(6): 446-61.

74. Petersen, O. W., Gudjonsson, T., *et al*. 2003. Epithelial progenitor cell lines as models of normal breast morphogenesis and neoplasia. *Cell Prolif Suppl*, 1:33-44.

75. Pink book; Course textbook. The human papilloma virus. Chapter 10/2011. Consulted 3.11.2013 http://www.cdc.gov/vaccines/pubs/pinkbook/downloads/hpv.pdf.

76. Quintana, L.F., Campistol, J.M., Alcolea, M.P., Banon-Maneus, E., Sol-Gonzalez, A., and Cutillas, P.R. 2009. Application of label-free quantitative peptidomics for the identification of urinary biomarkers of kidney chronic allograft dysfunction. *Mol Cell Proteomics*, 8:1658-1673.

77. Rasty, G., Hauspy, J., and Bandarchi, B. 2009. Assessment of sentinel lymph node in cervical cancer: review of literature. *J Clin Pathol*, 62:1062-1065.

78. Raza, K. YEAR. Application of Data Mining in Bioinformatics. *Indian Journal of Computer Science and Engineering* 1(2):114-118.

79. Ross, M.H., and Pawlina, W. 2006. Female reproductive system: Chapter 23. In histology: a text and Atlas. Lippincott, Williams and Wilkins, Philadelphia. 773-833.

80. Sahab, Z.J., Semaan, S.M., and Sang, Q.A. 2007. Methodology and Applications of Disease Biomarker Identification in Human Serum. *Biomarker Insights*, 2:21-43.

81. Schiffman, M., Castle, P.E, Jeronimo, J., Rodriguez, A.C., and Wacholder, S. 2007. Human papillomavirus and cervical cancer. *Lancet*; 370:890-907.

82. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., and Huttenhower, C. 2011. Metagenomic biomarker discovery and explanation. *Genome Biology*, 12(60):1-18.

83. Sehgal, A., and Singh, V. 2009. Human papilloma virus infection (HPV) and screening strategies for cancer. *Indian J Med Res*, 130: 234-240.

84. Smith, N., Moodley, M and Hoffman, M. 2003. Challenges to cervical cancer screening in the Western Cape Province. *South African Medical Journal*, 93(1):32-35.

85. Souter, J. 2012. Human papillomavirus vaccine. *Prof Nurs Today*; 16(4):16-18.

86. SouthAfrica.Info, 2012, Health care in South Africa, viewed 11 February 2013, from http://www.southafrica.info/about/health/health.htm.

87. Southern Africa Litigation Centre (SALC), 2012. Tackling Cervical Cancer: Improving Access to Cervical Cancer Services for Women in Southern Africa. ISBN 978-0-620-53607-3.

88. Srinivas, P.R., Srivastava, S., Hanash, S., and Wright, G.L. 2001. Proteoics in early detection of cancer. *Journal of Clinical Chemistry*, 47:1901-1911.

89. Stevens, A., and Lowe, J. 2005. Chapter 17: Female reproductive system, in human histology. *Elsevier Mosby*: New York, 345-372.

90. Suh, K.S., Sarojini, S., Youssif, M., Nalley, K., Milinovikj, N., Elloumi, F., Russell, S., Pecora, A., Schecter, E., Goy, A. 2012. Tissue Banking, Bioinformatics, and Electronic Medical Records: The Front-End Requirements for Personalized Medicine. *Journal of Oncology*, 1-12.

91. Tambor, V., Fučíková1, A., Lenčo1, J., kacerovský, M., Řeháček, V., Stulík, J., and Pudil, A. 2010. Application of Proteomics in Biomarker Discovery: a Primer for the Clinician. *Physiol. Res*, 59:471-497.

92. Tantipaiboonwong, P., Sinchaikul. S., Sriyam, S., Phutrakul, S., Tein-Chen, S.T. 2005. Different Techniques for Urinary Protein Analysis of Normal and Lung Cancer Patients. *Proteomics*, 5:1140-1149.

93. Tiffin, N., Kelso, J.F., Powell, A.R., Pan, H., Bajic, V.B., and Winston, A.H. 2005. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids*, 33(5):1543-1552.

94. Ueda, Y., Enomoto, T., Kimura, T., Miyatake, T., Yoshino, K., Fujita, M., and Kimura, T. 2010. Serum Biomarkers for Early Detection of Gynecologic Cancers. 2, 1312-1327; doi: 10.3390/cancers2021312.

95. Walker, A.R.P., Michelow, P.M., and Walker, B.F. 2002. Cervix cancer in African women in Durban, South Africa. *International Journal of Gynaecology and Obstetrics*, 79:45-46.

96. Wang, J., Gao, F., Mo, F., *et a*l. 2009. Identification of CHI3L1 and MASP2 as a biomarker pair for liver cancer through integrative secretome and transcriptome analysis. *Proteomics Clinical Applications* 20; 3(5):541‐551.

97. Weinberg, R.A. 1995. The retinoblastoma protein and cell cycle control. *Cell*, 81:323-330.

98. Wentzensen, N., and Von Knebel-Doeberitz, M. 2007. Biomarkers in cervical cancer screening. *Disease Markers*, 23:315-30

99. Wong, R.S.Y. 2011. Apoptosis in cancer: from pathogenesis to treatment. Journal of *Experimental & Clinical Cancer Research*, 30(87): 1-14

100. Wu, D., Rice, C.M., and Wang, X. 2012. Cancer bioinformatics: A new approach to systems clinical medicine. *BMC Bioinformatics*, 13(7)1:1-4.

101. Zarchi, M.K., Behtash, N., Chiti, Z., Kargar, S. 2009. Cervical Cancer and HPV Vaccines in Developing Countries. *Asian Pacific J Cancer Prev*, 10:969-974.

102. Zeimet, A. G., Riha, K. *et al*., 2000. New insights into p53 regulation and gene therapy for cancer. *Biochem Pharmacol*, 60(8):1153-63.

103. Zhang, X., Shi, L., Chen, G., and Yap, Y.L. 2011. Integrative Omics Technologies in Cancer Biomarker Discovery. *Landes Bioscience*, 129-137.

CHAPTER 2: Identification of Biomarkers Using an *in silico* Approach

2.1. Background

Advancement in DNA microarray technologies has made it possible for the differential expression levels of up to tens of thousands of genes to be effectively and efficiently measured under various conditions simultaneously. The information gained from these experiments is successfully implemented in gene function prediction, drug development, disease diagnosis and patient survival analysis (Raza and Parveen, 2012). Priority for the discovery of candidate biomarkers has always been granted to differential quantitative proteomic studies; however, candidate biomarkers can also be selected by *in silico* analysis starting with various publicly available databases. Through this route, the search can be established for candidate biomarkers that result from experiments and/or knowledge in scientific literature and/or the public domain (i.e. websites) (Rodríguez-Pérez *et al*., 2008).

2.1.1. Data Mining

Data mining (DM) is extracting or "mining" knowledge from large amounts of data. It is the science of discovering new interesting patterns and relationships in huge amounts of data. Data mining can defined as the process of discovering significant new correlations, patterns and trends by exploring large amounts of data stored in warehouses (databases). Thus Data Mining is also referred to as Knowledge Discovery in Databases (KDD). Mining of biological data assists in extracting useful knowledge and trends from massive datasets gathered in biology and in other life science areas such as medicine (Raza, 2010). There are numerous applications of data mining in bioinformatics such as gene discovery, protein function domain determination, functional motif detection, diagnosis of diseases, protein function inference, disease prognosis, optimization of disease treatment, reconstruction of protein and gene interaction networks, protein sub-cellular location prediction and data cleansing (Raza, 2010).

2.1.2. Microarray Data Mining

Microarray data mining is the application of Bioinformatic approaches in microarray data analysis in order to discover biological entities and pathways that define a phenotype such as human diseases. Microarray data mining has proved to be a productive approach to discover target genes associated with human diseases and has been increasingly used to detect diagnostic or prognostic marker genes. Supervised classification and unsupervised clustering are the two basic approaches broadly applied in microarray data mining. In the latter approach a group of genes that share coherent expression across a subset of conditions is determined using clustering methods such as hierarchical clustering, principal component analysis (PCA) and self-organising maps (SOM) (Kapushesky *et al*., 2011). A supervised analysis approach searches for genes that can distinguish between known samples and conditions. In a typical example of supervised analysis, the global gene expression profiles of disease tissues or fluids will be compared to those in normal tissues or fluids (e.g. cancer vs. healthy tissues/fluids) from which a list of target genes or biological pathways that are important in a particular disease will be identified (Kapushesky *et al*., 2011).

2.1.3. Gene Expression Profiling using Microarrays

Various techniques have been developed for using microarray gene expression data to study several aspects of cancer biology, with an accumulation of microarray data suitable for cancer target discovery. Many gene expression studies on numerous types of cancers are connected with entire datasets that can be downloaded from public repositories specifically dedicated to the dissemination of this valuable data. These include websites such as the Stanford Microarray Database (http://genome-www5.stanford.edu/MicroArray/SMD), the NCBI's Gene Expression Omnibus repository (http://www.ncbi.nlm.nih.gov/geo), the EBI's ArrayExpress (http://www.ebi.ac.uk/arrayexpress) and the MIT Cancer Genomics Program (http://www.broad.mit.edu/cancer/), making these databases valuable resources for target gene discovery (Desany and Zhang, 2004). Interesting targets are not solely defined by their expression patterns, other criteria such as type of molecule (e.g. kinase), subcellular localization (e.g. cell surface) and biological pathway (e.g. angiogenesis) are of importance in the decision to follow-up on a potential new target. To this end, resources such as the Gene Ontology Project (http://www.geneontology.org), the Kyoto Encyclopaedia of Genes and

Genomes Pathways Project (http://www.genome.ad.jp/kegg) attempt to place genes in the context of biological function, location and pathway (Desany and Zhang, 2004).

2.1.4. Digital Expression Profiling using EST and SAGE

Gene expression profiling is not only applicable with microarrays, however, digital expression based on either expressed sequence tags (ESTs) or serial analysis of gene expression (SAGE) is also complementary to microarrays and can be just as powerful. Both EST-derived expression and SAGE are centred on the principle that the frequency of sequence tags sampled from a pool of cDNAs is directly proportional to the expression level of the corresponding gene (Desany and Zhang, 2004). There are three key advantages of EST and SAGE over microarrays; firstly, the simple digital data format in sequence clone counts and frequencies enables direct and platform-independent data comparison among different data collections from multiple tissues. Secondly, since there is no need for designing any DNA chips, no prior knowledge of gene sequence is required and therefore many novel genes not covered by microarrays are represented (Desany and Zhang, 2004).

Lastly, since the expression levels are represented by mRNA abundance relative to all transcripts and it is thus independent of probe selection and hybridization biases. Through these advantages digital expression becomes a more quantitative measurement of gene expression than microarrays (Desany and Zhang, 2004). A significant fraction of these ESTs are derived from cancer tissues as a result of the large-scale efforts of the Cancer Genome Anatomy Project (CGAP; http://www.ncbi.nlm.nih.gov/ncicgap) at the National Cancer Institute (NCI) to generate EST libraries from tumour samples. ESTs present an attractive resource for differential expression analysis between normal and cancer tissues (Desany and Zhang, 2004).

2.2. Biological Databases

In the past decade, genome-wide gene expression assays, the majority using microarrays and more recently high throughput sequencing have become common tools in biomedical and biological research. Most assays are performed to answer specific questions, for example to determine which genes are differentially expressed in a particular disease state in comparison to a healthy condition in a tissue or cell type (Kapushesky *et al*., 2011). The accessibility of biological data is of paramount significance for bioinformatics applications and fortunately there are innumerable biological databases that gather data and organize it in such a way that their content is easily accessible. Biological databases are grouped into three categories depending on the type of stored data: (i) primary databases, which contain DNA and protein sequences, (ii) secondary databases, derive their information from a primary database and (iii) composite database, combine numerous sources from primary databases (Kapushesky *et al*., 2011).

This section of the study aimed at identifying proteins/genes implicated in cervical cancer, by extracting gene lists from various databases namely Oncomine, Gene Expression Atlas, and TiGER to name a few. The gene lists will be functionally characterised by assigning GO terms suitable for gene products that could be detected in bodily fluids. It is objectively targeted at prioritising these genes through literature mining using databases such as PubMed, iHOP and Google Scholar etc. It not feasible to cover all the available biological databases due to their high number, however, major databases of interest will be covered in this thesis.

2.2.1. Gene and Gene Expression Databases

2.2.1.1. Gene Expression Atlas

Gene expression atlas (GEA) (http://www.ebi.ac.uk/gxa/) is a database launched by the European Bioinformatics Institute (EBI). This database allows users to query gene expression under various biological conditions, including different cell types, developmental stages, physiological states, phenotypes and disease states. The database contains information about more than 200 000 genes from nine species and almost 4500 biological conditions studied in over 30 000 assays from over 1000 independent studies. GEA can help investigators determine which conditions or where in the organism is a gene of interest differentially expressed and which genes are differentially expressed in a condition or site for example in a disease or in an organ (Kapushesky *et al*., 2011).

2.2.1.2. Oncomine

Oncomine (http://www.oncomine.org) is a cancer microarray database and web-based data-mining platform aimed at facilitating discovery from genome-wide expression analyses. Oncomine contains 65 gene expression datasets comprising nearly 48 million gene expression measurements from over 4700 microarray experiments. Differential expression analyses comparing most major types of cancer with their respective normal tissues as well as a variety of cancer subtypes and clinical-based and pathology-based analyses are available for exploration. Data can be queried and visualized for a selected gene across all analyses or for multiple genes in a selected analysis (Rhodes *et al*., 2004). Oncomine is designed from a collection of microarray studies focusing on published literature using sophisticated data normalization and statistical methods to enable comparison of results across multiple platforms and experiments (Rhodes *et al*., 2007). This database also provides tools that enable organisation and visualisation to prioritise and identify certain patterns from the processed and integrated experimental information collected on literature that has been published. The objective of Oncomine is to provide the public with micro-array data that has been identified from literature studies that have composed data of gene expression patterns of different disease states, cancer stages and experimental populations and conditions with statistical support for all genes in the experiment (Rhodes *et al*., 2004).

Oncomine to date has accumulated over 18 000 cancer gene expression experiments and automated analysis has identified the genes, pathways, regulatory networks and functional networks that are activated and/ or repressed in human cancers (Rhodes *et al*., 2007). Oncomine produces different types of outputs, i.e. a search using a gene can produce different experiments, and one can search by cancer types and disease property. Amongst other types of outputs Oncomine also produces list of authors that have done similar studies confirming the users query (Rhodes *et al*., 2004).

2.2.1.3. Integrative Oncogenomics

The Integrative Oncogenomics (intOGen) database (http://www.intogen.org/) is a cancer analysis tool database designed to facilitate the integration, analysis, exploration and interpretation of oncogenomic data for the identification of genes and groups of genes involved in cancer development. This database aims at facilitating the detection of the most recurrent alterations that drive tumourigenesis. It collates, annotates and analyses high-throughput data regarding transcriptional, genomic and mutational changes taking place in tumours from different studies annotated with specific cancer types. Currently, intOGen contains 118 studies for mRNA expression profiling and 188 studies for genomic alterations, covering in total 64 different tumor topographies (Perez-Llamasy *et al*., 2011).

2.2.1.4. Cancer Genome Anatomy Project

The Cancer Genome Anatomy Project (CGAP) (http://cgap.nci.nih.gov/) of the National Cancer Institute (NCI) is an attractive starting point for cancer-specific gene discovery. CGAP is a collaborative network of cancer researchers with a common goal to decipher the genetic changes that occur during cancer formation and progression. This database sought to determine the gene expression profiles of normal, precancer and cancer cells, which ultimately leads to improved detection, diagnosis and treatment of patients. This database consists of expression information (mRNA) of thousands of known and novel genes in diverse normal and tumour tissues. By monitoring the electronic expression profile of many of these sequences making it possible to compile a list of genes that are selectively expressed in the cancers (Strausberg, *et al*., 1997).

2.2.1.5. C-It

C-It (http://c-It.mpi-bn.mpg.de) is a knowledge database focusing on uncharacterized genes to build a starting point for biologists to study genes with unknown functions. The database implements literature information from the PubMed database to identify genes that lack publication records. Based on the assumption that genes are likely to fulfil important functions when their expression is enriched in a certain tissue, C-It uses the tissue expression information of UniGene, ESTs profiles to identify tissue-enriched genes (Gellert *et al*., 2010).

C-It combines microarray and SAGE data to give users integrated access to comprehensive transcriptional profiles. The database is designed to include additional expression studies, which might provide more comprehensive coverage of gene expression patterns and tissue-enriched splicing isoforms. C-It is thus an excellent starting point to study uncharacterized genes (Gellert *et al*., 2010).

2.2.1.6. Tissue-specific Gene Expression and Regulation

TiGER (Tissue-specific Gene Expression and Regulation) is a web database that gives comprehensive information of human gene specificity using three types of data: the gene expression profile (EST), combinational gene regulation (based on transcription factor binding sites) and a cis-regulatory module (CRM) (http://bioinfo.wilmer.jhu.edu/tiger/). The database is also a good example for comparison of EST data and Microarray data. 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). For comparison of this research predicted result, only the record from gene expression profile will be used.

2.2.1.7. VeryGene

VeryGene (http://www.verygene.com/) is a web-accessible database for the annotation of human tissue-specific genes, with a primary focus on integration with disease association and drug targets. A significant effort was made to integrate tissue-specific genes (TSGs) from two large-scale data analyses with information on subcellular localization, Gene Ontology, Reactome terms, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, Mouse Genome Informatics (MGI), Mammalian Phenotype, disease association, and drug targeting (Yang *et al*., 2011). To date there are 3960 annotated TSGs derived from 127 normal human tissues and cell types, including 5672 gene-disease and 2171 drug-target relationships. This database can be used as a discovery tool by generating novel inferences and a potentially useful resource for many applications, for instance, screening for therapeutic targets or biomarkers by tissue, subcellular localization, or gene-drug relationship or looking for functional enrichment of similarly localized genes or genes participating in a common pathway/disease or *vice versa* (Yang *et al*., 2011).

2.2.1.8. Gene Expression Barcode 2.0

The Gene Expression Barcode 2.0 (GEB) is the first database to provide reliable absolute measures of expression for most annotated genes for 131 human and 89 mouse tissue types, including diseased tissue (http://barcode.luhs.org/). This is made possible by a novel algorithm that leverage information from Gene Expression Omnibus (GEO) and ArrayExpress public repositories to build statistical models that permit converting data from a single microarray into expressed/unexpressed calls for each gene (McCall *et al*., 2011). The database can create a gene expression barcode for a single microarray that provides information about the expression states of all genes. GEB has combined thousands of gene expression barcodes to create vast catalogs of transcriptome information spanning hundreds of cell types and tens of thousands of genes. These catalogs are easily accessible via a series of web tools that allow an investigator to readily access gene and/or cell type specific information (McCall *et al*., 2011).

2.2.1.10. Database for Annotation, Visualization and Integrated Discovery

DAVID (Database for Annotation, Visualization and Integrated Discovery) is a database that has numerous features (http://david.abcc.ncifcrf.gov/). DAVID is a publicly available 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 likelihood that the researcher will identify the biological process most pertinent to the biological phenomena under study (Khatri and Draghici, 2008). The annotation tool in DAVID provides several gene annotation options including; GenBank, Unigene, LocusLink, RefSeq, Gene Symbol, Gene Name, OMIM, Affymetrix description, Summary and Gene Ontology. Each of these tools can be used for various reasons or functions, for this study the Gene Ontology (GO) annotation tool will be of interest since Gene Ontology is a controlled vocabulary that is applied to the functions of genes and proteins. The functional classifications that are used in DAVID are those included in the Locus Report provided by NCBI (Dennis *et al*., 2003). This feature will categorize the genes that have been identified into three categories namely, Biological Processes, Molecular Function and Cellular Components. The cellular component category is of particular importance to this study as the objective is to identify genes that are expressed on the cell surface of the cervix as this will indicate their shedding into biological fluid for diagnostic purposes (Huang *et a*l., 2008).

2.2.1.11. Human Protein Atlas

The Human Protein Atlas (HPA) (http://www.proteinatlas.org/) uses high-resolution images to show protein expression profiles in 46 normal tissues, 20 cancer types, and 47 cell lines for the human species. The gene-centric manner of HPA enables the comparison of proteomic data (antibody) and genomic (microarray) data. The expression intensity is marked as “level of antibody staining” with Strong, Moderate, Weak and Negative levels. Only genes marked as “Strong” will be considered as specific/selective for the purpose of this project and only 46 normal human tissues will be used for comparison (Uhlen *et al*., 2010).

2.2.1.12. Cervical Cancer Gene Database

The Cervical Cancer gene Database (CCDB, http://crdd.osdd.net/raghava/ccdb) is a manually curated catalog of experimentally validated genes that are thought, or are known to be involved in the different stages of cervical carcinogenesis. The database have compiled 537 genes that are linked with cervical cancer causation processes such as methylation, gene amplification, mutation, polymorphism and change in expression levels, as evident from published literature. Each record contains details related to gene like architecture (exon–intron structure), location, function, sequences (mRNA/CDS/protein), ontology, interacting partners, and homology to other eukaryotic genomes, structure and links to other public databases, thus augmenting CCDB with external data (Agarwal *et al*., 2011). Also, manually curated literature references have been provided to support the inclusion of the gene in the database and establish its association with cervical cancer. In addition, CCDB provides information on microRNA altered in cervical cancer as well as a search facility for querying-several browse options and an online tool for sequence similarity searches, thereby providing researchers with easy access to the latest information on genes involved in cancer of the cervix (Agarwal *et al*., 2011).

2.3. Text Mining

Text mining (TM) is the computational discovery of new, previously unknown information by automatically extracting information from different written sources. There are two major steps involved in text mining, information retrieval (IR) and information extraction (IE). Information retrieval finds literature or abstracts associated to a specific topic with the help of general search engines or specifically designed IR searching tools such as google scholar, GoPubMed, iHOP, PolySearch and GeneWays just to mention a few (Yang *et al.,* 2009). There are two search methods in IR: rule-based or knowledge based and statistical or machine learning. The first approach uses patterns that rely on basic biological insights, for instance <cervical> and <cell surface>, to find literature or abstracts of interest on genes implicated in cervical cancer and simultaneously found on the surface of the cell. The second approach utilises synthetic parse trees (which can also be rule-based) or classifiers to classify the related biomedical literature (Yang *et al.,* 2009). A prerequisite for IE is named entity recognition (NER), which relies on tools or methods for automatic term recognition in order to extract entities such as genes, proteins, drugs or other molecules. Text mining has been broadly applied to identify disease-associated entities (genes/proteins) and to understand their roles in diseases. The goal of text mining is to filter knowledge and present information to users in a concise and an understandable format (Faro *et a*l., 2012).

2.3.1. Text Mining Databases

2.3.1.1. Universal Protein Knowledgebase

Universal Protein Knowledgebase (UniProt) (http://www.uniprot.org) is a database formed by multiple sources such as the Swiss-Prot, TrEMBL and PIR protein database activities, in order to provide the scientific community with a single, centralized, authoritative resource for protein sequences and functional information. The database provides a comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and query interfaces (Apweiler *et al*., 2004).

2.3.1.2. Human Gene Navigator

Human Gene Navigator (HuGENavigator) (http://www.hugenavigator.net/) is a searchable and a continuously updated knowledge base (KB) in human genome epidemiology, including information on population prevalence of genetic variants, gene-disease associations, gene-gene and gene- environment interactions, and evaluation of genetic tests (Yu *et al*., 2008).

2.3.1.3. GoPubmed

GoPubMed (www.gopubmed.org) is a web server knowledge-based search engine for biomedical texts. The database allows users to explore PubMed search results with the GO annotation approach, a hierarchically structured vocabulary for molecular biology. GoPubMed provides the following benefits: firstly, it gives an overview of the literature using abstracts by categorizing these abstracts according to their GO term annotations and thus allowing users to quickly navigate through the abstracts by category (Doms and Schroeder, 2005). Secondly, it automatically shows general ontology terms related to the original query, which often do not even appear directly in the abstract. Thirdly, it enables users to verify its classification since the GO terms are highlighted in the abstracts and as each term is labelled with an accuracy percentage. Lastly, exploring PubMed abstracts with GoPubMed is useful as it shows definitions of GO terms without the need for further reading of additional literature (Doms and Schroeder, 2005).

2.3.1.4. PolySearch

PolySearch (http://wishart.biology.ualberta.ca/polysearch) is a web accessible tool that is designed specifically for extracting and analyzing text-derived relationships between human diseases, genes/proteins, mutations, drugs, metabolites, pathways, tissues, organs and sub-cellular localizations. It also displays links and ranks text, as well as sequence data in multiple forms and formats (Cheng *et al*., 2008). A feature that distinguishes PolySearch from other biomedical text mining tools is the fact that it extracts and analyses not only PubMed data, but also text data from multiple databases (DrugBank, SwissProt HGMD, Entrez SNP, etc.). This integration of current literature text and database ‘factoids’ allows

PolySearch to extract and rank information that is not easily found in databases or in journals alone (Cheng *et al*., 2008). PolySearch supports >50 different classes of queries against nearly a dozen different types of text, scientific abstract or bioinformatic databases. PolySearch also exploits a variety of techniques in text mining and information retrieval to identify, highlight and rank informative abstracts, paragraphs or sentences. This database consists of seven basic components: (i) a web-based user interface for constructing queries; (ii) a collection of internal and external biomedical databases; (iii) a collection of biomedical synonyms (custom thesauruses and all entity lists); (iv) a general text search engine for extracting data from heterogeneous databases; (v) a schema for selecting, ranking and integrating content; (vi) a display tool for displaying and synopsizing results and (vii) a PCR primer-designing tool to facilitate SNP and mutation studies (Cheng *et al*., 2008).

2.3.1.5. Information Hyperlinked over Proteins

Information Hyperlinked over Proteins (iHOP) (http://www.ihop-net.org/) is an online text-mining service that provides a gene-guided network to access PubMed abstracts. The concept underlying iHOP is that by using genes and proteins as hyperlinks between sentences and abstracts, the information in PubMed can be converted into one navigable resource (Hoffmann and Valencia, 2004). Navigating across interrelated sentences within this network rather than the use of conventional keyword searches allows for stepwise and controlled acquisition of information. Moreover, this literature network can be superimposed onto experimental interaction data to facilitate the simultaneous analysis of novel and existing knowledge. The network presented in iHOP contains 28.4 million sentences and 110,000 genes from over 2,700 organisms, including the model organisms *Homo sapiens*, and many more (Hoffmann and Valencia, 2004).

Aims of this chapter

1) Explore several databases (as outlined above) for the extraction of potential biomarkers for the early diagnosis of cervical cancer using in silico methods

2) Refinement of the compiled gene list obtained from method 1 using literature mining tools.

2.4. Materials and Methods

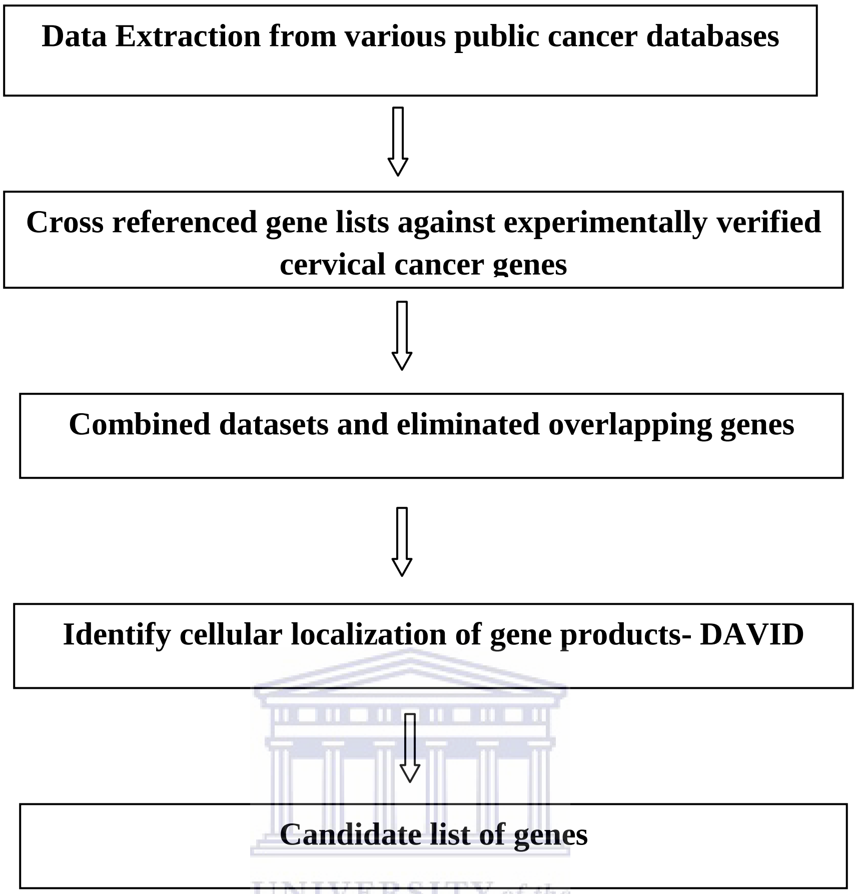

Data Extraction from various public cancer databases
Cross referenced gene lists against experimentally verified cervical cancer genes
Combined datasets and eliminated overlapping genes
Identify cellular localization of gene products- DAVID
Candidate list of genes

Figure 2.1: Outline of the *in silico* methodology for biomarker discovery

2.4.1. Extraction of Candidate Biomarkers

The focus of this research analysis approach was to retrieve and filter genes differentially expressed in cervical cancer compared to normal tissues, to a manageable gene list. A bioinformatics pipeline was used to interrogate different databases and in this study eight databases were mined to identify proteins highly specific to or strongly expressed in the cervix tissue and genes differentially expressed in cervical cancer. The pipeline was divided into three sections: i) Data-mining of publicly available databases, ii) Literature- mining, and iii) Gene enrichment analysis (Figure 2.1).

2.4.1.1. Oncomine Database

The first set of genes was searched on Oncomine with the following input query: 1. *Analysis type*: Cancer vs. normal, cervical cancer vs normal, differential analysis and outlier analysis. 2. *Cancer type*: Cervical cancer. 3. *Data type*: mRNA and 4. *Pathology subtyp*e: Stage and grade type. The second set of genes was searched with the above mentioned criteria, however, with the addition of Human Papillomavirus (HPV) infection status selected under molecular subtype.

2.4.1.2. Gene Expression Atlas Database

The GEA database was searched for genes differentially expressed in cervical cancer using the following parameters, <all genes >< up/down in ><homosapiens>< cervical cancer. The following GO terms were also used as a search criteria, cell surface, membrane, integral to membrane and plasma membrane to filter the up/down regulated genes identified.

2.4.1.3. intOGen Database

The Integrative Oncogenomics database was queried using the search criteria, all experiments were selected in the browser. The tumor type cervix uteri C53 (Adenocarcinoma and squamous cell carcinoma) was chosen and genes/modules was left as default “all”.

2.4.1.4. CGAP Database

The Cancer Genome Characterization Initiative database was searched using the gene finder function, using the default setting: tissue, function, location or keyword. The cervix tissue type was selected as *Homo sapiens*.

2.4.1.5. C-It Database

The database was searched for proteins enriched in cervix tissue (human data only). Literature information search parameters of fewer than five publications in PubMed and fewer than three publications with the Medical Subject Headings (MeSH) term of the searched tissue were used.

2.4.1.6. TIGER Database

The Tissue-specific Gene Expression and Regulation (TiGER) database was searched for proteins preferentially expressed on the cervix tissue based on ESTs by using “Tissue View”.

2.4.1.7. VeryGene Database

The database was searched for the cervix tissue using “Tissue View” for tissue-selective proteins.

2.4.1.8. Gene Expression Barcode 2.0 Database

Gene expression barcode was searched for genes expressed in >95% samples of cervical cancer tissue by selecting gene expression. The consensus input type was chosen, using the affymetrix human genome U 133A (HGU 133a).

2.4.2. Analysis of Gene Lists

A command line script using Ubuntu software was used to eliminate duplicates per list of genes from each of the databases and to combine the list of genes together using the following command line (cat FILE NAMES |sort -u > OUTPUT FILE NAME) in Ubuntu. A total of 27 datasets were combined and thereafter, the genes were submitted to DAVID for the same enrichment analyses.

2.4.3. Functional Characterisation using DAVID

The candidate gene lists was then submitted to The Database for Visualization and Integrated Discovery (DAVID) Version 6.7 for gene enrichment through the following steps.

Step 1: **Gene List Submission:** start analysis was selected and a gene list was uploaded. Gene identifiers were selected as "official_gene_symbol". In the list type "gene list" was chosen and the gene list was submitted. *Homo sapiens* were the selected as the organism from which the protein products were derived.

Step 2: A**nalyze gene list using DAVID tool**s: Functional annotation clustering was selected from the functional annotation tools. The classification stringency was set to high or left at default "medium". On the "options" setting the following were selected, display, fold change and Bonferroni analysis. Using the same setting "option" re-run was chosen.

Step 3: Choose annotation clusters: different clusters were searched by identifying the following GO terms: cell surface, secreted, secretory granules, extracellular matrix, extracellular space and extracellular membrane. The newly derived list of genes were exported and saved. Gene Ontology was selected and the cellular component with the highest percentage (100%) was chosen. The classification data was displayed with a count of the number of genes annotated to be "cell surface, secretory granules, extracellular matrix and extracellular space" as the output. When selected, the genes were displayed with gene identifier, gene name and class of species.

2.4.4. Comparison to Reference Lists

After gene enrichment in DAVID, the list of genes were cross-referenced against the genes from HPA and CCDB, to ascertain the ability of Bioinformatics to correctly identify genes that were implicated in cancer through experimental studies and to gauge its reproducibility/variability. CCDB was mined for the extraction of a reference gene list containing documented experimentally verified cervical cancer genes. Genes were individually entered into HPA to search for a moderate to strong association with cervical cancer expression. Genes were also cross-referenced with each other to identify overlapping genes. No genes were eliminated at this point, but all candidate genes were further subjected to literature mining to ascertain if genes were already experimentally verified as cervical cancer genes, despite not being found in the reference gene lists.

2.4.5. Literature Review of the Candidate Entities

A text mining approach was used to search each gene from the output obtained from DAVID against literature. Uniprot, PolySearch, Google Scholar, HuGENavigator, GoPubmed and iHOP were used to search for abstracts or journal articles using the "gene symbols" the Boolean term "AND" and terms that imply neoplastic cervical cancer tissue e.g. <cervical cancer> AND <gene name>. A search was then done through the relevant literature for any information or data that links the gene as a biomarker for cervical cancer. All genes found to have been validated or inferred as biomarkers for cervical cancer were recorded and these genes will be eliminated from further analysis. Subsequent to these mining approaches, a final list of putative genes was compiled and the genes were subjected to tissue specificity analysis.

2.5. Results and Discussion

2.5.1. Identification of Eligible Cancer Biomarkers

The approach presented here was designed to exploit several cancer databases to identify genes encoding proteins with differential expression that could be secreted into bodily fluids and subsequently be used as potential biomarkers for the early diagnosis of cervical cancer. The methods that were used are complementary to the extent that they query fundamentally different aspects of biological knowledge stored within these databases. A series of data mining steps was used to increase the stringency such that the huge number of entities (genes/proteins) present in various databases and in literature was reduced to a manageable size. At each step, the criteria, choice of tool, and databases were selected to reduce the list of identified hits. From the Oncomine platform, data mining of 5 microarray datasets: (i) Bachtiary cervix, (ii) Biewenga cervix, (iii) Bittner cervix, (iii) Pyeon Multi-Cancer and (iv) Scotto cervix for genes differentially expressed in cervical cancer compared with their expression in normal tissues led to the identification of a list of 16023 differentially expressed gene profiles. The output was five seed lists which varied from 1%, 5% to 10% fold expression derived from each dataset. This means that from the five datasets, a total of 15 seed lists were presented as an output.

The second set of genes consisted of three datasets, Bittner Cervix, Scotto Cervix and Pyeon Multi-cancer with a total of 49674 genes as outlined in section 2.4.1.1. The results was 3 seed lists which varied from 1%, 5% to 10% fold expression derived from each dataset. This means that from the three datasets, a total of 9 seed lists were presented as an output. Each dataset is titled according to the first author, which is used as a classifier. The selected datasets were categorised based on the different folds/levels of gene expression as compared to the normal as per condition of each study. Table 2.1 depicts a summary of all genes that were extracted from all datasets presented as percentage fold (1%, 5% and 10%) irrespective from the study it was obtained. Each dataset represent a study that has been conducted, under certain experimental conditions and each gene list extracted from each dataset represents the outcome of that study. It was therefore pivotal that the sampling strategies and conditions at which these studies were conducted were well understood. All the seed lists in Oncomine produced 24185 genes, after combining all the seed lists and eliminating redundant genes, a total of 16023 genes remained.

Table 2.1: Summary of the genes extracted from Oncomine

Fold Expression %	No. of genes Extracted
1%	10483
5%	53548
10%	99316

Conversely GEA a general microarray database generated an output based on all data available to support the query as outlined in section 2.4.1.2. This database can be queried for datasets from functional genomic experiments using the meta-analysis of microarray and high throughput sequencing data. Two sets of gene lists were extracted from GEA, first dataset consisting of all genes up/downregulated in cervical cancer and the second set included genes queried with various GO terms such as cell surface etc. (refer to section 2.4.1.2). This was done to increase the sample size and also include genes that may not appear in the background. The first set produced 13431 genes using the query term up/downregulated in cervical cancer, and set 2 resulted in 11996 genes using GO terms with a summary of the number of genes extracted from the various GO terms shown in table 2.2. In summary the total number of genes mined from GEA were 6696 following curation.

Table 2.2: Summary of gene set 2 mined from GEA based on GO terms

GO Terms	No. of genes extracted
Cell surface	2288
Membrane	5116
Integral to membrane	2288

Plasma membrane	2304

For clarification the set of genes extrapolated from GEA were not particularly from one study but from multiple experiments combined in a global repository. Thus, one gene list from GEA is extrapolated from multiple experiments. IntOGen allows for the extraction of genes involved in expression changes and copy number variations across multiple tumours and cancer types. This database generated 22686 genes. However, the actual number of genes extracted using this database was 836 genes after duplicates were removed. VeryGene, a curated database containing tissue specific enriched genes identified 20 tissue-selective proteins. C-It focuses on tissue-enriched gene variants and differentially expressed genes that are still uncharacterised and identified 1075 tissue-enriched proteins after filtering of genes according to the set parameters and the TiGER database identified 209 proteins preferentially expressed in tissue. The gene expression barcode database produced 836 genes found in cervical cancer and the CGAP database identified 8216 genes after removal of common genes. A summary of all the genes extracted from the various databases as well as the number of common genes that were removed in each database are indicated in table 2.3 with the total number of genes identified.

Table 2.3: Total number of genes identified from mining gene and protein databases

Databases	**No. of genes identified**	**No. of genes duplicated**	**No. of unique genes**
Oncomine	24185	8162	16023
Gene Expression Atlas	18458	11762	6696
intOGen	22686	21850	836
Gene Expression barcode	1099	263	836
CGAP	8337	121	8216
C-It	1601	526	1075
TIGER	281	72	209
VeryGene	20	0	20

Total number of combined genes	**Total number of eliminated genes**	**Final number of genes after curation**
34033	5398	28190

2.5.2. Gene Enrichment Analysis

A total of 28190 genes were uploaded to DAVID for enrichment bioinformatics analysis and the output was 113 genes. The enrichment analyses of GO terms including biological process, cellular component and molecular function were performed on the 113 genes by using the functional clustering annotation tools as highlighted by figure 2.2, 2.3 and 2.4 respectively. The default options with medium/high classification stringency were used, and finally cluster names were extracted from the most biologically relevant GO term assigned to that cluster.

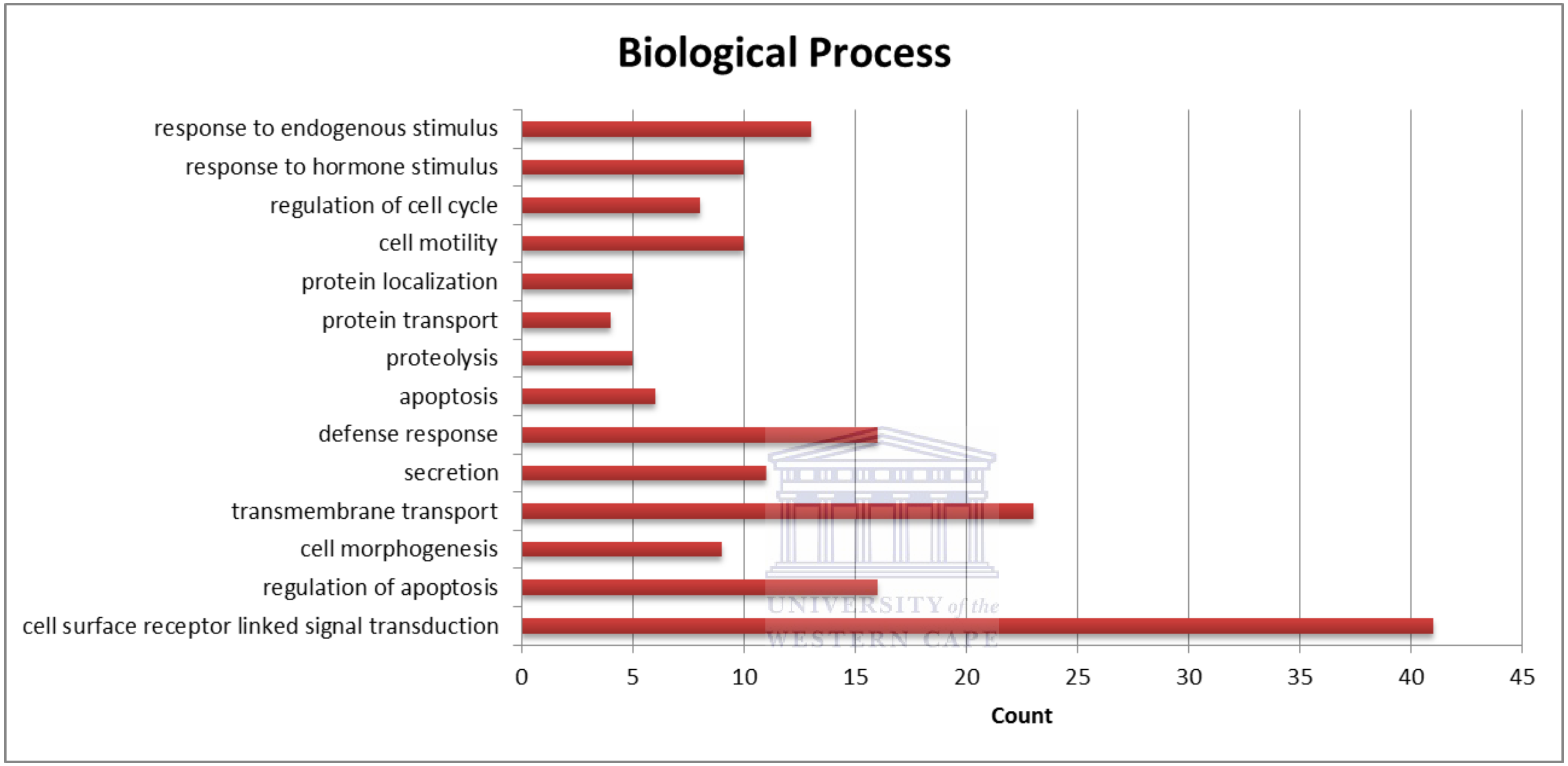


Figure 2.2: Functional Characterisation of genes in DAVID based on their biological process using GO analysis.

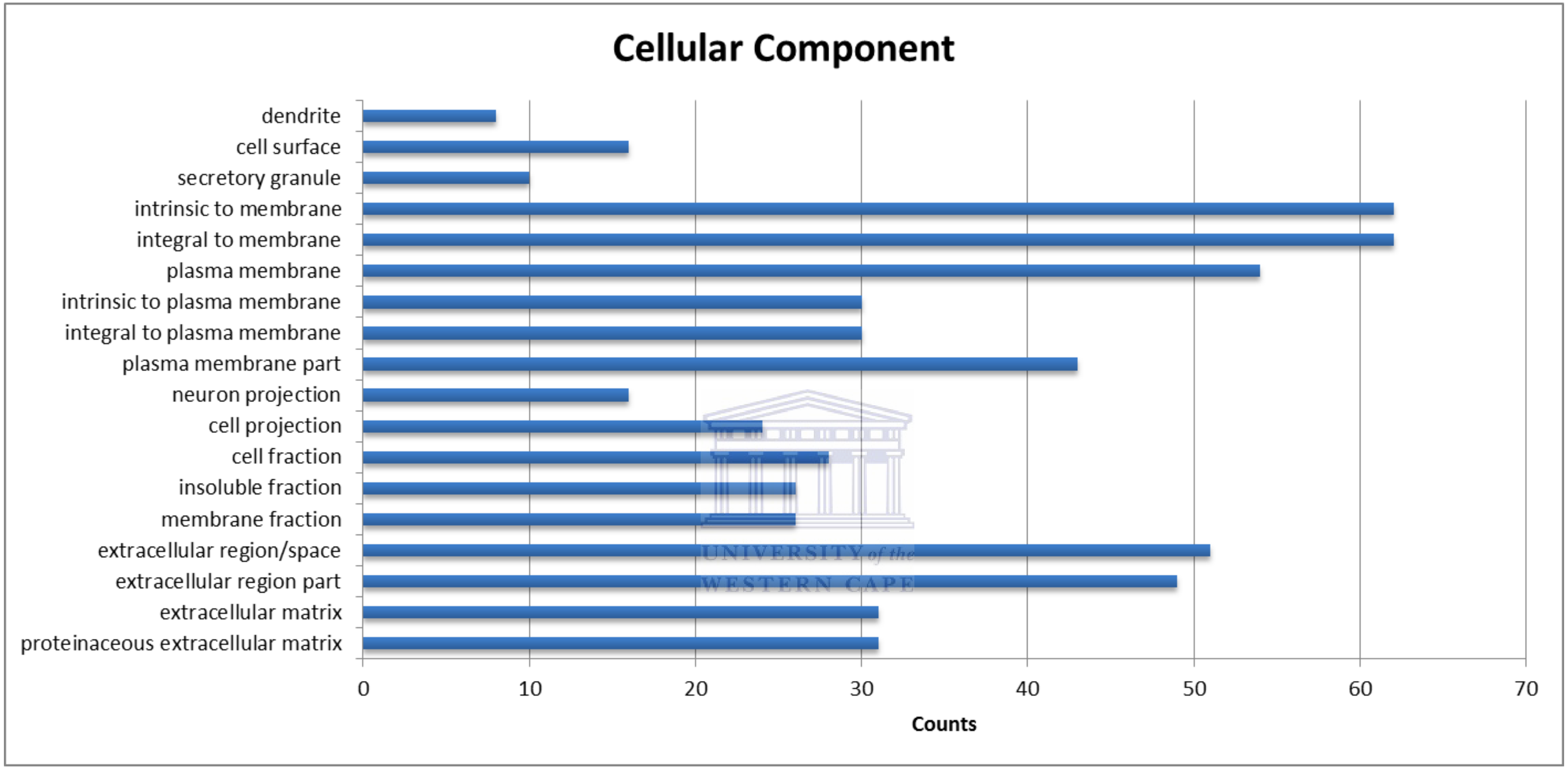


Figure 2.3: Functional Characterisation of genes in DAVID based on their cellular component using GO analysis.

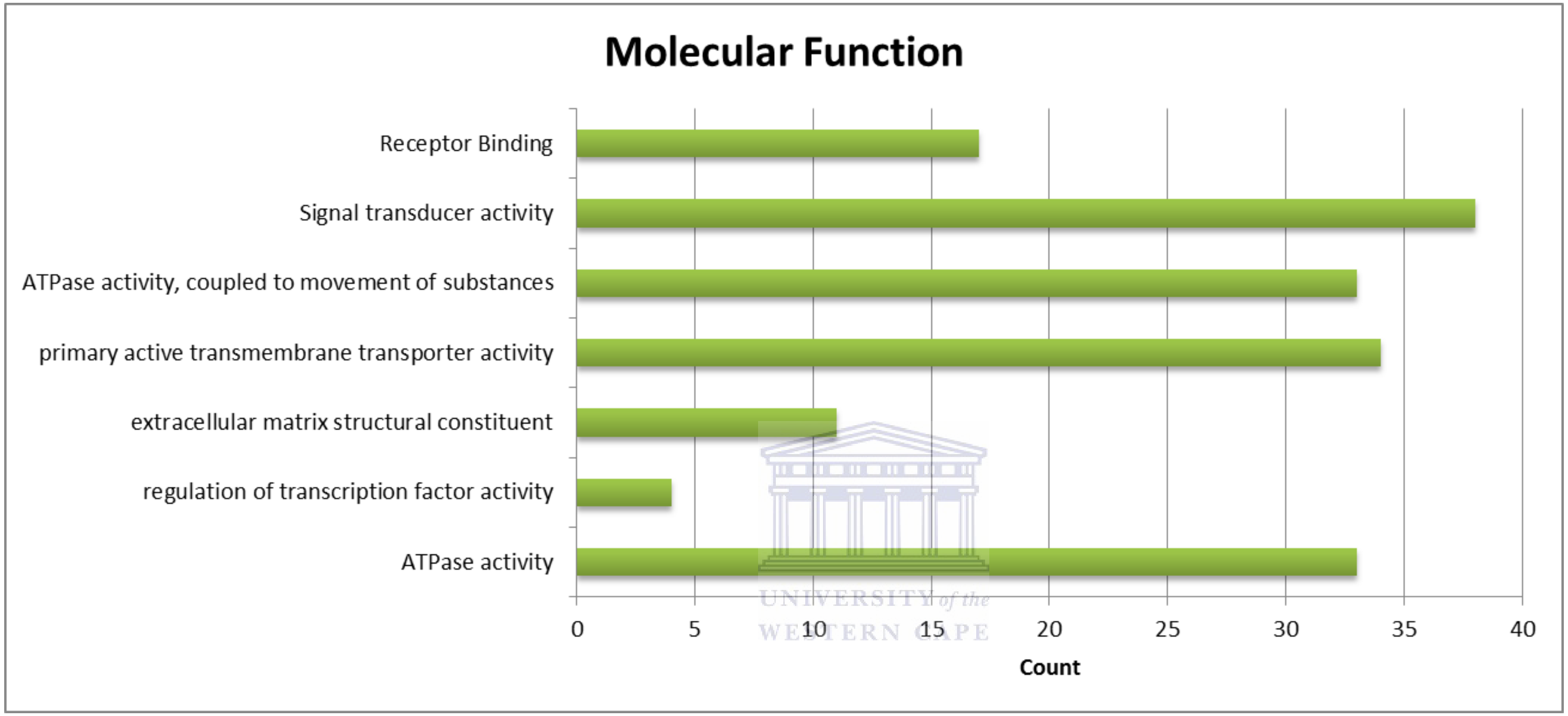


Figure 2.4: Functional Characterisation of genes in DAVID based on their molecular function using GO analysis

2.5.3. Literature Review of the Candidate Entities

Subsequent to functional characterisation in DAVID, the list of candidates were further investigated in order to select a subset of higher priority genes that will be further validated as putative biomarkers for cervical cancer. Further analysis and assessment of the resulting hits were performed retrospectively using various databases such as Uniprot, PolySearch, Google Scholar, HuGENavigator, GoPubmed and iHOP. To find links and cited articles to genes/proteins and identify the particular gene product if the gene name or synonym is known. The entities obtained were checked by carefully reading the associated literature references or original publications. This subset of candidates included genes that have not yet been inferred as putative biomarkers in cervical cancer or have not yet been experimentally validated to be connected to cervical cancer. From the 28190 genes that were investigated in DAVID, only 113 genes matched the stringent criteria imposed in this work and these were further investigated using literature mining. The genes were searched in literature using the selected GO terms, a total of 36 genes were identified and 7 genes were found to be common amongst the list and these were eliminated from the study as highlighted in table 2.4. Thus after review of literature, only 29 candidate genes remained and these were not experimentally linked to cervical cancer to date. As depicted in table 2.4, a number of these genes had been implicated in cervical cancer through various experimental studies.

Table 2.4: Gene Categorisation based on literature

Go Terms	No. of genes identified	No. of genes experimentally validated
Cell surface	14	2
Secretory granules	7	1
Extracellular space	7	3
Extracellular matrix	8	1

Biomedical literature in this research study ensured that extensive information was investigated prior making conclusions. The literature search confirmed the database mining route impressing confidence in the candidate genes, thus indicating that out of more than 28 000 genes extracted from databases, the study managed to prioritise these genes and also obtained a subset of genes that have been experimentally validated to be implicated in cervical carcinogenesis. After literature studies, a total of 29 genes remained as candidate, after experimentally validated genes were eliminated and only those genes showing no relation to cervical cancer were deemed as novel biomarkers.

2.6. Discussion and Conclusion

Cervical cancer continues to represent a major health problem for women from developing countries. Cervical cancer lethality occurs because most patients are first diagnosed in advanced stages. Even if early stages are successfully treated, advanced cervical cancer represents a major problem due to increased rates of recurrence and distant metastasis (Balacescu *et al*., 2014). Cancer is an intricate disease whereby many proteins, genes and molecular processes are involved. Genes and proteins do not work independently, but are organised into co-regulated units that perform a common biological function. The alteration of these functional elements leads to the development of a particular cancer phenotype and subsequently their study cannot be undertaken from the classical one-gene approach (Sanz-Pamplona *et al*., 2012). A systems biology approach, the analysis of the molecular relationship between the implicated genes and proteins as a whole, is required to understand the disease phenotype. Research on biomarkers will help in understanding diseases at the initial stages of development. Therefore the introduction of bioinformatics has improved ways in which new hypothesis is generated for knowledge discovery (Sanz-Pamplona *et al*., 2012).

In this chapter several *in silico* methodologies were employed to interrogate microarray databases in order to unearth the affluence of information that goes unnoticed in databases. Various bioinformatics tools were utilised to excerpt a list of genes based on input queries. The databases chosen for this study were based on several factors (i) the number of times they are curated (verify the accuracy of the information held by these databases) (ii) how often they are referenced as reliable source of data for biomarker discovery studies. The data that was retrieved from all databases was effectively refined to ascertain that each step is treated as an independent validation step thus enhancing the signal-to-noise ratio (Baron *et al*, 2011). This research aimed at identifying tissue-specific biomarkers by making use of publicly available gene and protein databases. According to Prassas *et al* (2012), mining protein expression databases for the identification of candidate biomarkers seems more relevant since serological biomarkers are protein-based. Using gene expression databases also has limitation since the considerable variation between mRNA and protein expression and gene expression does not account for post-translational modification events. Thus, mining both gene and protein expression databases minimizes the limitations of each platform (Prassas *et al*., 2012). The databases were searched for genes and proteins highly specific to or strongly expressed in cervical tissue. The search criterion was designed to accommodate the design of the databases. In the gene expression databases the criteria used were set for maximum stringency for candidate identification to identify a manageable number of candidates. Through the methods that were employed in this study, ~ 361122 genes and 1902 proteins highly specific to or strongly expressed in the cervical tissue and cervical cancer were filtered from the microarray databases and further refined to 113 genes using DAVID.

As with the candidate lists, most array or sequencing databases generate large sets of gene lists that may contain thousands of candidate genes (Huang *et al*., 2008). The integration and interpretation of these heterogeneous data in order to draw meaningful scientific inference from can be a challenging task (Huang *et al*., 2008). Furthermore, it can be difficult to evaluate the biasedness of results and whether gene lists from multiple databases are reproducible or whether they generate overlapping genes. Enrichment analysis allows one to take a data-driven approach to framing results in a functional context (Huang *et al*., 2008). Functional annotation allows for the clustering of putative genes according to their cellular component, molecular function and biological process by using sequence similarity techniques. Gene Ontology provides ontology of a large number of terms which are representative of the gene product properties or the functions of the gene products. The Gene Ontology covers three domains; Cellular component, the parts of the cell or its environment; Molecular function, the activities of the gene product at the molecular level; and Biological processes, molecular events pertaining to function of integrated living units. Most of the genes in a genome are annotated with the ontology terms relevant to the gene products. Each gene is associated with many ontology terms and each term is associated with more than one gene. Genes that are similar in their functioning, share many common ontology terms. A GO Enrichment analysis on a set of genes analyses the GO terms associated with the set and returns the enriched terms in the order of decreasing significance as described in section 2.4.3. In this study, Gene ontology cellular localization annotations of ‘extracellular space’, ‘cell surface ’, ‘secretory granules’ and ‘extracellular matrix’ (Figure 2.3) were selected to identify a protein as secreted or shed. Many groups in biomarker discovery use Gene Ontology protein cellular localisation annotations of ‘extracellular space’, ‘plasma membrane’ and so forth to identify a protein as secreted or shed, hence, this study was able to identify such biomarker candidates thus making this research valuable and this pipeline can be integrated in other biomarker discovery studies for other types of cancer. Proteins interacting with cancer-related proteins have a higher probability of being related with the cancer process than non-interacting proteins. Hence, the study of those proteins may be an efficient way to discover novel cancer genes and cancer biomarkers. The significance is given by the p-values, which is the probability that of a term occurring in the set by chance or a true hit. If the GO enrichment analysis on a particular set of genes returns many terms with very high significance i.e. very low p-values, then that set of genes is highly similar in its properties. When the clustering of the genes was analysed (which combined totalled 113), most of the genes were projected to be enriched for cell surface receptor-linked signal transduction as their primary biological process as shown in figure 2.2. The results for molecular function categorization were consistent with the biological process assigned to these genes, whereby the great majority of the genes were predicted to participate in signal transduction, 17 of the gene variants were allocated to receptor binding (Figure 2.4).

Aberrations in signal transduction have been linked to the characteristics of cancer such as increased proliferation and inhibition of apoptosis (Rowinsky, 2003), the latter being one of the categories the genes was allocated to by DAVID, while 12 of the genes were predicted to function in secretion (Figure 2.2). The majority of the genes were predicted to be intrinsic and integral to membrane and 16 of the genes were enriched for the cell surface (Figure 2.3). This is promising since the targeted biomarkers for this study were those that are easily detectable in bodily fluids. Many of the functional categories such as apoptosis, signalling and focal adhesion were compatible with similar Bioinformatics cancer studies, such as the analyses carried out by Romaschin *et al* (2009) and Zang *et al* (2011). The dissimilarities in this study compared to others can be attributed to this study’s focusing on retrieving secreted cancer genes as opposed to any significant genes in cervical cancer. Determining the subcellular localization of a protein can provide insights into how it functions and the pathways that are involved, as well as highlighting whether the protein could either provide a therapeutic target or act as a biomarker. In many biomarker studies secreted proteins are targeted because they are more likely to be present in body fluids and eventually measured by non-invasive assays (Klee and Sosa, 2007). After extensive literature studies, 29 genes remained as candidate biomarkers. Inclusion parameters were well defined for each selection method so as to attain a list of candidate genes that were differentially expressed in cervical cancer and can be located in biological fluids. Secreted or shed proteins have the highest chance of entering the circulation and being detected in the serum (Prassas *et al*., 2012).

2.7 References

1. Agarwal, S.M., Raghav, D., Singh, H., and Raghava, G.P.S. 2011. CCDB: a curated database of genes involved in cervix cancer. *Nucleic Acids Research*, 39:975-979.

2. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., and Yeh, L-S. L. 2004. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, 32:115-119.

3. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):5.

4. Balacescu, O., Balacescu, L., Tudoran, O., Todor, N., Rus, M., Buiga, R., Susman, S., Fetica, B., Pop, L., Maja, L., Visan, S., Ordeanu, C., Berindan-Neagoe, I., and Nagy, V. 2014. Gene expression profiling reveals activation of the FA/BRCA pathway in advanced squamous cervical cancer with intrinsic resistance and therapy failure. *BMC Cancer*, 14(246):1-14.

5. Baron, D., Dubois, E., Bihouée, A., Teusan, R., Steenman, M., Jourdon, P., Magot, A., Péréon, Y., Veitia, R., and Savagner, F. 2011. Meta-analysis of muscle transcriptome data using the MADMuscle database reveals biologically relevant gene patterns. *BMC genomics*, 12(1):113.

6. Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., and Wishart, D.S. 2008. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research*, 36:399-405.

7. Dennis, G. Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. and Lempicki, R.A. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, 4(5):3.

8. Desany, B., and Zhang, Z. 2004, Bioinformatics and cancer target discovery. *Drug discovery today*, 9(18):795-802.

9. Doms, A., and Schroeder, M. 2005. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res*. 33 (Web Server issue): W783–6. doi:10.1093/nar/gki470.

10. Faro, A., Giordano, D., and Spampinato, C. 2012. Combining literature text mining with microarray data: advances for system biology modelling. *Briefings in bioinformatics*, 13(1):61-82.

11. Gellert, P., Jenniches, K., Braun, T., and Uchida, S. 2010. C-It: a knowledge database for tissue-enriched genes. *Bioinformatics,* 26(18): 2328-2333.

12. Hoffmann, R., and Valencia, A. 2004. A gene network for navigating the literature. *Nature Genet*, 36:664.

13. Huang, D.W., Sherman, B.T., and Lempicki, R.A. 2008. Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. N*ature Protocol*. Accessed 09/09/2013. Available from http://david.abcc.ncifcrf.gov/manuscripts/protocol/np_manuscript.pdf

14. Kapushesky, M. Adamusiak, T., Burdett, T., Culhane, A., Farne, A., Filippov, A., Holloway, E., Klebanov, A., Kryvych, N., Kurbatova, N., *et al*. 2011. Gene Expression Atlas update -a value-added database of microarray and sequencing-based functional genomics experiments. *Gene Expression* (6):1

15. Khatri, P. and Draghici, S. 2008. Ontological analysis of gene expression data: current tools, limitations and open problems. *Bioinformatics*, 21(18):3587-3595.

16. Klee, E.W., and Carlos, C.P. 2007. Computational classification of classically secreted proteins. *Drug Discovery Today*, 12(5/6): 234-240.

17. Liu, X., Yu, X., Zack, D.J., Zhu, H., and Qian, J. 2008. TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics,* 9(271):1-7.

18. McCall, M.N., Uppal, K., Jaffee, H.A., Zilliox, M.J., and Irizarry, R.A. 2011. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Research*, 39:1011-1015.

19. Perez-Llamasy, C., Gundemy, G., and Lopez-Bigas, N. 2011. Integrative Cancer Genomics (IntOGen) in Biomart. Database, Article ID bar039.

20. Prassas, I., Chritoja, C.C., Makawita, S., and Diamandis, E.P. 2012. Bioinformatic identification of proteins with tissue-specific expression for biomarker discovery. *BMC Medicine*, 10(39):1-13.

21. Raza, K., 2010. Application of data mining in bioinformatics. *Indian Journal of Computer Science and Engineering*, 1(2):114-118.

22. Raza, K., and Parveen, R., 2012. Evolutionary algorithms in genetic regulatory networks model. *Journal of Advanced Bioinformatics Applications and Research*, 3(1): 271-280.

23. Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu J., Briggs, B.B., Barrette, T.R., Anstet, M.J., Kincead-Beal, C., Kulkarni, P., Varambally, S., Ghoshy, D., and Chinnaiyan, A.M. 2007. Oncomine 3.0: Genes, Pathways, and Networks in a Collection of 18,000 Cancer Gene Expression Profiles. *Neoplasia,* 9(2): 166-180.

24. Rhodes, D.R., Yu, J., Shanker, K., Deshpandez, N., Varambally, R., Debashis Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A.M. 2004. Oncomine: A Cancer Microarray Database and Integrated Data-Mining Platform. *Neoplasia*, 6(1):1-6.

25. Rodríguez-Pérez, M.A., Medina-Aunon, A., Encarnación-Guevara, S.M., Bernal-Silvia, S., Barrera-Saldaña, H., and Albar-Ramírez, J.P. 2008. In silico analysis of protein neoplastic biomarkers for cervix and uterine cancer. *Clin Transl Oncol*, 10:604-617.

26. Romaschin, A.D., Youssef, Y., Chow, T.F., Siu, M., DeSouza, L.V., Honey, J., Stewart, R., Pace, K.T., and Yousef, G.M. 2009. Exploring the pathogenesis of renal carcinoma: pathway and bioinformatics analysis of dysregulated genes and proteins. *Biological Chemistry*, 390:125-135.

27. Rowinsky, E.K. 2003. Signal Events: Cell Signal Transduction and its Inhibition in Cancer. *The Oncologist*, 8(3):5-17.

28. Sanz-Pamplona, R., Berenguer, A., Sole, X., Cordero, D., Crous-Bou, M., Serra-Musach, S., Guinó, E., Ángel Pujana, M., Moreno, V. 2012. Tools for protein-protein interaction network analysis in cancer research. *Clin Transl Oncol*, 14:3-14.

29. Strausberg, R. L., Dahl, C. A., and Klausner, R. D. 1997. New opportunities for uncovering the molecular basis of cancer. *Nat. Genet*, 17: 415-416.

30. Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Björling, L., Ponten, F. 2010. Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol*, 28(12):1248-50.

31. Yang, X., Ye, Y., Wang, G., Huang, H., Yu. D., Liang, S. 2011. VeryGene: linking tissue-specific genes to diseases, drugs, and beyond for knowledge discovery. *Physiol Genomics* 43:457-460.

32. Yang, Y., Adelstein, S.J., and Kassis, A.I. 2009. Target discovery from data mining approaches, *Drug discovery* today, 14(3):147.

33. Yu, W., Gwinn, M., Clyne, M., Yesupriya, A., and Khoury, M.J. 2008. A Navigator for Human Genome Epidemiology. *Nat Genet*, 40:124-125.

34. Zang, W-D., Liu, J., Wang, L.S., and Pan, T-W. 2011. Identifying genes related with non-small cell lung cancer via transcription factors- target genes relationships. *International Journal of Physical Sciences*, 6(28):6450-6457.

CHAPTER 3: *In Silico* Expression Analysis of Putative Biomarkers

3.1. Introduction

In order to extract most key features from the data, methods for visualization and analysis of large-scale data are of utmost importance (Hastie *et al*., 2009). Enrichment analysis allow for the biological interpretation of large gene and protein lists by investigation of the functional categories present in the data. Tools that allow construction of biological networks are utilised for visualizing relationships such as protein-protein interactions, and are valuable for extracting meaningful information from extensive data sets (Hastie *et al*., 2009). The large amounts of data generated by high-throughput strategies in genomics and proteomics often results in long lists of interesting genes, which are challenging to interpret. The biological knowledge stored in the vast number of databases described can be exploited to allow for a systematic functional analysis of these lists to summarize the most relevant properties. Bioinformatic tools for enrichment analysis have been successful in adding valuable information to large-scale biological studies (Huang et *al*., 2009).

Biological processes and functions within a cell are seldom dependent on a single gene, however most often made up of a group of genes and this is the principal basis behind enrichment analysis. Co-functioning genes are likely to be selected together if a certain process or function is atypical in a biological study (Huang et *al*., 2009). Once a certain functional term has been associated with a set of genes, a test has to be performed to determine if the enrichment based on the proportion of associated genes is significantly different than what would be expected by chance alone. Thus, an enrichment of a gene set uses a statistical model such as binomial, hyper-geometric or a chisquare or Fisher's exact test for equality of proportions to calculate overrepresented or enriched terms, where a p-value examines the significance of the enrichment (Draghici *et al*., 2003). To determine whether a functional term is overrepresented in a set of genes or not, a reference or background list is used for comparison and for determination of the degree of enrichment. A reference set can for example consist of the entire genome of the species being analyzed, or the complete set of genes with the potential of being part of the annotation category in question. When many categories are considered and hence a large number of tests are performed, a multiple-testing correction such as False Discovery Rate, Bonferroni or

Benjamini-Hochberg can be useful to correct for false rejections of the null hypothesis. Some examples of bioinformatic enrichment analysis tools include GoMiner, Onto-Express, GSEA and DAVID (Huang *et al*., 2009). Gene ontology (GO) terms have been the most primarily used annotation data to date, however, recently several of the new or updated tools have started to comprise a larger assortment of underlying information, such as data on KEGG pathways, Online Mendelian Inheritance in Man (OMIM) disease associations, protein domains and gene-expression result in their annotation databases (Huang *et al*., 2009). The differences between the many available methods lie mainly in their supported gene identifiers, choice of statistical model, reference data, annotation data, mapping between databases and many other aspects that can have a great impact on the results. Therefore it is highly important to be aware of the strengths and drawbacks of each method when deciding which one to use (Huang *et al*., 2009).

3.1.1. GeneHub-GEPIS

GeneHub-GEPIS (http://www.gepis.org/./) is a web application that performs digital expression analysis on human and mouse tissues based on an integrated gene database, using aggregated EST library information and EST counts. The application calculates the normalized gene expression levels across a large panel of normal and tumor tissues, thus providing rapid expression profiling for a given gene (Zhang *et al*., 2007). The backend GeneHub component of the application contains pre-defined gene structures derived from mRNA transcript sequences from major databases and includes extensive cross references for commonly used gene identifiers. ESTs are then linked to genes based on their precise genomic locations as determined by Genomic Mapping and Alignment Program (GMAP). In addition, the gene-centric design makes it possible to add several important features, including text-searching capabilities, the ability to accept diverse input values, expression analysis for microRNAs, basic gene annotation, batch analysis, and linkage between mouse and human genes (Zhang *et al*., 2007).

3.1.2. Genecards

Genecards (http://www.genecards.org/) is an integrated database of human genes that provides concise genome related information on all known and predicted human genes. It extracts and integrates a carefully selected subset of gene related transcriptomics, genetic, proteomic, functional and disease information, from dozens of relevant sources. The information is automatically mined and integrated from a variety of data sources, resulting in a web based card for each of the 7000 human genes that currently have an approved gene symbol published by the HUGO/ Genome Database (GDB) nomenclature committee (Stelzer *et al*., 2011). The aim of the database is to provide immediate current knowledge on a given gene. Source databases, mined to compile information stored by GeneCards, include SWISS-PROT, OMIM, Gene Atlas and GDB. This composite database aims to integrate information fragments, scattered over a variety of specialised databases into a coherent picture. Genecards is a freely accessible web resource that offers one hypertext card for each of the genes in the database and the recent version features both novel infrastructure and an improved search engine (Stelzer *et al*., 2011).

3.1.3. GeneMania

GeneMania (http://pages.genemania.org/) uses a heuristic algorithm derived from ridge regression to predict the function of a set of input genes. It functions by finding directly interrelated/interacting genes and uses functional association from multiple genomics and proteomics network data to link genes/proteins of interest in real-time (Mostafavi *et al*., 2008). Two genes are linked if their expression levels are similar across a specific condition in a gene expression study. The data is collected from publications within Gene Expression Omnibus (Mostafavi *et al*., 2008).

3.1.4. STRING

The database STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) aims to provide comprehensive, yet quality controlled collection of protein-protein associations for a large number of organisms (http://string-db.org/). The associations are derived from high throughput experimental data, from the mining of databases and literature, and from predictions based on genomic context analysis (Von Mering *et al*., 2005). STRING integrates and ranks these associations by benchmarking them against a common reference set, and presents evidence in a consistent and intuitive web interface. Importantly, the associations are extended beyond the organism in which they were originally described, by automatic transfer to orthologous protein pairs in other organisms, where applicable. STRING currently holds 730 000 proteins in 180 fully sequenced organisms (Von Mering *et al*., 2005). STRING specializes in three ways: (i) it provides uniquely comprehensive coverage, with >1000 organisms, 5 million proteins and >200 million interactions stored; (ii) it is one of very few sites to hold experimental, predicted and transferred interactions, together with interactions obtained through text mining; and (iii) it includes a wealth of accessory information, such as protein domains and protein structures, improving its day-to-day value for users (Franceschini *et al*., 2013)

3.2. Network Analysis

The majority of processes in the cell are dependent on various proteins working together in signalling cascades or larger complexes and co-operating inside organelles. The interaction partners of a protein can contribute to defining its function and it is known that co-expressed genes are more likely to interact and be involved in the same biological pathway than genes that are not expressed at the same time (Bader *et al*., 2003). Thus protein interactions and other relationships between biological molecules are important areas to study. Such interactions are often visualised by networks or more formally two-dimensional graphs consisting of vertices (nodes) connected pair wise by edges. The connections can express various forms of relationships such as proteins known to interact with one another or be co-expressed sharing a domain belonging to the same protein family or being evolutionary related (Pavlopoulos *et al*., 2008).

3.3. Methods and Materials

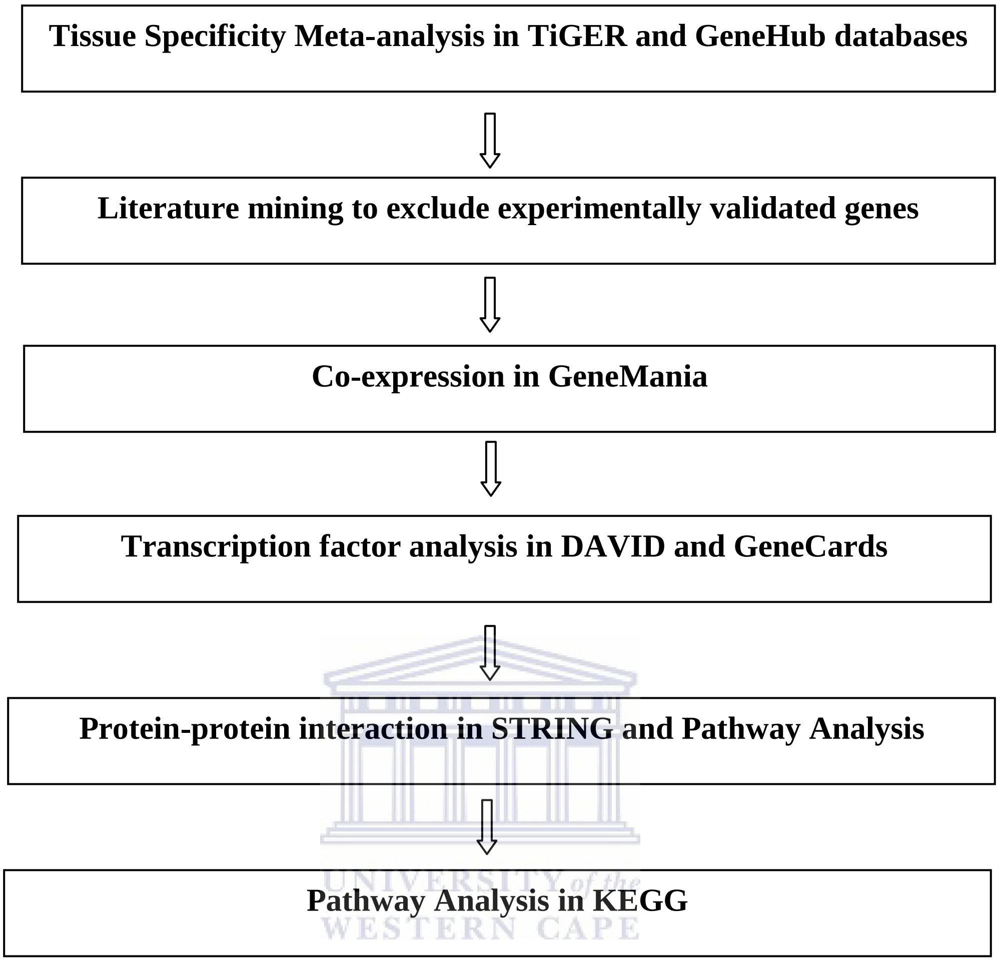

Tissue Specificity Meta-analysis in TiGER and GeneHub databases
Literature mining to exclude experimentally validated genes
Co-expression in GeneMania
Transcription factor analysis in DAVID and GeneCards
Protein-protein interaction in STRING and Pathway Analysis
Pathway Analysis in KEGG

Figure 3.1: Representation of the in silico enrichment analysis

3.3.1. Verification of *In Silico* Expression Profiles

The TIGER and GeneHub-GEPIS databases were used to manually verify the expression profiles of the proteins and genes identified as "potentially" being secreted for strength and specificity of expression (figure 3.1). These databases were chosen above others as they offer a gene expression chart based on ESTs. These databases were also used as a source of elimination. The rationale was to determine whether the genes were expressed in other types of female cancers apart from cervical cancer and whether the proteins were expressed in other female tissues besides the cervix. The female cancers/tissues chosen were: cervix, breast, ovary, uterus, colon, lung and skin. For each tissue, proteins with gene expression profiles showing similar values of expression or high expression in more than the selected tissue were eliminated. If the gene or protein was absent in the cervix but was present in the other tissues, that protein or gene was eliminated from further study. If high expression was observed in the tissue of interest (cervix), but not in the other tissues, the protein or gene was not eliminated. If a protein was observed to be highly expressed in another tissue distinct from the criterion set for elimination, that particular protein was not eliminated. For data accuracy the lack of specificity had to be observed in both databases before non-specific genes were officially eliminated. The TIGER database was used to manually check for the expression of each gene individually across different cancer types by using "Gene View" and GeneHub-GEPIS database: was used to search each gene individually by using "Search by Accession/Gene Symbol". In addition to accounting for tissue specificity, a co-expression analysis was carried out.

3.3.2. Co-Expression Analysis

To ascertain if these genes shared a similar expression pattern, a co-expression analysis was performed using GeneMania (http://www.genemania.org/). The genes of interest (GOI) were searched for co-expression in *Homo sapiens*.

3.3.3. Transcription Factor Analysis

To confirm if the putative genes had any connection to cancer through their regulatory network, a transcription factor (TF) analyses was carried out. All possible TFs regulating a specific gene were extracted from Genecards and DAVID and validated for an association with cervical cancer development using literature mining. The TFs predicted to be associated with 80% or more genes were carried forward for further analysis.

3.3.4. STRING Analysis

An analysis in STRING 9.0 (http://string-db.org/) was carried out with the intention to view the most common pathway via Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/pathway.html). By performing the analysis in STRING via KEGG it was easier to observe the actual number of GOI enriched for a specific pathway. STRING also grouped the genes of interest according to their biological process, molecular function and cellular component via their protein-protein interaction network.

3.4. Results and Discussion

3.4.1. *In silico* Tissue Specificity Expression Analysis

The 29 candidate genes were subjected to a cross cancer analysis in TiGER and GeneHub-GEPIS according to their specificity for cervix tissue. The figures 3.2-3.11 provide a graphical display of the individual candidate genes. In TiGER, the profile expression level is normalized with tissue-library size. Each value for a gene in a tissue is a ratio of observed ESTs to the expected one in that particular tissue. The expected number of ESTs is the product of total ESTs of the genes and the fraction of total ESTs in the tissue among all ESTs in 30 tissues. In GeneHub-GEPIS, Digital expression unit (DEU) is the number of matching clones per 1 million library clones and is directly proportional to the copy number of mRNA per cell. Statistical significance was measured by the Z- test ($p<0.025$). A cross-cancer analysis in TiGER and GeneHub-GEPIS was used to characterize the 29 remaining genes according to their specificity for the cervical cancer tissue. Figures 3.2-3.11 gives a graphical display of the 10 candidate genes that showed expression in the cervix tissue for both databases, with GeneHub-GEPIS displaying expression between cancerous and normal tissue. A drawback of using only functional association data of DEGs is that it does not take into account any physical interactions between genes or proteins (Glaab *et al*., 2012). This prompted further analysis of the 10 candidate genes by submitting these candidate genes to STRING to investigate gene-gene (protein-protein) interactions or possible co-expression and co-regulation in cancer-related metabolic pathways.

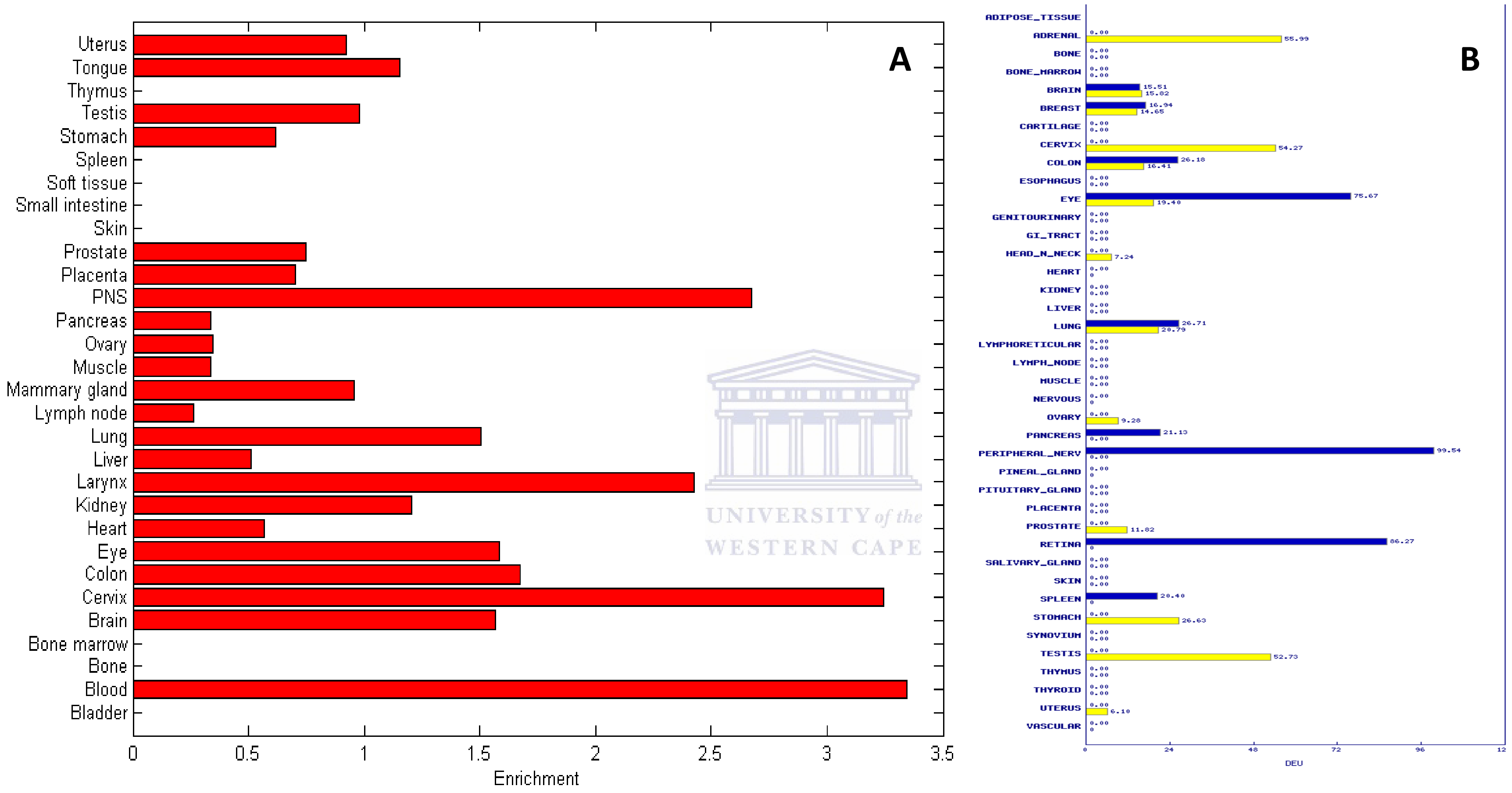


Figure 3.2: Expression profile of Gene 1, Adapted from TiGER (A) and GeneHub-GEPIS (B), 2013. Normal expression is shown in blue while over-expression in tumour tissue is shown in yellow.

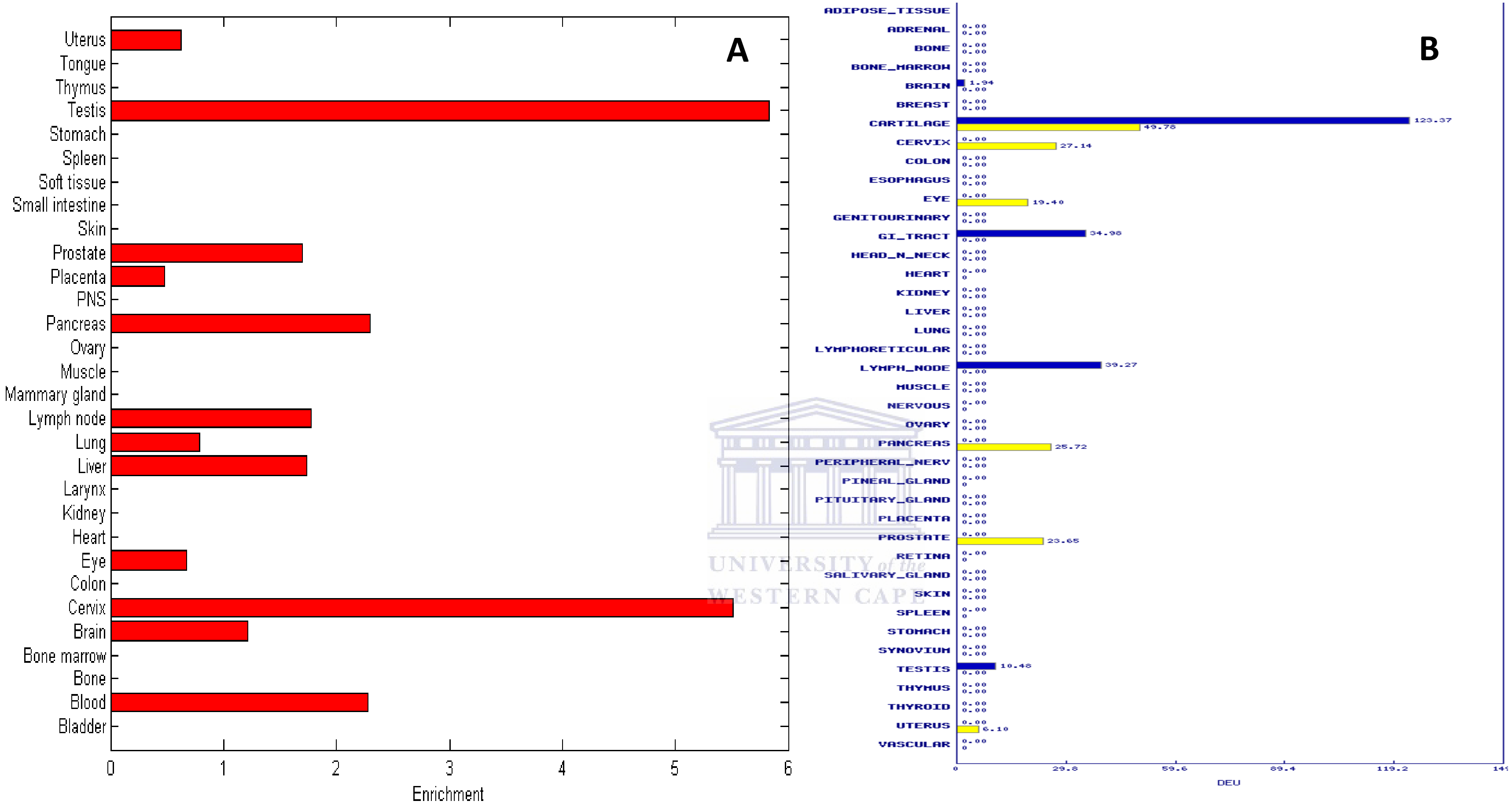


Figure 3.3: Expression profile of Gene 2, Adapted from TiGER (A) and GeneHub-GEPIS (B), 2013. Normal expression is shown in blue while over-expression in tumour tissue is shown in yellow.

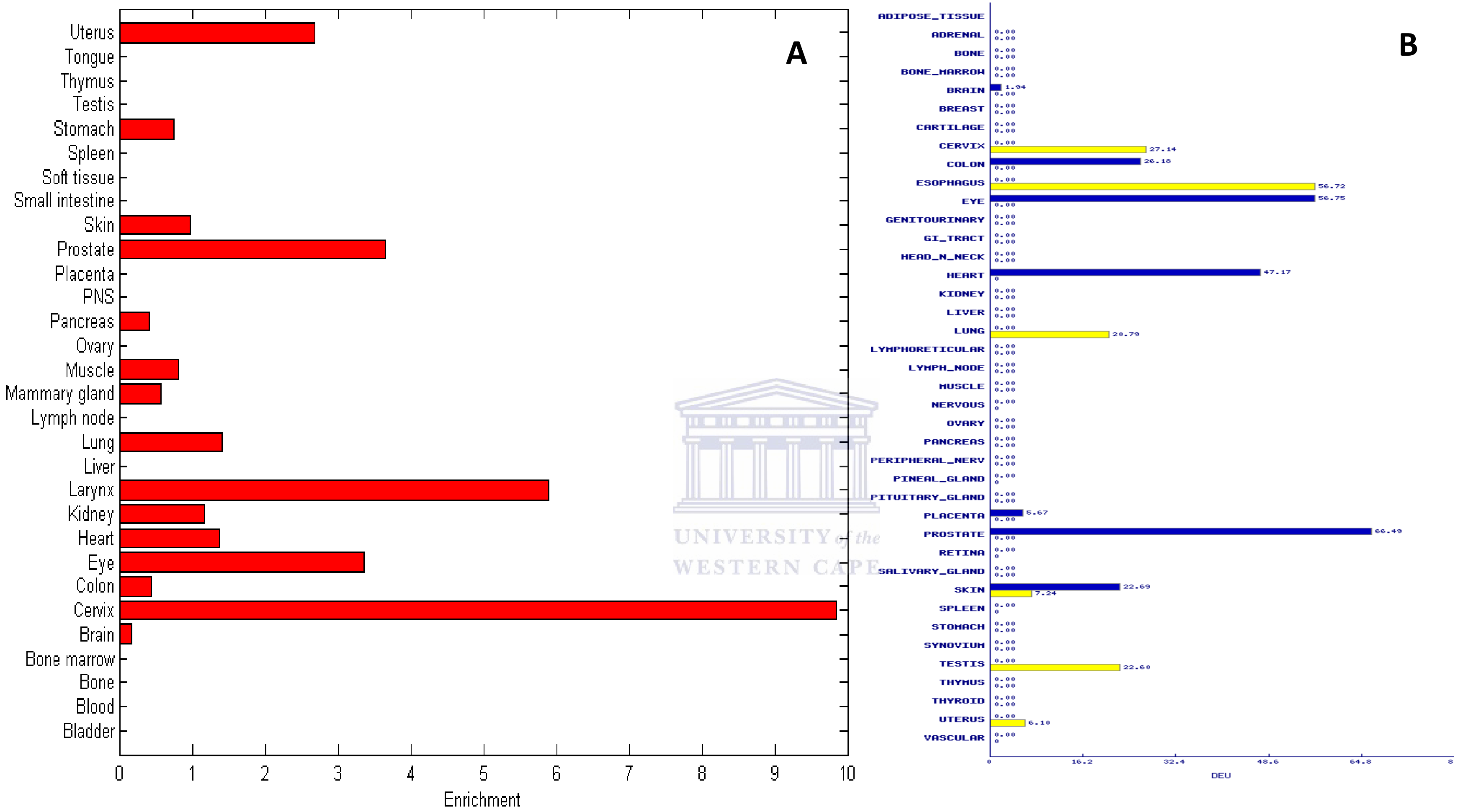


Figure 3.4: Expression profile of Gene 3, Adapted from TiGER (A) and GeneHub-GEPIS (B), 2013. Normal expression is shown in blue while over-expression in tumour tissue is shown in yellow.

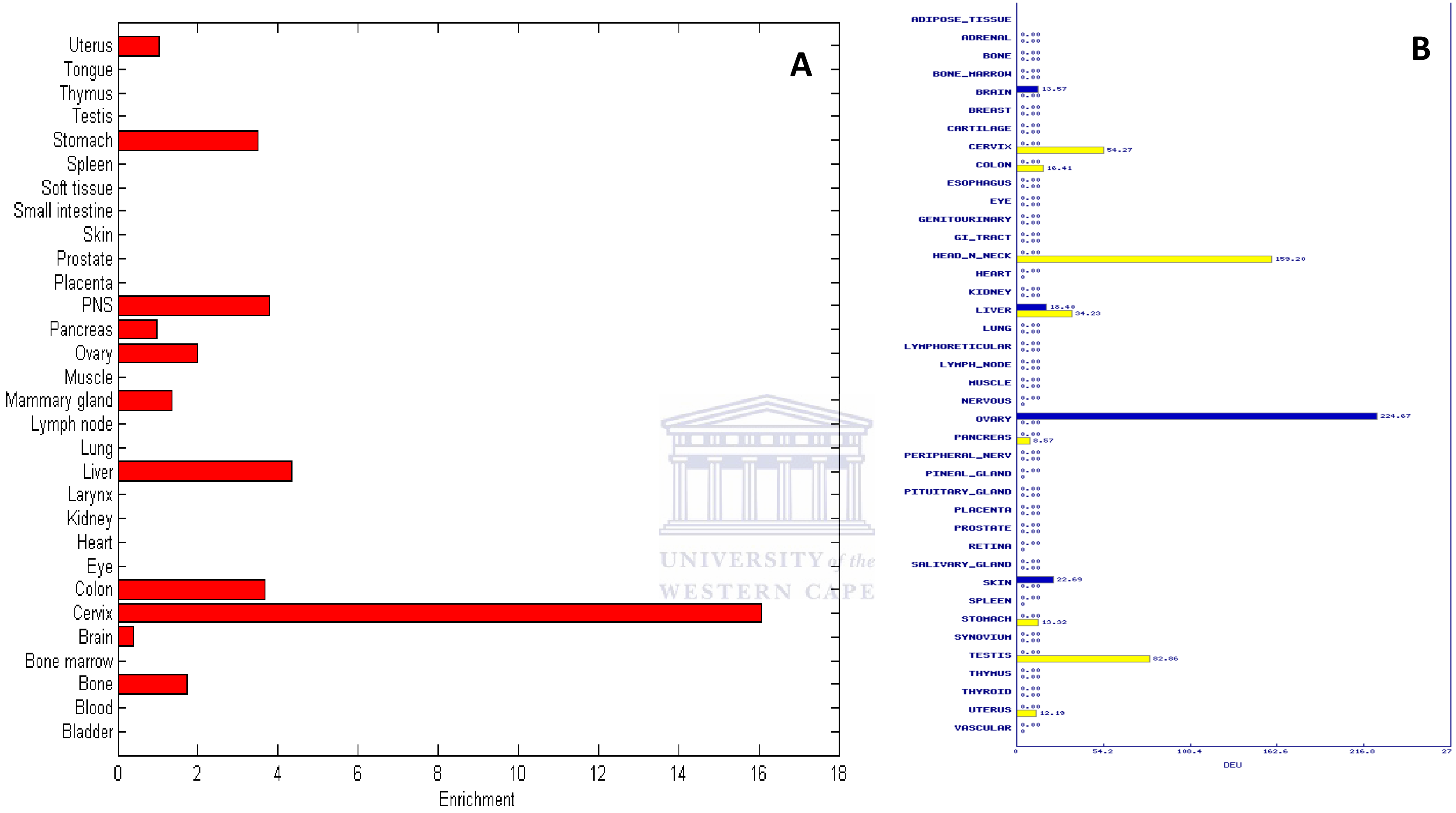


Figure 3.5: Expression profile of Gene 4, Adapted from TiGER (A) and GeneHub-GEPIS (B), 2013. Normal expression is shown in blue while over-expression in tumour tissue is shown in yellow.

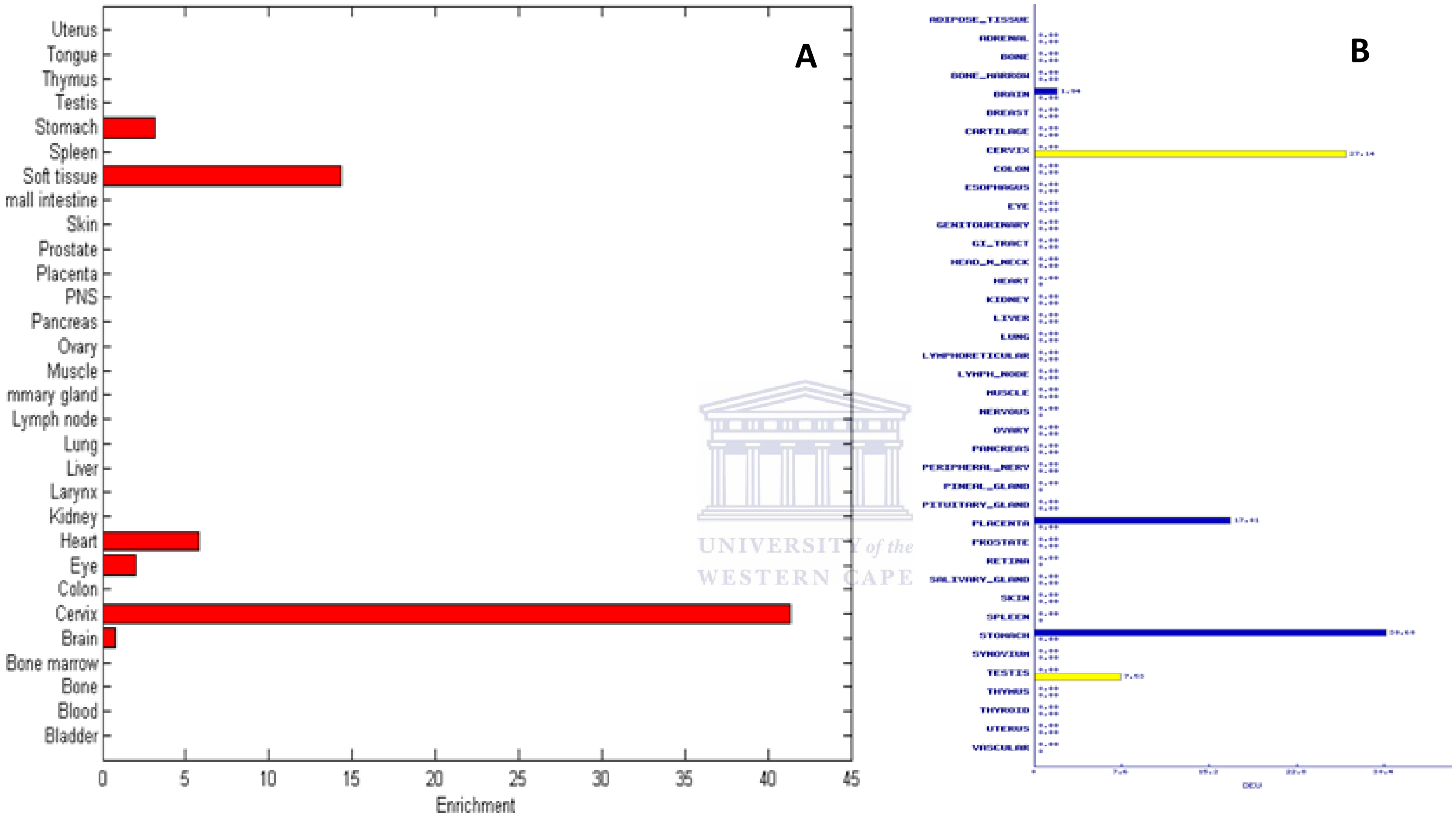


Figure 3.6: Expression profile of Gene 5, Adapted from TiGER (A) and GeneHub-GEPIS (B), 2013. Normal expression is shown in blue while over-expression in tumour tissue is shown in yellow.

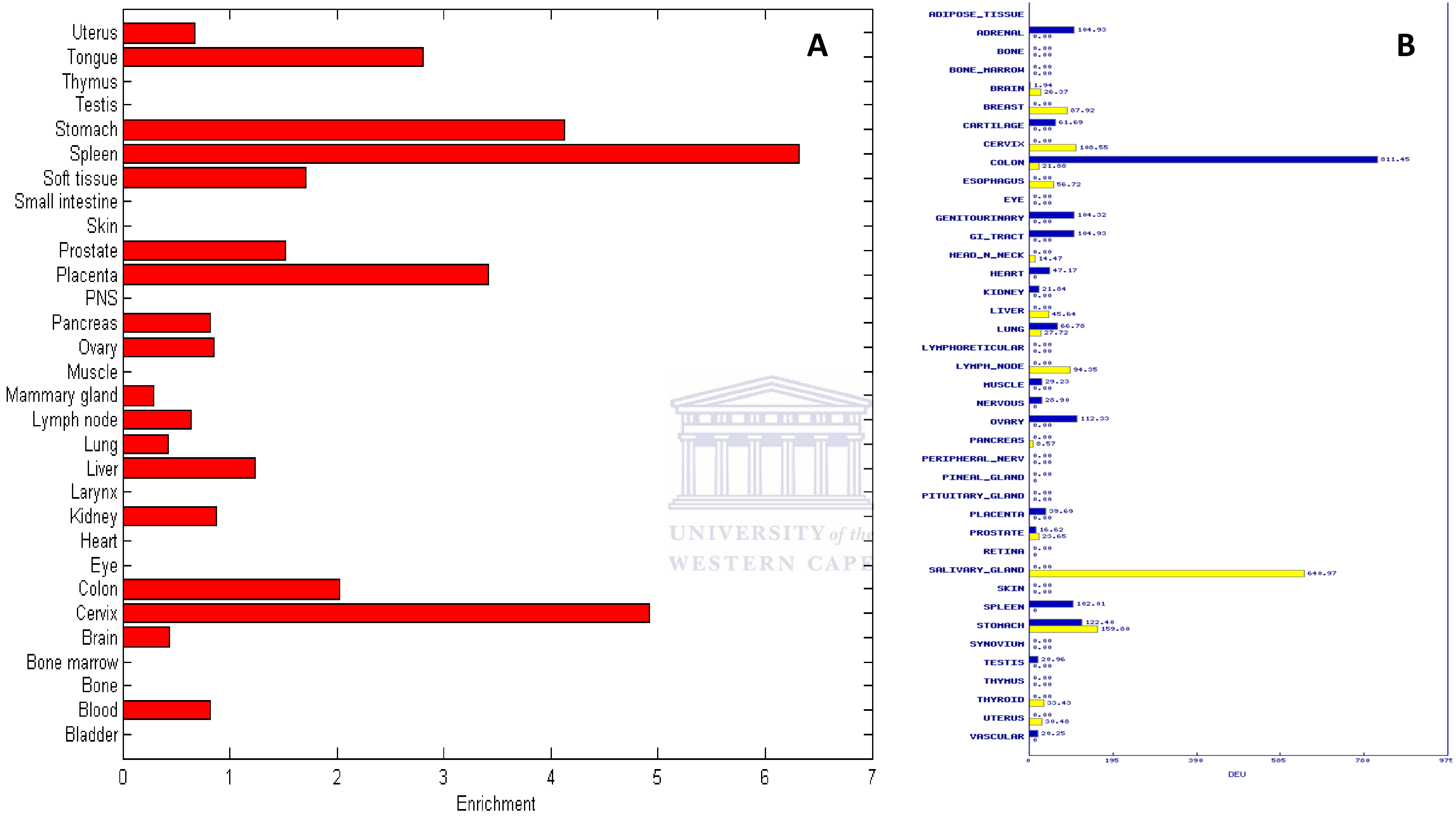


Figure 3.7: Expression profile of Gene 6, Adapted from TiGER (A) and GeneHub-GEPIS (B), 2013. Normal expression is shown in blue while over-expression in tumour tissue is shown in yellow.

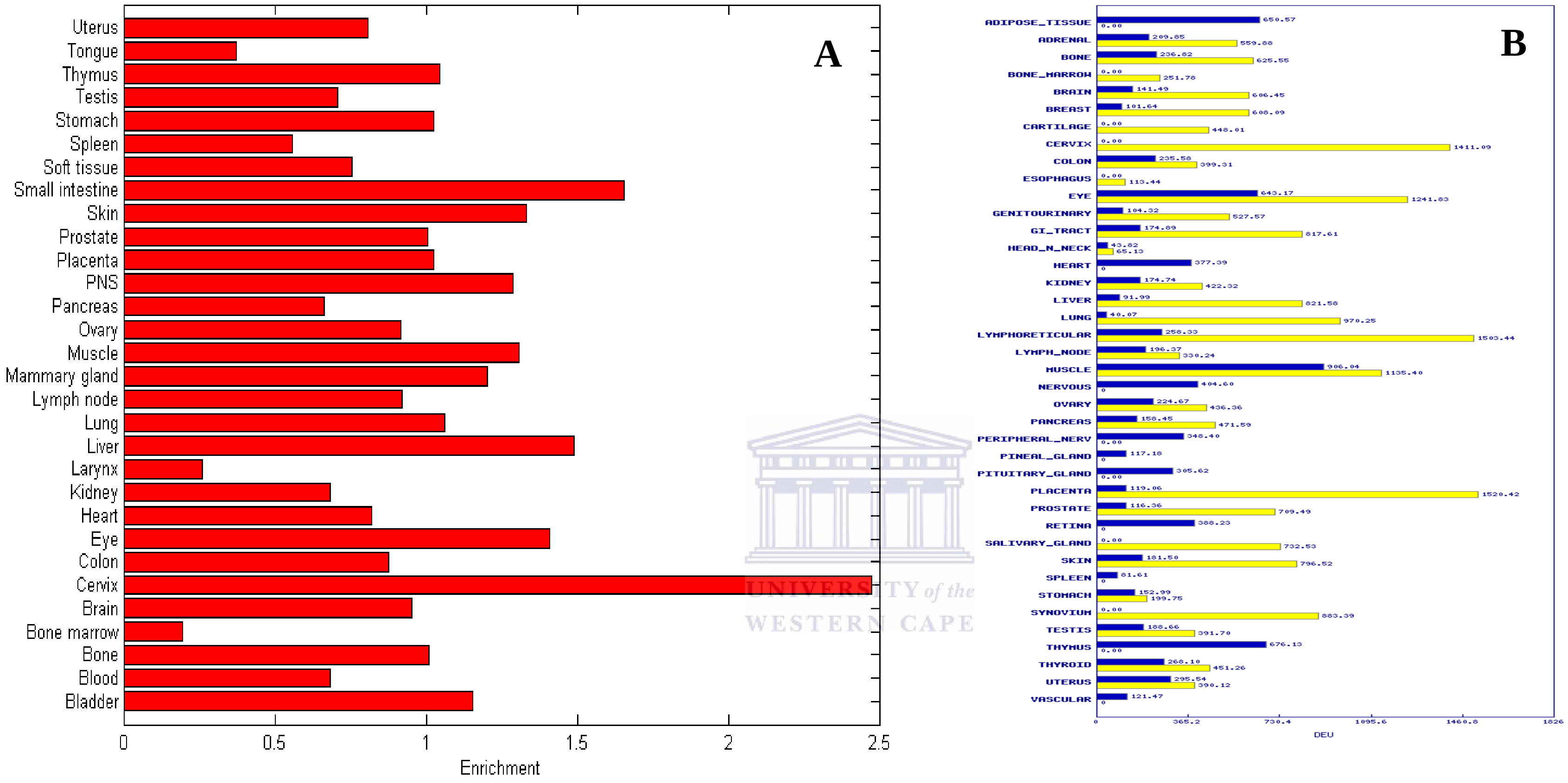


Figure 3.8: Expression distribution of Gene 7, Adapted from TiGER (A) and GeneHub-GEPIS (B), 2013. Normal expression is shown in blue while over-expression in tumour tissue is shown in yellow.

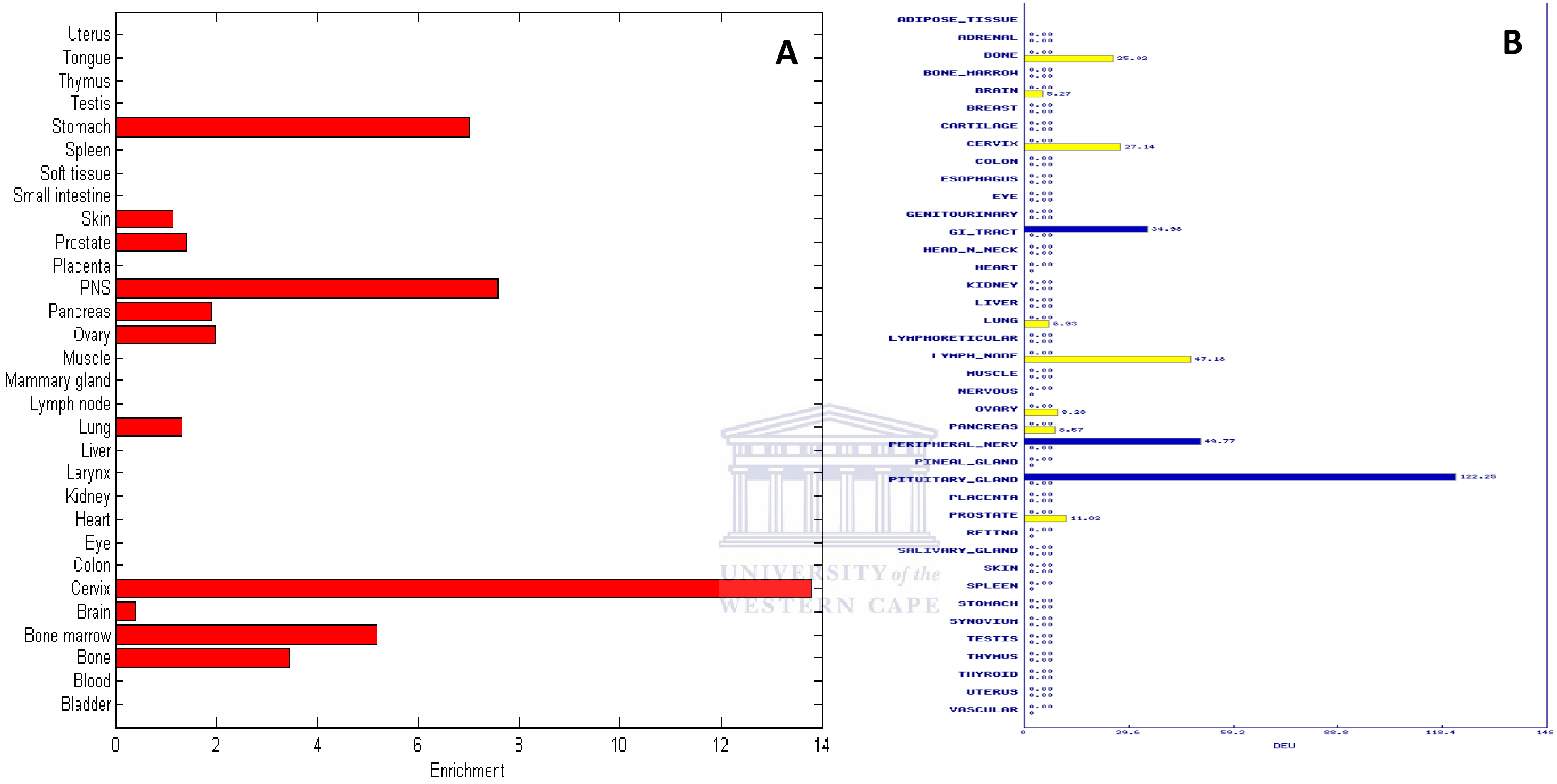


Figure 3.9: Expression distribution of Gene 8, Adapted from TiGER (A) and GeneHub-GEPIS (B), 2013. Normal expression is shown in blue while over-expression in tumour tissue is shown in yellow.

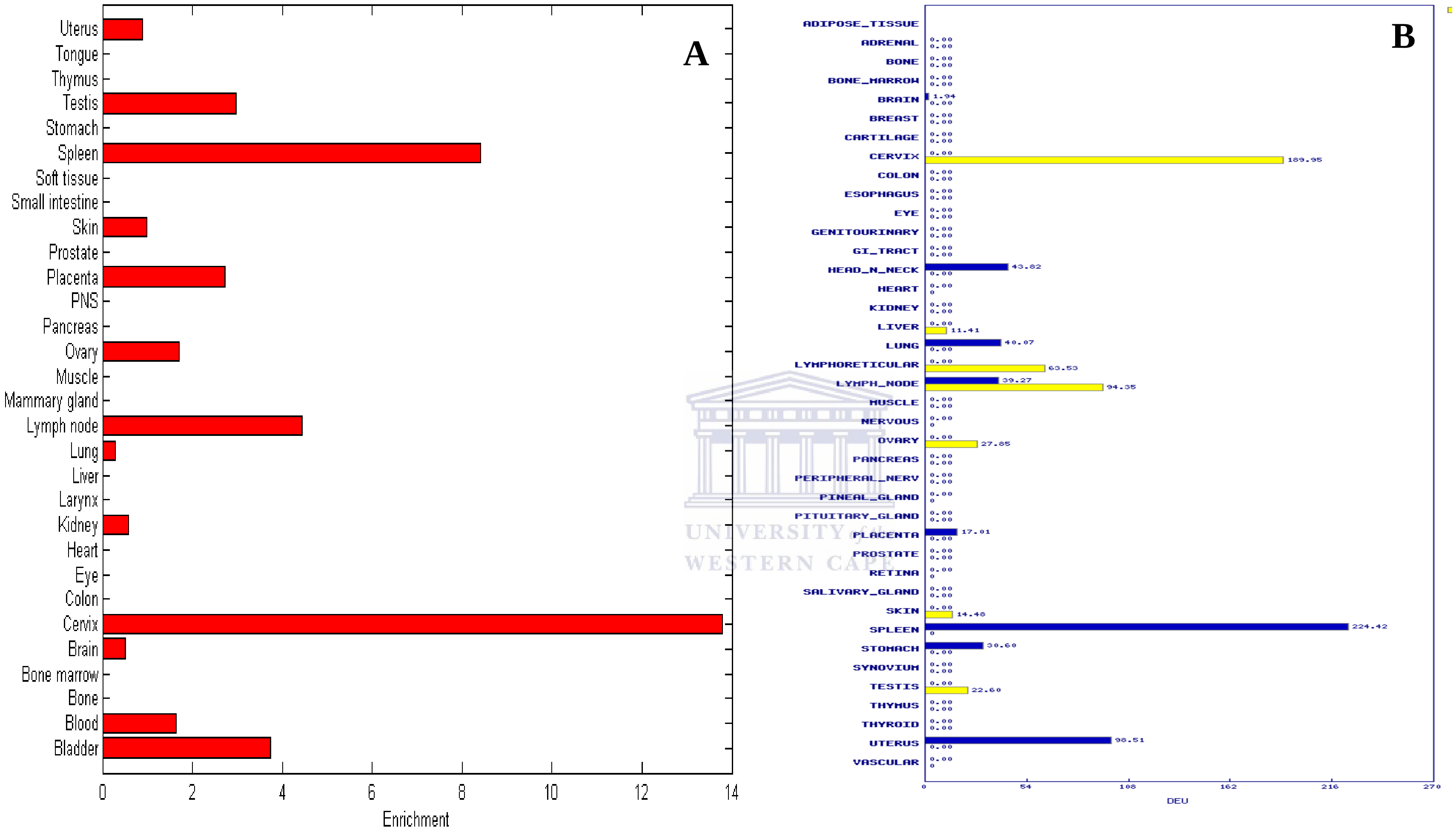


Figure 3.10: Expression distribution of Gene 9, Adapted from TiGER (A) and GeneHub-GEPIS (B), 2013. Normal expression is shown in blue while over-expression in tumour tissue is shown in yellow.

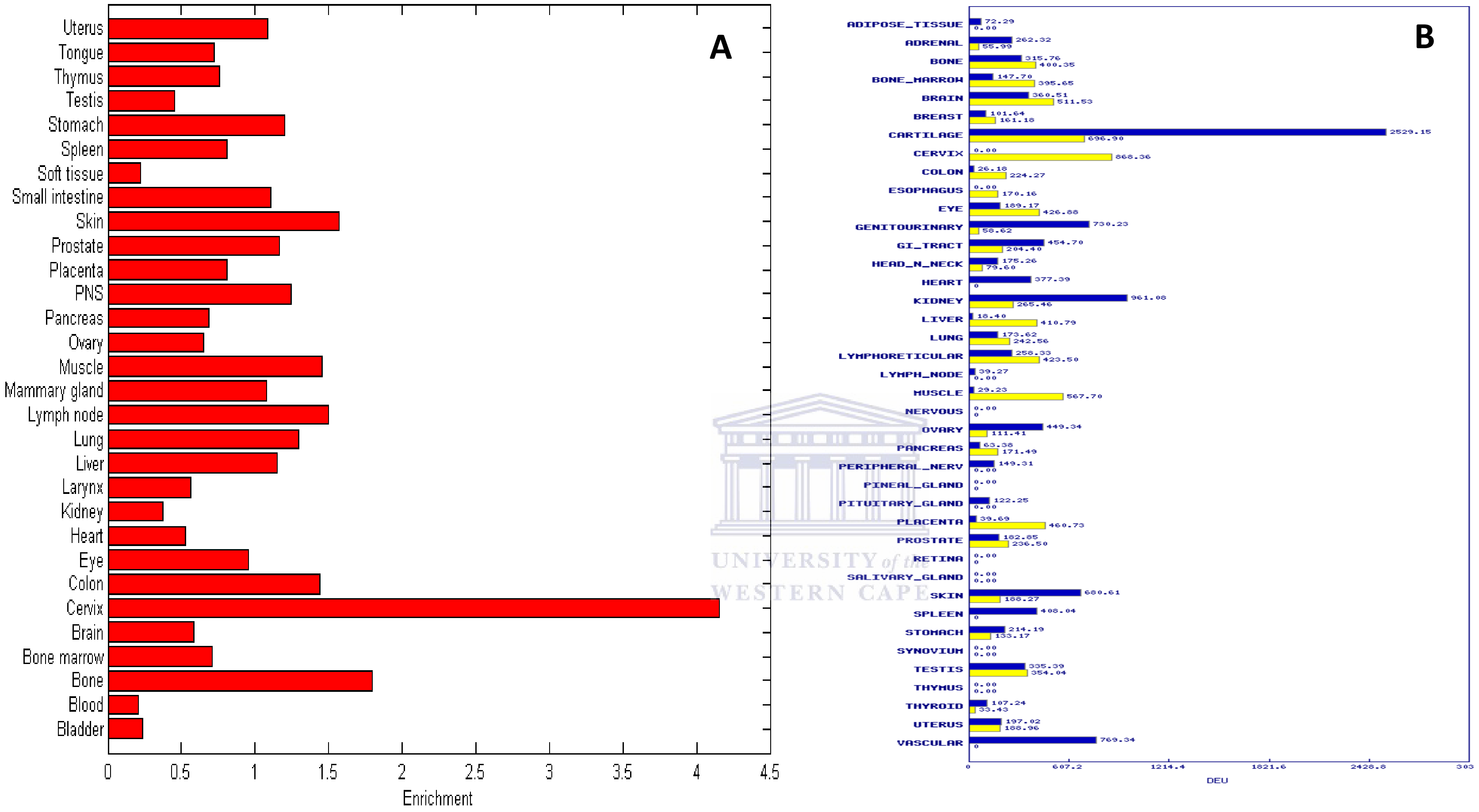


Figure 3.11: Expression distribution of Gene 10, Adapted from TiGER (A) and GeneHub-GEPIS (B), 2013. Normal expression is shown in blue while over-expression in tumour tissue is shown in yellow.

3.4.2. STRING Analysis

The results from the STRING database indicated that the ten genes were not all linked as measured by parameters such as co-expression in STRING, with only one interaction observed as illustrated in figure 3.12. However, 50% (5/10) of the genes were shown to be enriched for receptor binding when classified based on their molecular function. Due to the fact not much scientific inference was drawn from protein interactions between the genes of interest, further analysis were done by conducting a co-expression analysis in order to search for a possible link between the putative genes and a common pathophysiological process.

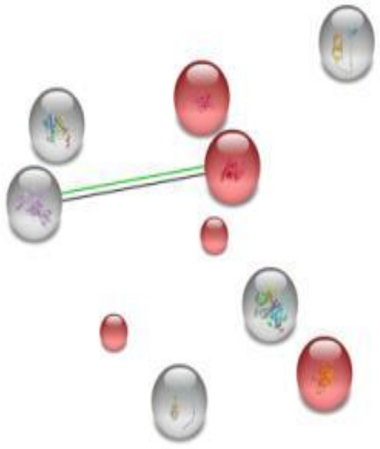


Figure 3.12: STRING showed that five (shown in red) out of the ten putative genes are implicated in receptor binding (gene names not shown) (Adapted from STRING, 2013).

3.4.3. Co-Expression Analysis

According to GeneMania, 90% (9/10) of the putative genes are co-expressed as depicted in figure 3.13. Therefore, it can be deduced that there is a strong probability that the candidate genes co-function in the same pathophysiological process/disease. Following co-expression analysis, the putative genes were further subjected to TF analysis. It has been documented that co-expression analysis may also reveal information regarding the regulatory system of the candidate genes. Genes that are shown to have similar expression patterns, generally they are also controlled by the same underlying regulatory system (Heyer *et al*., 1999).

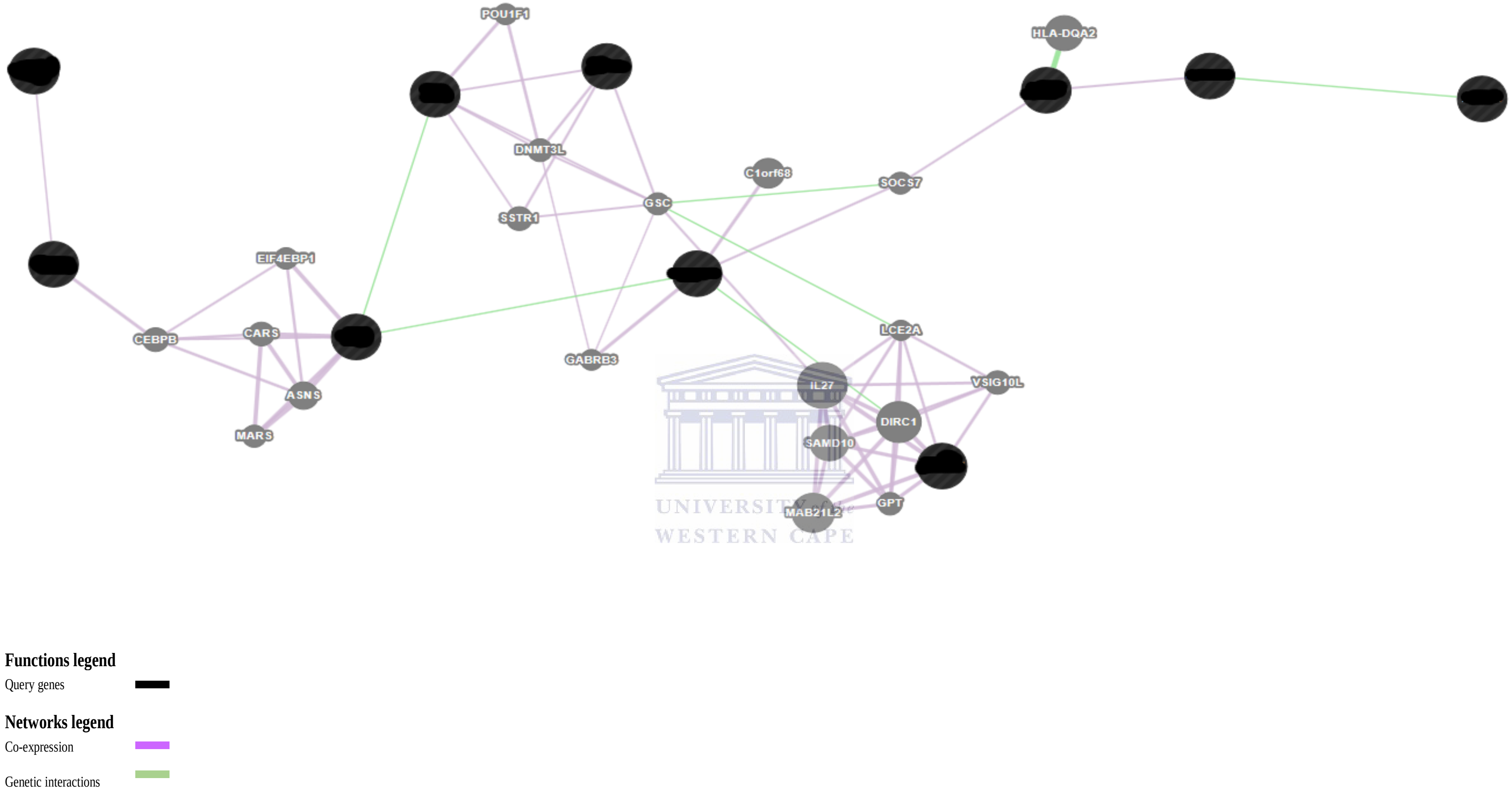


Figure 3.13: Co-expression analysis display of the putative genes, adapted from GeneMania, 2013.

3.4.4. Transcription Factor Analysis

Relevant TFs for the ten candidate genes were extracted from Genecards and DAVID and these were validated for their association with cervical cancer through published literature. Genecards is a web based database that extracts data from genetics, proteomics, transcriptomics, functional and disease information to provide a complete summary of a specific gene. It was discovered that the putative genes are associated with multiple TFs implicated in cervical carcinogenesis such as c-Myc, c-Jun, NFkB and E2F. However, 100% (10/10) of the putative genes were projected to be modulated by perixosome proliferator-activated gamma (PPARγ), while 80% (8/10) were predicted to be regulated by TAL1BETAITF2, p53 and CCAAAT enhancer binding protein alpha (C/EBPα). PPARγ is a ligand-dependent transcription factor that belongs to the nuclear hormone receptor family and is known to regulate cell differentiation and apoptosis (Han *et al*., 2003, Smith *et al*., 2001). With the TF analysis, the genes were analysed individually with their associated TFs, yet no real association between the genes of interest and their TFs could be visualized. Hence, a gene-gene interaction analysis considering the putative genes and their three TFs of interest (which together totalled 13 genes) was carried out in STRING to observe a network view of their interactions.

STRING analyses indicated that when a wider view of the 13 genes/proteins was considered, they were remarkably all re-centred around Proliferating cell nuclear antigen (PCNA) as shown in figure 3.14. The E7 oncoprotein of high-risk HPV has been shown to activate PCNA, causing its up-regulation in cervical intraepithelial neoplasia, the premalignant condition of cervical cancer (Branca *et al*., 2006). Since PCNA is up-regulated in pre-invasive cervical cancer and the candidate genes are linked to PCNA, they could potentially serve as early diagnostic molecular markers. Similarly, the relationship of p53 and the GOI were once again reaffirmed via PCNA. The p53 is known to induce and up-regulate PCNA (Paunesku *et al*., 2001). When PCNA and p53 are both up-regulated, they function in DNA repair. However, when p53 is down-regulated (or absent) due to degradation via the HPV E6 oncogene and PCNA is up-regulated due to E7 induction, cell proliferation occurs instead (Paunesku *et al*., 2001). Hence the two viral oncogenes, E6 and E7 function together to promote cervical carcinogenesis by inversely modulating PCNA and p53. PCNA and p53 in combination have been suggested as valuable diagnostic biomarkers in low-grade cervical intra-epithelial neoplasia (CIN - the first stage of cervical cancer when a diagnosis is ideal) (Goel *et al*., 2011). Henceforth, the genes directly interacting with p53 and PCNA could act as molecular signatures in the early diagnosis of cervical cancer. Although this information was promising in possibly implicating the putative genes in cervical cancer, the co-expression analysis demonstrated the significance of considering genes interacting with the putative genes. Hence, the ten putative genes, their interacting genes (extracted from GeneCards) and their TFs (which altogether totalled 79 genes) were once more assessed in STRING.

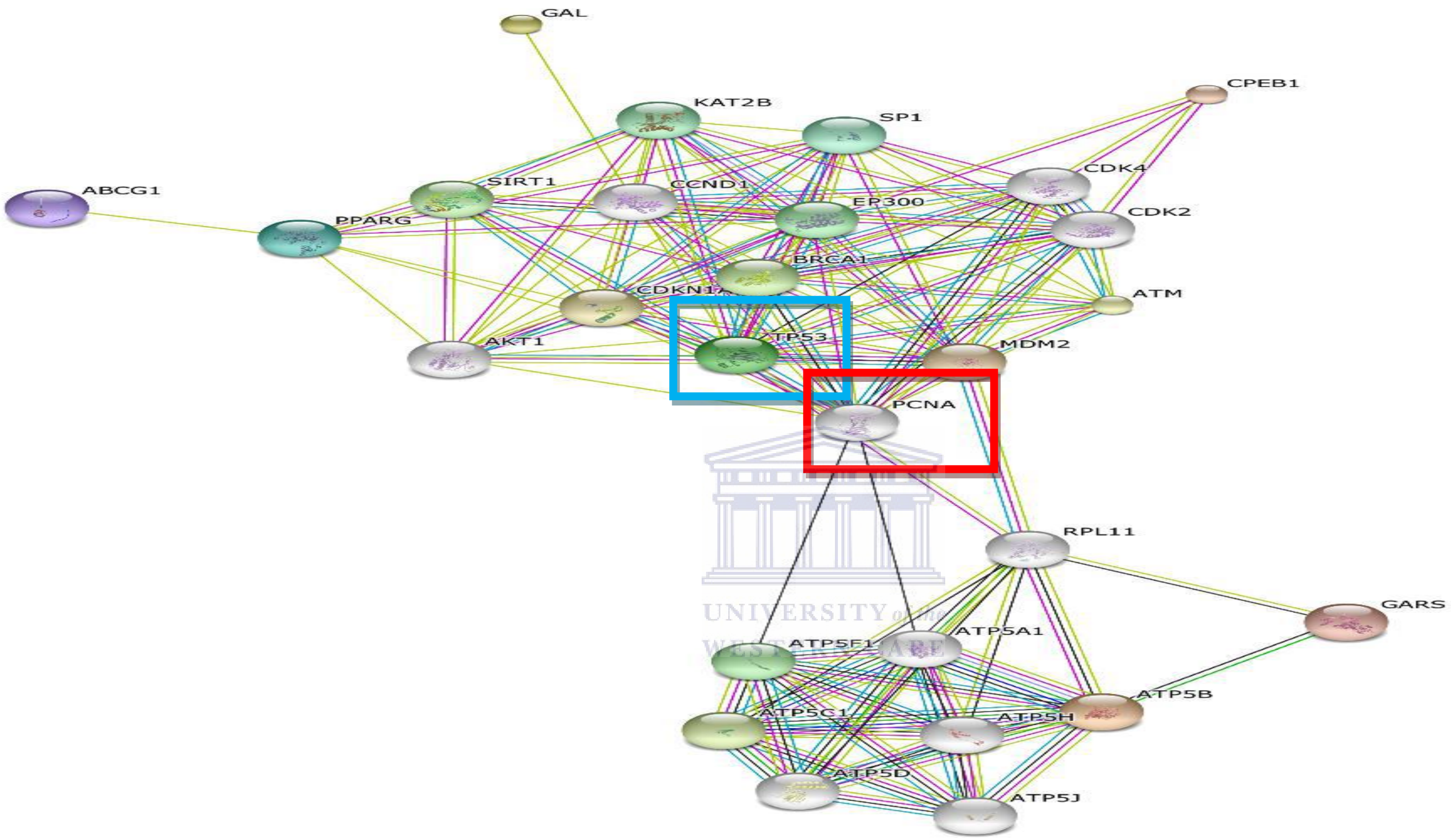
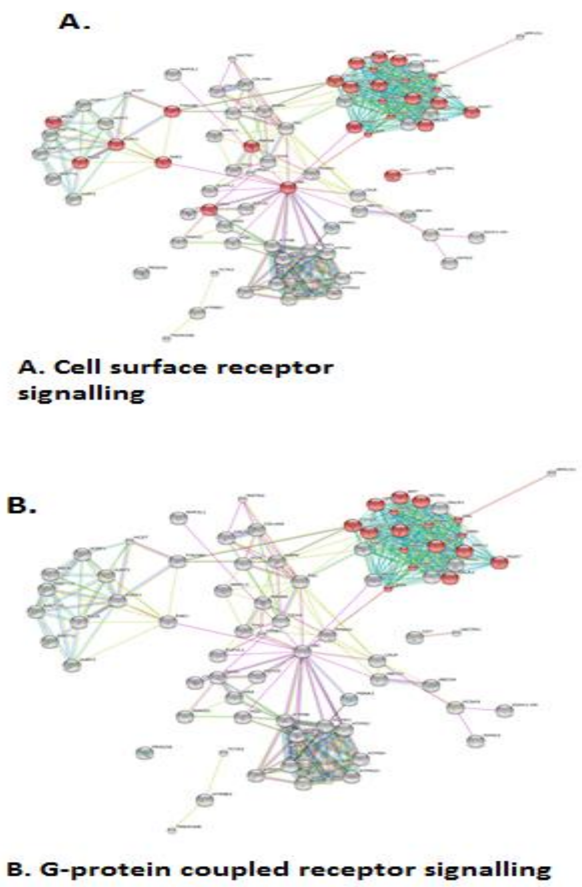
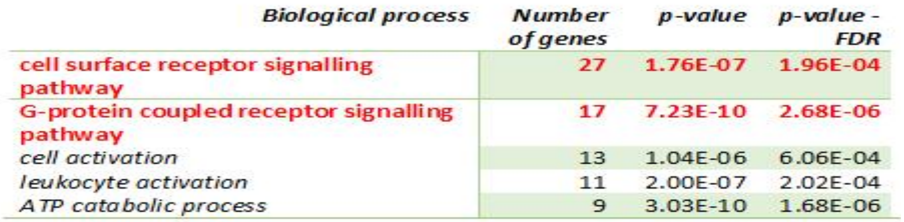


Figure 3.14: The ten putative genes and their corresponding 3 cervical cancer-associated transcription factors were all linked to Proliferating cell nuclear antigen (PCNA) (PCNA is shown in the red rectangle, while p53 is shown in the blue rectangle). Disconnected nodes are not shown (Adapted from STRING, 2013).

3.4.5. STRING analysis with GOI, interacting genes and TFs

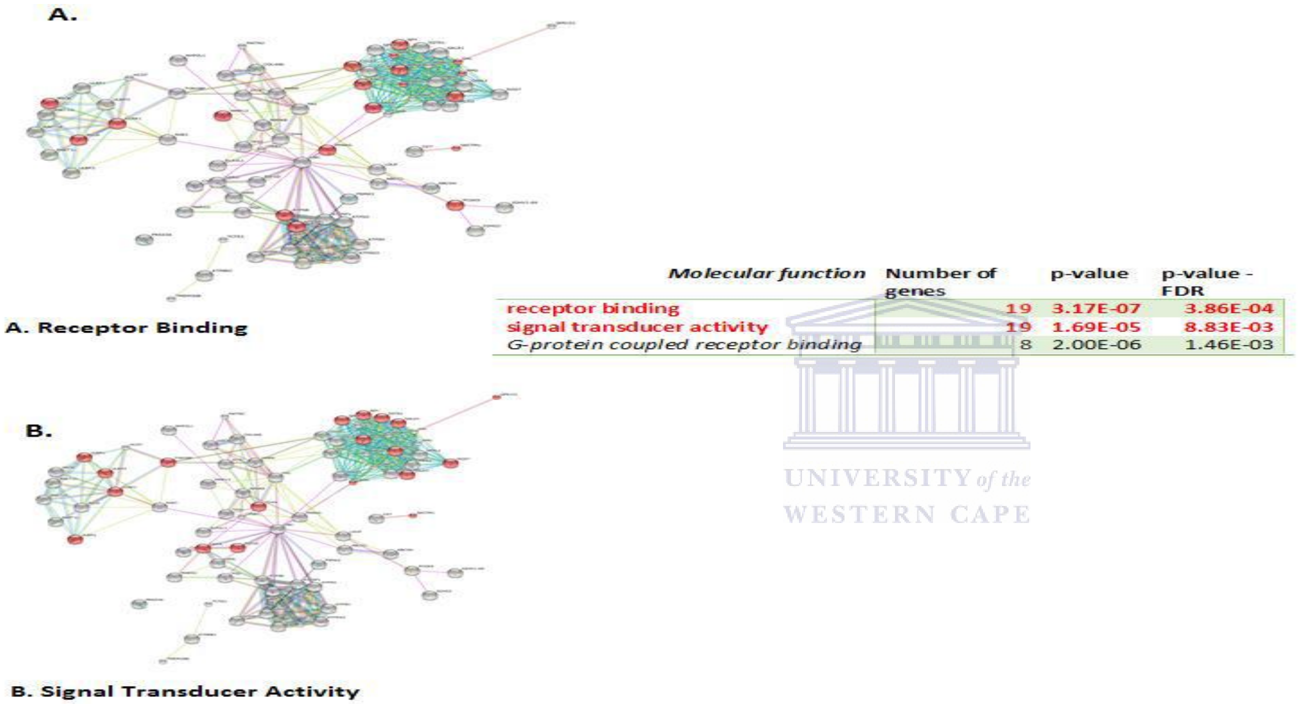
STRING explores and assigns functions of genes based on gene-gene interactions; it confirmed the results obtained by the GO analysis carried out in DAVID. The categories enriched with the highest number of genes were selected. STRING showed that of the 79 genes, 27 were enriched for cell surface receptor signalling pathways, while 17 genes were predicted to participate in G-protein coupled receptor (GPCR) signalling when clustered according to biological process (Figure 3.15). The molecular functions showed corresponding categorization with the majority of the genes involved in receptor binding (19 genes) and signal transducer activity (19 genes) (Figure 3.16). The highest proportion of genes was found in the plasma membrane component of the cell (Figure 3.17). The logical and systematic next step was to carry out a pathway analysis on the genes of interest.


A.
A. Cell surface receptor signalling
B.
B. G-protein coupled receptor signalling



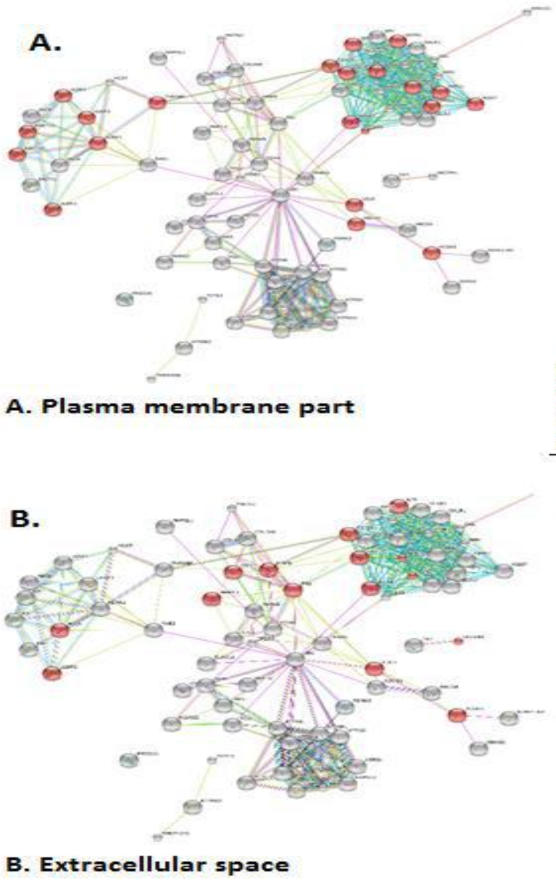
Biological process	Number of genes	p-value	p-value - FDR
cell surface receptor signalling pathway	27	1.76E-07	1.96E-04
G-protein coupled receptor signalling pathway	17	7.23E-10	2.68E-06
cell activation	13	1.04E-06	6.06E-04
leukocyte activation	11	2.00E-07	2.02E-04
ATP catabolic process	9	3.03E-10	1.68E-06

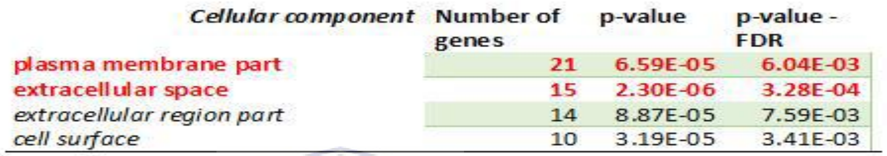
Figure 3.15: Distribution of genes according to their biological process. The images on the left show the genes (red) involved in cell surface receptor signalling (A) and G-protein coupled receptor signalling (B). The partial table on the right shows the number of genes predicted to participate in the most enriched biological process and their corresponding statistical values (Adapted from KEGG, 2013).



Molecular function	Number of genes	p-value	p-value - FDR
receptor binding	19	3.17E-07	3.86E-04
signal transducer activity	19	1.69E-05	8.83E-03
G-protein coupled receptor binding	8	2.00E-06	1.46E-03

Figure 3.16: Distribution of genes according to molecular function. The images on the left show the genes (red) involved in receptor binding (A) and signal transducer activity (B). The partial table on the right shows the number of genes predicted to participate in the most enriched molecular functions and their corresponding statistical values (Adapted from KEGG, 2013).





Cellular component	Number of genes	p-value	p-value - FDR
plasma membrane part	21	6.59E-05	6.04E-03
extracellular space	15	2.30E-06	3.28E-04
extracellular region part	14	8.87E-05	7.59E-03
cell surface	10	3.19E-05	3.41E-03

Figure 3.17: Distribution of genes according to their cellular component. The images on the left show the genes (red) localized in the plasma membrane (A) and the extracellular space (B). The partial table on the right shows the number of genes predicted to be found in the most enriched cellular components and their corresponding statistical values (Adapted from KEGG, 2013).

3.4.6. Pathway Analysis

When a KEGG pathway analyses was carried out in STRING, by inputting the ten putative genes, their directly interacting genes (which totalled 79 genes) and TFs, most of the genes were mapped to the Natural killer cell-mediated cytotoxicity (11 genes) and Oxidative phosphorylation pathways (9 genes) for FDR<0.01) as shown in figure 3.18, the most enriched pathways were selected. These enriched pathways were confirmed by DAVID, which additionally mapped 18 genes to signalling by GPCR. It was interesting to note that many of the genes in STRING were also mapped to the mitochondrial proton-transporting adenosine triphosphate (ATP) synthase complex. Signalling pathways (previously associated with the GOI) modulate ATP production via mitochondrial oxidative phosphorylation (Fosslien, 2008). Altered signal transduction is known to directly affect mitochondrial proteins, while mitochondrial dysfunction has been highlighted as one of the recurrent features of neoplastic cells (Solaini *et al*., 2010). Hence, many of the GOI were predicted to participate in diverse cancer metabolic pathways.

3.4.7. A Gene Profile for cervical cancer diagnosis

However, no single biomarker has to date been effective in the diagnosis or treatment of cervical cancer (Folgueira *et al*., 2005); most likely because cancers are complex diseases which are often also multigenic in nature. Several studies have revealed that considering a gene profile (a combination of many genes) may be more effective in implicating/diagnosing a specific disease, as opposed to using a single gene common to many biological processes (Folgueira *et al*., 2005). Although the number of genes enriched for these pathways seem small, if even one of the candidate genes can be positively implicated in a cervical cancer pathway and its co-expression with another candidate gene as well as transcriptional co-regulation by PPARγ can be verified, then a possible gene profile for cervical cancer can be established. Since the probability that a set of the same genes can be associated with all cancers is very low, combining the GOI as biomarkers in a cervical cancer diagnostic tool can be very promising. Bioinformatics have been invaluable in predicting genes and their functions that could previously not be identified using quantitative genetics data (Guan *et a*l., 2012). However, since microarray data and Bioinformatics algorithms in general may be affected by “noise” that may generate errors, this data must be empirically verified (Rajeevan *et al*., 2001).

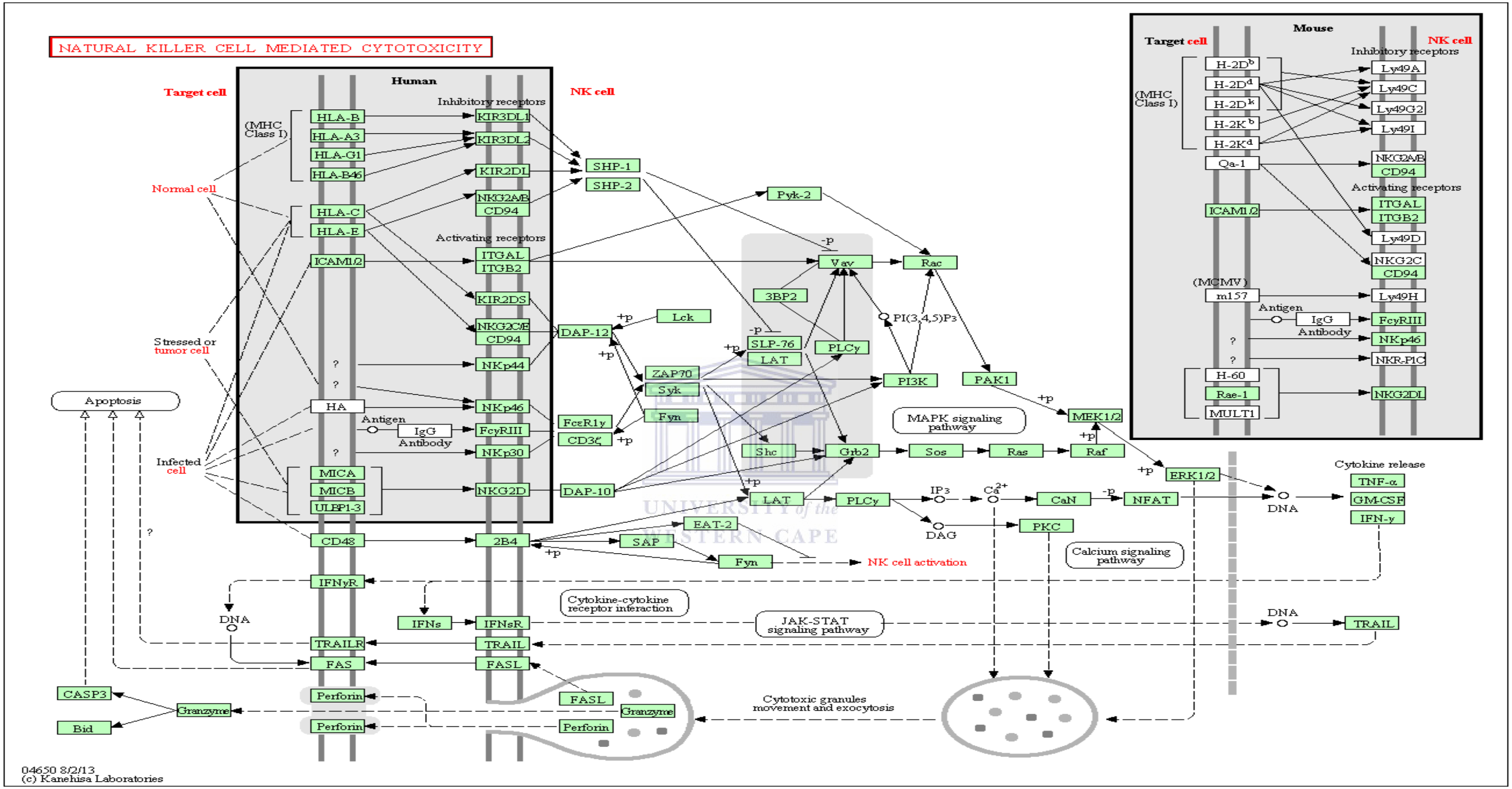


Figure 3.18: The Natural Killer cell-mediated cytoxocity was a one of the most enriched pathways (Adapted from KEGG, 2013).

3.5. Discussion and Conclusion

Understanding how genes are expressed and regulated in various tissues or under various conditions can help elucidate the molecular mechanisms of tissue development and function (Liu *et al*., 2008). The changes in gene expression can be identified or evaluated by *in silico* gene expression analysis. Gene expression profiling is a technology that is used to identify genes that are active in a sample of tissues or cells, for both 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 (Thomas *et al*., 2013). The availability of large amounts of sequence data, coupled with the advances in computational biology provides an ideal framework for *in silico* gene expression analysis (Murray *et al*., 2007). Although *in silico* gene identification remains a difficult task, public (freely available) and private (subscription) expressed sequence tag (EST) databases represent an important source for biomarker or target discovery (Terstappen and Reggiani, 2001). These databases contain short sequence information of expressed genes; this leads to their identification and is indicative of the encoded proteins (Marra *et al*., 1998).

The analysis of gene expression patterns derived from large EST databases have become a valuable tool in the discovery of prognostic and diagnostic markers. Sequence data derived from a variety of cDNA libraries provides a wealth of information for identifying genes that can be used for the development of pharmaceutical products as well as potential diagnostic biomarkers (Fannon, 1996). The tissue specificity of a gene is a measure of the relative distribution of gene expression across major tissue types in the human body. Tissue-specific genes enable an assay to detect small increases in the serum protein levels which can be unambiguously attributed to a neoplastic lesion or disease onset in the affected organ (Vasmatzi *et al*., 2007). According to Vasmatzi *et al* (2007), the discovery of novel serum biomarkers must not only identify differentially expressed genes encoding products with selected cellular localization but should also identify genes with high specificity in the tissue type of interest. The 29 remaining genes were further investigated for tissue-specificity in TiGER and GeneHub-GEPIS to identify genes specific to cervical cancer.

After the cross-cancer analysis, only ten putative genes with sufficient specificity remained (Figures 3.4-3.13). Guan *et al* (2012) demonstrated that prediction performance is significantly improved when incorporating tissue-specific networks as opposed to global functional data. By accounting for tissue-specificity one can identify more accurate candidate disease genes (Guan *et al*., 2012) and bypass redundant laboratory work (Zang *et al*., 2004). Bioinformatics analysis investigating potential protein-protein interactions and cellular pathways were performed for the ten candidate genes. This was done to identify if there are any commonalities between the genes products and if they are involved in a cancer pathway or pathways related to cancer such as apoptosis and cell cycle regulation. Based on the protein interactions between the genes; some genes can be indirectly implicated in pathways based on the fact that these genes interact with proteins involved in those particular pathways. Proteins have the ability to form a wide range of direct and indirect interactions with each other that can be conceptualized as networks. Analysing genes as a network increases its statistical power in human genetics and can assist in predicting diseased phenotypes (Szklarczyk *et al*., 2011).

The interactions that are generated can be conceptualized as networks; this allows the genome to be seen as more than a static collection of distinct genomic functions (Skrabanek *et al*., 2008). STRING provides a source for all functional links between proteins. STRING's main strengths are that it has a unique comprehensiveness; it provides a confidence score and an interactive user interface (Szklarczyk *et al*., 2011). In STRING analysis only one interaction was observed for the ten putative genes and since not much information was drawn from the protein-protein interaction studies, further *in silico* analysis was conducted. Network knowledge can give rise to understanding the biological function and dynamic behaviour of cellular systems, generating biological hypothesis about putative biomarkers, therapeutic targets or deregulated pathways in cancer. Cancer-related proteins have a higher ratio of promiscuous structural domains, making them more prone to interact with other proteins. In fact, they have a large number of interacting proteins and occupy a central position in the networks (Sanz-Pamplona *et al*., 2012). Gene co-expression network analysis identifies groups of genes highly correlated to each other in expression levels across multiple samples. Genes that are functionally related to each other are believed to express similarly and thus have high correlations between their expression profiles. So, functionally similar genes group together in gene co-expression networks. In this study 90% of the genes were shown to be co-expressed. Co-expression studies are conducted because there is evidence that co-expressed genes may be functionally related, for instance, genes encoding the various subunits of a complex protein will have similar expression patterns (Heyer *et al.*, 1999). According to Mostafavi *et al* (2008), two genes are linked if their expression levels are similar across a specific condition in a gene expression study. The process of regulation of gene transcription is controlled by a group of regulators called transcription factors (TFs). TFs facilitate the final steps in the relay of information from the cell surface to the nucleus and the gene. This is accomplished by the interaction of the TFs with specific DNA elements, these elements are usually situated upstream of the sequence that encodes the gene (Eckert *et al*., 2013). Transcription factors are cellular components that regulate gene expression and their activities subsequently control cell function and cellular response to the environment (Vaquerizas, 2009).

A constructive approach to establishing a gene regulatory network is to identify the regulatory components such as the TFs that may induce the expression of a set of co-expressed genes underlying a biological process or diseased phenotype (Whitfield *et al*., 2012). If the TFs regulating the candidate genes are known to be oncogenic, the genes could possibly be associated with cancer. In order to establish a connection between the putative genes through a common regulatory element, an analysis was conducted so as to determine if all the ten putative genes shared a common transcription factor. According to Jung *et al* (2005), PPARγ shown to be associated with most of the candidate genes, is down-regulated in newly transforming cervical neoplasia. The genes PPARγ regulates can therefore possibly serve in the early diagnosis of cervical cancer. Interestingly, a link between PPARγ and C/EPB was established by demonstrating that a decrease in PPARγ also facilitated a decrease in the levels of C/EPB. C/EBPα is a differentiation-inducing transcription factor. Furthermore, loss of function or inactivation of C/EBPα is commonly associated with squamous cell line cancers and most of the cells affected in cervical cancer are squamous cells (Koschmieder, 2009). This could therefore further contribute to the specificity of the candidate biomarkers for cervical cancer (which was already demonstrated by the tissue specificity analysis).

Similarly, the HPV oncoproteins E6 has been shown to neutralize p53, effectively reducing its levels and its subsequent ability to effectively suppress tumour formation (Tommasino *et al*., 2003). The rationale of this study is to find biomarkers that can detect cervical cancer in its pre-invasive stage and the HPV oncogenes E6 and E7 are responsible for many of the changes that occur when the neoplastic tissue forms in the cervix. Therefore establishing which genes and/or transcription factors are targeted by these oncogenes, it will be possible to use those genes/transcription factors or a combination thereof to design a diagnostic tool specific for cervical cancer detection. By performing a network-based analyses based on gene variants and their interacting genes, one can explore how sub-network level features contribute to the phenotype of a complex disease such as cancer (Okser, 2013). Signalling pathways are activated by extracellular proteins (ligands), which bind their specific cell surface receptors which dimerize or oligomerize at the cell surface to begin the intracellular signalling phase (Darnell Jnr, 2002). GPCR's constitute the largest family of cell surface molecules responsible for converting extracellular stimuli into intracellular signals and have emerged as key players in tumourigenesis and metastasis (Dorsam and Gutkind, 2007).

Events such as signal transduction are often targeted by oncogenes to sustain growth signals for uncontrolled proliferation (Chial, 2008). Hence, it is highly likely that the GOI are targeted by the HPV E6 and E7 oncogenes for dysregulation. TFs participate at the ends of signal transduction and stress-response pathways by up or down regulating certain genes. All primary and 39 modifier genes leading to cancer partake in at least one of these two pathways (Nebert, 2002). Since the annotated molecular function of the GOI also predicted their participation in one of the critical cancer pathways (signal transduction), this further strengthens their possible implication in cancer-related pathways. However, this does not mean that they can be easily detected in bodily fluids, which is necessary for a biomarker based test to be non-invasive. Hence, their cellular localization was also considered. More specifically, most genes were predicted to be secreted into the extracellular space or to be localized in the extracellular region and on the cell surface. This is very promising since the targeted genes are those that facilitate the entry of cervical cancer biomarkers into the circulatory system. Nevertheless, events such as cell surface binding and signal transduction are normal biological processes that do not generally translate into cancer. Although, it is worthwhile to reiterate that 90% of the GOI's were shown to be co-expressed, while all of the

GOI were regulated by the same element showing possible co-regulation and co-expression. As previously stated, the general consensus is that there is likely to be a relationship between co-expression and co-regulation and that co-expressed and/or co-regulated genes are likely to function in the same metabolic pathway (Emmert-Streib and Glazko, 2011). By considering pathways one can observe gene sets that function in a coordinated manner to define a biological process (Emmert-Streib and Glazko, 2011). KEGG stores higher order functional information in the form of pathways showing graphical representations of cellular processes and metabolic processes (Kanehisa and Goto, 2000). Natural killer cells are known to target foreign and cancer cells for apoptosis via multiple mechanisms. If any of these genes from this pathway are mutated or deregulated it could therefore negatively influence their ability to induce cancer cells to undergo apoptosis. Similarly the genes involved in oxidative phosphorylation could make for ideal targets for oncogenes since a wide spectrum of oxidative phosphorylation deficit has been associated with tumourigenesis (Solaini *et al*., 2010).

This study managed to discover and identify putative biomarkers to be further validated using molecular methodologies. Combining biological data mining, text mining and *in silico* gene enrichment techniques proved to be effective in classifying genes and linking them to cervical cancer. The advent of microarray based technology has helped study the expression patterns of more than 40,000 genes at a time. Several groups have used microarray based technology to look for differentially expressed genes in the different stages of cervical tumourigenesis. Few studies have followed up and validated the microarray data in a large number of genes (Rajkumar *et al*., 2011). However, *in silico* methods need to be coupled with *in-vitro* work to confirm the in-silico approach and also establish the confidence of the identified putative markers in a biological system.

3.6 References

1. Bader, G.D., Betel, D., and Hogue, C.W. 2003. BIND: the Biomolecular Interaction Network Database, *Nucleic Acids Res*, 31:248-250.

2. Branca, M., Ciotti, M., Giorgi, C., Santini, D., Di Bonito, L., Costa, S., Benedetto, A., Bonifacio, D., Di Bonito, P., Paba, P., Accardi, L., Syrjanen, S., Favalli, C., and Syrjanen, K. 2006. Up-regulation of proliferating cell nuclear antigen (PCNA) is closely associated with high-risk human papilloma virus (HPV) and progression of cervical intraepithelial neoplasia (CIN), but does not predict disease outcome in cervical cancer. *European Journal of Obstetrics and Gynecology and Reproductive Biology,* 130(2):223-231.

3. Chial, H. 2008. Proto-oncogenes to Oncogenes to Cancer. Nature Education. Available from http://www.nature.com/scitable/topicpage/proto-oncogenes-to-oncogenes-to-cancer-883.

4. Darnell Jnr, J.E. 2002. Transcription Factors as Targets for Cancer Therapy. *Nature Reviews*, 2:740-749.

5. Dorsam, R.T., and Gutkind, S. 2007. G-protein-coupled receptors and cancer. *Nature Reviews Cancer*, 7:79-94.

6. Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C., Krawetz, S.A. 2003. Global functional profiling of gene expression. *Genomics*, 81:98-104.

7. Eckert, R.L., Adhikary, G., Young, C.A., Jans, R., Crish, J.F., Xu, W., and Rorke, E.A. 2013. AP1 transcription factors in epidermal differentiation and skin cancer. *Journal of Skin Cancer*, 1-9.

8. Emmert-Streib, F., and Glazko, G.V. 2011. Pathway Analysis of Expression Data: Deciphering Functional Building Blocks of Complex Diseases. *PLoS Biology*, 7(5).

9. Fannon, M. R. 1996. Gene expression in normal and diseased states- identification of therapeutic targets. *Trends in Biotechnology*, 14(8):294-298.

10. Folgueiro, M.A.A.K., Carraro, D.M., Brentani, H., da Costa Patrao, D.F., Barbosa, E.M., Netto, M.M., Caldeira, J.R.F., Katayama, M.L.H., Saoros, F.A., Oliveira, C.T.,

Reis, L.F.L., Kaiano, J.H.L., Camargo, L.P., Vencio, R.Z.N., Snitcovsky, I.M.L., Makdissi, F.B.A., e Silva, P.J.D., Goes, J.C.G.S., and Brentani, M.M. 2005. Gene Expression Profile Associated with Response to Doxorubicin-Based Therapy in Breast cancer. *Clinical Cancer Research*, 11(20):7434-7443.

11. Fosslien, E. 2008. Cancer Morphogenesis: Role of Mitochondrial Failure. *Annals of Clinical and Laboratory Science,* 38(4):307-330.

12. Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L.J. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research,* 41: 808–815.

13. Glaab, E., goel Baudot, A., Krasnogor, N., Schneider, R., and Valencia, A. 2012. Enrichment: network-based gene set enrichment analysis. *Bioinformatics,* 28:451-457.

14. Goel, M., Somani, K., Mehrotra, A., Singh, U., and Mehrota, R. 2011. Immunohistochemical Expression of Cell Proliferating Nuclear Antigen (PCNA) and p53 Protein in Cervical Cancer. *Journal of Obstetrics and Genecology of India*, 62(5): 557-561.

15. Guan, Y., Gorenshteyn, D., Burmeister, M., Womg, A.K., Schimenti, J.C., Handel, M.A., Bult, C.J., Hibbs, M.A., and Troyanskaya, O.G. 2012. Tissue-Specific Functional Networks for Prioritizing Phenotypes and Disease Genes. *PloS Computational Biology*, 8(9).

16. Han, S., Inoue, H., Flowers, L.C., and Sidell, N. 2003. Control of COX-2 Gene Expression through Perixosome Proliferato-Activated Receptor in Human Cancer Cells. *Clinical Cancer Research*, 9:4623-4635.

17. Hastie, T., Tibshirani, R.., Friedman, J.H. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2 editions.

18. Heyer, L.J., Kruglyak, S., and Yooseph, S. 1999. Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome Research*, 9:1105-1115.

19. Huang, D.W., Sherman, B.T. and Lempicki, R.A. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Access*, 37(1):13.

20. Jung, T.I., Baek, W.K., Suh, S.I., Jang, B.C., Song, D.K., Bae, J.H., Kwon, K.Y., Bae, J.H. Cha, S.D., Bae, I., and Cho, C.H. 2005. Down-regulation of peroxisome proliferator-activated receptor gamma in human cervical carcinoma. *Gynecological Oncology*, 97(2):365-373.

21. Kanehisa, M., and Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acid Research,* 28(1):27-30.

22. Koschmieder, S., Halmos, B., Levantini, E., and Tenan, D.E. 2009. Dysregulation of the C/EBP Differentiation Pathway in Human Cancer. *Journal of Clinical Oncology*, 27(4):619-629.

23. Liu, X., Yu, X., Zack, D.J., Zhu, H., and Qian, J. 2008. TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, 9(271):1-7.

24. Marra, M., Hiller, L., and Waterston, R.H. 1998. Expressed sequence tags-ESTablishing bridges between genomes. *Trends in Genetics*, 14: 4-7.

25. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. and Morris, Q. 2008. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, 9(S4). Accessed 21/10/2013. Available from http://genemania.org/pdf/Mostafavi.pdf.

26. Murray, D., Doran, P., MacMathuna, P. and Moss, A.C. 2007. *In silico* gene expression analysis- an overview. *Molecular Cancer*, 6(50).

27. Nebert, D.W. 2002. Transcription factors and cancer: an overview. *Toxicology*, 181(182):131-141.

28. Okser, S., Pahikkala, T., and Aittokallio. 2013. Genetic Variants and their interactions in disease prediction-machine learning and network perspectives. *BioData Mining*, 6(5).

29. Paunesku, T., Mittal, S., Protic, M., OrYhon, J., Korelov, S.V., Joachimiak, A., and Woloschak, G.E. 2001. Proliferating cell nuclear antigen (PCNA): ringmaster of the genome. *International Journal of Radiation Biology*, 77(10):1007-1021.

30. Pavlopoulos, G.A., Wegener, A.L., and Schneider, R. 2008. A survey of visualization tools for biological network analysis. *BioData Mining* 1:12.

31. Rajeevan, M.S., Ranamukhaarachchi, D.G., Vernon, S.D., and Unger, E.R. 2001. Use of Real-Time Quantitative PCR to Validate the Results of cDNA and Differential Display PCR Technologies. *Methods*, 25:443-451.

32. Rajkumar, T., Sabitha, K., Vijayalakshmi, N., Shirley, S., Bose, M.V., Gopal, G., and Selvaluxmy, G. 2011. Identification and validation of genes involved in cervical tumourigenesis. *BMC Cancer*, 11(80):1-14.

33. Sanz-Pamplona, R., Berenguer, A., Sole, X., Cordero, D., Crous-Bou, M., Serra-Musach, S., Guinó, E., Ángel Pujana, M., Moreno, V. 2012. Tools for protein-protein interaction network analysis in cancer research. *Clin Transl Oncol*, 14:3-14.

34. Skrabanek, L., Saini, H.K., Bader, G.D., and Enright, A.J. 2008. Computational prediction of protein-protein interactions. *Molecular Biotechnology*. 38:1-17.

35. Smith, W.M., Zhou, X.P., Kurose, K., Gao, X., Latif, F., Kroll, T., Sugano, K., Cannistra, S.A., Clinton, S.K., Maher, E.R. Prior, T.W., and Eng, C. 2001. Opposite Association of two PPARG variants with cancer: overrepresentation of H449H in endometrial carcinoma cases and underrepresentation of P12A in renal cell carcinoma cases. *Human Genetics*, 109(2):146-151.

36. Solaini, G., Sqarbi, G., and Baracca, A. 2010. Oxidative phosphorylation in cancer cells. *Biochimica et biophysica acta*, 1807(6):534-542.

37. Stelzer, G., Dalah, I., Stein, T.I., *et al*. 2011. In-silico human genomics with GeneCards. *Human Genomics*, 5(6):709-717.

38. Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L.J., and von Mering, C. 2011. The STRING database in 2011: functional networks of protein globally integrated and

scored. *Nucleic Acid Research*, 39. Accessed on 20/09/2013. Available from http://nar.oxfordjournals.org/content/39/suppl_1/D561.full.pdf+html.

39. Terstappen, G.C., and Reggiani, A. 2001. In silico research in drug discovery. *Trends in Pharmacological Sciences*. 22(1): 23-26.

40. Thomas, M., Poignee-Heger, M., Weisser, M., Wessner, S., and Belousov, A. 2013. An optimized workflow for improved gene expression profiling for formalin-fixe, paraffin-embedded tumor samples. Journal of Clinical Bioinformatics. 3(10):1-11.

41. Tommasino, M., Accardi, R., Caldeira, S., Dong, W., Malanchi, I., Smet, A., Zehbe, I. 2003. The role of TP53 in cervical carcinogenesis. *Hum Mutat*, 21:307-312.

42. Vaqeurizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. 2009. A census of human transcription factors: function, expression, evolution. N*ature Reviews Genetics,* 10:252-263.

43. Vasmatzis, G., Klee, E.W., Kube, D.M., Therneau, T.M., and Kosari, F. 2007. Quantitating tissue specificity of human genes to facilitate biomarker discovery. *Bioinformatics*, 23(11):1348-1355.

44. Von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., and Bork, P. 2005. STRING: known and predicted protein–protein associations integrated and transferred across organisms. *Nucleic Acids Research*, Vol. 33, Database issue D433–D437.

45. Whitfield, T.W., Wang, J., Collins, P.J., Partridge, P.J., Christopher, E., Aldred, S.F., Trinklein, N.D., Myers, R.M., and Weng, Z. 2012. Functional analysis of transcription factor binding sites in human promoters. *Genome Biology*, 13:R50

46. Zang, Y., Eberhard, D.A., Frantz, G.D., Dowd, P., Wu, T.D., Zhou, Z., Watanabe, C. Luoh, S-M., Polakis, P., Hillan, K.J., Wood, W.I., and Zang, Z. 2004. GEPIS- quantitive gene expression profiling in normal and cancer tissues. *Bioinformatics*, 20(15):2390-2398.

47. Zhang, Y., Luoh, S-M., Hon, L.S., Baertsch, R., Wood, W.I., and Zhang, Z. 2007. GeneHub-GEPIS: digital expression profiling for normal and cancer tissues based on an integrated gene database. *Nucleic Acids Research*, 35:152-158.

CHAPTER 4: Differential Expression Analysis of Putative Biomarkers Using Molecular Techniques

4.1. Background

Gene expression analysis is increasingly important in various biological research fields. Understanding the patterns of expressed genes is expected to provide insight into the complex regulatory networks and will probably lead to identification of genes implicated in disease. Quantifying gene expression levels can yield valuable clues about the function of a gene, for instance, accurate measurements of gene expression can identify the type of cells or tissues where a particular gene is expressed, reveal individual gene expression levels in defined biological states and detect alterations in gene expression levels in response to specific biological stimuli (Fraga *et al*., 2008). This section of the study is aimed at validating the ten putative genes identified in chapter 2 by *in silico* expression analysis. This was accomplished by means of molecular expression profiling using qPCR in cervical cancer cell lines, other types of cancer as well as non-cancerous cell lines.

4.1.1. Quantitative real-time PCR (qPCR)

Real-time polymerase chain reaction (PCR) has gained popularity in the past few years and has become the most widely used technique in modern molecular biology. This technique depends on fluorescence-based detection of amplicon DNA and permits the kinetics of PCR amplification to be monitored in real time, making it possible to quantify nucleic acids with extraordinary ease and precision (Guescini *et al*., 2008). Quantitative real-time PCR (qPCR) has become a very versatile technique to examine expression changes of one or more genes of interest in various pathological states such as cancer. This method offers a broad range of advantages over standard methods such as Northern blot and semi-quantitative PCR due to its specificity, sensitivity, simplicity, costs and high-throughput. Thus, it has become the most emerging tool for absolute and relative quantification of mRNA transcription levels (Jacob *et al*., 2013). This technique is very sensitive for the detection and quantification of gene expression levels particularly for low abundance mRNA in tissues that have low mRNA concentrations and to elucidate small changes in mRNA expression levels (Pfaffl *et al*., 2002). PCR can be broken into three major steps (Figure 4.1): exponential phase, log-linear phase and plateau phase. During the first 10-15 cycles (linear ground phase), the PCR is at the initiation stage and emission of fluorescence at each cycle has not risen above background and during this time the baseline fluorescence is calculated. At the exponential phase, the amount of fluorescence has reached a threshold where it is significantly higher than the background levels. In the log-linear phase, the PCR reaches its optimal amplification period with the PCR product doubling after every cycle in ideal reaction conditions. Lastly, the plateau stage is reached when reaction components become limited and the fluorescence intensity is no longer useful for data calculation (Wong and Medrano, 2005).

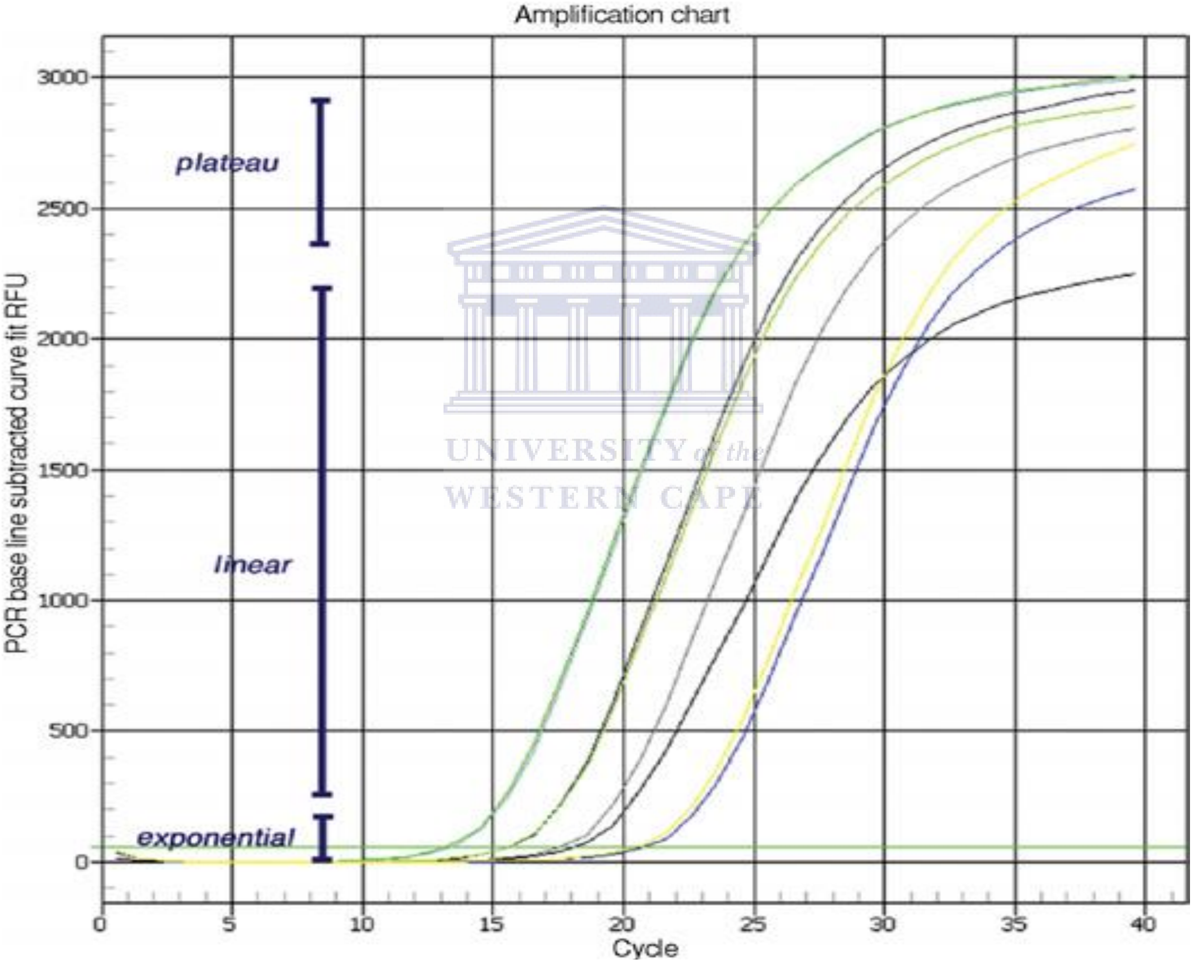


Figure 4.1: Graph illustrating the phases of the PCR amplification curve, adapted from Fraga *et al.*, 2008.

4.1.2. Quantitation Strategies in qPCR

There are two strategies used to quantify gene expression in qPCR viz.: absolute quantification and relative quantification. Absolute quantitation relies on a standard curve which is generated by using serially diluted standards of known concentration. The standard curve produces a linear relationship between the cycle threshold (Ct) and initial amounts of total RNA or cDNA, allowing the determination of the concentration of unknowns based on their Ct values. The Ct is defined as the number of cycles required for the fluorescent signal to cross the threshold (i.e. exceeds background level). Ct levels are inversely proportional to the amount of target nucleic acid in the sample (i.e. the lower the Ct level the greater the amount of target nucleic acid in the sample) (Wong and Medrano, 2005). In relative quantitation strategy changes in gene expression are measured based on either an external standard or a reference sample known as a calibrator. When a calibrator is used, the results are expressed as a target reference ratio (Wong and Medrano, 2005). This method depends on the comparison between expression of a target gene versus a reference gene and the expression of the same gene in target sample versus reference samples (Yuan *et al.*, 2006). This chapter aimed at testing the ten target genes identified in chapter 2 by *in silico* methodology, by generating expression profiles across an array of cancer cell lines in order to identify putative biomarker significantly expressed in cervical cancer.

This was achieved by utilising various molecular techniques such as cell culture, RNA extraction, synthesis of cDNA and qPCR. In order to calculate the expression of a target gene in relation to a suitable reference gene, various mathematical models have been established. These models can ascertain relative expression levels either without real-time PCR efficiency correction as shown in equation 4.1 or with kinetic PCR efficiency correction in equations 4.2 and 4.3 (Pfaffl, 2001). Existing models are capable of determining a single transcription difference between one control and one sample, e.g. LightCycler Relative Quantification Software, or permit a group-wise comparison of up to one hundred samples, e.g. Relative Expression Software Tool (REST) and REST-XL (Pfaffl, 2001).

Equation 4.1

$$R = 2^{-[\Delta Cp\text{ sample} - \Delta Cp\text{ control}]}$$

$$R = 2^{-\Delta\Delta Cp}$$

Equation 4.2

$$\text{Ratio} = \frac{(\text{Etarget})\ \Delta^{\text{Cp target (control – sample)}}}{(\text{Eref})\quad \Delta^{\text{Cp ref (control – sample)}}}$$

Equation 4.3

$$\text{Ratio} = \frac{(E\text{target})^{\Delta \text{C}p \text{ target (}MEAN\text{ control – }MEAN\text{ sample)}}}{(E\text{ref})^{\Delta \text{C}p \text{ ref (}MEAN\text{ control – }MEAN\text{ sample)}}}$$

The relative expression ratio of a target gene is calculated on the basis of its actual real-time PCR efficiency (*E*) or on a static *E* of 2, and the crossing point difference (ΔC*p*) of one unknown sample versus one control. The relative computation procedure, using either REST or REST-XL, is based on the mean C*p* of the investigated groups (Equation 4.3) (Pfaffl, 2004).

4.1.3. Amplification Efficiency

An imperative consideration when doing relative quantification is amplification efficiency of the PCR reaction. Techniques that have been previously used to calculate gene expression have been established on the hypothesis that the amplification efficiency of the reaction is ideal, indicating the doubling of the PCR product concentration during each cycle within the exponential phase of the reaction (Gibson *et al*., 1996). Even so, various PCR reactions lack perfect amplification efficiencies and calculations without an appropriate correction factor may substantially miscalculate the initial concentration (Liu and Saint, 2002). Generally, the reactions amplification efficiency is calculated using data collected from a relative standard curve with equation 4.4 (Rasmussen, 2001).

Equation 4.4 Efficiency

(E) = [10(-1/slope)]-1

4.1.4 Data Evaluation

The Pfaffl model (2002) calculates gene expression by combining gene quantification and normalisation into a single calculation. The model makes use of amplification efficiencies of the target and reference genes which are also used for normalisation to correct for differences between two assays. The Pfaffl method makes use of excel based software known as REST® which automates data analysis using this model. REST uses a pairwise Fixed Reallocation Randomisation test to calculate result significance and is also able to indicate if the reference gene is suitable for normalisation. Expression ratios are calculated by the REST software based on Ct values of the target gene relative to the reference gene and based on the results, the software generates a plot of the expression ratios for all samples involved (Wong and Medrano, 2005).

The aim of this chapter was to validate the putative biomarkers that were identified using an *in silico* approach described in chapter 2. This was accomplished by using various molecular techniques including cell culture, mRNA extraction cDNA synthesis and qPCR.

4.2. Materials and Methods

Cell Culture Media and Reagents

Dulbecco's Minimal Essential Medium (DMEM)	Invitrogen
Dimethyl Sulphoxide (DMSO)	Sigma
Eagle's Minimal Essential Medium (EMEM)	Sigma
Fetal Bovine Serum (FBS)	Invitrogen
Leibovitz's L-15 Medium	Sigma
McCoy's 5a Medium Modified	Sigma
MCDB	Sigma
Phosphate Saline Buffer (PBS)	Invitrogen
Roswell Park Memorial Institute Medium (RPMI) 1640	Sigma

Materials and Suppliers

Agarose	Whitehead Scientific
Bovine Serum Albumin	Roche
Cell culture media and reagents	Invitrogen
Diethylpyrocarbonate (DEPC)	Sigma
Ethanol	Merck
Ethylene Diamine Tetra-acetic acid (EDTA)	Merck
Gel Loading Dye (6X)	Merck
Hydrochloric Acid	Fermentas

KAPA Taq extra hotstart readymix	KAPABiosystems
KAPA SYBR FAST qPCR kit	KAPABiosystems
Nuclease free water	Merck
Nucleospin® Tripep kit	Fermentas
Oligonucleotides	Macherey-Nagel
Sodium Dodecyl Sulphate (SDS)	Inqaba Biotech
SYBR® Safe DNA Gel stain	Promega
SYBR® Fast Master Mix (2x) ABI Prism	Invitrogen
Sodium Hydroxide	Lasec
Transcriptor First Strand cDNA Synthesis Kit	Merck
TEMED (N, N, N’, N’-Tetra methylethylene-diamine)	Roche
Tris [Hydroxymethyl] aminoethane (Tris)	Sigma
Tris (2-carboxyethyl) phosphine (TCEP)	Merck
Trypsin	Sigma

4.2.1. Cell Culture Approach

4.2.1.1. Cell Lines and Media

All the cell lines that were used during the course of this research were purchased from ATCC (American Type Culture Collection), shown in table 4.1. The cells were cultured in their respective growth medium per instructions of the supplier. Cell lines used in this study are adherent and semi-adherent.

Table 4.1: List of cell lines used in this research study

Cell line	Origin	Growth medium
HeLa	HPV18 cervical cancer	DMEM + 10% FBS
CaSki	HPV16 cervical cancer	DMEM + 10% FBS
HT-3	Cervix Carcinoma	McCoy's 5a + 10% FBS
A549	Lung carcinoma	EMEM + 10% FBS
MCF-7	Breast adenocarcinoma	DMEM + 10% FBS
KMST-6	Normal skin cell line	DMEM + 10% FBS
SK-OV-3	Ovarian Carcinoma	McCoy's 5a + 10% FBS
CAOV3		DMEM + 10% FBS
T-84	Colorectal adenocarcinoma	DMEM-F12 +10% FBS

4.2.1.2. Starting Cell Culture from frozen Cells:

Frozen cryovials were placed in a water bath at 37ºC, for 1-2minutes, until defrosted. Slowly, drop by drop, cells were diluted in pre-warmed media (table 4.1) in a 15 mL tube and then centrifuged for 5-10 minutes. The supernatant was removed (as much as possible); making sure the pellet was not disturbed so as to remain intact. The pellet was then ressuspended in the respective media and the entire tube's content transferred to a 25 cm^2 Flask (T25). The flask was then incubate at 37ºC with 5% CO$_2$ for 24 hours, after which all the medium was removed and replaced with new medium to ensure no DMSO remained in the cells.

4.2.1.3. Maintenance of Human Cells

A schedule of cell maintenance, feeding and passaging was adopted to maintain appropriate cell density, nutrient concentration and pH levels in cultures. Cells were best passaged when growing logarithmically, at 70 to 80 % confluency. A schedule used for routine maintenance, every 2 days medium was changed if the confluency was below 40-50% and every day until 70-80 % confluency was reached upon which the cells were passaged.

4.2.1.4. Sub-Cultivating

All the media (table 4.1) was pre-warmed to 37ºC. The culture was inspected for contamination. If no contamination was present and the cells were at 70-80% confluence, the medium was aspirated with a sterile Pasteur pipette and the cells were washed with 5 mL of PBS to remove any residual medium (~15 seconds). After the PBS was aspirated with a sterile Pasteur pipette, 1 mL of 0.05% trypsin-EDTA (TE) was added, evenly dispersed over the surface by gently rocking the flask. The flask was then placed in the incubator with the cap screwed tightly. After 2 minutes the flask was taken to the microscope to check the progress of the detachment. When the cells were detached, 5 mL of new media was added, rinsing the surface of the flask to inactivate the trypsin. The cells were collected by centrifugation at 3000 xg.

4.2.1.5. Changing Medium

Cells were fed with their respective media; the culture was inspected under the microscope to ensure no contamination was present. The old medium was aspirated from the flask with a sterile Pasteur pipette and 5 mL of new medium was added. The freezing down of cells was subsequent to trypsinisation of cells. Centrifugation was done at 3000 xg for 3 min. The supernatant was then discarded and the pellet resuspended in 90% complete medium and 10% DMSO. The resuspended cells were aliquoted as 1ml fractions into cryotubes and stored at -150^{o}C.

4.3. Analytical Laboratory Techniques

4.3.1. Extraction of RNA

Cells were prepared prior to RNA extraction as described in section 3.2.1; the cell pellet was washed with PBS and then collected by centrifugation at 3000 xg. The nucleospin kit was used to isolate RNA and proteins respectively according to the manufacturer instructions. However for the purposes of this research only RNA extraction will be the focus. The cells were lysed by adding buffer RP1 (350 µl) and 3.5 µl, of 20mM Tris (2-carboxyethyl) phosphine (TCEP) to the cell pellet and were vortex vigorously for a minute. The cells were washed with 350 µl 70% ethanol and were transferred to a new Eppendorf; the pellet was collected by centrifugation at 11 000 xg for 30sec. The silica membrane was desalted by adding 350 µl MDB and the pellet was collected by centrifugation at 11 000 xg. A 1: 100 stock solution of recombinant deoxyribonuclease (rDNAse) in reaction buffer of rDNAse was added to the RNA silica membrane of the column. Thereafter the DNA was digested by adding 95 µl DNase reaction mixture; the cells were incubated at room temperature for 15 min. The silica membrane was washed and dried by adding 200 µl RA2 (wash 1), then 600 µl RA3 (wash 2) and lastly 250 µl RA3 for 30 secs and 2min at 11 000 xg respectively. Furthermore, the RNA was eluted using 60 µl RNAse free water and centrifuged at 11000 xg for 1 minute. The concentration and quality of RNA was assessed using the Nanodrop ND-1000 spectrometer (NanoDrop Technologies) and all RNA samples were stored at -20^{0}C.

4.3.2. Agarose Gel Electrophoresis of RNA

Agarose gels were prepared by dissolving agarose powder in 1X TBE buffer prepared with Diethylpyrocarbonate (DEPC) water by heating until agarose was dissolved. The solution was cooled and 0.8X SYBR safe gel stain was added. The solution was then poured into a gel tray and the tray was placed into the electrophoresis tank filled with 1X TBE buffer. RNA samples were prepared by mixing 1 part of the RNA with 6X loading dye and heated for 2-5 minutes at 95^{0}C. After loading the gel was electrophoresed for 60 minutes at 100 V and the RNA was visualised using the UVP system from Bio-Rad.

4.3.3. cDNA Synthesis

The cDNA was synthesized using the Transcriptor First Strand cDNA synthesis kit (Roche) according to the manufacturer's instructions. All the reagents were kept on ice throughout the experiment. The template primer mixture was prepared with the following reagents in a sterile, nuclease-free, thin walled PCR tube as shown in table 4.2 to a final volume of 13µL.

Table 4.2: Reagents for Template Primer Mix

Reagent	Final Concentration
RNA	1 µg
Anchored-oligo (dT) 18 Primer	2.5 µM
PCR grade H_2O	Variable

Once all the reagents in table 4.2 were mixed, the reaction mixture was heated for 10 minutes at 65^0C in a pre-heated thermal block cycler to denature the template. The reaction mixture was immediately cooled and the remaining reagents were added as specified in the following table 4.3.

Table 4.3: Reagents for final cDNA Synthesis Mixture

Reagents	Final Concentration
Transcriptor Reverse Transcriptase reaction buffer	1 X
Protector RNAse inhibitor	20 U
Deoxynucleotide Mix	1 mM each
Transcriptor Reverse Transcriptase	10 U

Once all the reagents in table 4.3 were added, they were carefully mixed and the tube centrifuged briefly to collect the sample to the bottom of the tube. The tube was placed in the thermal block cycler and the RT reaction mixture was incubated for 33 minutes at 55^0C. The Transcriptor Reverse Transcriptase was inactivated by heating the mixture for 5 minutes at 85^0C and the tube was placed on ice in order to stop the reaction. The cDNA was quantified using the Nanodrop ND-1000 Spectrophotometer (Nanodrop Technologies). The cDNA was stored at -20^0C.

4.3.4. Primer Design

Gene specific oligonucleotides were designed for cDNA amplification and oligonucleotide sequences are shown in table 4.4. Each oligonucleotide was designed to be 20bp long using the NCBI primer design algorithm (www.ncbi.nlm.nih.gov). The oligonucleotide sequences were sent for synthesis to Inqaba biotech (http://www.inqababiotec.co.za/). Once the oligonucleotides were synthesized, a 100 µM stock solution of the oligonucleotides was prepared by re-suspending the pellet into 1x TE buffer (10 mM Tris, pH 7.5 to 8.0, 1 mM EDTA). The concentrated oligonucleotides stocks were stored at -20^0C.

Table 4.4: Oligonucleotide sequences for PCR amplification of cDNA

Primer	Forward Sequences (5'-3')	Reverse Sequence (5'-3')	length(bp)
Gene 1	GTTCTTCGATGAGCCCACCA	GCAGACTTTTCCCCGGTACA	193
Gene 2	CTAGAGGACCTGGGGACACG	CGTTGTAGGCACGGTTGTTG	288
Gene 3	GGCCACTTCGTCCACCTACT	TTCCAATTGGTCCAGGTCGT	295
Gene 4	GAACCTGGAGCGGATTACCC	AGATACACCTCCACCAGGCT'	84
Gene 5	TCCAGATATTGCCAGGGATGC	CCTCATAGGTAGCCACAGCAG	192
Gene 6	CCTGCTTCCTTTAGCGTGAAC	GGTCCTTGTCACTGGCTCTT	290
Gene 7	TAGCTCTGACTGGGCTGACT	TAGCTCTGACTGGGCTGACT	289

Gene 8	GCCAAGGAAAAACGAGGCTG	AGGCCATTCTTGTCGCTGAA	98
Gene 9	TCTCTGAGCAGGAATCCTTTGT	GCTACAGCGATGAAGCAGCA	261
Gene 10	TGATGAGATTGGCGTGGCTT	AGGATACCTGGCCTCCACAT	190
PTEN	CTCAGCCGTTACCTGTGTGT	AGGTTTCCTCTGGTCCTGGT	129
HpRT-1	TGCTCGAGATGTGATGAAGG	TCCCCTGTTGACTGGTCATT	
GAPDH	ACCCACTCCTCCACCTTTG	CTCTTGTGCTCTTGCTGGG	

4.3.5. PCR Amplification of cDNA

In order to verify reverse transcription and accessibility of cDNA for PCR, PCR was performed as shown in table 4.5. In this PCR, sequences from the 5’ and 3’ end of the candidate genes were amplified. The amplified PCR products were visualized for the expected size on a 2 % Agarose gel. The gel was stained with SYBR safe DNA stain and 1X TBE was used as the electrophoresis buffer (see section 4.3.2).

Table 4.5: Standard PCR Reaction Composition

Reagent	**Final Concentration**
2X KAPA Taq Extra Hotstart ReadyMix with dye (2 mM $MgCl_2$ at 1X)	1 X
Forward Primer	10 µM
Reverse Primer	10 µM
PCR-grade water	Variable
Template DNA	250 ng

The reaction mixture was cycled with the following parameters as shown by table 4.6

Table 4.6: PCR cycling protocol

Step	Temperature	Duration	Cycles
Initial Denaturation	95 °C	5 min	1
Denaturation	95 °C	30 sec	20-40
Annealing	Primer Specific	30 sec	
Extension	72 °C	45 sec	
Final Extension	72 °C	10 min	1
Hold	4 °C	α	1

4.3.6. Quantitative Real-Time PCR (qPCR) Protocol

Expression levels of the ten selected genes were assessed with quantitative real-time PCR (qPCR). PTEN was added to the panel as a positive control as it is a gene that has been shown to be expressed in cervical cancer. Two housekeeping genes, GAPDH and HPRT were used as reference genes. All reactions were performed on the LightCycler® 480 System (Roche Applied Science) instrument. The reactions were prepared as outlined in table 4.7:

Table 4.7: Reagents for a standard qPCR reaction

Reagents	Final Concentration
SYBR GreenI Master Mix (10X)	1X
Forward Primer (100 µM)	10 µM
Reverse Primer (100 µM)	10 µM
cDNA	250 ng
PCR Grade dH2O	Variable
Final Volume	20 µl

The reactions were performed using the selected candidate genes; a positive control (PTEN), a calibrator (a cocktail of cDNA from all cell lines) and a no-template control (water). An 18 µl aliquot of reaction mastermix was transferred to each well of the white 96 well plates. A 2 µl aliquot of cDNA (250 ng) from various cancer cell lines was then added as the PCR template. A negative control was set up for each run containing 2 µl of PCR-grade water as a substitute for cDNA. The white 96 well plates were sealed with clear sealing foil for the LightCycler® 480 system and cycled on the LightCycler® 480 instrument according to the parameters in table 4.8. The evaluating parameters selected for data analysis were fluorescence (d[F1]/dT), melting temperature (Tm) and crossing point (Cp). The Second Derivative Maximum algorithm was employed for Cp determination where Cp was measured at the maximum increase of fluorescence.

Table 4.8: qPCR Run Protocol

Detection Format	**Block Type**	**Reaction Volume**	
SYBR® Green	96 well	20 µl	
Program Name	**Cycles**	**Analysis Mode**	
Pre-incubation	1	None	
Amplification	45	Quantification	
Melting Curve	1	Melting Curves	
Cooling	1	None	
Program Name	**Target (°C)**	**Acquisition Mode**	**Hold (hh:mm:ss)**
Pre-Incubation	95	None	00:10:00
Amplification	95	None	00:00:30
	Primer dependent	None	00:00:20

	72	Single	00:00:01
Melting Curve	95	None	00:00:30
	65	None	00:01:00
	97	Continuous	5-10 acquisitions/^{0}C
Cooling	40	None	00:00:30

Specificity of real-time PCR primers was determined by amplification plots, melting temperature, and melting curve analysis using LightCycler Software, Version 1.5 (Roche Diagnostics). The standard or calibration curves were generated by the LightCycler software using serially diluted cDNA standards (250 ng to 0.25 ng) was quantified in each real-time PCR run and each dilution was amplified in duplicate. Data on expression levels for housekeeping genes were obtained in the form of crossing points or cycle threshold (Cp/Ct). Data acquisition was done through pairwise comparison using the geometric mean. The PCR efficiencies were calculated using the REST® software and all Ct values were taken into consideration according to the following equation: E=10[-1/slope] (Pfaffl 2001). The expression levels of all genes were determined relative to the housekeeping genes using the following equation:

$$\mathbf{R}= \frac{(\mathbf{E}_{\text{target}})^{\Delta\text{Cptarget (MEAN control- MEAN sample)}}}{(\mathbf{E}_{\text{reference}})^{\Delta\text{Cpreference (MEAN control- MEAN sample)}}}$$

4.4. Results and Discussion

The relative expression levels of the ten candidate genes identify through *in silico* methodologies were determined across an array of cancer cell lines using qPCR. Housekeeping genes (HKGs) / endogenous controls were selected to compensate for the inevitable sample variations observed in qPCR experiments. These variances such as preparation of RNA from biological samples, even the conversion step can affect the reliability of the qPCR results, thus, the consistency and reliability of qPCR experiments is significantly improved by including endogenous controls (Pfaffl, 2004). There are several alternative techniques used to normalize qPCR experiments, however the most commonly used strategy is normalization to an internal reference or a housekeeping gene. Housekeeping genes are widely used as reference genes since their expression is assumed to be stable. The ideal reference gene must be constitutively expressed and unregulated regarding the experimental conditions, treatment, and stage of the disease. It should be expressed at a similar level as the target gene (Nestorov *et al*., 2013).

Therefore the selection of reliable housekeeping genes would correct sample-sample variations; hence display constant expression in all tissues under variable conditions, allowing a comparative analysis across all samples (Kubista *et al*., 2006). The most recommended and mostly used reference genes in qPCR studies are GAPDH (glyceraldehyde-3-phosphate dehydrogenase), BA (B-Actin), HPRT1 (hypoxanthine phosphoribosyltransferase1), and UBC (ubiquitin C), to highlight a few (Nestorov *et al*., 2013). However for this research only GAPDH and HPRT1 were used as reference genes since they are also present in all nucleated cell types and essential for cell survival (Pfaffl, 2001).

4.4.1. Amplification Curve and Melting Curve Analysis

To investigate the expression levels of the ten target genes, eleven cell lines, of which three were cervical cancer as shown in table 4.2.1 were used and this was carried out as described in section 4.3.6. Before proceeding with qPCR analysis, it is imperative to verify that the PCR amplification is specific under the actual qPCR conditions. This was done in the Roche

LightCycler® 480 using the melting curve analysis feature. It is good laboratory practise to perform melting curve analysis after each real-time PCR run as a quality control step, as it is used to distinguish target amplicons from PCR artifacts such as primer-dimers or misprimed products (Fraga *et al*., 2008). The KAPA SYBR FAST qPCR kit optimized for LightCycler® 480 uses SYBR® Green I dye chemistry to detect and quantitate the accumulation of an amplicon. SYBR Green represents the simplest and the most economical choice for real-time PCR product detection. This fluorogenic intercalating dye emits a strong fluorescent signal upon binding to double-stranded DNA (dsDNA) while unbound dye in solution exhibits little undetectable fluorescence. SYBR Green I can be used with any pair of primers, for any target, with no need for any additional fluorescence-labelled oligonucleotide (Nestorov *et al*., 2013).

The major drawback of using this dye is that both specific and nonspecific PCR products are detected, because SYBR green will bind to any double-stranded DNA in the reaction, including primer-dimers and other nonspecific reaction products, leading to overestimation of the target sequence concentration (Nestorov *et al*., 2013). Melting peaks are generated by plotting the negative derivative (-dF/dT) of the melting curve. From figure 4.2, it can be observed from the amplification curves that there was no non-specific background amplification. Furthermore, the melting curve analysis shows specific, single, narrow, and distinctive melting peaks (Nolan *et a*l, 2006). As indicated by the figure only the target products were amplified. Each amplification product for the target genes demonstrated a specific and characteristic melting curve. No primer dimerization or nonspecific products were generated for the applied number of amplification cycles for the respective target genes.

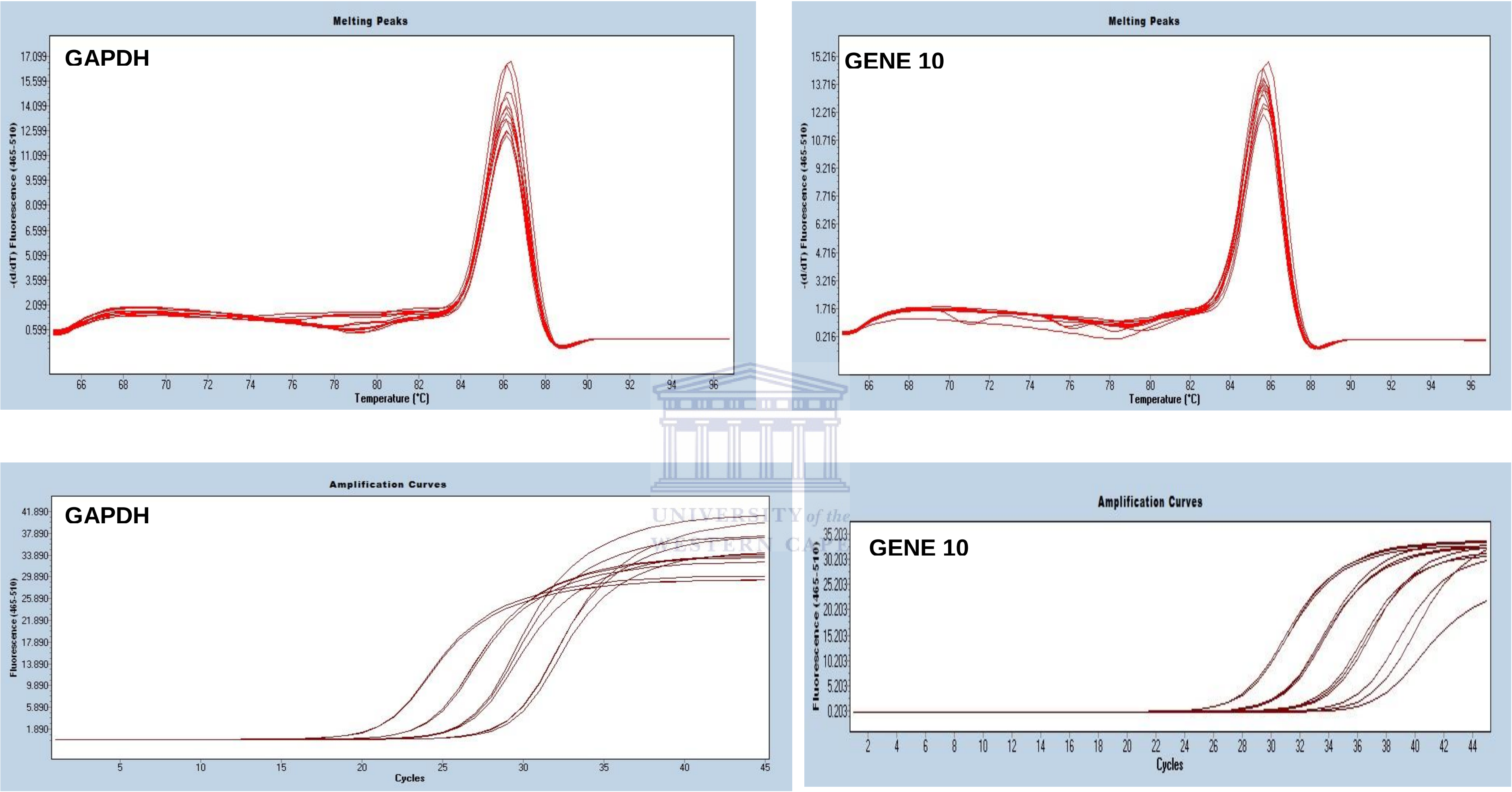


Figure 4.2: Amplification curve and melting peak of GAPDH and target gene 10.

4.4.2. Generation of an Absolute Standard Curve

The standard curve method is the most common approach in determining relative quantification, where the standard curve is generated for both the target and reference gene of choice (Sharkey *et al*., 2004) (Livak, 1997). The standard or calibration curves were generated by the Roche LightCycler software. The quantity of each target gene is mainly determined by means of a standard curve and subsequently expressed relative to a reference gene and they were used to calculate the efficiency of qPCR (Figure 4.3). To generate a standard curve, the serially diluted cDNA standard (250 ng to 0.25 ng) was quantified in each real-time PCR run. Each dilution was amplified in duplicate or triplicate. For each standard, the concentration was plotted against the cycle number at which the fluorescence signal increased above the threshold value (C*t*) or crossing point (C*p*). The gradient generated by each standard curve was used in the equation: Efficiency (E) = $10^{-1/\text{slope}}$ - 1 to determine the reaction efficiency (Rasmussen, 2001), with table 4.9 showing the efficiency and gradient of the housekeeping gene (GAPDH) and candidate genes.

Table 4.9: qPCR efficiencies and standard curve gradients of target genes

	Efficiency (E)	**Gradient**
GAPDH	1.99	-3.35
Gene 1	1.94	-3.47
Gene 2	1.92	-3.52
Gene 3	1.93	-3.50
Gene 4	1.92	-3.52
GENE 5	1.97	-3.52
Gene 6	1.97	-3.39
Gene 7	1.95	-3.45
Gene 8	1.94	-3.48

Gene 9	1.92	-3.52
GENE 10	1.95	-3.45
PTEN	1.94	-3.47

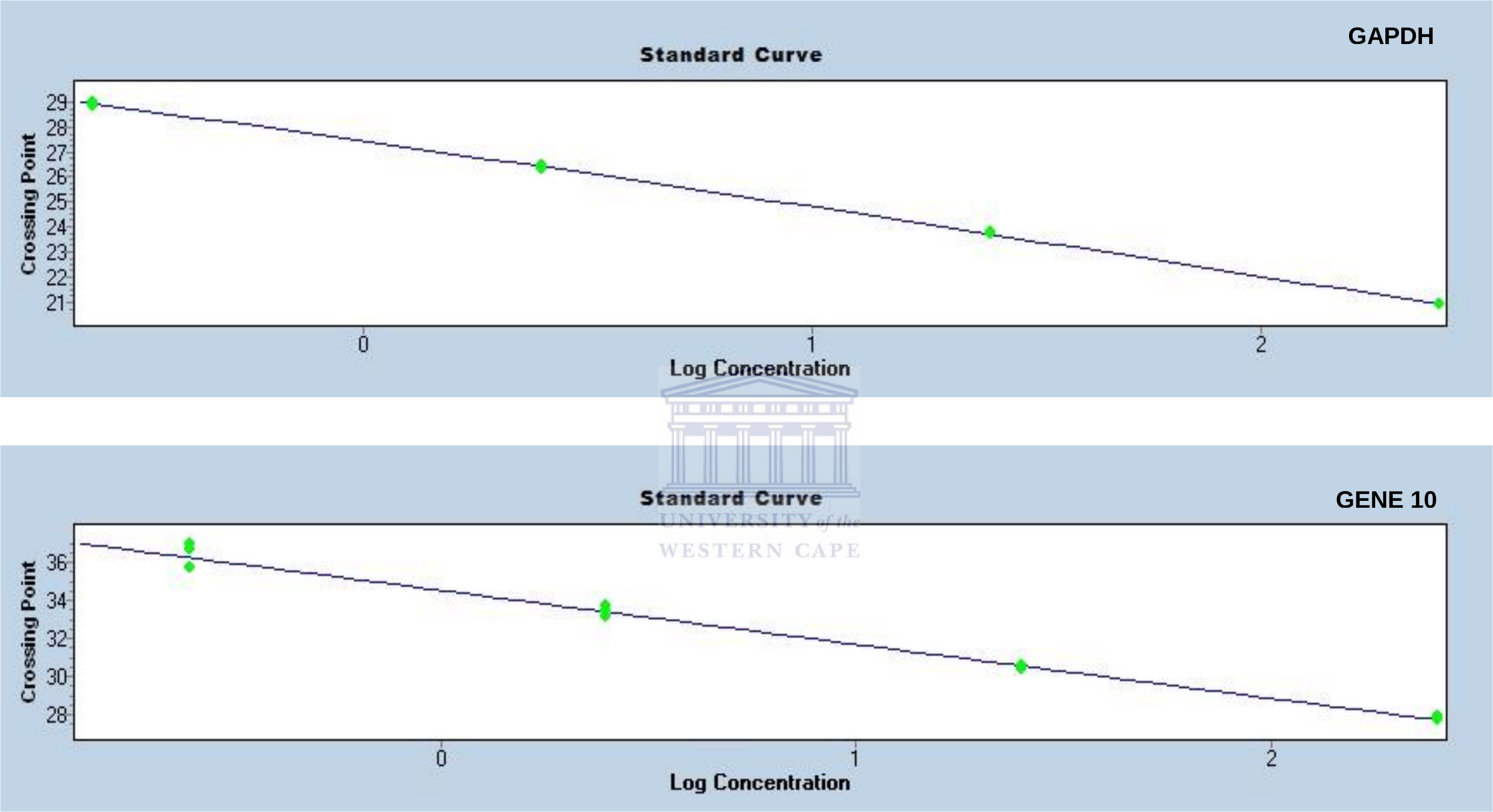


Figure 4.3: Standard curves of the housekeeping gene (GAPDH) and one of the target genes (gene 10)

4.4.3. qPCR data Analysis and Quantification of Gene Expression

Gene expression was quantified using the Pfaffl model which combines gene quantification and normalization and was calculated with the aid of Microsoft Excel based application, Relative Expression Software Tool (REST-348) - Version 1 (Pfaffl *et al*., 2002). For this mathematical model, it is essential to determine the crossing point (Cp) value of each transcript. Given that Cp values decrease linearly with an increasing target quantity, Cp values could be used as a quantitative measurement of the input target number (Heid *et al*., 1996). This method involved comparing the Cp values of the investigated transcripts with a control. The Cp values of both the control and the genes of interest were normalized to an appropriate housekeeping gene. REST-384 calculates relative expression using the statistical model Pair Wise Fixed Reallocation Randomization Test. For each sample, Cp values for the reference and target genes were randomly reallocated to the control and sample groups. Differences in gene expression levels between control and samples were evaluated in group means for statistical significance by randomization tests (Pfaffl *et al*., 2002). Descriptive statistics such as the sample means, minimal (Min) and maximal (Max) values, standard deviation (SD), and coefficient of variance expressed as a percentage (CV%), of the derived Cp values were computed for each investigated gene to determine intra-sample variation.

An analysis was done to evaluate the specificity of the putative genes for cervical cancer by analysing their expression patterns in three different cervical cancer cell lines; six different cancer cell lines and one non-cancerous cell line (refer to table 4.2.1). The normal skin fibroblast cell line, KMST-6 was used as a non-cancerous cell line since there is no non-cancerous cell line for the cervix tissue. The ten putative genes were relatively measured in the various cancer types against the KMST-6 cell line which served as a control using the Pfaffl method as depicted in figure 4.4. Differential expression of the genes was observed across all cancer cell lines, with gene 5 and gene 8 being significantly highly expressed in cervical cancer cell lines: Hela, Caski and HT-3 in comparison to other cancer cell lines. Furthermore, gene 10 is also highly differentially expressed in the cervical cancer when compared to other cancer types. Genes 5, 8 and 10 thus have the highest potential to be biomarkers for cervical cancer. This study was designed to identify genes that were differentially expressed during cervical cancer development, meaning genes that were either up-regulated or down-regulated.

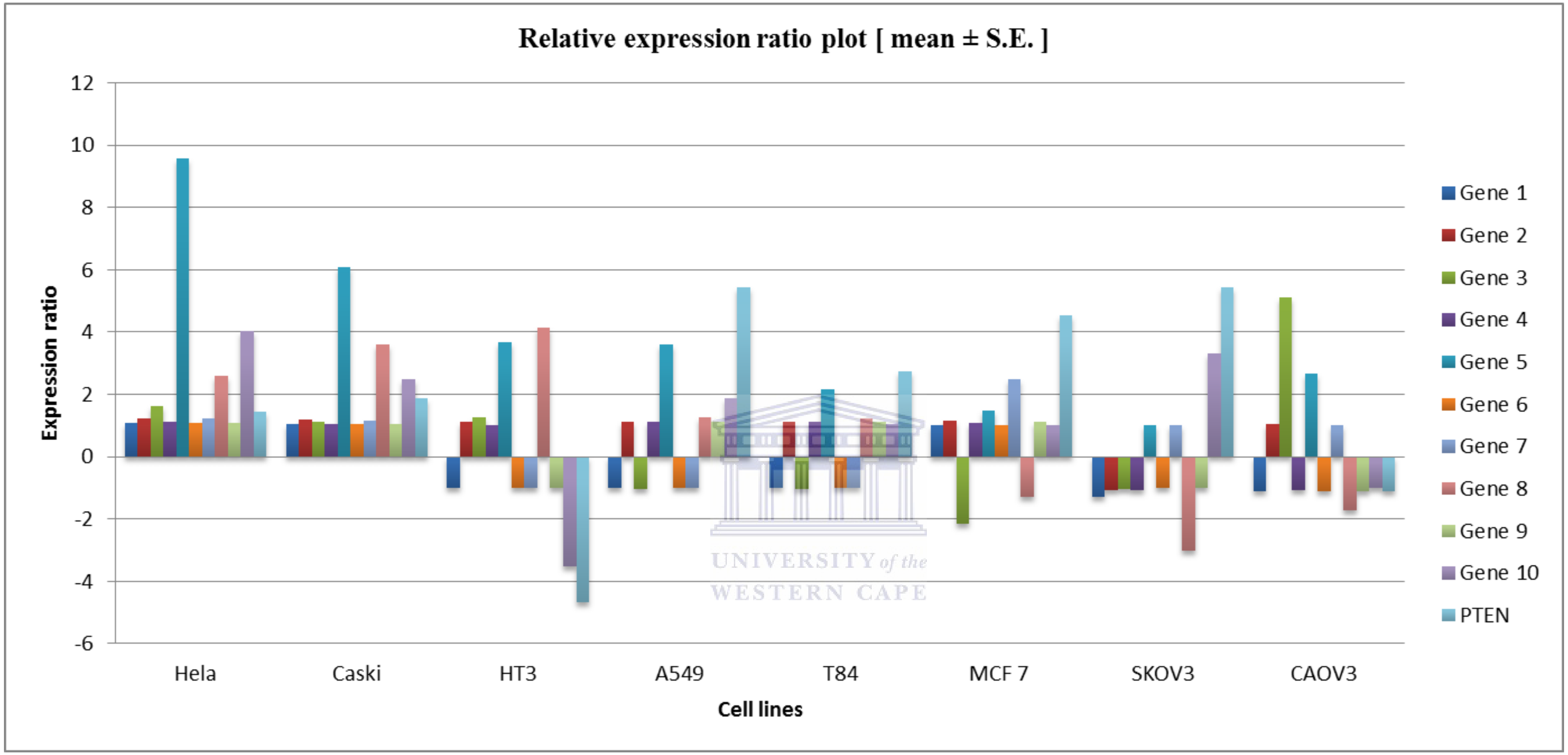


Figure 4.4: Relative expression ration plot of the ten putative genes across various cancer cell lines.

The data obtained from the *in silico* tissue specificity analysis in section 3.4, corresponds with the data acquired using qPCR analysis. Genes 5, 8 and 10 were shown to be highly specific to the cervix tissue as highlighted in figures 3.8, 3.11 and 3.13 respectively. The relative expression plot showed that the other putative genes were highly expressed in other cancer cell lines in comparison to cervical cancer cell lines. Table 4.10 represents a microcosm illustration of the relative expression ratio plot depicting the fold expression ratios of the putative genes across all cell lines.

Table 4.10: Fold expression ratios of 10 putative genes

	Hela	**Caski**	**HT3**	**A549**	**T84**	**MCF 7**	**SKOV3**	**Caov3**
Gene 1	1.080	1.032	-1.014	-1.014	-1.014	1.009	-1.277	-1.098
Gene 2	1.241	1.186	1.133	1.133	1.133	1.159	-1.091	1.046
Gene 3	1.608	1.112	1.275	-1.035	-1.035	-2.168	-1.042	5.129
Gene 4	1.106	1.057	1.009	1.107	1.107	1.082	-1.091	-1.073
Gene 5	9.569	6.088	3.681	3.597	2.163	1.495	1.005	2.686
Gene 6	1.080	1.032	-1.014	-1.014	-1.014	1.009	-1.018	-1.098
Gene 7	1.213	1.159	-1.014	-1.014	-1.014	2.485	1.005	1.022
Gene 8	2.600	3.612	4.132	1.272	1.243	-1.307	-3.017	-1.707
Gene 9	1.080	1.032	-1.014	1.107	1.107	1.133	-1.018	-1.098
Gene 10	4.033	2.495	-3.531	1.883	1.033	1.009	3.311	-1.001
PTEN	1.459	1.883	-4.659	5.452	2.726	4.532	5.428	-1.124

The seven putative genes showed to be differentially expressed in different cancer cell lines, however from table 4.10 it could be deduced that gene 7 was more significantly expressed in MCF7 cell lines, which is a breast cancer cell line. In contrast gene 3 showed a significant expression in CAOV3, an ovarian cancer cell lines. Gene 3 and gene 7 could serve as potential biomarkers for ovarian and breast cancer respectively. The fact that some of the putative genes showed up-regulation in other cancer types is positive; hence this study is a part of a bigger research for biomarker discovery for various cancers such as breast, ovarian, lung and prostate cancer. These candidates will be further validated in the respective cancers where they showed differential expression as potential biomarkers.

Phosphatase and tensine homologue (PTEN) has been implicated in biomarker studies as an indicator for cervical cancer. PTEN a tumour-suppressor gene is involved in cellular differentiation, reproduction and apoptosis, as well as cellular adhesion and mobility. The loss or down-regulation of PTEN plays an important role in the multiple steps of tumourigenesis and progression of malignancies (Qi *et al*., 2014). According to Grigore *et al*, PTEN expression in squamous cell carcinoma is lower compared with normal cervical epithelium. It might be used as a marker for early diagnosis and prognosis of cervical cancer. In this study PTEN was included as a positive control isolated from the candidate genes obtained from the *in-silico* pipeline (Chapter 2). It's inclusion as part of the candidate genes rendered as evidence of the pipeline to identify novel as well as existing biomarkers for cervical cancer. This meant that the putative biomarkers should have a fold change expression ratios similar to the positive control and even more. However from the data obtained it can be deduced that this biomarker was not significantly expressed in cervical cancer cell lines. All the candidate biomarkers were over-expressed in comparison to the positive control in cervical cancer, thus suggesting these genes as possible biomarkers for further validation. PTEN was over-expressed in other cancer cell line, with significant expression in SKOV3 (ovarian cancer), MCF-7 (breast cancer) and A549 (lung cancer). However this is not surprising because according to Loures *et al*, PTEN inactivation may play an important role in the pathogenesis of a variety of human malignancies.

A study done by Eijsink *et al* indicated that a loss of PTEN expression frequently occurs in early-stage cervical cancer. Qi *et al* (2014) established that PTEN plays an important role in the occurrence and development of cervical cancer and the PTEN protein expression phenotype can be considered as an indicator for the pathophysiological behaviour of cervical cancer. The fact that our putative biomarkers showed a significant expression to the PTEN biomarker indicates that these genes could also play a role in early stage cervical cancer, especially, genes 5, 8 and 10 since they showed substantial differential expression in cervical cancer cell lines. From the literature studies conducted in chapter 2, there were no known associations of these putative genes with cervical cancer and it was not clear what their involvement is in cervical cancer. However, the top three candidate biomarkers, gene 5, 8 and 10 will be discussed according to various publications.

Gene 5 is an 855 amino acid secreted protein that localizes to the extracellular space/ matrix, is not attached to membranes and contains three peptidase S1 domains. It is a Proteolytic enzyme or protease, which is a protein that performs a common biochemical reaction, the hydrolysis of peptide bonds. Proteases act as highly specific processing enzymes and perform a selective and limited cleavage of specific substrates. These proteolytic processing events are essential for the regulation of multiple events such as cell cycle progression, tissue morphogenesis and remodelling, cell proliferation and migration, ovulation, angiogenesis, haemostasis, apoptosis, and autophagy. Consistent with these diverse and essential roles of proteases in living organisms, structural changes in these enzymes or alterations in their expression patterns underlie many pathological conditions such as metabolic diseases, neurodegenerative disorders, cardiovascular alterations, arthritis, and cancer (Cal *et al*., 2005).

Most of the well-characterized members of the S1 family of serine proteases are either secreted enzymes or exocytosed from secretory vesicles into the extracellular environment. A structurally distinct group of S1 serine proteases, termed broadly as the membrane-anchored serine proteases, has emerged that are synthesized with amino- or carboxy-terminal extensions that serve to anchor their serine protease catalytic domains directly at the plasma membrane. Additional membrane-anchored serine proteases of the S1 family each possess an amino-terminal signal peptide and enter the secretory pathway. Surface localization studies demonstrate that membrane-anchored serine proteases normally localize to the cell surface and are differentially distributed on apical or basolateral surfaces of polarized cells in patterns unique for each protease (Antalis *et al*., 2011).

Gene 8 is a 29-amino acid COOH-terminally, highly conserved but unique neuroendocrine peptide originally isolated from intestine. This gene mediates biological effects by interacting with high-affinity cell surface receptors. Expression of gene 8 peptides has been detected in pheochromocytoma, pituitary adenoma, neuroblastic tumours, gastrointestinal cancer, squamous cell carcinoma, brain tumours, melanoma, breast cancer and embryonal carcinoma. In several cancers and tumour cell lines expression of gene 8 receptors has been shown as well. Expression of peptide or receptors has been correlated with tumour stage or subtypes of pituitary adenoma, neuroblastic tumours, colon carcinoma and squamous cell carcinoma. Gene 8 and its receptors are promising targets for diagnosis and treatment of several types of tumours. There are several explanations for a possible influence of the circulating levels of gene 8 on cancer growth. For instance, circulating levels may be influenced by cancer growth as a result of altered expression of gene 8 in cancer tissues. As gene 8 is an inhibitory factor in regulating cell proliferation the protection mechanism would be increased with the cancer growth. Gene 10 is an aminoacyl-tRNA synthetase (ARS) that links the amino acid glycine to its corresponding tRNA prior to protein translation and is one of three bifunctional ARS that are active within both the cytoplasm and mitochondria (McMillan *et al.*, 2014). Gene 10 catalyzes the attachment of glycine to tRNA-gly in the cytoplasm and mitochondria and thus plays an essential role in protein synthesis. Park *et al* (2012) have implicated secreted extracellular gene10 in immune surveillance against cancer. Gene 10 is secreted by macrophages in response to FAS ligand that is released from tumor cells and acts as a cytokine with a distinct role against specific tumor cells, this enzyme is essential for protein synthesis in all cells and tissues. ARS are being increasingly recognized as having important secondary functions that include regulation of transcription, translation, splicing and apoptosis. ARS mutations have been implicated in a wide range of human diseases (McMillan *et al*., 2014).

4.5. Conclusion

A critical step in the biomarker discovery pipeline is validation of biomarkers. Before any biological entity can be stated as a biomarker all required tests must be done in order to prove that the entity is fit for such a purpose. There are some limitations in array technologies even though they are comprehensive and relatively accurate in analysing gene expression and have been used in numerous human malignancy studies. Microarray results are influenced by various external factors such as RNA extraction. Consequently, differentially expressed genes in such preliminary discovery efforts need to be confirmed using alternative methods such as qPCR (Hu *et al*., 2006). Even though microarray technology is growing broadly, qPCR is still one of the best methods used extensively for gene expression studies in tissue samples (Pfaffl, 2001) and as such was used in this study. Quantitative real time PCR has the capability to quantify rare transcripts and small changes in gene expression. The difference in gene expression has the ability to shed light on the role of a gene or gene product in a particular process. Changes in gene expression of a particular gene or a group of genes can be indicative of a diseased state as the body tries to maintain homeostasis (Pfaffl, 2001).

This research study intended to validate the ten putative genes identified using an *in silico* approach in cervical cancer cell lines in comparison to a non-cancerous cervical cell line and relative to other cancer types. However, at present there are no non- cancerous cervical cell line, therefore normal fibroblast cell lines were used instead. A postulation established for this study was based on the theory that genes differentially expressed in cervical cancer compared to normal, non-diseased state might shed light into the progression and diagnosis of this disease. Further evaluation of the expression levels of these genes in other cancer types may lead into understanding which genes could possibly be explored as putative biomarkers for diagnostic and therapeutic purposes in cervical cancer and if new discoveries can be uncovered to better understand the role of these genes in other cancers. This aim was achieved by evaluating the expression levels of these genes using qPCR. The gene expression levels of the ten putative genes were evaluated across various cancer cell lines represented in table 4.2. The results from qRT-PCR were analysed using the REST-384© software package through Pfaffl model (Pfaffl, 2001). The software tested for significant results by means of randomisation test and target genes with a p-value less than 0.05 were deemed as showing significant gene expression in cancer cells relative to the calibrator (KMST-6).

A relative expression ratio plot was generated using the software as shown in figure 4.3. The gene expression levels for all cancer types were measured relative to a normal fibroblast cell. Three genes: 5, 8 and 10 showed a significant differential expression in cervical cancer comparative to other cancer types. It was noted that five genes: 1, 6, 7, 9, and 10 were down-regulated in HT3 (cervical cancer), with gene 10 being significantly down-regulated. This correlates with the design of the study to identify biomarkers differentially expressed in cervical cancer. Furthermore, it was interesting to observe that the expression of these putative genes had an expression ratio significantly above PTEN, a gene reported as a biomarker for cervical cancer. This suggested that these genes might play a role in cervical cancer development and be stronger indicators of disease onset and or progression compared to PTEN. The genes that showed slight expression in cervical cancer cannot be disregarded to have no significant role in cervical cancer. As it has been documented cancer is a heterogeneous disease and is very complex; hence some genes are expressed at basal levels at one point during the course of cancer development (Mishra and Verma, 2010).

Studies performed on cell lines can be limiting due to that analysis on tissues is outside a biological system. The assumption that some genes could be expressed at one point or stage of the disease state at which the tissue was isolated may not necessarily present the holistic process of cancer. These genes can be further defined in other cancer types, thus these genes were further investigated in other cancer types to evaluate if they will display a similar profile as they have in cervical cancer. The expression ratios exhibited by the putative genes revealed the need to understand mechanism at which these genes are associated to other cancers. As depicted in figure 4.3 it was clear that although these genes were differentially overexpressed across all cancer types, none were specific to one type of cancer. Therefore, the other seven genes (1, 2, 3, 4, 6, 7, and 9), showed overexpression in other cancer types in comparison to cervical cancer. Gene 3 was over-expressed in an ovarian cancer cell line (CAOV3).

It is evident that the genes presented in this study were expressed ubiquitously across all cancer types that were investigated. Some genes showed upregulated expression in cervical cell lines while other genes remained unchanged. The genes highlighted yellow in table 4.10 were of interest as they exhibited overexpression above PTEN. The study brought about an understanding that the red highlighted genes (gene 3 and 7) showed upregulated expression in ovarian and breast cancer respectively. It also managed to prioritise genes significantly overexpressed in cervical cancer and may be pursued further as biomarkers for cervical cancer. These studies serve as basis for future investigations to determine whether these candidate genes can be exploited as potential biomarkers for diagnosis of cancers in general, as well as, for specific cancers. Thus, they could also serve as potential drug targets for cancer treatment.

4.6. References

1. Antalis, T.M., Bugge, T., and Wu, Q. 2011. Membrane-anchored serine proteases in health and disease. *Prog Mol Biol Transl Sci*, 99: 1–50.

2. Cal, S., Quesada, V., Llamazares, M., Díaz-Perales, A., Garabaya, C., and López-Otín, C. 2005. Human Polyserase-2, a Novel Enzyme with Three Tandem Serine Protease Domains in a Single Polypeptide Chain. *Journal of biological chemistry*, 280(3): 1953–1961.

3. Eijsink, J.J., Noordhuis, M.G., ten Hoor, K.A., *et al*. 2010. The epidermal growth factor receptor pathway in relation to pelvic lymph node metastasis and survival in early-stage cervical cancer. *Hum Pathol*, 41:1735-1741.

4. Fraga, D., Meulia, T., and Fenster, S. 2008. Real-Time PCR. *Current Protocols Essential Laboratory Techniques*, 10.3.1-10.3.34.

5. Guescini, M., Sisti, D., Rocchi, M.B.L., Stocchi, L., and Stocchi, V. 2008. A new real-time PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition. *BMC Bioinformatics*, 9(326):1-12.

6. Grigore, M., Teleman, S., Ungureanu, D., and Mares, A. 2013. Molecular markers in cervical screening – a promise for the future. *Revista Română de Medicină de Laborator*, 21(2/4): 231-239.

7. Heid, C. A., Stevens, J., Livak, K.J., and Williams, P.M. 1996. Real-time quantitative PCR. *Genome Research*, 6:986-994.

8. Hu, N., Qian, L., Hu, Y., Shou, J-Z., Wang, C., Giffen, C., Wang, Q-H., Wang, Y., Goldstein, A.M., Emmert-Buck, M. and Taylor, P.R. 2006. Quantitative real-time RT-PCR validation of differential mRNA expression of SPARC, FADD, Fascin, COL7AI, CK4, TGM3, ECMI, PPL and EVPL in esophageal squamous carcinoma. *BMC Cancer*, 6(33), pp. doi: 10.1186/1471.

9. Jacob, F., Rea Guertler, R., Naim, S., Nixdorf, S., Fedier, A., Hacker, N.F., and Heinzelmann-Schwarz, V. 2013. Careful Selection of Reference Genes Is Required for Reliable Performance of RT-qPCR in Human Normal and Cancer Cell Lines. *PLOS ONE*, 8(3):1-8.

10. Kubista, M., Andrade, J.M., Bengtsson, M., Forootan, A., Jonák, J., Lind, K., Sindelka, R., Sjöback, R., Sjögreen, B., Strömbom, L., Ståhlberg, A., Zoric, N. 2006. The real-time polymerase chain reaction. *Mol Aspects Med*, 27: 95-125.

11. Livak, K. J. 1997. ABI Prism 7700 Sequence Detection System, User Bulletin 2. PE Applied Biosystems, Foster City, CA.

12. Loures, L.F., Cândido, E.B., Vidigal, P.V.T., Seabra, M.A.L., Cunha de Marco, L.A., and da Silva-Filho, A.L. 2014. PTEN expression in patients with carcinoma of the cervix and its association with p53, Ki-67 and CD31. Rev Bras Ginecol Obstet; 36(5):205-10.

13. Nestorov, J., Matic, G., Elakovic, I., Nikola Tanic, N. 2013. Gene expression studies: how to obtain accurate and reliable data by quantitative real-time RT PCR. *J Med Biochem*; 32 (4):325-338.

14. Nolan, T., Hands, R. E., and Bustin, S. A. 2006. Quantification of mRNA using realtime RT-PCR. *Nature Protocols*, 1:1559-1582.

15. Park, M.C., Kanga, T., Jina, Min Han, J., Bum Kim, S., Park, Y.J., Cho, K., Park, Y.W., Guo, M., He, W., Yang, X.L., Schimmel, P., and Kim, S. 2012. Secreted human glycyl-tRNA synthetase implicated in defense against ERK-activated tumorigenesis. *PNAS*, 640–647.

16. Pfaffl, M.W. 2001. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research*, 29(9):2003-2007.

17. Pfaffl, M.W. 2004. Quantification strategies in Real Time PCR. In A-Z of Quantitative PCR. Ed S.A Bustin. International University Line. California, pp.87-112.

18. Pfaffl, M.W., Horgan, G.W., and Dempfle, L. 2002. Relative expression software tool (REST©) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. *Nucleic Acids Research*, 30(9):1-10.

19. Qi, Q., Ling, Y., Zhu, M., Zhou, L., Wan, M., Bao, Y., and Liu, Y. 2014. Promoter region methylation and loss of protein expression of PTEN and significance in cervical cancer. *Biomedical Reports*, 2: 653-658.

20. Rasmussen, R. P. 2001. Quantification on the LightCycler. In Meuer, S., C. Wittwer, K. Nakagawara. (Eds.). Rapid cycle real-time PCR, methods and applications. pp. 21-34. Springer Press. Heidelberg.

21. Sharkey, F. H., Banat, I. M., and Marchant, R. 2004. Detection and quantification of gene expression in environmental bacteriology. *Applied and Environmental Microbiology*, 70:3795-3806.

22. Wong, M.L., and Medrano, J.F. 2005. Real-time PCR for mRNA quantitation. *BioTechniques*, 39(1):1-11.

23. Yuan, J.S., Reed, A., Chen, F., and Stewart Jr, C.N. 2006. Statistical analysis of real-time PCR data. *BMC Bioinformatics*, 7(85):1-1.

CHAPTER 5: General Discussion and Future Directions

Cancer is emerging as an under-recognised global threat to human development. It is estimated that there will be 22.2 million new cases of cancer and 12.7 million cancer-related deaths worldwide in 2030(Ginsburg, 2013). More than half of new cases and two-thirds of cancer deaths will occur in low and middle-income countries, where access to early, accurate diagnosis and quality care are woefully lacking. The increasing burden of cancer has especially harsh consequences for women due to gender discrimination, cultural taboos and stigma, all of which conspire to limit women's choices to seek care even when it is available (Ginsburg, 2013). In South Africa however, women, especially black women, present with advanced cancer–too late for them to be cured. Not only the World Health Organisation, but also the Cancer Association of South Africa (CANSA) stresses prevention and early detection as the best way to fight the disease (Maree and Wright, 2010). Early diagnosis of cancer is difficult because of the lack of specific symptoms in early disease and the limited understanding of aetiology and oncogenesis (Cho, 2007).

One of the main reasons for high death rates in cancer patients is due to the lack of well-validated and clinically useful biomarkers with adequate sensitivity and specificity to detect this disease at early stages. However, cervical cancer is curable when detected early before the development of metastasis. Development of cervical cancer is a multi-step process initiated by persistent infection with high-risk human papillomavirus, which in a limited number of cases progresses via cervical intraepithelial neoplasia to invasive cervical cancer that usually spans 10-20 years, which offers a period of several years to detect the tumour in an early stage and to interfere with the natural cause of the disease.

All South African women have access to cervical cancer screening at primary health clinics however the screening uptake is only approximately 20%, thus most women present with late cervical cancer. A study conducted by Van Schalkwyk *et al* (2008), revealed that women were willing to be diagnosed but were failed by the health care system, with their initial interaction with the health care system not resulting in prompt diagnosis or treatment and they repeatedly had to go back before being diagnosed (Issah *et al*., 2011). Early detection of CC can therefore significantly reduce the mortality associated with this malignancy.

However, current screening methods including Pap smear test, colposcopy, pathological and preoperative diagnosis either lack the required sensitivity and specificity or are costly and invasive. Some biomarkers such as the CEA levels and tumour-associated gene mutations have only shown some prognostic or predictive value (Cho, 2007). There is therefore an urgent need for developing new diagnostic/screening tests and identifying putative biomarkers to diagnose, predict and monitor the progress of CC and eventually find more efficient drug targets for this disease.

The aim of this study was to discover biomarkers that could aid in the early diagnosis of cervical cancer. A diagnostic system that can detect biomarkers that are found in bodily fluids can serve as a less invasive, inexpensive and specific and sensitive method for detection. The *in silico* approach used in this study was successful in identifying genes that are associated with cervical cancer as well as those who have not yet been associated with the disease thus verifying that data obtained from this pipeline was reliable. The methodology employed in chapter two allowed for genes to be proficiently prioritised based on various categorical levels. The results indicated that this pipeline can save time and can be incorporated in biomarker discovery studies successfully. A study done by Prassas *et al* (2012), used a similar bioinformatic approach to identify proteins with tissue-specific expression for biomarker discovery, their study managed to identify previously studied cancer biomarkers. This study identified ten putative biomarkers and they were verified in each step of elimination and were confidently selected for further validation.

Databases house vast amounts of information therefore for biomarker identification *in silico* approaches were carefully selected to uncover the wealth of information. Furthermore each tool employed in this study was able to prioritize and discriminate between data that was required for the next step of the study and eliminate those that were not. This required the user to establish parameters of inclusion or exclusion and these were defined in Chapter 2. The study made use of multistep processes which involved data mining, literature mining, refinement and annotation tools that have led to the identification of putative biomarkers for cervical cancer. The identification of candidate biomarkers on a small scale as opposed to wet-bench techniques could lead to more biomarkers being identified in shorter periods of time (Magni *et al*., 2010). Understanding genes on the transcriptional level gives new insight into how the genes are regulated; it can also implicate them in various pathways and can indicate similar protein-protein interactions. The *in silico* pipeline proved to be valuable in extracting complex candidate gene lists. Most of the genes were predicted to be enriched for various cancer-associated processes by Gene Ontology (DAVID), protein-protein interactions (STRING) and pathway analysis (KEGG). The GOI were predicted to be transcriptionally coupled to PPARγ (down-regulated in pre-invasive cervical cancer) and to be interacting with PCNA, thereby showing potential value for providing an early diagnosis of cervical cancer. Ninety percent (90%) of the GOI were co-expressed, while many genes participated in cancer-related pathways. This could implicate their possible role in a coordinated cervical cancer pathway. Lastly, their potential tissue specificity and sensitivity was also demonstrated by various *in silico* analyses. This section of the study successfully implicated the ten cell surface gene products in cancer based on their regulatory elements, interacting proteins and pathway analysis. The *in silico* biomarker discovery should be a continuous study in a quest to identify potential novel genes since microarray databases are frequently updated. Further bioinformatics studies should be done between the putative biomarkers and known clinical biomarkers for cervical cancer. Gene expression profiles should also be established by combining the candidate genes with their interacting proteins for early diagnosis. Further investigation of the GOI and their regulatory elements through projects such as the ENCODE Consortium could also provide valuable insights and help to map a possible cervical cancer pathway.

It is well accepted that *in silico* prediction of biomarkers should not be separated from *in-vitro* validation in a biological system. With that in mind, a molecular approach was employed to confirm gene expression patterns of the ten putative biomarkers which were identified using an *in silico* approach. An important aspect of cancer aetiology lies in the stepwise accumulation of genetic and epigenetic changes. These changes contribute to the development of cancer and they vary between different cancer types and stages, tissues and individual (Sadikovic *et al*., 2008). Bioinformatics approach in conjunction with experimental validation provided a powerful combination to carry out this research study. qPCR significantly simplifies and accelerates the process of producing reproducible and reliable quantification of target genes transcription.

The study revealed that PTEN was differentially expressed in CC cell lines and it is therefore being investigated as a potential biomarker for the prognosis of CC. However, the PTEN biomarker was significantly overexpressed in A549 (lung cancer), MCF7 (breast cancer) and SKOV3 (ovarian cancer), thus making this biomarker not to be specific and sensitive for cervical cancer. The study also revealed that the ten putative genes were differentially expressed in the cancer cell lines as compared to the non-cancerous cell line (KMST6). Three genes were shown to be potential biomarkers for cervical cancer, gene 5, gene 8 and gene 10. Gene 3 was upregulated in ovarian cancer and gene 7 was overexpressed in breast cancer. In conclusion, the work presented in this thesis has revealed several candidate putative biomarkers that need to be tested on samples from cervical cancer patients in combination with known cervical cancer biomarkers used in clinical settings.

Further studies could include confirming the expression of protein products of differentially expressed gene candidates in media of various human cell lines (cancer and non-cancerous) using western blots. Furthermore, the presence of these proteins could be investigated in cervical cancer patient samples using ELISA. This study may open avenues into nano-diagnostics by using nano-devices. Biomarkers can exist in small quantities in biological fluids and may be masked by other proteins. The application of nanotechnology can ensure the detection of these biomarkers on a nano-scale. By combining nanotechnology in the form of gold nanoparticles, the study can be used to develop a lateral flow device for the detection of CC as shown in Figure 5.1. This has the potential for a diagnostic test to be developed that is more cost effective and can rapidly be adopted into clinical practice. Also, it may be possible to test for these biomarkers in serum or other bodily fluids thus avoiding invasive diagnostic tests.

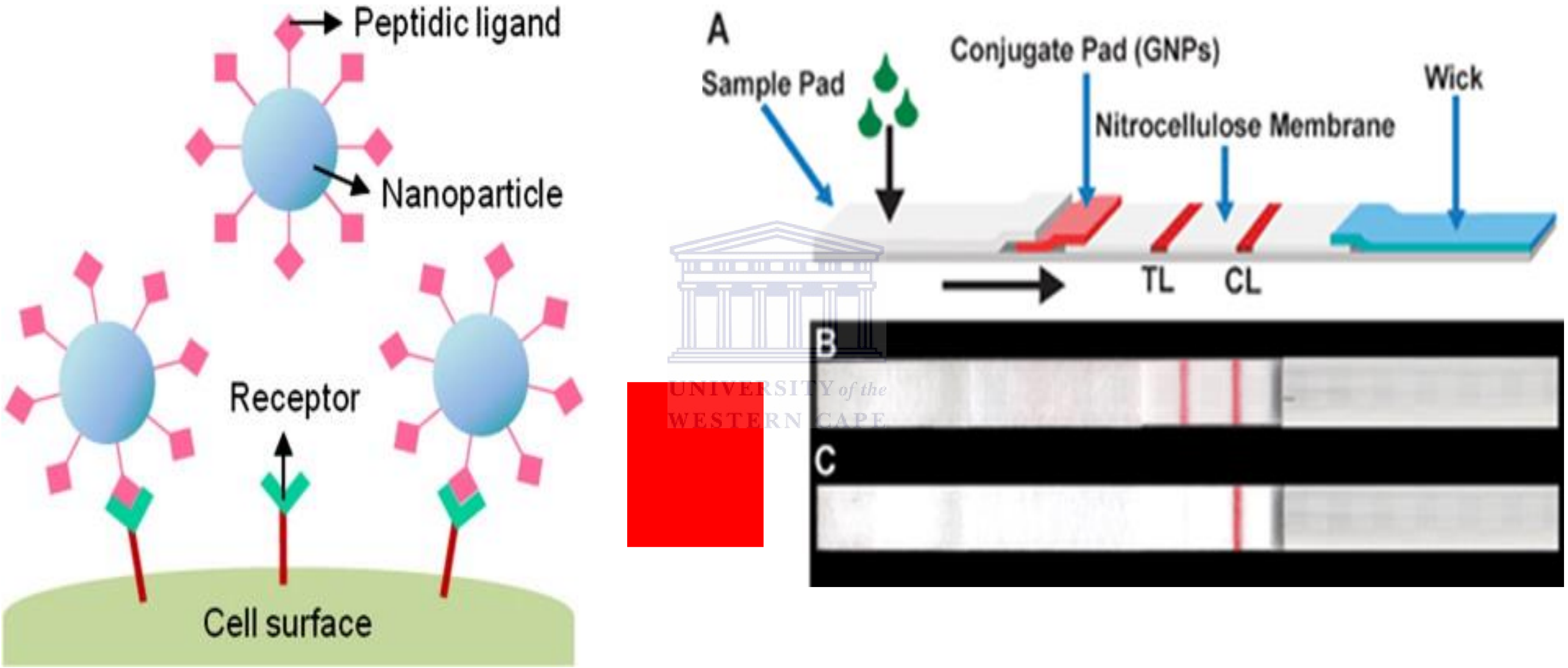


Figure 5.1: A nanotechnology lateral flow assay where ligands will bind to the biomarkers for a positive result.

References

1. Cho, C.S.W. 2007. Contribution of oncoproteomics to cancer biomarker discovery. *Molecular Cancer*, 6:25:1-13.

2. Ginsburg, O.M. 2013. Breast and cervical cancer control in low and middle-income countries: Human rights meet sound health policy. *Journal of Cancer Policy*, 1-7.

3. Issah, F., Maree, J.E., and Mwinituo, P.P. 2011. Expressions of cervical cancer-related signs and symptoms. *European Journal of Oncology Nursing*, 15: 67-72.

4. Magni, F., Van Der Burgt, Y.E.M., Chinello, C., Mainini, V., Gianazza, E., Squeo, V., Deelder, A.M., and Kienle, M.G. 2010. Biomarkers discovery by peptide and protein profiling in biological fluids based on functionalized magnetic beads purification and mass spectrometry. *Blood Tranfusion*, 8:92-97.

5. Maree, J.E.M., and Wright, S. C.D. 2010. How would early detection be possible? An enquiry into cancer related knowledge, understanding and health seeking behaviour of urban black women in Tshwane, South Africa. *European Journal of Oncology Nursing*, 14:190-196.

6. Prassas, I., Chritoja, C.C., Makawita, S., and Diamandis, E.P. 2012. Bioinformatic identification of proteins with tissue-specific expression for biomarker discovery. *BMC Medicine*, 10(39):1-13.

7. Sadikovic, B., Al-Romiah, K., Squire, J.A., and Zielenska, M. 2005. Cause and consequences of genetic and epigenetic alteration sin human cancer. *Current Genomics*, 9(6):394-408.

8. Van Schalkwyk, S.L., Maree, J.E., and Wright, S.C.D. 2008. Cervical cancer: the route from signs and symptoms to treatment in South Africa. *Reproductive Health Matters* 16 (32):9-17.