

**Identification of novel microRNAs as potential biomarkers for the early diagnosis of ovarian cancer using an *in-silico* approach**



**UNIVERSITY of the  
WESTERN CAPE**

**Zahra Latib**

**Student number: 3438895**

A thesis submitted in fulfilment of the requirements for the degree of  
Magister Scientiae, in the Department of Biotechnology, University of  
the Western Cape.

**Supervisor:** Dr Marshall Keyster

**Co-supervisor:** Dr Ashley Pretorius

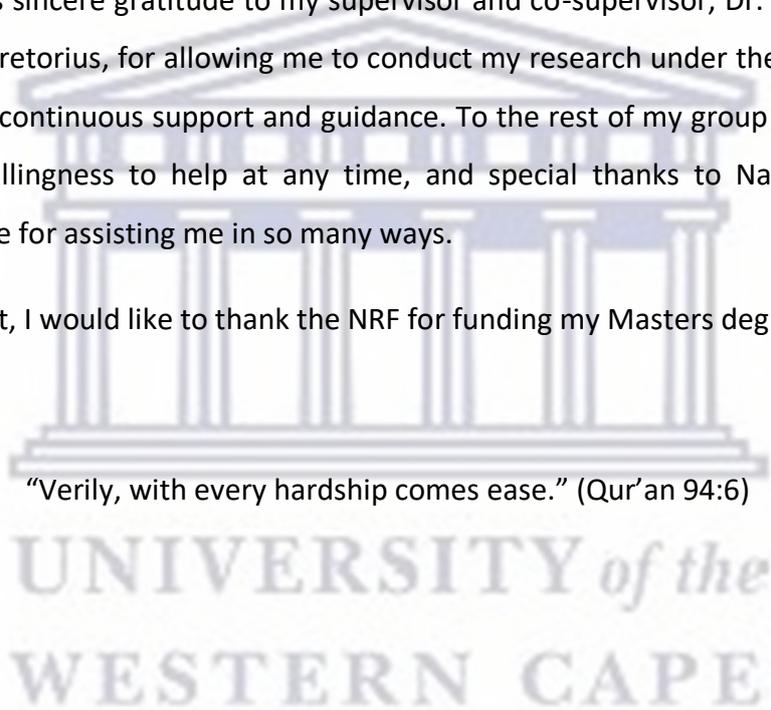
November 2019

## Acknowledgments

First and foremost, all praises to Almighty Allah for granting me the opportunity to pursue my Masters degree at the University of the Western Cape, without my religion I would not have made it thus far. I would like to thank my father and late mother for their love, guidance and motivation which has really encouraged me to follow my dreams. Their support was and always will be the greatest gift anyone has ever given me. Thanks are also due to my family and friends who have supported me through some of the most challenging days, leading me to where I am today.

I wish to express sincere gratitude to my supervisor and co-supervisor, Dr. Marshall Keyster and Dr. Ashley Pretorius, for allowing me to conduct my research under their supervision as well as for their continuous support and guidance. To the rest of my group members, thank you for your willingness to help at any time, and special thanks to Nasr Eshibona and Chipampe Lombe for assisting me in so many ways.

Last but not least, I would like to thank the NRF for funding my Masters degree.

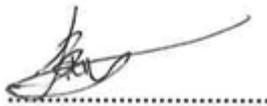


## Declaration

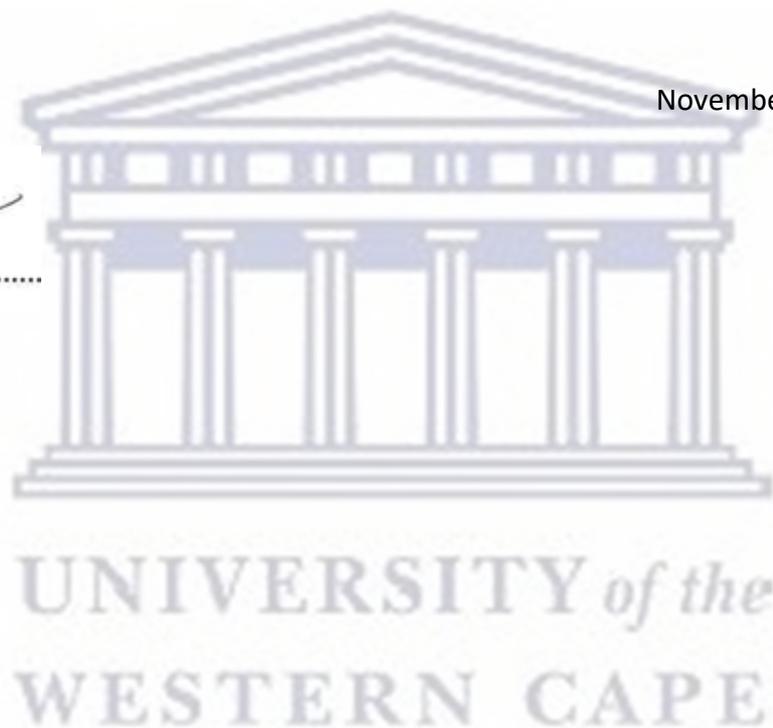
I declare that “Identification of novel microRNAs as potential biomarkers for the early diagnosis of ovarian cancer using an *in-silico* approach” is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Zahra Latib

November 2019



Signature



Dedication



*To my family*

UNIVERSITY *of the*  
WESTERN CAPE

## Abstract

Ovarian cancer (OC) is the most fatal gynaecologic malignancy that is generally diagnosed in the advanced stages, resulting in a low survival rate of about 40%. This emphasizes the need to identify a biomarker that can allow for accurate diagnosis at stage I. MicroRNAs (miRNAs) are appealing as biomarkers due to their stability, non-invasiveness, and differential expression in tumour tissue compared to healthy tissue. Since they are non-coding, their biological functions can be uncovered by examining their target genes and thus identifying their regulatory pathways and processes.

This study aimed to identify miRNAs and genes as candidate biomarkers for early stage OC diagnosis, through two distinct *in silico* approaches. The first pipeline was based on sequence similarity between miRNAs with a proven mechanism in OC and miRNAs with no known role. This resulted in 9 candidate miRNAs, that have not been previously implicated in OC, that showed 90-99% similarity to a miRNA involved in OC. Following a series of *in silico* experimentations, it was uncovered that these miRNAs share 12 gene targets that are expressed in the ovary and also have proven implications in the disease. Since the miRNAs target genes contribute to OC onset and progression, it strengthens the notion that the miRNAs may be dysregulated as well. Using TCGA, the second pipeline involved analysing patient clinical data along with implementing statistical measures to isolate miRNAs and genes with high expression in OC. This resulted in 26 miRNAs and 25 genes being shortlisted as the potential candidates for OC management. It was also noted that targeting interactions occur between 15 miRNAs and 16 genes identified through this pipeline. In total, 35 miRNAs and 37 genes were identified from both pipelines.

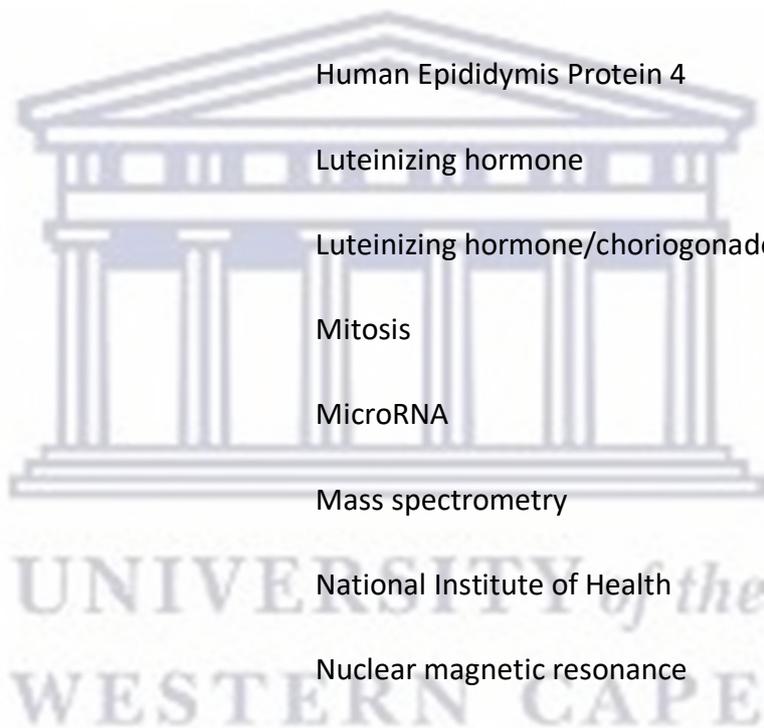
To rank all the identified candidates based on strength and priority, those with targeting interactions were subjected to trident to assess their triplex-forming potential. All candidates had similar triplex scores, except miR-2-14 and ACTB, identified through the second pipeline, which had the strongest interaction and was therefore deemed a top priority candidate.

Since bioinformatics offers a predictive outcome, all identified candidates need to undergo molecular validation to ensure not only their dysregulated expression in OC but also the modulating effect that these miRNAs have on their target genes.

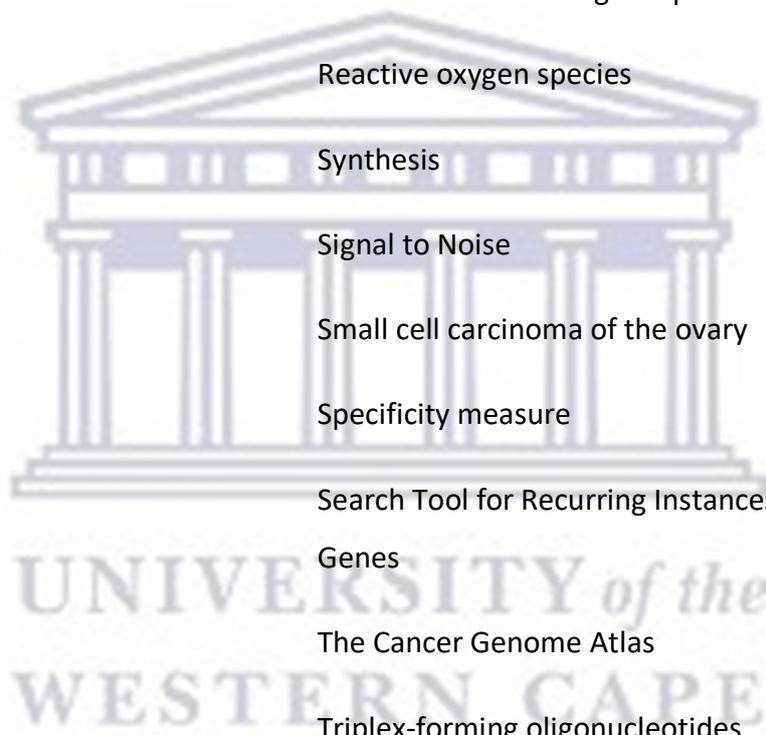
## List of abbreviations

2DE	Two-dimensional gel electrophoresis
ACTB	Actin beta
Ago2	Argonaute protein
ALT	Alternative Lengthening of Telomeres
AP-1	Activating protein 1
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like Alignment Tool
CA-125	Cancer antigen 125
CAMs	Cell-cell adhesion molecules
CFL1	Cofilin 1
CLL	Chronic lymphocytic leukaemia
CSCs	Cancer stem cells
CUP	Cancers of unknown primary origin
CYP19A1	Cytochrome P450 Family 19 Subfamily A Member 1
DAVID	Database for Annotation, Visualization and Integrated Discovery
DBP	DNA-binding protein
E	Energy score
ECM	Extracellular matrix
EGR2	Early growth response 2
ER	Endoplasmic reticulum

ESR1	Oestrogen receptor 1
E-value	Expect value
FTL	Ferritin light chain
G1	Gap 1
G2	Gap 2
GAD	Genetic Associations Database
H	Heuristic score
HE4	Human Epididymis Protein 4
LH	Luteinizing hormone
LHCGR	Luteinizing hormone/choriogonadotropin receptor
M	Mitosis
miRNA	MicroRNA
MS	Mass spectrometry
NIH	National Institute of Health
NMR	Nuclear magnetic resonance
OC	Ovarian Cancer
P	Genomic position
PaGenBase	Pattern gene database
PCR	Polymerase chain reaction
PGR	Progesterone receptor
pI	Isoelectric point
POC	Point of Care



Pre-miRNA	Precursor miRNA
Pri-miRNA	Primary precursor miRNA
PTEN	Phosphatase and Tensin homologue
qPCR	Quantitative PCR
Rb	Retinoblastoma-associated
RBP	RNA-binding protein
RISC	RNA-induced silencing complex
ROS	Reactive oxygen species
S	Synthesis
S/N	Signal to Noise
SCCO	Small cell carcinoma of the ovary
SPM	Specificity measure
STRING	Search Tool for Recurring Instances of Neighbouring Genes
TCGA	The Cancer Genome Atlas
TFOs	Triplex-forming oligonucleotides
TMBIM6	Transmembrane BAX Inhibitor Motif Containing 6
TTS	Triplex targeting sites
TVS	Transvaginal ultrasonography
UPS	Ubiquitin proteasome system
VEGF	Vascular endothelial growth factor
WNT5a	Wnt Family Member 5



## List of figures

### Chapter 1

**Figure 1.1:** The six hallmarks of cancer (Hanahan and Weinberg, 2011).

**Figure 1.2:** Anterior view of the female reproductive system (Cancer Association of South Africa, 2016).

**Figure 1.3.1:** Worldwide incidence rates of OC in 2008 (Chornokur *et al.*, 2013).

**Figure 1.3.2:** Percentages of OC cases by stage of diagnosis and their respective 5-year survival rates from 2007 to 2013 (National Cancer Institute, no date).

**Figure 1.4:** MicroRNA biogenesis overview pathway (Winter *et al.*, 2009).

### Chapter 2

**Figure 2.1:** Flow chart representing the outline of the *in-silico* methodology employed for miRNA identification in this chapter.

**Figure 2.2:** Flow chart depicting the outline of the *in-silico* methodology employed for prioritising the miRNA target genes.

**Figure 2.3:** Venn diagrams depicting the intersection of the ovary-specific genes with each miRNAs' target genes.

**Figure 2.4:** Biological processes, molecular functions and cellular components that the candidate genes are involved in. Yellow blocks indicate the functions of the respective genes.

**Figure 2.5:** Network from STRING depicting the interactions of the candidate proteins with each other.

### Chapter 3

**Figure 3.1:** Flow charts representing the outline of the *in-silico* methodology employed for both miRNA and gene identification in this chapter.

**Figure 3.2:** Distribution curve of the top 100 highly expressed miRNAs, based on their S/N values.

**Figure 3.2.1:** Box plot of the top 100 miRNAs, based on their S/N values.

**Figure 3.3:** Distribution curve of the top 100 highly expressed genes, based on their S/N values.

## **Chapter 4**

**Figure 4.1:** Triplex forming interactions between TFOs and DNA polypurine strand. Solid lines indicate Watson-Crick hydrogen bonding whereas dotted lines represent either Hoogsteen or reverse Hoogsteen hydrogen bonding (Maldonado *et al.*, 2017).

**Figure 4.2:** Flow chart depicting the outline of the *in-silico* methodology employed for identifying the candidates capable of triplex-formations.

## **List of tables**

### **Chapter 2**

**Table 2.1.1:** Number of mature human miRNAs obtained from miRBase.

**Table 2.1.2:** Number of miRNAs implicated in OC obtained from multiple databases.

**Table 2.2:** Number of miRNAs before and after the removal of duplicates via CD-HIT.

**Table 2.3:** Table showing the clusters of similar sequences with the percentage of similarity calculated by CD-HIT and BLAST. MiRNAs in bold are validated with implications in OC, and the miRNAs below them share more than 90% similarity.

**Table 2.4:** Shows the miRNAs implicated in OC, after literature mining. The miRNAs in bold indicate OC-implicated and the normal text relates to the potential novel OC miRNAs.

**Table 2.5.1:** Number of gene targets shared between each of the potentially novel miRNAs (in normal text) and their respective similar OC-implicated miRNA (in bold).

**Table 2.5.2:** Number of genes expressed specifically in the ovary.

**Table 2.6:** Genes with implications in OC.

**Table 2.7:** Targeting interactions between the novel OC miRNAs and target genes.

### **Chapter 3**

**Table 3.1:** Candidate miRNAs identified in this study, in no particular order.

**Table 3.2:** Candidate genes identified in this chapter, as well as their expression information.

**Table 3.3:** Candidate miRNAs that target the candidate genes identified.

### **Chapter 4**

**Table 4.1:** Candidate miRNAs and target genes identified in chapter two that have triplex forming abilities.

**Table 4.1.1:** Triplex interactions identified between each candidate miRNA and its target genes for both sense and antisense strands. Table indicates the Energy score (E), Heuristic score (H), Genomic position (P) and binding structure.

**Table 4.2:** Candidate miRNAs and target genes identified in chapter three that have triplex forming abilities.

**Table 4.2.1:** Triplex interactions identified between each candidate miRNA and its target genes for both sense and antisense strands. Table indicates the Energy score (E), Heuristic score (H), Genomic position (P) and binding structure.

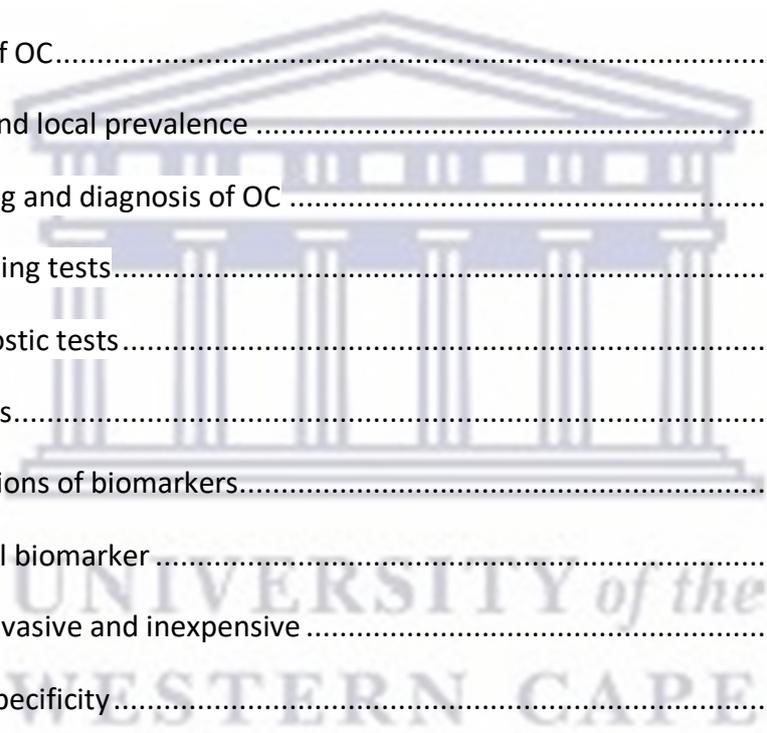
### **Chapter 5**

**Table 5.1:** Prioritization of all candidate miRNAs based on their ability to form triplexes with their target genes.

## Table of Contents

<b>Acknowledgments</b> .....	<b>i</b>
<b>Declaration</b> .....	<b>ii</b>
<b>Dedication</b> .....	<b>iii</b>
<b>Abstract</b> .....	<b>iv</b>
<b>List of abbreviations</b> .....	<b>v</b>
<b>List of figures</b> .....	<b>viii</b>
<b>List of tables</b> .....	<b>ix</b>
<b>Chapter 1</b> .....	<b>1</b>
<b>Literature review</b> .....	<b>1</b>
1.1. Cancer overview .....	1
1.1.1. Cancer and the cell cycle.....	1
1.1.2. Tumour biology.....	1
1.1.3. Hallmarks of cancer.....	3
1.1.3.1. Self-sufficiency in growth signals.....	3
1.1.3.2. Insensitivity to antigrowth signals.....	4
1.1.3.3. Tissue invasion and metastasis.....	4
1.1.3.4. Limitless replicative potential.....	5
1.1.3.5. Sustained angiogenesis .....	6
1.1.3.6. Evading apoptosis .....	6
1.2. Ovarian cancer .....	7
1.2.1. Physiology and anatomy of the ovary.....	7
1.2.2. Histological subtypes of OC.....	8
1.2.2.1. Epithelial carcinoma.....	8
1.2.2.2. Germ cell carcinoma .....	9

1.2.2.3. Stromal carcinoma .....	9
1.2.2.4. Small cell carcinoma.....	9
1.2.3. Aetiology of OC .....	9
1.2.3.1. Genetic factors and family history.....	9
1.2.3.2. Reproductive and hormonal factors .....	10
1.2.3.3. Age.....	10
1.2.3.4. Lifestyle and environmental factors .....	11
1.2.4. Symptoms of OC .....	11
1.2.5. Stages of OC.....	11
1.2.6. Global and local prevalence .....	12
1.2.7. Screening and diagnosis of OC .....	13
1.2.7.1. Screening tests.....	13
1.2.7.2. Diagnostic tests.....	14
1.3. Biomarkers.....	15
1.3.1. Applications of biomarkers.....	15
1.3.2. The ideal biomarker .....	16
1.3.2.1. Non-invasive and inexpensive .....	16
1.3.2.2. High specificity.....	16
1.3.2.3. High sensitivity.....	17
1.3.3. Current OC biomarkers .....	17
1.3.3.1. CA-125 .....	17
1.3.3.2. HE4 .....	17
1.3.3.3. VEGF .....	18
1.3.3.4. Ova1 .....	18
1.3.4. MicroRNAs as biomarkers .....	18
1.3.4.1. Biogenesis and function of miRNAs .....	18



1.3.4.2. Circulating miRNAs.....	20
1.3.4.3. Suitability of miRNAs as cancer biomarkers.....	20
1.3.4.4. MiRNAs in OC.....	21
1.3.5. Methods for biomarker discovery .....	23
1.3.5.1. Genomic approach.....	23
1.3.5.2. Proteomic approach.....	24
1.3.5.3. Metabolomic approach .....	24
1.4. Bioinformatics.....	25
1.4.1. Advantages of bioinformatics.....	26
1.4.2. Biomarker discovery .....	26
1.5. Problem identification .....	27
<b>Chapter 2 .....</b>	<b>29</b>
<b>Identification of miRNAs and target genes as biomarkers for the early stage diagnosis of OC via a sequence similarity approach .....</b>	<b>29</b>
2.1 Introduction.....	29
2.1.1 Biological databases.....	32
2.1.1.1 MiRNA discovery databases .....	32
2.1.1.1.1 MiRBase.....	32
2.1.1.1.2 DbDEMC 2.0.....	32
2.1.1.1.3 Mir2disease .....	33
2.1.1.1.4 MiRandola.....	33
2.1.1.1.5 MiRCancer .....	33
2.1.1.2 Tools for sequence similarity analysis.....	33
2.1.1.2.1 CD-HIT.....	33
2.1.1.2.2 BLAST .....	34
2.1.1.3 MiRNA gene target database .....	34

2.1.1.3.1 MiRDip .....	34
2.1.1.4 Gene annotation databases .....	35
2.1.1.4.1 PAGENBASE.....	35
2.1.1.4.2 DAVID .....	35
2.1.1.5 Protein network database .....	36
2.1.1.5.1 STRING .....	36
2.1.2 Aims.....	37
2.1.3 Objectives .....	37
2.2 Methodology .....	39
2.2.1 Data mining .....	40
2.2.1.1 MiRBase.....	40
2.2.1.2 dbDEMC 2.0 .....	40
2.2.1.3 MiR2Disease .....	40
2.2.1.4 MiRandola.....	40
2.2.1.5 MiRCancer .....	40
2.2.2 Duplication removal via CD-HIT.....	41
2.2.3 Sequence similarity analysis.....	41
2.2.3.1 CD-HIT.....	41
2.2.3.2 BLAST .....	41
2.2.2 Text-mining.....	42
2.2.5 Gene identification .....	44
2.2.5.1 miRNA target genes .....	44
2.2.5.2 Ovary-specific genes .....	44
2.2.6 Intersecting genes.....	44
2.2.7 Functional annotation.....	44
2.2.7.1 Disease .....	44

2.2.7.2 Gene function .....	45
2.2.8 Protein-protein interactions.....	45
2.3 Results and discussion .....	46
2.3.1 Data mining .....	46
2.3.2 Duplication removal via CD-HIT.....	47
2.3.3 Sequence similarity analysis.....	47
2.3.4 Text-mining.....	49
2.3.5 Gene identification .....	51
2.3.5.1 miRNA target genes .....	51
2.3.5.2 Ovary-specific genes .....	52
2.3.6. Intersecting genes.....	54
2.3.7 DAVID Functional annotation.....	56
2.3.7.1 Disease .....	56
2.3.7.2 Gene function .....	59
2.3.8 Protein-protein interactions.....	61
2.4 Conclusion .....	63
<b>Chapter 3 .....</b>	<b>64</b>
<b>Identification of miRNAs and genes as biomarkers for the early stage diagnosis of OC using patient clinical data from TCGA.....</b>	<b>64</b>
3.1 Introduction.....	64
3.1.2 Aim .....	65
3.1.3 Objectives .....	65
3.2 Methodology .....	66
3.2.1. MiRNA data extraction.....	68
3.2.1.1. Statistical parameters to isolate candidates .....	68
3.2.2. Gene data extraction .....	68

3.2.2.1. Statistical parameters to isolate candidates .....	68
3.2.3 MiRNA-gene targeting interactions .....	69
3.3 Results and discussion .....	70
3.3.1 MiRNA data extraction and candidate identification .....	70
3.3.2 Gene data extraction and candidate identification.....	74
3.3.3 MiRNA-gene targeting interactions .....	78
3.4 Conclusion .....	80
<b>Chapter 4 .....</b>	<b>82</b>
<b>Prioritization of candidate miRNAs and genes based on the triplex-forming potential between interacting miRNAs and genes. ....</b>	<b>82</b>
4.1 Introduction.....	82
4.1.2 Aims.....	85
4.1.3 Objectives .....	85
4.2 Methodology .....	86
4.2.1 MiRNA sequence extraction.....	87
4.2.2 Promoter sequence extraction of miRNA target genes.....	87
4.2.3 Trident tool .....	87
4.3 Results and discussion .....	89
4.4 Conclusion .....	105
<b>Chapter 5 .....</b>	<b>106</b>
<b>General discussion and future prospects .....</b>	<b>106</b>
<b>Appendix A.....</b>	<b>112</b>
<b>References .....</b>	<b>117</b>

# Chapter 1

## Literature review

### 1.1. Cancer overview

The human body is composed of trillions of cells and cancer can arise when one abnormal cell begins to divide and proliferate uncontrollably. Cancer can develop in any part of the body and while each cancer type has its own distinct characteristics, the overall processes that give rise to cancers are similar. Cells have specific rules on cell division; however, a cancerous cell follows its own program for proliferation. A mass of these abnormal cells forms a tumour which can either stay in its original tissue (benign) or invade other tissues (malignant). The cells from malignant tumours can establish new tumours throughout the body via the blood or lymphatic system, a process called metastasis (National Institutes of Health, 2007).

#### 1.1.1. Cancer and the cell cycle

The cell cycle has four consecutive phases, namely mitosis (M), Gap 1 ( $G_1$ ), synthesis (S) and Gap 2 ( $G_2$ ). According to Williams *et al.* (2011), the most important phases are the S phase, when DNA replication takes place and the M phase, when the cell divides to yield two identical daughter cells. Following the M phase is the  $G_1$  phase. This is the period between mitosis and the start of DNA replication, and in this phase the cell is metabolically active and constantly growing. After the  $G_1$  phase is the S phase, followed by the  $G_2$  phase. During the  $G_2$  phase, the cell continues to grow and proteins are produced to prepare for mitosis. A cell cycle checkpoint protects cells by not allowing damaged or incomplete chromosomes to be replicated and passed on to daughter cells. At the checkpoint in  $G_1$  phase, cell cycle arrest is overseen by p53, which is a protein activated in response to DNA damage. In many cancers, the gene that encodes p53 (TP53) is mutated therefore allowing the damaged DNA to be replicated and passed on to daughter cells rather than being corrected (Cooper, 2000).

#### 1.1.2. Tumour biology

Tumours that have become dedifferentiated and lost their tissue-specific traits are anaplastic as it is not possible to use histopathological criteria to determine their origin. These tumours are classified as cancers of unknown primary origin (CUP), showing the difficulty in determining the tumours original site of development in the patient. The progression of tumours is a multi-step process towards increasing aggressive and invasive behaviour. The first step is usually hyperplasia and this is when tissues have growths consisting of an abnormal number of cells. The cells may appear normal and may be in the correct tissue but their proliferation is no longer regulated. Metaplasia is an excessive proliferation of cells that are not normally found in that specific tissue; however, the cells appear to be normal. The next step in tumour progression is dysplasia and this occurs due to proliferation of abnormal cells. The cells giving rise to dysplastic tumours are considered to be abnormal cells as they have undergone changes in size, shape and organization. The last step before a tissue can become cancerous is termed neoplasia, and this is when the abnormal cells divide in the incorrect tissue (Weinberg, 2014).

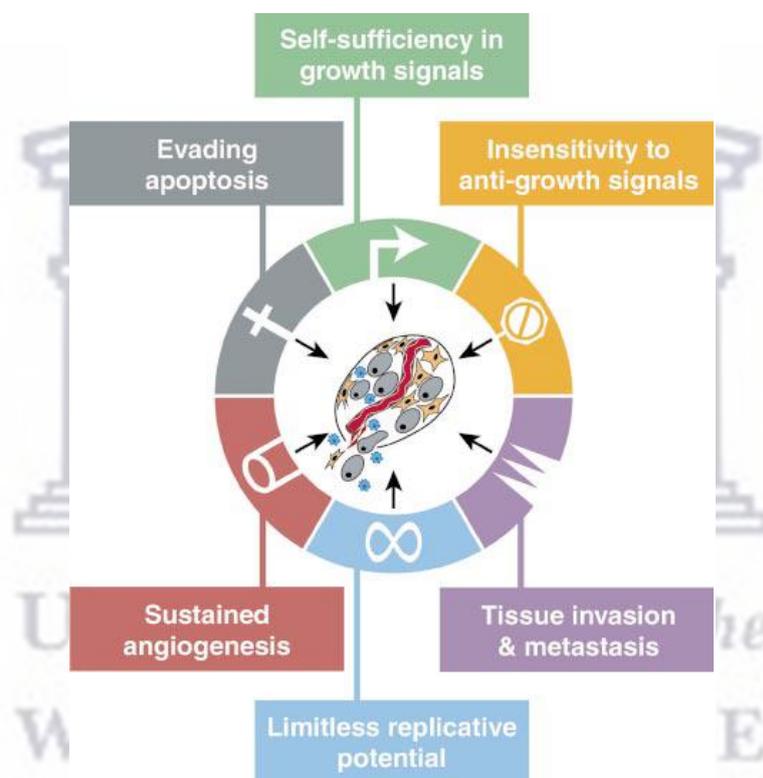
Cancer cells thrive and proliferate by prolonged growth signals (oncogenes) and avoiding anti-growth signals (tumour suppressors). Proto-oncogenes stimulate growth and a mutation in a proto-oncogene that allows it to be permanently activated leads to the cell growing out of control, and thus giving rise to cancer. The mutated proto-oncogene stimulating prolonged growth and division becomes known as an oncogene. Tumour suppressor genes either halt the cell cycle, repair DNA errors, or induce apoptosis. A mutation in a tumour suppressor gene allows for damaged DNA to be replicated, which could therefore give rise to cancer as well (Chow, 2010). Apoptosis is the process whereby cells follow a course towards death once specific stimuli is received. However, cancer cells avoid apoptosis which contributes to malignant transformation (Wong, 2011). In the 1970s, Kerr *et al.* (1972), connected apoptosis to the removal of “potentially malignant cells, hyperplasia and tumour progression”. Therefore, a reduction in apoptosis or resistance to it plays an important role in cancer (Wong, 2011).

Tumours can remain benign or they become malignant through the process of metastasis. A malignant tumour can metastasize by travelling through either the lymphatic system or the blood system (Hejmadi, 2010) and constructing secondary tumours in a distant tissue site (Leber *et al.*, 2009). Cancer cells that are not in close proximity to blood vessels create new

blood vessels, known as angiogenesis, in order to metastasize and invade nearby tissues (Neal and Berry, 2006).

### 1.1.3. Hallmarks of cancer

The hallmarks of cancer involve six major biological abilities acquired during the multistep advancement of tumours. These steps arise from genetic alterations enabling the transformation of normal cells into malignant cells (Hanahan and Weinberg, 2011). Figure 1.1 shows the six major hallmarks which will be discussed further.



**Figure 1.1:** The six hallmarks of cancer (Hanahan and Weinberg, 2011).

#### 1.1.3.1. Self-sufficiency in growth signals

Normal cells control the production and release of growth-promoting signals that moves them from a state of dormancy into an active state of proliferation (Hanahan and Weinberg, 2000), guaranteeing homeostasis in cell number and maintenance of tissue function. In cancer cells however, the growth and proliferative signalling pathways contain alterations in

their growth ligands, receptors or cytosolic signalling molecules, thus resulting in their uncontrolled growth and proliferation (Fouad and Aanei, 2017).

Without growth signals, healthy cells cannot proliferate thus demonstrating their dependence on these signals. Cancer cells thrive by creating a positive feedback signalling loop in which they synthesize their own growth factors, therefore reducing their dependence on exogenous growth signals and disrupting the crucial homeostatic mechanism (Hanahan and Weinberg, 2000).

In many cancers, growth factor receptors are overexpressed, enabling cancer cells to become hyperresponsive to ambient levels of growth factors that would not usually trigger a response (Hanahan and Weinberg, 2000).

### **1.1.3.2. Insensitivity to antigrowth signals**

Anti-growth signals maintain cellular and tissue homeostasis within healthy tissues. These signals block proliferation by either forcing cells out of their active state of proliferation, whereby they may reoccur if future signals allow, or by permanently relinquishing their ability to proliferate. In order for cancer cells to survive and thrive, they work to avoid and diminish these signals (Hanahan and Weinberg, 2000).

Tumour suppressors limit proliferation and cell growth and are generally deactivated or downregulated in cancer. Two key tumour suppressors in cancer encode the p53 and Rb (retinoblastoma-associated) proteins. The Rb protein receives signals from extracellular and intracellular sources and decides whether a cell will proceed into the cell cycle or not. p53 receives stimuli from stress and abnormality sensors within the cell and will halt the cell cycle if there is excessive DNA damage or unfavourable conditions. The Rb and p53 pathways are flawed in cancer cells therefore allowing uncontrolled proliferation (Hanahan and Weinberg, 2011).

### **1.1.3.3. Tissue invasion and metastasis**

The main aspect of a malignant tumour is the ability to invade nearby and distant tissues to form secondary tumours. For cancer cells to reach distant tissues, they have to (i) invade via

the extracellular matrix (ECM), along with the basement membrane and stromal cells, (ii) invade into blood or lymphatic vessels, (iii) endure transportation in circulation, (iv) exit the blood or lymphatic vessels at the parenchyma of distant tissues, (v) withstand and control foreign environments to form micro-metastases which may (vi) develop into macro-metastases (Fouad and Aanei, 2017).

Invasive and metastatic capabilities of a cell results in the change of proteins that link cells to their surroundings within a tissue. Such proteins include cell-cell adhesion molecules (CAMs) as well as integrins. The most commonly observed cancerous modification in cell-to-environment involves E-cadherin, which is an interaction molecule expressed on epithelial cells. The interaction between E-cadherins on neighbouring cells results in the stimulation of anti-growth signals. In most epithelial cancers, E-cadherin function is absent through various mechanisms that include mutational inactivation and transcriptional repression. This serves as evidence that E-cadherin acts as a tumour suppressor in epithelial cancers by inhibiting invasion and metastasis, thus making its elimination a crucial aspect for cancer cells (Hanahan and Weinberg, 2000).

#### **1.1.3.4. Limitless replicative potential**

The ability to replicate infinitely is a phenotype that is acquired during tumour progression and is vital for malignant growth conditions (Hanahan and Weinberg, 2000). This is a clear distinction from healthy cells, that can only pass through a restricted number of consecutive cell cycles due to two proliferation barriers namely (a) senescence, a generally irreversible entry into a viable but non-proliferative form, and (b) cell death when in crisis state (Hanahan and Weinberg, 2011).

The ends of chromosomes, called telomeres, has become a counting device for cell generations. With each cell cycle, telomeric DNA from the ends of all chromosomes are lost, thus making it possible to count the replicative generations. This loss of telomeric DNA is due to the inability of DNA polymerases to fully replicate the 3' ends of the chromosomal DNA during each S phase. Consecutive cycles result in destroyed telomeres and therefore unprotected chromosomal ends, which may result in death of the affected cell. Malignant cells with the potential to replicate limitlessly, protect and maintain their telomeres by one

of two mechanisms. They either upregulate their expression of the telomerase enzyme that adds hexanucleotide repeats onto telomeric DNA ends, or they activate Alternative Lengthening of Telomeres (ALT) which is a mechanism to maintain telomeres via “recombination-based interchromosomal exchanges of sequence information” (Hanahan and Weinberg, 2000).

#### **1.1.3.5. Sustained angiogenesis**

Blood vessels supply oxygen and nutrients to the cells which are essential for their function and survival, therefore compelling cells to exist within 100 µm of a capillary blood vessel. Angiogenesis is the process whereby new blood vessels are produced and is carefully regulated. Cancerous cells originally lack angiogenic abilities (Hanahan and Weinberg, 2000), therefore hindering their motive to expand as they require blood or lymphatic vessels as a transport to invade surrounding tissues. Members from the hypoxia-inducible transcription factor (HIF) family control the expression of genes implicated in angiogenesis, cell survival and metabolism, therefore making hypoxia an angiogenic trigger. Since hypoxia is a trait of tumours, it is understandable that they would have increased levels of HIF, correlating with a poor prognostic outcome (Fouad and Aanei, 2017).

Well known inducers of angiogenesis are members from the vascular endothelial growth factor (VEGF) family. They encode ligands involved in the generation of new blood vessels during stages of growth and postnatal development, as well as in homeostatic survival of endothelial cells. VEGF expression can also be upregulated due to hypoxia, thus further indicating how cancer cells ensure their survival via angiogenesis (Hanahan and Weinberg, 2011).

#### **1.1.3.6. Evading apoptosis**

Cell programmed death (apoptosis) is a natural way to ensure cell number and maintain homeostasis. When triggered by various signals, apoptosis takes place in a series of well-defined steps. Cellular membranes are disturbed, nuclear and cytoplasmic skeletons are destroyed, cytosol is released, chromosomes are degenerated and the nucleus is broken down. This entire process takes place within 30-120 minutes. There are two classes that make

up the apoptotic machinery, sensors and effectors. Sensors supervise the extracellular and intracellular environments by looking out for normal or abnormal conditions that dictate whether a cell should die or not. These signals regulate effectors which implement apoptosis (Hanahan and Weinberg, 2000).

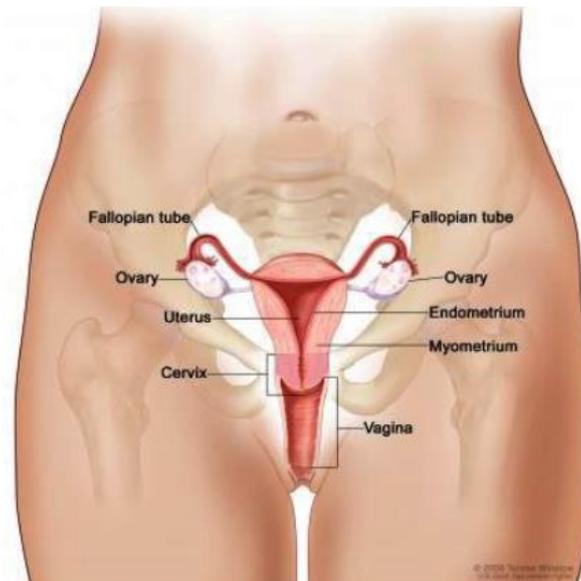
Tumour cells employ a variety of mechanisms in order to restrict and avoid apoptosis. The most common mechanism relates back to loss of TP53 function, therefore eliminating this important damage sensor from inducing apoptosis (Hanahan and Weinberg, 2011).

## **1.2. Ovarian cancer**

Ovarian cancer (OC) is the most common gynaecologic disease and has a 5-year survival rate of approximately 40% (Whittemore *et al.*, 1992). In 2009, the American Cancer Society stated that out of all gynaecologic malignancies, OC has the highest case-to-fatality ratio. This high fatality rate is mostly due to the fact that OC is usually diagnosed at an advanced stage and at this point the cancer has already metastasized within the peritoneal cavity (Lengyel, 2010). The risk of women developing OC in their lifetime is 1 in 71 and the possibility of dying from it is 1 in 95 (Razi *et al.*, 2016). There is a huge difference in overall cancer survival rates between developed and developing countries. A reason for this could be due to inadequate access to diagnostic and therapeutic procedures because of the expense that these procedures hold (Redaniel *et al.*, 2009). This would thus emphasize the need for cheaper diagnostic methods and therapies.

### **1.2.1. Physiology and anatomy of the ovary**

Females have two ovaries that form part of their reproductive system, as seen in Figure 1.2. The ovaries produce the eggs required in order to conceive a child and are connected to the uterus via the fallopian tubes (Torpy, 2011). The ovaries are located within the pelvic cavity and are approximately 2 to 3 cm in length.



**Figure 1.2:** Anterior view of the female reproductive system (Cancer Association of South Africa, 2016).

The ovaries are made up of two different types of cells, namely germ cells and somatic cells. Germ cells give rise to the oocytes (eggs) while somatic cells make up the granulosa, thecal and stromal cells (Richards *et al.*, 2010).

### **1.2.2. Histological subtypes of OC**

There are various types of OCs and they can be categorized according to the structures from which the tumours arise (Chen *et al.*, 2003). The four major categories for ovarian tumours are (i) Epithelial carcinomas, (ii) Germ cell carcinomas, (iii) Stromal carcinomas, and (iv) small cell carcinoma. For each category, there are many subcategories however only the main categories will be briefly overviewed.

#### **1.2.2.1. Epithelial carcinoma**

Epithelial OC is the most fatal gynaecological malignancy and according to Fagotti *et al.* (2010), patients with this type of OC have a 5-year survival rate of approximately 39%. This type of tumour arises from the cells lining the ovary and is the most common form of OC. As a result of tumour growth, cancerous cells shed into the peritoneal fluid and are able to spread to other parts in the peritoneal cavity.

### **1.2.2.2. Germ cell carcinoma**

Germ cell ovarian tumours originate in the cells forming the eggs. They make up about 20-25% of all ovarian tumours, but only 5% are malignant while the rest are benign (Shaaban *et al.*, 2014).

### **1.2.2.3. Stromal carcinoma**

Ovarian stromal tumours occur in the connective tissue cells inside the ovary and they comprise about 8% of all OCs (Chen *et al.*, 2003). A cancerous stroma produces molecules essential for tumour biology and a large ovarian stroma is linked to low survival rates in advanced stage of the disease (Davidson *et al.*, 2014).

### **1.2.2.4. Small cell carcinoma**

Small cell carcinoma of the ovary (SCCO) is a scarce, malignant and very aggressive tumour. It accounts for 0.1% of all OCs and is linked to a poor prognosis and high fatality rate. SCCO is an undifferentiated neoplasm and its origin is yet unknown (Orioni *et al.*, 2013).

## **1.2.3. Aetiology of OC**

There are various factors associated with OC development; specific factors may increase the chance of a woman getting OC while other factors may reduce the likelihood. Having certain risk factors does not mean that the woman is destined to develop OC and some women who have this disease may have no known risk factors.

There are various factors believed to reduce the risk of OC, and the biggest preventative strategies are using oral contraceptives and tubal ligation (McLemore *et al.*, 2009). The section that follows will focus on some of the risk factors for OC.

### **1.2.3.1. Genetic factors and family history**

The most crucial genetic risk factor for OC is a genetic mutation in the inherited BRCA1 and/or BRCA2 genes, which are accountable for approximately 85% of hereditary OCs (Toss *et al.*, 2015). Mutations in the BRCA1 and BRCA2 genes are predominantly associated with breast

cancer development, however since mutations in these genes are also involved in OC, it is safe to say that women who have had breast cancer are at increased risk for OC (Pruthi *et al.*, 2010).

Lynch Syndrome is an autosomal dominant condition resulting from errors in the mismatch repair region genes, MLH1, MSH2, MSH6 or PMS2 (Kastrinos, 2009). Mutations in these genes results in an increased susceptibility to cancer, and studies have shown that women with Lynch Syndrome have a high risk of developing OC (Lu *et al.*, 2013).

A study conducted by Negri *et al.* (2002), demonstrated that the risk of a woman developing OC increases if someone in her family has had ovarian or breast cancer. This thus shows that family history also acts as a genetic factor in the development of OC.

### **1.2.3.2. Reproductive and hormonal factors**

In 1971, Fathalla proposed that frequent ovulation increases the risk of DNA mutations due to the rupture and repair cycle of the ovarian surface epithelium. Since pregnancy suppresses ovulation, it lowers the risk of OC (Salehi *et al.*, 2008). According to Hunn and Rodriguez (2012), pregnancy reduces the risk of OC by one-third, with the reduction increasing with each additional pregnancy. It was also proposed by Rostgaard *et al.* (2003) that pregnancy removes premalignant and damaged cells from the ovary.

The risk of OC increases with the use of hormone replacement therapy (HRT). Various studies found a slight increase in OC risk for long-term users of oestrogen replacement therapy and noted that progestins might sporadically further increase the risk (Lacey, 2002; Riman, 2002; Daniilidis and Karagiannis, 2007).

### **1.2.3.3. Age**

Age is considered an important risk factor for OC as it usually affects older women. This disease is most prevalent in women aged between 50 and 70 years old, with 70% of cases being diagnosed in women who are older than 55 years (Sundar *et al.*, 2015).

There is also an increase in risk for women who started menstruating before the age of 12 years (McLemore *et al.*, 2009).

#### **1.2.3.4. Lifestyle and environmental factors**

It has been confirmed by Olsen *et al.* (2013) that obesity increases the risk for OC and according to Feng (2015), the risk is increased by 30%. Adipose tissue (fat tissue) produces high levels of oestrogen, which supports growth of ovarian surface epithelial cells. Therefore, an increase in adipose tissue would result in an increase in the production of oestrogen (Leitzmann *et al.*, 2009) and thus an over-proliferation of ovarian surface epithelial cells.

Many studies found a correlation between talcum powder usage and an increase in risk for OC development. The talc particles may either become fixed on the ovarian surface epithelial or it can be absorbed into the pelvic cavity where it is found in inclusion cysts. The foreign body in the inclusion cyst forms a granuloma which starts an inflammatory response. It is proposed that this inflammatory response leads to DNA damage which initiates the events needed for tumorigenesis (McLemore *et al.*, 2009).

There have been many studies attempting to determine dietary factors influencing OC risk, and results have been either conflicting or inconclusive. Therefore, further studies will be needed in order to conclude if dietary factors have an effect on the risk of OC (Hunn and Rodriguez, 2012).

#### **1.2.4. Symptoms of OC**

The early symptoms of OC are extremely mild therefore it is usually diagnosed in the advanced stages (Russo *et al.*, 2009). The symptoms are non-specific to OC and these symptoms include abdominal discomfort, nausea, indigestion, fatigue and frequent urination. The absence of early symptoms accounts for the late diagnosis of OC which then results in the low survival rate. This therefore further emphasizes the need for an early diagnostic tool (Burges and Schmalfeldt, 2011).

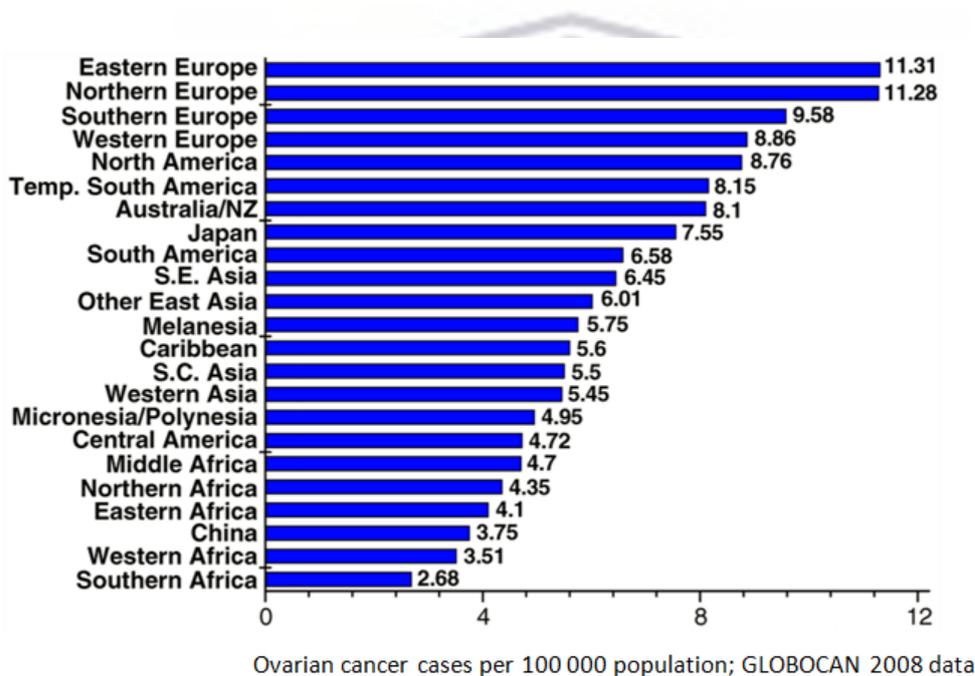
#### **1.2.5. Stages of OC**

There are four main stages in OC. The first stage is when the cancer is still enclosed within the ovaries and the second stage occurs when the cancer involves one or both ovaries and the tumour spreads to other regions within the pelvic cavity. Stage three OCs involve one or both

ovaries and the spread of the tumour to other regions within the peritoneum and/or spread to the lymph nodes. The fourth and final stage of OC involves the tumour spreading to distant regions in the body (Sahdev, 2016).

### 1.2.6. Global and local prevalence

In 2013, OC was ranked the sixth most common cancer and the seventh cancer-related cause of death amongst women globally. As seen in Figure 1.3.1, the highest incidence rate for OC was reported to be in Europe with more than 8 cases per 100 000, while the lowest incidence rate was said to be in Africa with less than 5 cases per 100 000.



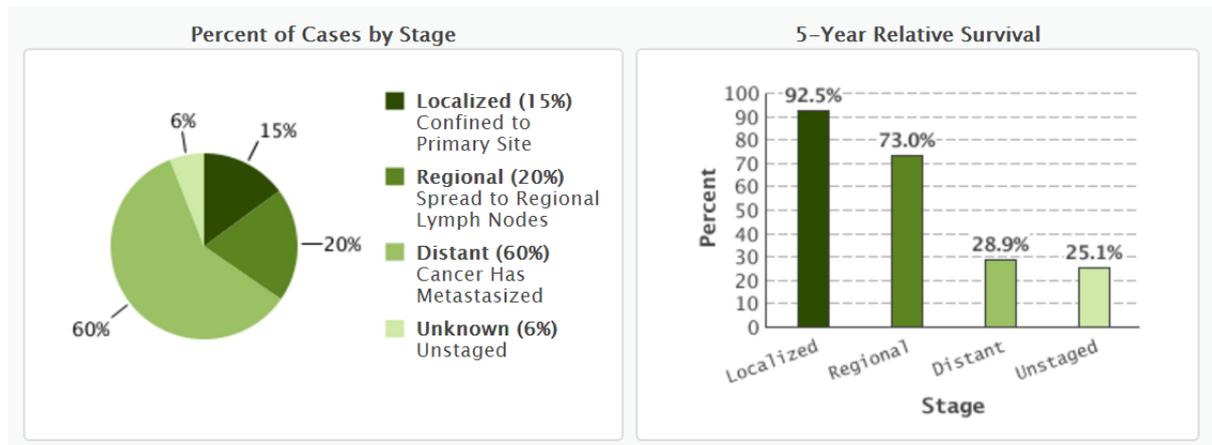
**Figure 1.3.1:** Worldwide incidence rates of OC in 2008 (Chornokur *et al.*, 2013).

While the incidence rate for OC in African women may be the lowest, the mortality rate is the highest. This is a result of racial-related health disparity as Africa has inadequate access to accurate diagnostic and therapeutic methods (Chornokur *et al.*, 2013).

Death rates per 100 000 individuals in South Africa for OC have increased by 41% from 1990 to 2015 (CNS reporter, 2016). In South Africa, 1 in 460 women have a lifetime risk of developing OC, which is the 9<sup>th</sup> cancer-related cause of death in South African women (South African Medical Research Council, no date). In 2001, the National Cancer Registry recorded

529 cases of OC in South Africa. Of these 529 cases, 207, 225, 69, and 16 were in white, black, coloured and Asian patients, respectively (Smith and Guidozi, 2009).

As mentioned earlier, the symptoms for OC are extremely elusive, thus resulting in advanced-stage diagnosis and low five-year survival rates. According to Chornokur *et al.* (2013) the survival rates for OC can increase to over 90% if it is caught in stage I. Figure 1.3.2 shows the percentage of cases from 2007 to 2013 and the relative survival rates by stage at diagnosis.



**Figure 1.3.2:** Percentages of OC cases by stage of diagnosis and their respective 5-year survival rates from 2007 to 2013 (National Cancer Institute, no date).

As observed in Figure 1.3.2, when OC is diagnosed at an early stage, preferably stage I, the survival rate is above 90%. This thus emphasizes the need for early detection of OC which would also positively impact the effectiveness of treatment. The reason for OC usually being diagnosed late is due to the ineffectiveness of current screening methods available.

### 1.2.7. Screening and diagnosis of OC

The main purpose of OC screening tests is for early detection of the disease or risk factors in seemingly healthy patients that are possibly at risk. Diagnostic tests confirm the existence or absence of OC and are performed on patients who either display the relevant symptoms or patients with a positive screening test result.

#### 1.2.7.1. Screening tests

Transvaginal ultrasonography (TVS) and the cancer antigen 125 (CA-125) are the two most considered methods for OC screening. A clinical trial performed in the United States reported that the usage of these tests did not decrease the mortality risk, whereas a trial performed in the United Kingdom recorded benefits of using these tests (Doubeni *et al.*, 2016).

TVS scans a woman's reproductive organs by using a probe that is inserted into the vagina. The probe emits sound waves that reflect off the body structures while a computer captures the waves and converts them into a picture (Cancer Association of South Africa, 2016). TVS is accurate in identifying abnormalities in the mass and physical structure of the ovaries; however, it is unable to distinguish between benign and malignant tumours. It has sensitivity to OC but a low positive predictive value when used alone for screening; therefore, it is usually used together with serum markers (e.g., CA-125) (Van Nagell and Hoff, 2014).

CA-125 is a glycoprotein of high molecular weight, expressed in large amounts by cancerous ovarian cells. The levels of CA-125 are found to be higher in cancerous cells than in healthy cells, thus it may assist doctors in establishing diagnosis. However, it is common when using CA-125, to obtain a false positive result as it is elevated in some benign conditions as well in other cancers (Rauh-Hain *et al.*, 2011). This thus lacks the sensitivity and specificity for OC, but it is still an important means for monitoring OC progression, reoccurrence as well as the effectiveness of treatment (Cancer Association of South Africa, 2016).

Since TVS and CA-125 are insufficient when used as a screening method alone, it has been observed that screening for OC is more effective when these methods are performed together (Van Nagell and Hoff, 2014).

#### **1.2.7.2. Diagnostic tests**

There are various diagnostic tests that can be performed to confirm the presence of OC. These include, imaging tests, blood tests and biopsies.

Imaging tests such as ultrasounds, computed tomography, positron emission tomography and magnetic resonance imaging have the ability to conclude if an abnormal pelvic mass is present, by using waves to generate images of the ovary (Doubeni *et al.*, 2016).

Blood tests have the ability to detect OC at an early stage based on the identification of certain proteins and DNA in a patient's blood, caused by or as a by-product of the cancerous cells (Robertson, 2005). However, some proteins may not be present during early stages of OC.

Surgery is the only way to accurately diagnose OC. This involves removing the ovary completely and analysing the cells. A biopsy of the ovary is not generally performed as the cancer may possibly spread, producing an even more advanced cancer (Chen and Berek, 2017).

There is currently no technology-based diagnostic technique that can accurately diagnose OC, therefore, performing surgery is still the preferred way to establish accurate diagnosis and staging of OC (Burgess and Schmalfeldt, 2011).

### **1.3. Biomarkers**

A biomarker is defined as a biological molecule identified in the blood, tissues and body fluids that is an indication of normal or irregular processes (Henry and Hayes, 2012). There are various types of biomarkers such as DNA, RNA, miRNAs, proteins, peptides and chemical modifications (Goossens *et al.*, 2015). They allow for the differentiation of individuals with a disease from those without the disease. The changes could be attributed to mutations, transcriptional alterations, and post-translational modifications. Biomarkers can be non-invasive and easily accessible from circulations, excretions or secretions, or they can be derived from tissues, which would involve biopsies (Henry and Hayes, 2012).

A biomarker can serve many functions in cancer. Biomarkers can either serve as a (i) diagnostic biomarker, which establishes a certain diagnosis and confirms the presence of cancer, (ii) prognostic biomarker, which conveys information about a probable cancer outcome, regardless of treatment used or it can be a (iii) predictive biomarker, which shows the effectiveness of treatment in cancer patients with the biomarker compared to patients without the biomarker (Ballman, 2015).

#### **1.3.1. Applications of biomarkers**

Since the development of cancer is an intricate process, it goes without saying that many changes will be made to cells, their contents and specific pathways. Biomarkers allow for the early detection of cancer as well as identifying a patient at risk. Over the years, cancer treatment has become relatively specific and target-oriented, with highly characterized targets in various cancers. Unfortunately, most targets have only been identified in advanced and metastatic cancers therefore restricting the effectiveness of treatment. There is a rationale that if targets can be found preferably in stage I, treatments will most likely be more effective (Negm *et al.*, 2002). Thus, there is a need to identify early stage cancer biomarkers which can serve as targets in treatment.

Biomarkers play important roles: (i) before diagnosis, in risk assessment; (ii) at time of diagnosis, in identifying the stage and progression of disease; and (iii) after diagnosis, in monitoring treatment effects and disease reoccurrence (Ludwig and Weinstein, 2005).

Techniques for OC biomarker discovery can be performed on patient tissue samples, blood, urine, and other body fluids. However, it is impractical to rely on tissues to identify biomarkers for early detection. Asymptomatic women who have no reason to believe that they have OC would not deem invasive surgeries, such as biopsies, to be necessary. Therefore, the most efficient approach would be to screen for biomarkers that can be identified from bodily fluids. Since OC is extremely fatal, biomarkers for its early detection must be of high sensitivity and high specificity (Stephen *et al.*, 2013).

### **1.3.2. The ideal biomarker**

The ideal OC biomarker would be used in a general screening process and it would allow for the diagnosis of women without symptoms. It should also have the following qualities:

#### **1.3.2.1. Non-invasive and inexpensive**

It would be preferred to identify the biomarker via bodily fluids instead of tissues as it is easily accessible from body fluids and will not require invasive surgery. It should also be performed using an easy laboratory test which reduces the cost compared to surgery (Fathi *et al.*, 2013).

#### **1.3.2.2. High specificity**

The biomarker should be highly specific to OC, with a specificity of 99.6% which will allow for a positive predictive value. It would be an even greater advantage if the biomarker is able to differentiate between subtypes and causes of this disease (Fathi *et al.*, 2013; Zhang *et al.*, 2011).

### **1.3.2.3. High sensitivity**

The biomarker should be highly sensitive for OC, which could help in early detection. The sensitivity should be more than 75% to avoid a false-negative result (Fathi *et al.*, 2013; Zhang *et al.*, 2011).

### **1.3.3. Current OC biomarkers**

There is yet no effective screening technique for the early diagnosis of OC (Rastogi *et al.*, 2016). There are various biomarkers used however none of them display the high specificity and the sensitivity required. The biomarkers currently used are discussed below:

#### **1.3.3.1. CA-125**

CA-125 is a high molecular weight glycoprotein and is expressed by approximately 80% of OCs. It lacks sensitivity and specificity as it is only elevated in 50-60% of stage I OCs and can be expressed at high levels in some benign conditions, pregnancy, menstruation as well as other cancers (Moore *et al.*, 2010). Regardless, it is still useful in monitoring treatment and cancer reoccurrence. Using CA-125 and TVS together is the standard method for detecting OC (Rastogi *et al.*, 2016).

#### **1.3.3.2. HE4**

Human Epididymis Protein 4 (HE4) is an over-expressed protein in OC. When compared to CA-125, HE4 has a higher sensitivity for stage I OC and a lower false-positive result. Many OCs that do not express CA-125, do in fact express HE4 therefore, combining these tests for OC diagnosis is an excellent strategy. While the specific function of HE4 currently remains unknown, it has been shown to be absent from normal ovarian surfaces and highly expressed on OC surfaces (Coticchia *et al.*, 2010).

### **1.3.3.3. VEGF**

VEGF is an angiogenic factor and thus is responsible for producing blood vessels via a process called angiogenesis. Cancers require an increase in blood supply in order to metastasize and invade other areas. VEGF levels are high in OC and also contribute to ascites (build-up of fluid in the peritoneal cavity). Many studies have shown that although high levels of ascites are a prognostic factor for OC, no differences in ascites levels were observed between control, benign and OC samples (Coticchia *et al.*, 2010). Therefore, VEGF is not specific and sensitive to OC.

### **1.3.3.4. Ova1**

Ova1 is a five-biomarker panel composed of second-generation CA-125 and other inflammatory and transport proteins. Even though this test has high sensitivity for OC, it is not used to establish diagnosis but rather for determining the possibility of malignancy in patients with ovarian tumours (Ueland, 2017).

## **1.3.4. MicroRNAs as biomarkers**

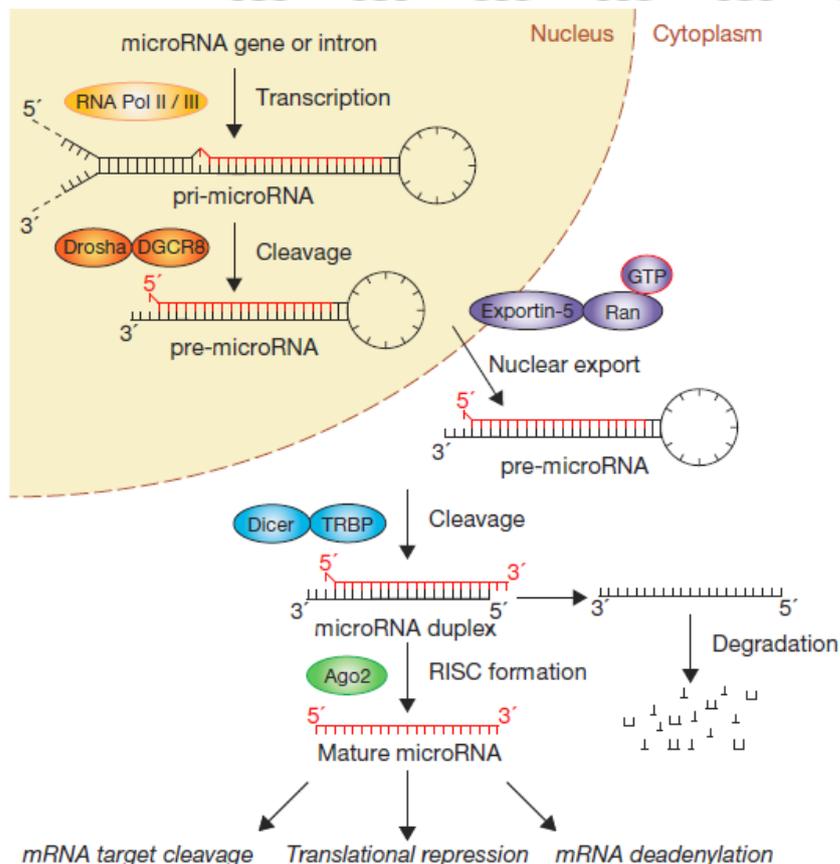
MicroRNAs (miRNAs) are short, non-coding, single-stranded RNA molecules, with a length of 17-25 nucleotides that are generally conserved across species. There are various studies establishing that miRNAs are present in different tissues and cell types, and their dysregulation is linked to many diseases, including cancer (Wang *et al.*, 2015). According to MacFarlane and Murphy (2010), miRNAs are predicted to constitute 1 to 5% of the human genome and they regulate approximately 30% of protein-coding genes.

### **1.3.4.1. Biogenesis and function of miRNAs**

MiRNAs are important regulators of gene expression as they control various cellular and metabolic pathways (MacFarlane and Murphy, 2010). Figure 1.4 illustrates the overview of the miRNA biogenesis pathway. In the nucleus, the stem-loop precursors from the transcribed miRNA gene serves as the primary precursor miRNA (pri-miRNA). The 3' poly-A-tail and 5' 7-methylguanosine cap of the pri-miRNA is cleaved by the miRNA processing complex,

comprised of RNase III Drosha along with its cofactor DGCR8, generating the precursor miRNA (pre-miRNA). The pre-miRNA is transported to the cytoplasm via exportin-5 (nuclear transport receptors) and Ran-GTP (nuclear protein), where it is cleaved by Dicer, producing a miRNA duplex (miRNA:miRNA\*; with the asterisk denoting the passenger strand). The miRNA duplex is loaded onto an Argonaute protein (Ago2) by the Dicer-TRBP complex, forming the RNA-induced silencing complex (RISC). (Wang *et al.*, 2015). The guide strand separates from the passenger strand which is subsequently degraded. The functional miRNA (guide strand) guides RISC to its target mRNAs and silences it through mRNA cleavage, translational repression or deadenylation. Although this is the standard pathway of miRNA biogenesis, some steps may be left out or replaced for certain miRNAs (Winter *et al.*, 2009).

At the 5' end of the mature miRNA, lies the "seed region" which is a 6-8 nucleotide sequence that has near-perfect complementarity to the mRNA target (Rolle *et al.*, 2016). The level of complementarity doesn't have to be an exact match, which therefore allows for one miRNA to target multiple mRNAs and multiple miRNAs to target a single mRNA (Barca-Mayo *et al.*, 2012).



**Figure 1.4:** MicroRNA biogenesis overview pathway (Winter *et al.*, 2009).

#### **1.3.4.2. Circulating miRNAs**

Although majority of miRNAs have been identified within cells, there are also miRNAs circulating in the extracellular environment. These circulating miRNAs are highly stable, can withstand unfavourable conditions for an extended period of time and have differential expression between cancerous and normal samples (Zhang *et al.*, 2015). Their availability in bodily fluids such as blood, saliva and urine, is appealing as it renders them a non-invasive biomarker (Allegra *et al.*, 2012).

It was recently discovered that these circulating miRNAs are contained within exosomes, along with proteins and nucleic acids, and can be taken up by surrounding or distant cells when exosomes circulate. Exosomes are membrane-bound vesicles present in nearly all biological fluids and play important roles in conveying information between cells. Due to the presence of specific surface proteins, the tissue or cell of origin of the exosome can be determined (Zhang *et al.*, 2015). Exosomes can be released from many different types of cells including T-cells, B-cells, epithelial cells, and tumour cells. For this reason, major interest has been in exosomal miRNAs as diagnostic biomarkers in various diseases (Tian *et al.*, 2017).

#### **1.3.4.3. Suitability of miRNAs as cancer biomarkers**

Cell communication is vital for tumour formation as tumour cells need to interact with each other and other healthy cells in order to subsist, thrive and metastasize. Communication between different tumour cells within the same patient (inter-tumour) allows for a heterogeneous population of cells to cooperate and survive in an unreceptive environment. Exosomes play an important role in the communication of cells, making the miRNAs transported and conveyed in tumour-derived exosomes of special interest (Thind and Wilson, 2016).

MiRNA alterations can be attributed to chromosomal abnormalities, genomic mutations, epigenetic changes as well as changes in miRNA biogenesis. Certain miRNA genes may be located at regions known to be mutated in cancer thus changing their function (Lan *et al.*, 2015). Circulating miRNAs released from cancer cells can prompt tumorigenesis in the

recipient cells, by acting as either oncogenes (oncomiRs) or tumour suppressive miRNAs. There have been numerous oncomiRs identified, promoting cancer onset, invasion and migration (Wang *et al.*, 2015). OncomiRs are predominantly overexpressed in cancer while tumour suppressive miRNAs are underexpressed (Svoronos *et al.*, 2016). Aberrant levels of miRNA expressions have been evaluated in many cancers and the studies suggest that they serve as good diagnostic biomarkers and predictors of the patient's response to treatment (Cheng, 2015).

MiRNAs are suitable biomarkers as they are easily available in bodily fluids and have a high degree of specificity and sensitivity. They are also incredibly stable and resistant to boiling, pH changes, cycles of freeze-thawing and chemical or enzymatic fragmentation (Larrea *et al.*, 2016). The demand for biomarkers that can accurately diagnose early stage cancers is crucial as the patient's survival and prognosis depends on the tumour stage at the time of detection, with early diagnosis almost always correlating to a better prognosis (Lan *et al.*, 2015).

The first case of miRNA association with a human cancer was discovered in 2002 by Calin and co-workers. They found that in chronic lymphocytic leukaemia (CLL), a commonly deleted chromosomal region (13q14) encodes miR-15 and miR-16. The deletion of the region containing the miRNA-encoding genes resulted in the deletion or downregulation of these miRNAs in CLL and also indicated that these genes are the targets of inactivation by allelic loss in CLL (Calin *et al.*, 2002). Since then, miRNAs have been linked to several other cancers including OC.

#### **1.3.4.4. MiRNAs in OC**

Various high-throughput technologies and studies have discovered miRNA upregulation or downregulation in OC when compared to healthy ovaries and the different cell types (Deb *et al.*, 2017). There are many miRNAs dysregulated in OC onset and progression, most of which are downregulated due to genetic and epigenetic processes. Various studies have investigated the potential of miRNAs in OC diagnosis by comparing different expression profiles in the ovarian surface epithelium with cancerous ovaries (Katz *et al.*, 2015).

It was reported that in OC patients, the under-expression of Dicer is strongly associated with advanced stage OC, and decreased expression of Drosha correlates with suboptimal surgery.

These findings imply that the abnormal processing of Dicer and Drosha contributes to tumorigenesis and undesirable clinical outcomes (Nakamura *et al.*, 2016).

As stated in the previous section, miRNAs may either act as oncomiRs or tumour suppressive. miRNAs. As an example, peritoneal dissemination is one of the key characteristics in the metastasis of OC and it has been reported that integrin  $\alpha 5$  is a crucial molecule in this process. Ohyagi-Hara *et al.* (2013) discovered that miR-92a prevents peritoneal dissemination by inhibiting integrin  $\alpha 5$ , therefore acting as a tumour suppressor. Table 1.1 shows the current circulating miRNA biomarkers used for the diagnosis of OC, and where the miRNA is sampled from.

**Table 1.1:** Current diagnostic miRNA biomarkers in OC (Nakamura *et al.*, 2016).

Source	Upregulated miRNA	Downregulated miRNA
Serum (exosome)	miR-21, miR-141, miR-200a, miR-200b, miR-200c, miR-203, miR-205, miR-214	
Serum	miR-21, miR92, miR-93, miR-126, miR-29a, miR-182, miR-200a, miR-200b, miR-200c, miR-21, miR-221, miR-7, miR-429, miR-141	miR-155, miR-127, miR-99b, miR-132, miR-26a, let7-b, miR-145, miR-25, miR-93, miR-145
Whole blood	miR-30c-1	miR-342-3p, miR-181a, miR-450-5p

Plasma	miR-205, miR-16, miR-21, miR-191 (endometriosis associated OC) miR-16, miR-191, miR-4284 (serous OC), miR-191-5p, miR-206, miR-548a-3p, miR-320a, miR-574-3p, miR-590-5p, miR-34c-5p, miR-106b-5p, miR-1274a, miR-625-3p, miR-720, miR-200b	let-7f, miR-19a-3p, miR-30a-5p, miR-645, miR-150-5p, miR-106b, miR-126, miR-150, miR-17, miR-20a, miR-92a
Urine	miR-30-5p	
Serum/plasma		let-7i-5p, miR-122, miR-152, miR-25-3p

### 1.3.5. Methods for biomarker discovery

As stated above, there are various types of biomarkers and there are also many different technologies allowing for the evaluation of changes in molecular profiles between healthy and disease samples. The discovery of biomarkers is crucial in clinical research as well as targeted therapies (Hu *et al.*, 2011; Mäbert *et al.*, 2014). There are various approaches to biomarker discovery including genomics, proteomics, metabolomics and bioinformatics. These approaches will be further discussed, with bioinformatics being the main focus of the next section.

#### 1.3.5.1. Genomic approach

Genomic approaches for biomarker discovery are based on the measuring of gene expression via various technologies including micro-arrays and polymerase chain reaction (PCR). Micro-

array analysis has uncovered many different biomarkers (Ilyin *et al.*, 2004) and measures gene expression based on the principle that complementary sequences will bind to each other. Since mRNA is less stable, the transcripts are converted to cDNA. The unknown DNA is fragmented, fluorescent markers are attached and they are then allowed to interact with probes of the DNA chip. DNA fragments that are complementary to the probe will bind and they can be identified based on their fluorescent emission. The fluorescence pattern is measured on a computer and certain gene expression levels are thus recorded. Although this method is fast, specific and sensitive, it is limited due to high costs involved (Govindarajan *et al.*, 2012).

### **1.3.5.2. Proteomic approach**

Proteomic approaches attempt to isolate, classify and characterize a range of proteins to obtain information about protein abundance, location, modifications and interactions. For biomarker discovery, proteomic approaches have certain advantages over other methods as it accounts for post-translational modifications which can affect protein function and activity. The final amount of protein can also vary greatly from the amount of mRNA transcribed (Ilyin *et al.*, 2004).

Various separating techniques have been established which can be divided into gel-based and non-gel-based categories. Two-dimensional gel electrophoresis (2DE), the most commonly used gel-based technique, separates proteins according to their isoelectric point (pI) and molecular weight. For biomarker discovery, the proteins from healthy and disease samples are stained to allow for the detection and identification of differentially expressed proteins (Hudler *et al.*, 2014).

### **1.3.5.3. Metabolomic approach**

Cancer metabolism has been one of the main areas of focus in biomarker discovery, since 1956 when Otto Warburg showed that cancer cells depend on anaerobic metabolism. The study also proposed that cells have a higher rate of glycolysis as well as an overproduction of lactic acid, even under normal oxygen levels. Altered cancer metabolism results in higher levels of reactive oxygen species (ROS), enabling cell survival and proliferation. Cancer cells

develop adaptations enabling them to maintain ROS levels below toxic levels, as excessive ROS production triggers apoptosis (Mäbert *et al.*, 2014).

Metabolomics has developed into an approach that complements genomic and proteomic technologies, as the analysis of metabolites are just as crucial as genome and proteome analyses for understanding cellular functions. Since there is no single technique that can analyse different types of molecules, a combination of techniques is required in this field (Jain, 2010). A metabolomic-based approach for biomarker discovery generally comprises of two platforms; (i) either nuclear magnetic resonance (NMR) or mass spectrometry (MS) along with (ii) separation techniques. Various metabolite biomarkers including fatty acids, amino acids, and lipids, have been discovered for many different types of cancers (Mäbert *et al.*, 2014).

The limitations of using metabolomic approaches arise from the errors in study designs and experimental procedures. Along with the advantages that each technology has for biomarker discovery, they also have their own set of disadvantages. This makes bioinformatics for biomarker discovery more appealing.

#### **1.4. Bioinformatics**

The definition of bioinformatics is stated by Luscombe *et al.* (2001) as “a management information system for molecular biology” that has many applications. It uses computational techniques to comprehend and categorize biological information. Since biological data are being produced at an exponential rate, computers allow for the storage of this data as they can handle an immense amount of data. Bioinformatics allows for the access and submission of existing and new work respectively. The information is stored in databases thus making it easily accessible. There are various databases focusing on protein sequences, molecular structures, genomic and nucleotide sequences, gene expression as well as integrated databases which allows for the incorporation of relevant information.

Bioinformatics has been applied many times in OC. For example, Xue *et al.* (2015), used bioinformatics to investigate the molecular interaction of the NSC319726 gene in OC, and discovered that the gene might play an effective role in OC by targeting certain genes implicated in the oocyte meiosis pathway. Another study conducted by Du *et al.* (2015), used

bioinformatics to identify differentially expressed genes as well as OC-related genes acting as potential therapeutic targets.

It is evident that bioinformatics is extremely useful in identifying novel interactions which could aid in OC diagnosis and treatment.

#### **1.4.1. Advantages of bioinformatics**

Bioinformatics allows individuals to access, sort out, analyse, predict, and store biological data (Bayat, 2002), making the accessibility of information much easier and cheaper. Before experimental work is performed, bioinformatics can be employed to select candidates that are of priority (Thébault *et al.*, 2015), which is also cost efficient as laboratory reagents will only be bought once the list of potential candidates are narrowed down. Applying bioinformatics in cancer research aids in understanding the mechanisms behind the disease. It also allows for the identification and validation of novel biomarkers (Wu *et al.*, 2012). Bioinformatics also has additional advantages such as fast sequencing abilities and it allows for enormous storage capability (Mishra, 2016).

#### **1.4.2. Biomarker discovery**

Biomarker discovery depends on the idea that certain molecular species displaying the highest degree of variation across phenotypes may be identified as potential biomarkers. The traditional method for biomarker discovery involves analysing a particular gene or protein that is aberrantly expressed in disease tissues when compared to normal tissues. Since traditional methods mostly focus on gene expression levels, it does not necessarily provide information on the interactions of these markers, on a gene or protein level. Using bioinformatics for biomarker discovery entails analysing the list of potential markers as well as their signalling and interactions in order to form a deeper understanding and analysis of the proposed biomarkers. Because bioinformatics offers a predictive outcome, the proposed results must be validated molecularly before it can reach the clinical setting (Azuaje, 2013).

With the use of bioinformatics, a vast number of biomarkers have been discovered, but only a small number of them have been established in the clinical setting. This is because most biomarkers discovered are found to be irreproducible (Wang *et al.*, 2015). Bioinformatics is

also only a prediction network and cannot replicate the exact conditions within a cell, therefore molecular testing will still need to be performed in order to validate the results obtained.

## 1.5. Problem identification

OC is the most common gynaecologic malignancy and accounts for approximately 4% of all cancers diagnosed in women (Brain *et al.*, 2014). It is known as the 'silent killer' as the early symptoms are extremely mild and very easy to ignore. More than 70% of women are usually diagnosed in the advanced stages but by this point the cancer has already metastasized to other regions distant from the ovary (Zhang *et al.*, 2011). Due to the absence of early symptoms, the five-year survival rate of this disease is about 40% (Whittemore *et al.*, 1992). Early detection of OC is therefore important as it could help increase survival rates. However, current diagnostic biomarkers are unsuccessful due to the low sensitivity and specificity for OC (Zhang *et al.*, 2011). Since there are no adequate biomarkers for OC diagnosis, it would be of crucial significance to identify a sensitive and specific biomarker for OC to be able to distinguish between normal and cancerous ovaries.

MiRNAs show great promise as biomarkers as they are proven to have differential expression profiles between cancer types according to diagnosis and the developmental stage of the tumour, therefore being highly sensitive and specific. They are extremely stable and easily accessible in body fluids such as the blood, saliva, and urine, thus rendering them non-invasive (Wang *et al.*, 2015).

Due to the importance of this topic, the current study was employed to identify potential miRNAs and their target genes that can serve as biomarkers for the early diagnosis of OC, using *in silico* methodologies. Specific study aims are as follows:

- (i) Identification of miRNAs and their target genes based on a sequence similarity approach.
- (ii) Identification of miRNAs and genes using patient clinical data extracted from TCGA, followed by the implementation of statistical parameters to isolate the candidates.

- (iii) Analysis of triplex-forming potential between candidate miRNAs and target genes identified from each pipeline.
- (iv) Prioritization of candidate list for future experimentation.



## Chapter 2

### Identification of miRNAs and target genes as biomarkers for the early stage diagnosis of OC via a sequence similarity approach

#### 2.1 Introduction

OC is a fatal malignancy with a 5-year survival rate of about 40%. This low survival rate arises from the lack of early symptoms which results in an almost always late stage diagnosis. It is proposed that if OC is diagnosed in the early stages, the survival rates will increase to 80% (Burges and Schmalfeldt, 2011). The symptoms of OC are extremely vague and can be confused for something minor. These symptoms include bloating, loss of appetite, pelvic pain and increased urinary frequency (Sundar *et al.*, 2015).

If a patient has these symptoms, they will be required to undergo a full physical examination, to evaluate for pelvic and abdominal masses. This approach however lacks accuracy as a mass could be missed or mistaken for another condition. If a mass is detected, the patient would then be recommended to undergo transvaginal ultrasonography to determine the ovarian architecture and vascularity, differentiate between cystic and solid masses, and identify ascites (Doubeni *et al.*, 2016).

Biomarkers are also commonly used to diagnose and detect OC, however due to their lack of sensitivity and specificity to OC, they are failing. This emphasizes the need to identify more sensitive and specific biomarkers. As stated in the previous chapter, miRNA dysregulation has been linked to a variety of diseases including OC, thus making them and their sequences appealing for diagnostic purposes (Fathi *et al.*, 2013).

Roughly 60% of human genes are governed by miRNAs, involved in crucial processes such as the immune system, cell cycle, development, differentiation, proliferation, metabolism and inflammation. It is proposed that overexpressed miRNAs may act as oncogenes by downregulating tumour suppressor genes, and underexpressed miRNAs function as tumour suppressors by negatively regulating its oncogenes. For example, the *RAS* oncogenes (*H-*, *K-*,

and *N-RAS*) contain binding sites in their 3'UTR for miRNAs from the let-7 family. These miRNAs are generally downregulated in various tumours and were shown to negatively regulate the *RAS* oncogenes, therefore acting as tumour suppressors (Kinose *et al.*, 2014).

Since miRNAs exert their function by regulating target genes, identifying the functions of these targets are crucial in understanding their biological role (Wong and Wang, 2014; Hammond, 2015).

Functional genomics comprises of genome-wide methods to annotate functions of genes and proteins as well as their interactions. The data from DNA sequencing, gene expression and protein functions are combined to model powerful networks that control gene expression, cell differentiation, and cell cycle progression. Technological advancements such as the accessibility of full genome sequences allows for studying cells at a systems level. Knowledge regarding gene function and regulatory pathways can be expanded by identifying the abundance of transcripts in diverse cell types under a variety of conditions (Bunnik and Le Roch, 2013).

Approximately one-fifth of all OC cases are due to hereditary conditions, with 65-85% of these cases resulting from a germline mutation in the BRCA genes causing defective DNA repair. Carriers with a BRCA1 or BRCA2 mutation have a 54% and 84% increased lifetime risk of developing OC and breast cancer respectively. Currently, there are at least 16 genes known to be implicated in the hereditary OC mechanism (Toss *et al.*, 2015). Sporadic OCs arise from genetic mutations that are not inherited, in genes such as p53, Ki-ras and *erbB-2* (Angioli *et al.*, 1998). Identifying mutations in genes linked to OC is a crucial step for diagnostic and therapeutic potential (Toss *et al.*, 2015).

MiRNA loci that are clustered together may contain members of either the same or different families. The cluster of sequences, even from different families, all share similar targeting properties and therefore share a similar sequence (Marco *et al.*, 2013). In 2008, Lu *et al.* discovered evidence that miRNAs deriving from clustered miRNA genes are more inclined to share similar functional roles and disease implications. This finding is fundamental in identifying novel disease-associated miRNAs (Lu *et al.*, 2008; Kamanu *et al.*, 2013), and serves as the basis behind employing a sequence similarity approach in this chapter. The understanding is that if a novel miRNA with no links to OC shares a certain degree of similarity

in sequence to a miRNA proven to have implications in OC, then it will most likely have the same dysregulated function.

Due to the large amounts of biological data stored online, many bioinformatic approaches implement data mining to discover novel compounds and molecules (Zaki *et al.*, 2007). Data mining involves the process of extracting knowledge from great amounts of data. In bioinformatics, this is employed to identify novel significant patterns and relationships. A few applications of data mining in bioinformatics includes gene identification, detecting protein functional domains and disease diagnosis and prognosis. The results from data mining falls into one of two categories: (i) supervised learning and (ii) unsupervised learning (Raza, 2012). In supervised learning, the sample data is analysed from a source with the correct classification already assigned (Sathya and Abraham, 2013). Classification, estimation and prediction are examples of tasks in supervised learning. Classification organizes data into various predefined classes, estimation provides a value for an unknown continuous variable, and prediction is similar to classification and estimation however, data is classified according to future estimated behaviour. Unsupervised learning involves no variable being selected as the target, instead the aim is to establish an association between all variables. Examples of unsupervised learning are: (i) association rules, (ii) clustering and (iii) description and visualization. Association rules group certain variables together, clustering gathers a population into clusters or subgroups, and description and visualization represents the data through visual techniques (Raza, 2012).

Databases are crucial for storing the vast amounts of data generated, while still being constantly updated and compared to other data. Databases comprising of gene sequences are split into two types: primary databases and secondary databases. Primary databases contain direct experimental results whereas secondary databases combine data from primary databases as well as other data such as gene variants and sequences information. There are various types of databases containing information on genetic diseases, gene sequence, mutations and gene and protein expression levels (Bianco *et al.*, 2013).

## **2.1.1 Biological databases**

A biological database comprises of organized data that can be accessed, maintained and updated. Sequence and structure databases are two of the broad categories, with sequence databases applying to both nucleic acids and proteins, and structure databases being applicable to proteins only. There are various databases available for both study and research by industries and academic institutions (Babu, 1997).

Since the discovery of miRNAs, a variety of bioinformatic tools have been created to study their physiological roles and predict their regulated targets (Stępień *et al.*, 2018). There are many different online databases to recover not only miRNAs expressed in different species, but also differentially expressed miRNAs in a variety of diseases, including OC.

### **2.1.1.1 MiRNA discovery databases**

#### **2.1.1.1.1 MiRBase**

MiRBase is a publicly available online database containing information on all published mature miRNAs, such as miRNA sequence data, annotation, as well as predicted gene targets. MiRBase is the dominant repository for miRNA information, thus making it a reliable and efficient source of obtaining both precursor and mature miRNAs in various species (Griffiths-Jones *et al.*, 2006). The database was created in 2002 and was previously called the miRNA registry. It collects data submitted from authors as well as publications identifying novel miRNAs (Kozomara and Griffiths-Jones, 2011).

#### **2.1.1.1.2 DbDEMC 2.0**

DbDEMC 2.0 is an integrated database that presents miRNAs differentially expressed in human cancers, validated via high-throughput approaches (Yang *et al.*, 2016). This database combined expression profiles from 48 miRNA microarray data sets in peer-reviewed articles to provide insight into the differential expression levels of disease-linked miRNAs. DbDEMC is also preferred as it provides data on miRNA expression for a variety of cancer cell lines (Yang *et al.*, 2010).

### **2.1.1.1.3 Mir2disease**

Mir2disease is a manually curated database containing miRNAs that are dysregulated in many different human cancers. To establish this database, over 600 literature papers were consulted and 1939 associations between 299 miRNAs and 94 diseases were reported. Mir2disease also provides the approach employed to detect the miRNAs respective expression patterns, including qRT-PCR, microarrays or northern blotting (Jiang *et al.*, 2009).

### **2.1.1.1.4 MiRandola**

MiRandola is a manually curated database comprising of extracellular circulating miRNAs, with an option for users to contribute submissions. Depending on their extracellular form, miRandola classifies these circulating/extracellular miRNAs into four categories: (i) miRNA-Ago2, (ii) miRNA-exosome, (iii) miRNA- High-density lipoprotein and (iv) miRNA-circulating (Russo *et al.*, 2012).

### **2.1.1.1.5 MiRCancer**

MiRCancer is an online database consisting of miRNA-cancer associations identified through textmining, followed by manual validation. The database works by implementing 75 constructed rules to identify the miRNA-cancer associations in PubMed. In 2013, miRCancer recognized 878 relationships between 236 miRNAs and 78 human cancers through consulting more than 26 000 PubMed articles (Xie *et al.*, 2013).

## **2.1.1.2 Tools for sequence similarity analysis**

### **2.1.1.2.1 CD-HIT**

Due to large scale genome projects, the sizes of databases that contain biological sequences are increasing at a rapid rate. This strengthens the call for bioinformatic tools that can organize and analyse data effectively. Since biological sequences are related and may share homology, clustering and determining a representative sequence is an efficient way to solve many sequence analysis problems (Huang *et al.*, 2010).

CD-HIT-EST is a tool that clusters either DNA or RNA according to a user-specified sequence identity level, based on a greedy incremental clustering algorithm. This algorithm involves the sorting of sequences in order of decreasing length, with the longest sequence becoming the representative of the first cluster (Li and Godzik, 2006). The remaining sequences are compared to the representative sequence and if their similarity is above a specified threshold, it is grouped in that cluster as a duplicate and if not, that sequence becomes the representative of a new cluster (Manconi, *et al.*, 2016).

CD-HIT-EST-2D compares two nucleotide datasets, db1 and db2. The sequences in db1 are arranged in order of decreasing length and each sequence in db2 is compared to each sequence in db1, starting with the longest one. If the similarity is greater than a given threshold, the sequence is grouped with its similar one in db1. At the end of the search, CD-HIT produces two files; (i) a report of the similar sequences between db1 and db2, and (ii) a list of sequences in db2 that are not comparable to any sequence in db1 (Li and Godzik, 2006).

#### **2.1.1.2.2 BLAST**

Basic Local Alignment Search Tool (BLAST) is a program that can be used online or as stand-alone tool to search sequence similarities. It works by identifying short matches between two sequences and seeks to perform alignments from these matches. BLAST also provides statistical information to interpret the biological significance of the data, in the form of an E-value (expect value). There are various different types of BLAST programs to compare protein and/or nucleotide sequences (McGinnis and Madden, 2004) and according to Babu (1997), all have been “designed for speed, with a minimal sacrifice of sensitivity”.

#### **2.1.1.3 MiRNA gene target database**

##### **2.1.1.3.1 MiRDip**

Algorithms for miRNA target gene prediction, usually pair the miRNA seed region to a similar mRNA sequence. Many factors complicate this binding, especially the imperfect miRNA-mRNA binding that occurs, therefore single base-pair mismatches should be considered. MiRDip is a free online data portal with combined data from various databases, that allows

for the visualization and interpretation of miRNA-target gene networks. Databases predicting miRNA target genes consider various traits of miRNA:mRNA target binding. These traits include seed sequence match, conservation, site accessibility, free energy of the miRNA:mRNA duplex, contribution of multiple binding sites, local ALU content, local mRNA sequence, ribosomal shadow, position effects and 3' pairing (Shirdel *et al.*, 2011).

Since miRNAs fall under non-coding RNAs and therefore do not get transcribed into proteins, their biological function is examined through the identification of their gene targets along with their function (Ling *et al.*, 2013). The general pipeline employed in miRNA research involves identifying their gene targets using various prediction tools, such as MiRDip. MiRDip contains approximately 152 million human miRNA predicted target genes retrieved from 30 independent sources. It provides an integrative score allocated to each target to increase accuracy of the predicted interaction (Tokar *et al.*, 2017).

#### **2.1.1.4 Gene annotation databases**

##### **2.1.1.4.1 PAGENBASE**

Pattern gene database (PaGenBase) combines gene patterns from literature and data mining. Pattern genes can be defined as a group of genes exhibiting specific expression patterns under various physiological conditions. Housekeeping, selective/specific, and repressed genes are three categories currently attracting great attention. Housekeeping genes are believed to preserve basal cellular functions as they are ever-present in all tissues under all types of physiological conditions and developmental stages. Specific/selective genes are expressed preferentially under certain conditions, whereas repressed genes are unanimously expressed except under specific conditions. Identifying gene patterns serves as a gateway to understanding gene functions and exploring molecular mechanisms leading to pathogenesis. PaGenBase comprises of pattern genes identified through the comparison of their expression levels under serial conditions, including in various tissues or developmental stages (Pan *et al.*, 2013).

##### **2.1.1.4.2 DAVID**

Database for Annotation, Visualization and Integrated Discovery (DAVID) is a publicly available tool, established in 2003, that annotates biological meaning and function to large genes lists (Huang *et al.*, 2009). This database originally comprised of four core elements; annotation tools, GoCharts, KeggCharts, and DomainCharts. (i) Annotation Tools are used to assign functions to the input list of genes. (ii) GoCharts presents differentially expressed genes according to their biological processes, molecular functions and cellular component functions. Biological processes entail various extensive functions, whereas molecular functions define tasks achieved by individual gene products. Cellular component functions comprise of genes with functions in subcellular structures and locations. (iii) KeggCharts exhibits the distribution of differentially expressed genes between various KEGG biochemical pathways. This function enables for the implication of genes in various diseases through identifying and analysing the pathways linked to the specific disease. (iv) DomainCharts depicts the distribution of differentially expressed genes between families of protein domains (Dennis *et al.*, 2003).

Currently, the DAVID annotation tool has over 40 categories including protein-protein interactions, functional domains, disease links, homologies, gene tissue expression, and many more. The expanded annotation tool allows for a more comprehensive analysis due to various biological features being available in a single space. The clustering tool, a new feature in functional annotation, runs on a novel algorithm that evaluates the relationships between the annotated terms based on the degrees of their co-association genes, with the intent to group similar, redundant and heterogenous annotation content from the same or different resources into annotation groups (Huang *et al.*, 2007).

### **2.1.1.5 Protein network database**

#### **2.1.1.5.1 STRING**

Search tool for recurring instances of neighbouring genes (STRING) is an online web-server that identifies interacting genes/proteins to the query gene/protein. STRING also retrieves all genes occurring within potential operons for the query gene, as it is noted that genes constantly occurring in each other's vicinity in potential operons within the genome tend to encode functionally interacting proteins (Snel *et al.*, 2000).

Protein-protein interactions may occur (i) directly through physical binding, or (ii) indirectly - due to a shared substrate in a metabolic pathway, by controlling each other transcriptionally, or by interacting in bigger multi-protein assemblies. Methods predicting these functional associations are based on the notion that functionally associated proteins are encoded by genes sharing common selection pressures (Von Mering *et al.*, 2003).

While most of the protein interactions in STRING are imported from other databases, a large amount predicted interactions are also produced *de novo*, based on systematic genome comparisons. Fully sequenced genomes are imported and searched for three types of associations; namely conserved genomic neighbourhood, gene fusion events, and co-occurrence of genes across genomes. The goal is to discover pairs of genes which seem to be under similar selective pressures during evolution, so that they may be deemed as functionally associated (Von Mering *et al.*, 2005).

### 2.1.2 Aims

The aim of this chapter is to identify novel miRNAs via a sequence similarity approach, followed by the identification and characterization of their target genes to serve as biomarkers for the early diagnosis of OC, using various *in silico* approaches.

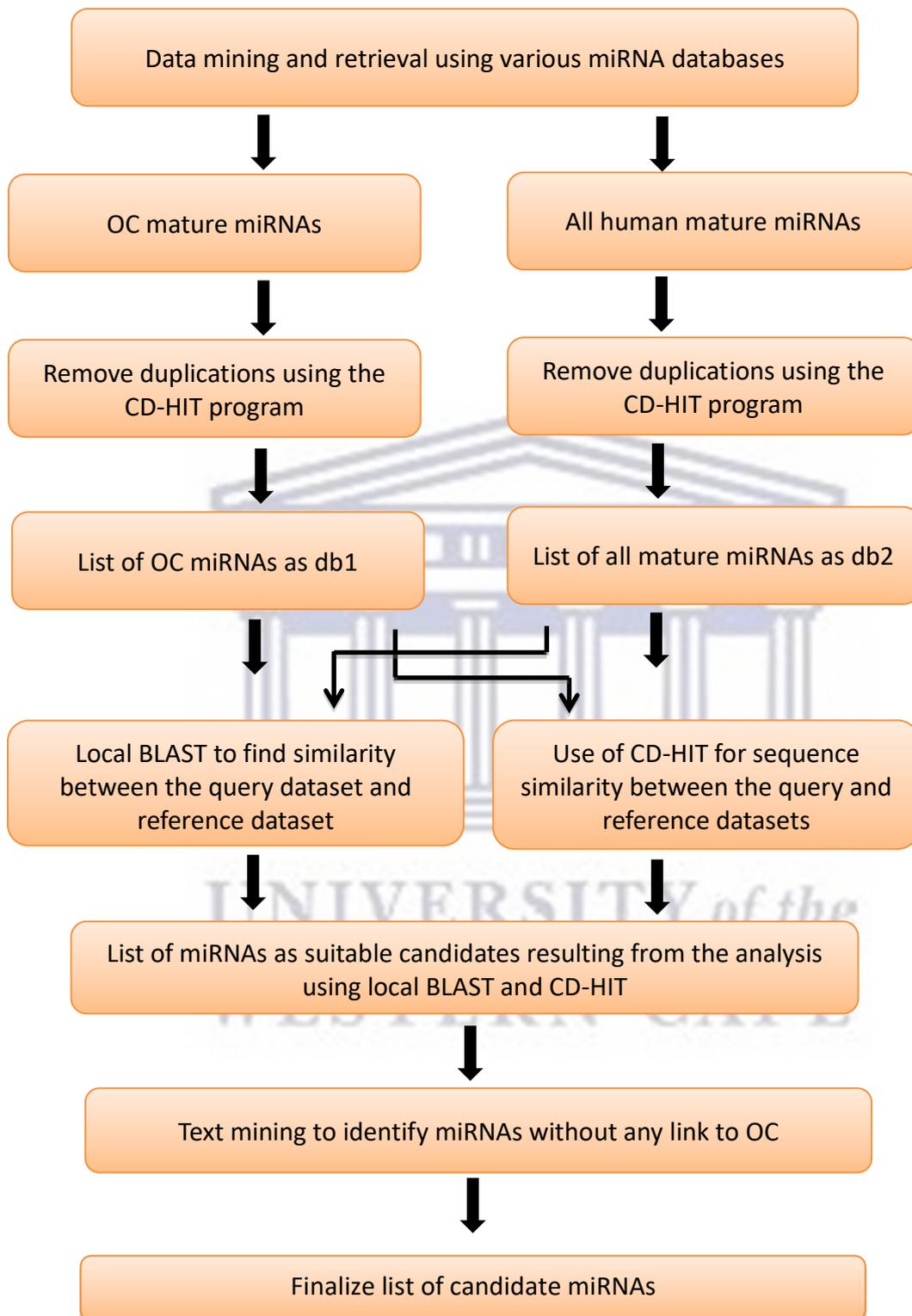
### 2.1.3 Objectives

- Retrieve miRNAs implicated in OC along with their sequences to create db1
- Retrieve all human mature miRNAs and their sequences to create db2
- Sequence similarity analysis via local BLAST and CD-HIT programs, using db1 and db2
- Create a list of miRNAs with high similarity to miRNAs already linked to OC
- Curation through textmining of the list of miRNAs to ensure no implication in OC
- Identify the target genes of the potentially novel miRNAs using mirDIP
- Extraction of all ovary-specific genes from PaGenBase
- Venn diagram construction for each miRNA's target genes that intersect with those expressed in the ovary
- Functional annotation to prioritize genes with direct links to OC and determining of their function using DAVID

- Discover protein interactions of the prioritized genes with each other using STRING
- Create a final list of miRNAs and their target genes as candidate biomarkers for OC diagnosis



## 2.2 Methodology



**Figure 2.1:** Flow chart representing the outline of the *in-silico* methodology employed for miRNA identification in this chapter.

## **2.2.1 Data mining**

To identify all mature miRNAs expressed in the human body, miRBase was utilized. For the extraction of OC-implicated miRNAs, dbDEMC, mir2disease, miRandola and miRCancer was employed.

### **2.2.1.1 MiRBase**

MiRBase was launched using the URL [www.mirbase.org/](http://www.mirbase.org/), and the “browse” tab was selected with *Homo sapiens* being specified. The option to view high confidence miRNAs was selected, and all mature sequences were downloaded. The miRNAs were stored in fasta format as “allmirna.fasta”.

### **2.2.1.2 dbDEMC 2.0**

Under the “browse” tab on the dbDEMC homepage (<http://www.picb.ac.cn/dbDEMC/>), OC was selected, all experiment IDs were ticked, and the miRNA sequences were downloaded.

### **2.2.1.3 MiR2Disease**

On the mir2disease homepage (<http://www.mir2disease.org/>), OC was searched under “search by disease name”, “malignant neoplasm of the ovary” was selected, and the sequences were downloaded.

### **2.2.1.4 MiRandola**

The full MiRandola database (<http://mirandola.iit.cnr.it/index.php>) was downloaded and only miRNA sequences linked to OC were kept.

### **2.2.1.5 MiRCancer**

Under the “download” tab on the MiRCancer homepage (<http://mircancer.ecu.edu/index.jsp>), the “miRCancerOctober2017” dataset was downloaded, and miRNAs implicated in OC were kept.

The miRNA sequences from all four databases with implications in OC were combined and duplications were removed in excel. The miRNAs were stored in fasta format as “ovmirnas.fasta”.

### 2.2.2 Duplication removal via CD-HIT

CD-HIT-EST was used to remove duplicates by clustering query datasets that met a specified similarity threshold. On the CD-HIT web server homepage ([http://weizhongli-lab.org/cdhit\\_suite/cgi-bin/index.cgi?cmd=Server%20home](http://weizhongli-lab.org/cdhit_suite/cgi-bin/index.cgi?cmd=Server%20home)), the CD-HIT-EST tab was selected and “allmirna.fasta” was uploaded as the query. The sequence identity was set to 0.99 and all other parameters were left as default. The output fasta file and cluster file were downloaded.

The page was then reset and “ovmirnas.fasta” was uploaded as the query. The parameters specified were the same as before.

### 2.2.3 Sequence similarity analysis

#### 2.2.3.1 CD-HIT

In order to identify sequences that share similarity, CD-HIT-EST-2D was run on Ubuntu software with the following command line:

```
cdhit-est-2d -i ovmirna \fasta.txt -i2 allmirna \fasta.txt -O result -c0.9 -n8
```

Where: “-i” is db1 with the validated OC miRNAs, “-i2” is db2 with all human mature miRNAs, “-O” is the output name, “-c” is the sequence identity threshold set to 0.9 and “-n” is the word size specified as 8.

#### 2.2.3.2 BLAST

BLASTN was used in this section to search nucleotide-nucleotide similarity. The ncbi-blast-2.7.1+win64.exe program was downloaded from NCBI and installed. A database was created out of the “allmirna.fasta” file using the following command line:

```
makeblastdb.exe -in allmirna.fasta.txt -parse_seqids -dbtype nucl -input_type fasta -out C:\blast\allmirna.out
```

Where: “- in” is the input file, “-parse\_seqids” enables parsing of the sequence id and “-dbtypenucl” is the particular type of input, which in this case is nucleotides.

The OC miRNAs were queried against the newly created database of all human mature miRNAs, using the following command line with a specified e-value of 1e-3 and word size of 7:

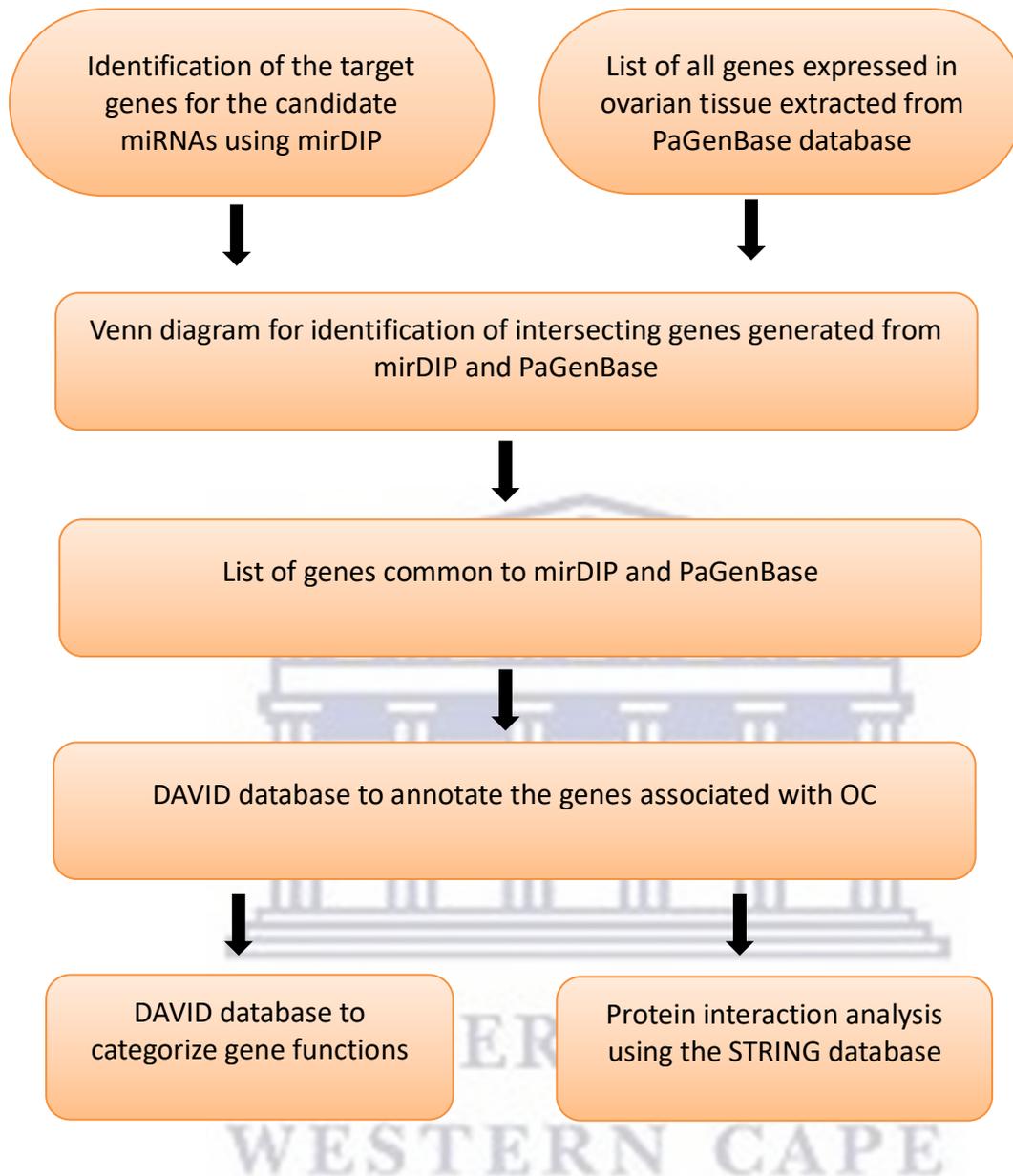
```
blastn.exe -db C:\blast\allmirna.out -evalue 1e-3 -word_size 7 -query ovmirna.fasta.txt -out C:\blast\result.out
```

Where: “-db” is the database created by previous command line, “-evalue” describes the number of hits expected to be seen by chance within the database and was set to 1e-3, “word\_size” is specified to 7, meaning that there should be at least 7 identical nucleotides between two sequences for it to be deemed as similar, “-query” is the input file to be queried against the database created, and “-out” is the output result file name and destination.

The results from and CD-HIT-EST-2D and BLAST were analysed and the miRNAs in common were kept for further analysis.

### **2.2.2 Text-mining**

To determine if any of the miRNAs common to the BLAST and CD-HIT-EST-2D results had known implications in OC, literature papers were consulted. Those miRNAs with clear implications were eliminated while the ones that showed no association with OC were kept as candidates for further analysis.



**Figure 2.2:** Flow chart depicting the outline of the *in-silico* methodology employed for prioritising the miRNA target genes.

## **2.2.5 Gene identification**

### **2.2.5.1 miRNA target genes**

To identify the gene targets of each miRNA, the miRDip database was utilized. On the miRDip homepage (<http://ophid.utoronto.ca/mirDIP/>) under the miRNA-gene matrix tab, the option to search miRNAs was selected. The novel candidate miRNA along with its corresponding validated miRNA were inputted into the search box and a score class of medium was selected. The gene targets shared between the validated and candidate miRNAs were kept for further analysis. This process was repeated for each candidate miRNA and its validated reference.

### **2.2.5.2 Ovary-specific genes**

PaGenBase was used to extract the genes with expression specific to the ovary. On the PaGenBase homepage (<http://bioinf.xmu.edu.cn/PaGenBase/index.jsp>), under the search tab; *Homo sapiens*, tissue and specific genes were selected as the conditions and “ovary” was typed in as the sample. The Specificity Measure (SPM) was left on default as 0.9. Once the query was submitted “ovary” was selected as the sample once more and the results were downloaded.

### **2.2.6 Intersecting genes**

To identify if the miRNAs target any genes expressed in the ovary, Venn diagrams of the miRNA target genes and ovary expressed genes were constructed using a bioinformatics intersection tool (<http://bioinformatics.psb.ugent.be/webtools/Venn/>). The miRNA target genes and ovary genes were uploaded separately and the result was downloaded. This was performed for each miRNA's target genes. All intersecting genes for each miRNA targets were combined with duplications being removed and the common genes kept for further analysis.

## **2.2.7 Functional annotation**

### **2.2.7.1 Disease**

The DAVID database was employed to shortlist the common genes that have a proven mechanism in OC. The genes were copied and pasted into the DAVID database (<https://david.ncifcrf.gov/>) using their official gene symbols and selecting *Homo sapiens* for both the input list and background species. The “disease tab” was expanded and the results for “Genetic Associations Database (GAD) disease” were analysed. Only the genes with implications in OC were kept for further analysis.

### **2.2.7.2 Gene function**

The shortlisted genes were once again inputted into the DAVID database for functional analysis, using the same parameters as before. The functional annotation clustering file was downloaded and all biological processes, molecular functions and cellular components that the shortlisted genes are involved in were identified and kept for further analysis. Literature papers were consulted to identify the roles that these processes play in cancer onset and progression.

### **2.2.8 Protein-protein interactions**

To identify the protein-protein interactions between the shortlisted genes, STRING was used. On the homepage (<https://string-db.org/>), the option to query multiple proteins was selected and the shortlisted genes were pasted into the space provided, with *Homo sapiens* selected as the organism. All parameters were left as default and the protein network was downloaded and analysed.

## 2.3 Results and discussion

### 2.3.1 Data mining

**Table 2.1.1:** Number of mature human miRNAs obtained from miRBase.

Database name	Number of miRNAs
miRBase	2634

As the table above indicates, there were 2634 mature human miRNAs extracted from miRBase that were deemed to have been identified with high confidence. Accuracy of the sequences needed to be ensured since the basis of the algorithms employed, rely on sequence similarity. A miRNA is said to be of “high confidence” when it has multiple sequencing reads, which serves as validation and support for its sequence being accurate (Kozomara *et al.*, 2018).

**Table 2.1.2:** Number of miRNAs implicated in OC obtained from multiple databases.

Database name	Number of miRNAs
dbDEMC 2.0	188
miR2disease	8
miRandola	4
miRCancer	12
	212

To increase the number of miRNAs with implications in OC, four different databases were used. Table 2.1.2 shows the number of miRNAs with dysregulations in OC obtained from each database. All four databases obtained their miRNAs via data mining and are updated by constantly consulting new literature (Jiang *et al.*, 2009; Russo *et al.*, 2012; Xie *et al.*, 2013; Yang *et al.*, 2016). All miRNAs were combined to a total of 212 and duplications were removed

in excel to eliminate multiple records of the same miRNA housed in more than one database. This reduced the total to 201 non-redundant miRNAs implicated in OC.

### 2.3.2 Duplication removal via CD-HIT

While the previous section 2.3.1 removed duplicates in excel based on the name of the OC-linked miRNA, this section focused to remove duplicates based on sequence as it is known that miRNAs arising from various precursors may share identical sequences (Griffiths-Jones *et al.*, 2006).

**Table 2.2:** Number of miRNAs before and after the removal of duplicates via CD-HIT.

	Before duplication removal	After duplication removal
<b>Mature miRNAs</b>	2634	2593
<b>OC-implicated miRNAs</b>	201	198

Many databases may contain multiple records of identical sequences. These duplicates impair data quality and lead to both redundancies and inconsistencies. This therefore makes the removal of duplicates a fundamental process. CD-HIT-EST uses a sequence similarity threshold to identify duplicates (Chen *et al.*, 2017), which was set to 99% to remove miRNAs that share identical sequences. As can be seen from Table 2.2 above, 41 mature miRNAs and 3 OC-linked miRNAs shared identical sequences to other miRNAs in their respective datasets and were thus termed as duplicates and clustered under their representative sequence. The 198 unique OC-implicated miRNAs are shown in Table A.1 Appendix A.

### 2.3.3 Sequence similarity analysis

**Table 2.3:** Clusters of similar miRNA sequences with the percentage of similarity calculated by CD-HIT and BLAST. MiRNAs in bold are validated with implications in OC, and the miRNAs below them share more than 90% similarity.

MiRNA family	Clusters	Percent similarity		MiRNA family	Clusters	Percent similarity		
		CD-HIT	BLAST			CD-HIT	BLAST	
<b>1</b>	<b>&gt;miR-1-1</b>			10	<b>&gt;miR-1-10</b>			
	>miR-1-1a	+/91.30%	+/+ 91%		>miR-1-10a	+/95.65%	+/+ 96%	
<b>2</b>	<b>&gt;miR-1-2</b>			11	<b>&gt;miR-1-11</b>			
	>miR-1-2a	+/90.91%	+/+ 95%		>miR-1-11a	+/90.91%	+/+ 95%	
	>miR-1-2b	+/90.91%	+/+ 90%	12	<b>&gt;miR-1-12</b>			
	>miR-1-2c	+/95.45%	+/+ 95%		>miR-1-12a	+/90.91%	+/+ 91%	
	>miR-1-2d	+/95.45%	+/+ 95%		13	<b>&gt;miR-1-13</b>		
	>miR-1-2e	+/90.91%	+/+ 91%			>miR-1-13a	+/90.91%	+/+ 95%
<b>3</b>	<b>&gt;miR-1-3</b>			14	<b>&gt;miR-1-14</b>			
	>miR-1-3a	+/95.24%	+/+ 95%		>miR-1-14a	+/95.45%	+/+ 95%	
	>miR-1-3b	+/95.00%	+/+ 95%	15	<b>&gt;miR-1-15</b>			
	>miR-1-3c	+/94.74%	+/+ 95%		>miR-1-15a	+/90.91%	+/+ 95%	
	>miR-1-3d	+/90.48%	+/+ 94%		16	<b>&gt;miR-1-16</b>		
>miR-1-3e	+/90.91%	+/+ 91%	>miR-1-16a	+/95.45%		+/+ 95%		
<b>4</b>	<b>&gt;miR-1-4</b>			17	<b>&gt;miR-1-17</b>			
	>miR-1-4a	+/90.48%	+/+ 95%		>miR-1-17a	+/95.45%	+/+ 95%	
<b>5</b>	<b>&gt;miR-1-5</b>			18	<b>&gt;miR-1-18</b>			
	>miR-1-5a	+/91.30%	+/+ 95%		>miR-1-18a	+/95.45%	+/+ 95%	
<b>6</b>	<b>&gt;miR-1-6</b>			19	<b>&gt;let-1-19</b>			
	>miR-1-6a	+/95.65%	+/+ 96%		>let-1-19a	+/90.91%	+/+ 95%	
<b>7</b>	<b>&gt;miR-1-7</b>							
	>miR-1-7a	+/90.91%	+/+ 95%					
<b>8</b>	<b>&gt;miR-1-8</b>							
	>miR-8a	+/95.65%	+/+ 96%					
<b>9</b>	<b>&gt;miR-1-9</b>							
	>miR-9a	+/95.65%	+/+ 96%					

Table 2.3 above indicates only the results that were produced by both the CD-HIT and BLAST algorithms. The analysis identified 28 miRNAs clustered under 19 OC miRNAs with similarities between 90-99%. The nomenclature employed for miRNAs identified in this chapter relates to: **miR** – miRNA; **1** – identified through sequence similarity (pipeline 1); **1a to 19a** – numerical order of identification, with the letters differentiating between miRNAs identified from the same cluster.

Incorporating various methods to conduct sequence similarity analysis is integral for exploratory research in bioinformatics (Eser *et al.*, 2014). BLAST and CD-HIT produce clusters based on different alignment criteria, therefore the miRNAs in common to both algorithms results are deemed to have a higher level of accuracy since both programs state they are more than 90% similar to their validated reference miRNA.

The “+” and “+/+” symbols in front of each percentage indicates similarity between the forward strand of the query sequence, whereas “-” and “+/-” depicts similarity involving the reverse complement of the query sequence (Wheeler and Bhagwat, 2007). For this reason, only the miRNAs with forward strand similarity were analysed and shown in Table 2.3. MiRNAs that had similarity involving their reverse complement were ignored in this study since that is not the genuine sequence of the miRNA.

Since the identified miRNAs share high similarity to sequences proven to have implications in OC, it is highly plausible that they will have the same dysregulated function in the above-mentioned disease. In 2013, Pearson stated that this type of analysis is important in identifying novel molecules. These miRNAs were text-mined to discover their potential novelty in the pathogenesis of OC.

#### **2.3.4 Text-mining**

Literature papers were consulted and it was found that 19 out of the 28 miRNAs had implications in OC, and the remaining 9 miRNAs with no implications became the candidate miRNAs for this chapter to be implicated in the disease. The text-mining results are tabulated in Table 2.4 below.

**Table 2.4:** Shows the miRNAs implicated in OC, after literature mining. The miRNAs in bold indicate OC-implicated and the normal text relates to the potential novel OC miRNAs.

MiRNA family	Clusters	Implicated?	MiRNA family	Clusters	Implicated?
1	<b>&gt;miR-1-1</b>		10	<b>&gt;miR-1-10</b>	
	>miR-1-1a	No		>miR-1-10a	Yes
2	<b>&gt;miR-1-2</b>		11	<b>&gt;miR-1-11</b>	
	>miR-1-2a	No		>miR-1-11a	Yes
	<b>&gt;miR-1-2b</b>	No	12	<b>&gt;miR-1-12</b>	
	>miR-1-2c	No		>miR-1-12a	Yes
	>miR-1-2d	Yes		13	<b>&gt;miR-1-13</b>
>miR-1-2e	Yes	>miR-1-13a	Yes		
3	<b>&gt;miR-1-3</b>		14	<b>&gt;miR-1-14</b>	
	>miR-1-3a	No		>miR-1-14a	Yes
	<b>&gt;miR-1-3b</b>	No	15	<b>&gt;miR-1-15</b>	
	>miR-1-3c	No		>miR-1-15a	Yes
	>miR-1-3d	No		16	<b>&gt;miR-1-16</b>
>miR-1-3e	Yes	>miR-1-16a	Yes		
4	<b>&gt;miR-1-4</b>		17	<b>&gt;miR-1-17</b>	
	>miR-1-4a	No		>miR-1-17a	Yes
5	<b>&gt;miR-1-5</b>		18	<b>&gt;miR-1-18</b>	
	>miR-1-5a	Yes		>miR-1-18a	Yes
6	<b>&gt;miR-1-6</b>		19	<b>&gt;let-1-19</b>	
	>miR-1-6a	Yes		>let-1-19a	Yes
7	<b>&gt;miR-1-7</b>				
	>miR-1-7a	Yes			
8	<b>&gt;miR-1-8</b>				
	>miR-1-8a	Yes			
9	<b>&gt;miR-1-9</b>				
	>miR-1-9a	Yes			

Annotating function to a sequence is commonly based on their similarity to sequences of known function (Klasberg *et al.*, 2016). As mentioned in section 2.1.8, miRNA biological functions and target pathways can be deduced by identifying the functions of their gene targets (Ling *et al.*, 2013). The 9 novel OC miRNAs were carried forward to identify and analyse their target genes.

### 2.3.5 Gene identification

The identified miRNAs are novel for OC and to determine their potential function, their target genes were functionally annotated. If the target genes are implicated in OC, by virtue of this, the regulating miRNA can also be implicated, in the absence of experimental evidence for the miRNA. Identifying target genes expressed specifically in the ovary, as their dysregulation would most likely lead to OC, was of interest as it provides additional support in the regulating miRNA's potential.

#### 2.3.5.1 miRNA target genes

**Table 2.5.1:** Number of gene targets shared between each of the potentially novel miRNAs (in normal text) and their respective similar OC-implicated miRNA (in bold).

	<b>miRNA</b>	<b>Number of target genes shared</b>
<b>Cluster 1</b>	<b>&gt;miR-1-1</b>	
	>miR-1-1a	17957
<b>Cluster 2</b>	<b>&gt;miR-1-2</b>	
	>miR-1-2a	12645
	>miR-1-2b	14419
	>miR-1-2c	14285
<b>Cluster 3</b>	<b>&gt;miR-1-3</b>	
	>miR-1-3a	9822
	>miR-1-3b	10654
	>miR-1-3c	10337

	>miR-1-3d	10121
<b>Cluster 4</b>	>miR-1-4	
	>miR-1-4a	17892

The Table above indicates the relative target genes shared between the novel miRNAs and their respective OC-implicated miRNA reference. Gene targets in common to both the novel candidate miRNA and their reference miRNA were specified based on the notion that if two miRNAs share the same set of target genes, they will most likely be involved in the same pathways (Bhajun *et al.*, 2015).

From the results obtained, a high number of gene targets were identified for each miRNA cluster, with cluster 1 having the highest. The high number of gene targets is accounted for by the fact that one miRNA can target thousands of genes and one gene can be targeted by a large number of different miRNAs. This also arises in overlapping of gene targets between miRNAs, with different miRNAs exerting diverse functions on the same target gene (Peter, 2010). MiRNAs may act as either oncogenes or tumour suppressors based on their effects on target genes (Makondi *et al.*, 2019).

MiRDip has four specific confidence classes; *very high*, *high*, *medium* and *low* confidence, corresponding to the results from the top 1%, top 5%, top 1/3 and the remaining predictions, respectively (Tokar *et al.*, 2017). Medium score was selected in this section to increase the number of genes used for identifying overlapping genes in the next section.

### 2.3.5.2 Ovary-specific genes

**Table 2.5.2:** Number of genes expressed specifically in the ovary.

Database used	Number of ovary-specific genes
PaGenBase	245

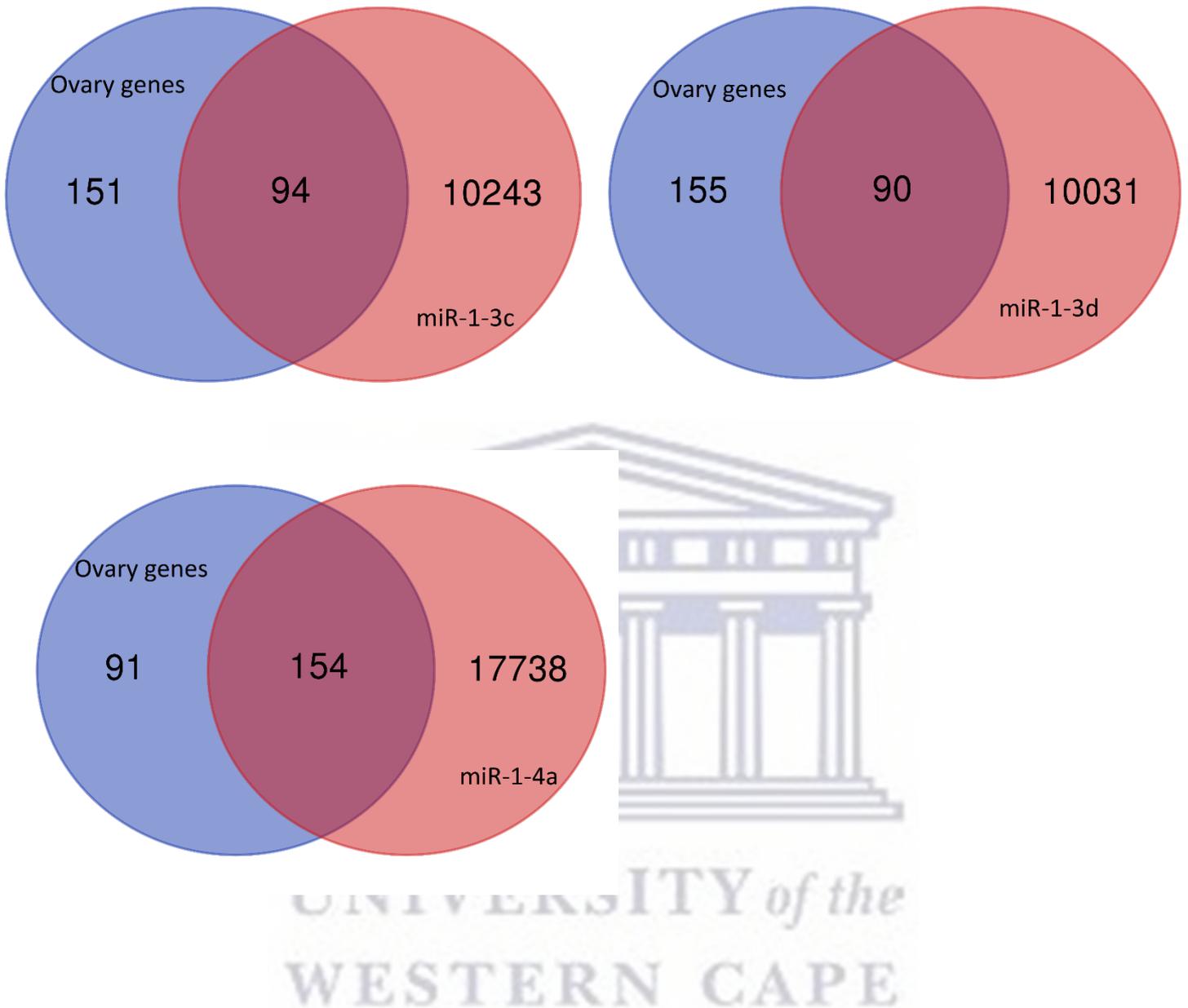
One of the many uses of the PaGenBase database is to identify tissue-specific genes that can serve as potential biomarkers for specific diseases/conditions (Pan *et al.*, 2013), hence the

selection of ovary-specific genes. According to the database, there are 245 genes specifically expressed in the ovary. The SPM (specificity measure) was set to 0.9 and more and is a parameter that measures the specificity of a gene's expression in a certain sample, with a higher SPM indicating more sample specificity (Pan *et al.*, 2013).



### 2.3.6. Intersecting genes





**Figure 2.3:** Venn diagrams depicting the intersection of the ovary-specific genes with each miRNAs' target genes.

Figure 2.3 displays Venn diagrams depicting the intersection of genes for each of the 9 miRNAs identified target genes, with the ovary-specific genes. From the results obtained, it is evident that there are candidate miRNA target genes expressed in the ovary. Various studies have demonstrated that many genes with either specific or preferential expression in the ovary govern crucial molecular elements of ovarian function (Hennebold *et al.*, 2000). This motivates that these miRNAs could be putative biomarkers for OC, as their dysregulation

could result in the alteration of the target gene's expression (Paranjape *et al.*, 2009) which could potentially lead to OC.

The total combined number of intersecting genes was 1035 and once duplications were removed, this total was reduced to 158. These 158 genes, shown in Table A.2 Appendix A, were taken to DAVID to identify the ones with OC associations, i.e. onset and/or progression.

### 2.3.7 DAVID Functional annotation

#### 2.3.7.1 Disease

**Table 2.6:** Genes with implications in OC.

Gene name	Disease
FOS	OC, alcohol consumption, Alzheimer's Disease ...
WISP1	OC, asthma, Bone Mineral Density ...
WNT5A	OC, amyotrophic Lateral Sclerosis   Anoxia ...
AMHR2	OC, estradiol, female infertility...
BNC2	OC, diabetes, type 1, Tobacco Use Disorder...
CYP19A1	OC, abortion, habitual   Infertility...
EGR2	OC, alzheimer's disease, Bone Mineral Density...
ESR1	OC, abdominal aortic aneurysm, abortion...
ESR2	OC abdominal aortic aneurysm, Abortion...
LHCGR	OC, abortion, habitual   Infertility...
PDGFB	OC, amyotrophic Lateral Sclerosis   Anoxia...
PGR	OC, abdominal aortic aneurysm, Abortion...

Out of the 158 genes targets expressed in the ovary, 12 genes were returned from DAVID as having implication in OC as well as showing high expression in the ovary based on their SPM value. The fact that the candidate miRNAs target these genes strongly suggests these miRNAs may have mechanisms in OC as well.

The table below indicates the targeting interactions between the regulating novel OC miRNA and its target genes.



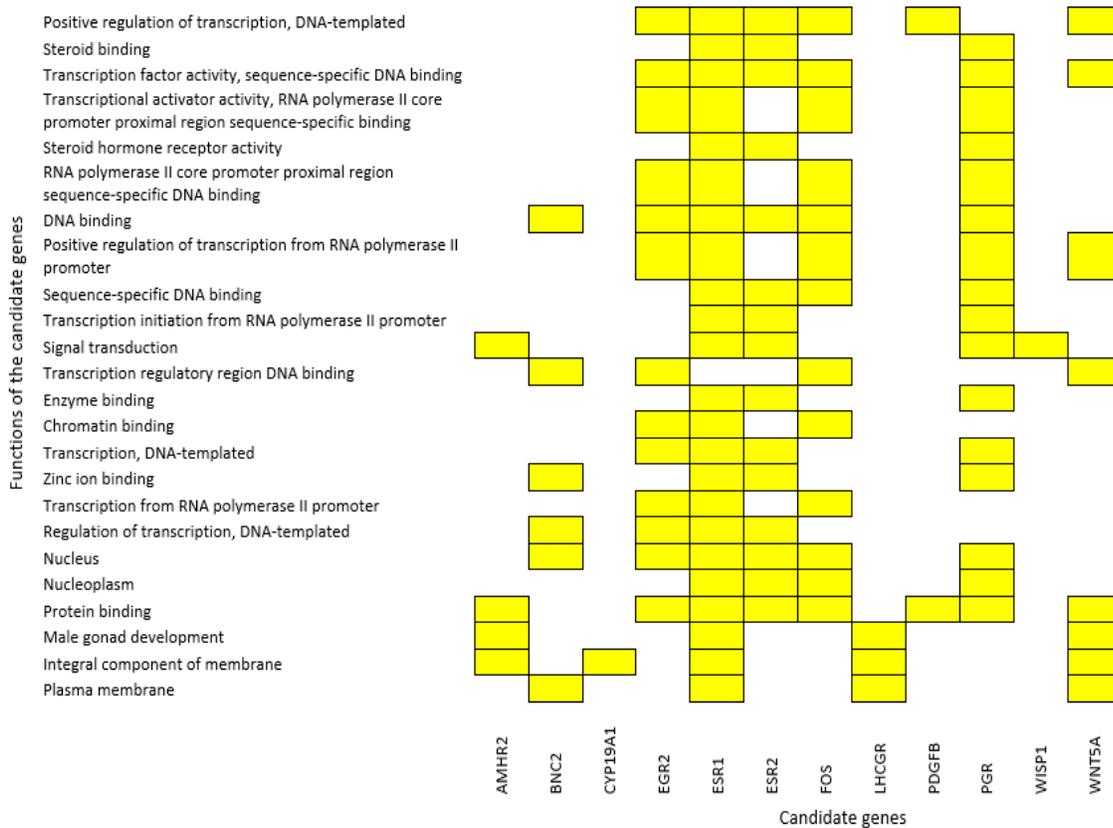
**Table 2.7:** Targeting interactions between the novel OC miRNAs and target genes.

	Target genes											
	FOS	WISP1	WNT5A	AMHR2	BNC2	CYP19A1	EGR2	ESR1	ESR2	LHCGR	PDGFB	PGR
miR-1-1a	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
miR-1-2a	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
miR-1-2b	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
miR-1-2c	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
miR-1-3a		✓	✓	✓	✓			✓	✓			✓
miR-1-3b		✓	✓	✓	✓			✓	✓	✓	✓	✓
miR-1-3c		✓	✓	✓	✓			✓	✓			✓
miR-1-3d		✓	✓		✓			✓	✓			✓
miR-1-4a	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Novel OC miRNAs

In order to understand the roles of these gene within a cell, their functions need to be uncovered.

### 2.3.7.2 Gene function



**Figure 2.4:** Biological processes, molecular functions and cellular components that the candidate genes are involved in. Yellow blocks indicate the functions of the respective genes.

From the results shown in Figure 2.4, most of the genes are involved in protein binding. Proteins may bind either DNA or RNA, and are involved in a variety of crucial processes including DNA packing, transcription, replication, modification and repair (Hudson and Ortlund, 2014; Peled *et al.*, 2016). DNA binding proteins (DBPs) holds an essential role as transcription factors governing gene expression, and alterations within these DBPs contributes greatly to tumorigenesis (Liu *et al.*, 2001). RNA binding proteins (RBPs) plays a fundamental part in gene regulation, and act as important coordinators in maintaining genome integrity. They are key players in many post-

transcriptional events including mRNA splicing, stability, localization and polyadenylation, which ultimately affects gene expression levels within each cell (Wang *et al.*, 2019). Their dysregulation has become clear in various cancer types thus influencing the function and expression of oncoproteins and tumour suppressor proteins (Pereira *et al.*, 2017).

Early growth response 2 (EGR2) and Progesterone receptor (PGR), genes regulated by the identified miRNAs, are transcription factors with proven mechanisms in OC. EGR2 is upregulated in response to hormones, growth factors, cytokines and environmental stimulants and is highly associated with OC survival and recurrence. It has also been implicated in the Phosphatase and Tensin homologue (PTEN) - induced apoptotic pathway (Delfino and Rodriguez-Zas, 2013; Jin *et al.*, 2017). Progesterone suppresses ovulation, which not only decreases the proliferative effect of oestrogen, but also inhibits inflammation and cancer infiltration. It has also been proven to initiate apoptosis in tumour cells. PGR expression is reported to be significantly lower in tumour ovarian tissues when compared to healthy tissues and serves as a prognostic biomarker in OC (Mungenast and Thalhammer, 2014).

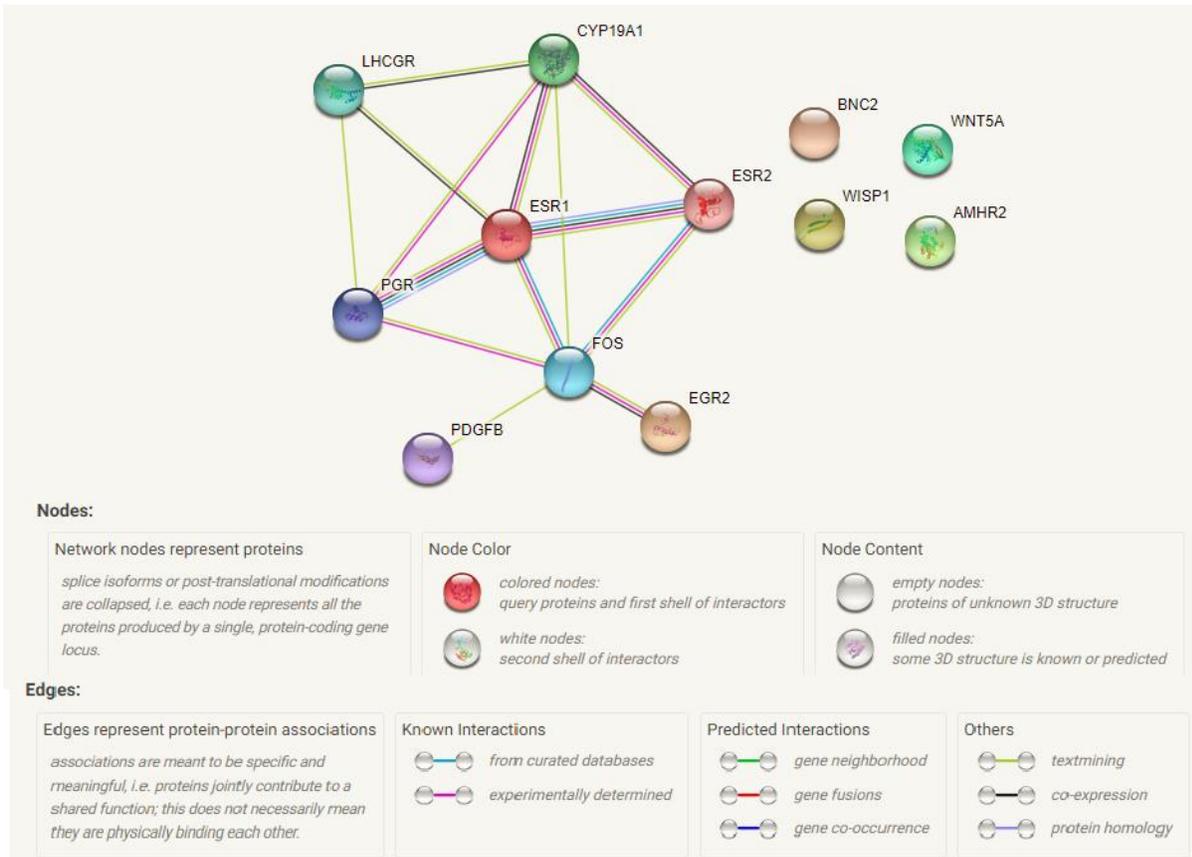
Luteinizing hormone/choriogonadotropin receptor (LHCGR) regulates ovulation by binding luteinizing hormone (LH) and hCG. Its expression is reported to be downregulated in a cancerous ovary and also correlates to a poor overall survival (Zhong *et al.*, 2019).

Some of the candidate genes are localized to the cell membrane. Membrane proteins are an essential component of biological membranes mediating vital cellular functions (Kampen, 2011). Cell membrane proteins are altered throughout normal cellular processes as well as during tumorigenesis. Membrane proteins present in normal cells may partially or completely disappear in cancer cells, while over expressed or newly synthesized proteins may be present (Grimm *et al.*, 2011). Cytochrome P450 Family 19 Subfamily A Member 1 (CYP19A1) is localized to the membrane and involved in the synthesis of oestrogen (Khayeka-Wandabwa *et al.*, 2019). In 2008, a study by Goodman *et al.*, revealed that variants of CYP19A1 influences susceptibility to OC.

It is clear that the candidate genes are not only implicated in OC, as seen in the literature, but also involved in processes leading to cancer onset and metastasis, as seen from the DAVID

analysis. Since the candidate miRNAs target these genes and therefore control these processes, it strengthens their potential as being promising biomarkers for early stage OC diagnosis.

### 2.3.8 Protein-protein interactions



**Figure 2.5:** Network from STRING depicting the interactions of the candidate proteins with each other.

Out of the twelve candidate proteins, eight have evidence-based interactions with each other. Oestrogen Receptor 1 (ESR1), the central gene in the network, is known to be highly expressed in OC (Giannopoulou *et al.*, 2018). Oestrogen provides signalling systems for cell division and differentiation by binding to its receptor, leading to transcriptional activation of oestrogen-responsive genes. Included in these genes are proto-oncogenes such as c-fos (Mungenast and Thalhammer, 2014).

FOS protein forms the transcription factor activating protein (AP-1) along with Jun protein. Each protein is differentially expressed and regulated; therefore, each cell type has a complex mixture of AP-1 dimers with slightly dissimilar functions (Hess, 2004; Hein *et al.*, 2009). Numerous studies have implicated AP-1 transcription factors in major cancer-related pathways including inflammation, differentiation, cell migration, metastasis and angiogenesis. Depending on the cellular context, both the upregulated and downregulation of AP-1 promotes tumorigenesis (Garces de los Fayos Alonso *et al.*, 2018).

The proteins not involved in the network were still considered to be candidates, due to several other lines of evidence. Lack of interactions between these proteins could possibly be attributed to the need for studies yet to characterize such interactions. It is also likely that the proteins may interact through another intermediary protein not part of the input list and network. One of the proteins not connected to the network is Wnt Family Member 5 (WNT5a), which is known to have high expression in OC when compared to healthy, borderline and benign controls (Ford *et al.*, 2014). The Wnt signalling pathway is involved in both the development and homeostasis of tissues by regulating their endogenous stem cells. Abnormal Wnt signalling is a key contributor to cancer onset and progression through affecting the behaviour of cancer stem cells (CSCs). CSCs are responsible for tumour establishment as well as disease relapse due to their drug-resistant properties (Duchartre *et al.*, 2016). Various lines of evidence have indicated interactions between the Wnt signalling pathway and miRNAs in cancer development. MiRNAs were found to activate or inhibit the Wnt pathway at several steps. Conversely, Wnt activation increases miRNA expression by binding to its promoter and initiating transcription (Peng *et al.*, 2016).

MiRNAs regulating the same gene are expected to consequently regulate the same target pathway (Kehl *et al.*, 2017). This further strengthens the potential of these miRNAs as all of their candidate target genes are implicated in OC, with eight out of twelve having interactions with each other.

## 2.4 Conclusion

It has been proven that miRNAs serve crucial functions such as RNA silencing, post-transcriptional regulation and gene regulation (Cai *et al.*, 2018). They act as suitable biomarkers for both diagnosis and prognosis due to their differential expression across various cancer stages and types (Zhang *et al.*, 2017; Du *et al.*, 2018).

This chapter identified nine novel OC miRNAs based on a sequence similarity analysis as well as twelve target genes that are expressed within the ovary. The candidate miRNAs were produced by both BLAST and CD-HIT as having 90-99% similarity to their validated references, and were further deemed as novel candidates based on their lack of involvement in OC.

The candidate target genes of these miRNAs were reduced and characterized based on them having, (i) selective expression in the ovary, and (ii) clear implications in OC. The candidate genes are involved in various molecular functions, biological processes and cellular components including protein, chromatin, enzyme and zinc binding, plasma and integral membrane components, regulating transcription and signal transduction. This indirectly indicates the various pathways that the novel miRNAs target and control, all of which could contribute to cancer onset and metastasis (Liu *et al.*, 2014). STRING analysis indicated that more than 50% of the candidate genes interact with each other, and since each candidate miRNA targets the same gene as another, they most likely target the same dysregulated OC pathway (Kehl *et al.*, 2017).

MiRNAs comprise a complex regulatory network due to the fact that multiple miRNAs can target the same gene, and therefore determine the expression levels of the gene (Peter, 2010). Based on the shared targeting interactions involving multiple candidate miRNAs and a single gene, it is highly plausible that the combination of miRNAs is what drives the target genes into carcinogenesis.

## Chapter 3

### Identification of miRNAs and genes as biomarkers for the early stage diagnosis of OC using patient clinical data from TCGA

#### 3.1 Introduction

Cancer comprises of dynamic genomic aberrations in the form of somatic mutations, copy number variations, epigenetic alterations, and altered gene expression levels. Due to tumour heterogeneity, each cancer type has its own different molecular profile, thus requiring unique diagnostic and therapeutic strategies (Tomczak *et al.*, 2015; Wang *et al.*, 2016).

Clinical data acts as a central resource in healthcare progression, by allowing for the development of novel knowledge based on individual clinical experiences (Grossmann, 2010). Many investigators have incorporated The Cancer Genome Atlas (TCGA) as a resource to not only support their studies, but also to aid in interpreting molecular testing of individual patients in a clinical setting (Liu *et al.*, 2018).

Launched by the National Institute of Health (NIH), TCGA is a comprehensive atlas of genomic cancer profiles that helps in producing new cancer therapies, diagnostic techniques, and preventative strategies. TCGA involves various centres responsible for the collection and processing of samples, followed by high-throughput and accurate bioinformatics analysis. It utilizes various platforms to generate many data types, including gene, exon, and miRNA expression, single nucleotide polymorphisms, and loss of heterozygosity for more than 30 cancer types and 10 cancer tissues (Chu *et al.*, 2015; Tomczak *et al.*, 2015). Since TCGA has a large amount of clinical data stored, statistical measures are important in filtering data and identifying those that are statistically significant. From there, biologically relevant data can be uncovered (Antunović *et al.*, 2011).

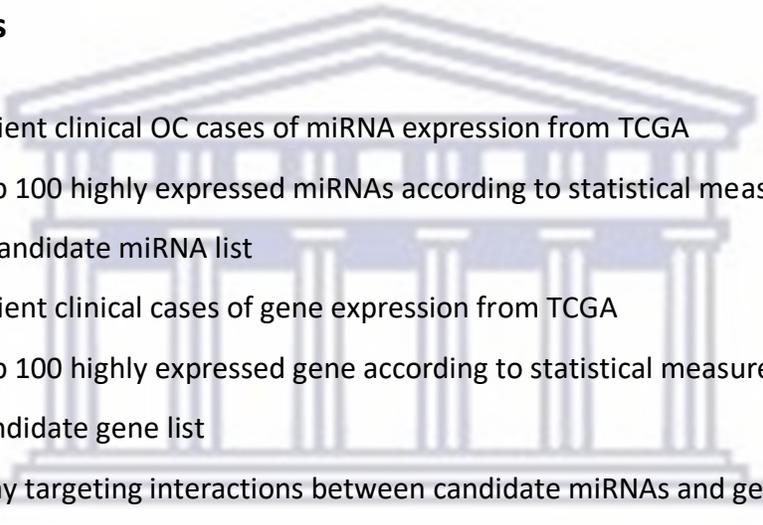
As mentioned previously, using miRNAs and genes as biomarkers for OC diagnostics offers great promise (Wang *et al.*, 2015). The immense knowledge and potential that clinical data provides in identifying such biomarkers, was the reason behind clinical data being the basis of this chapter.

### 3.1.2 Aim

The aim of this chapter is to identify miRNAs and genes, extracted from TCGA clinical data, with the potential to serve as biomarkers for early stage OC diagnosis.

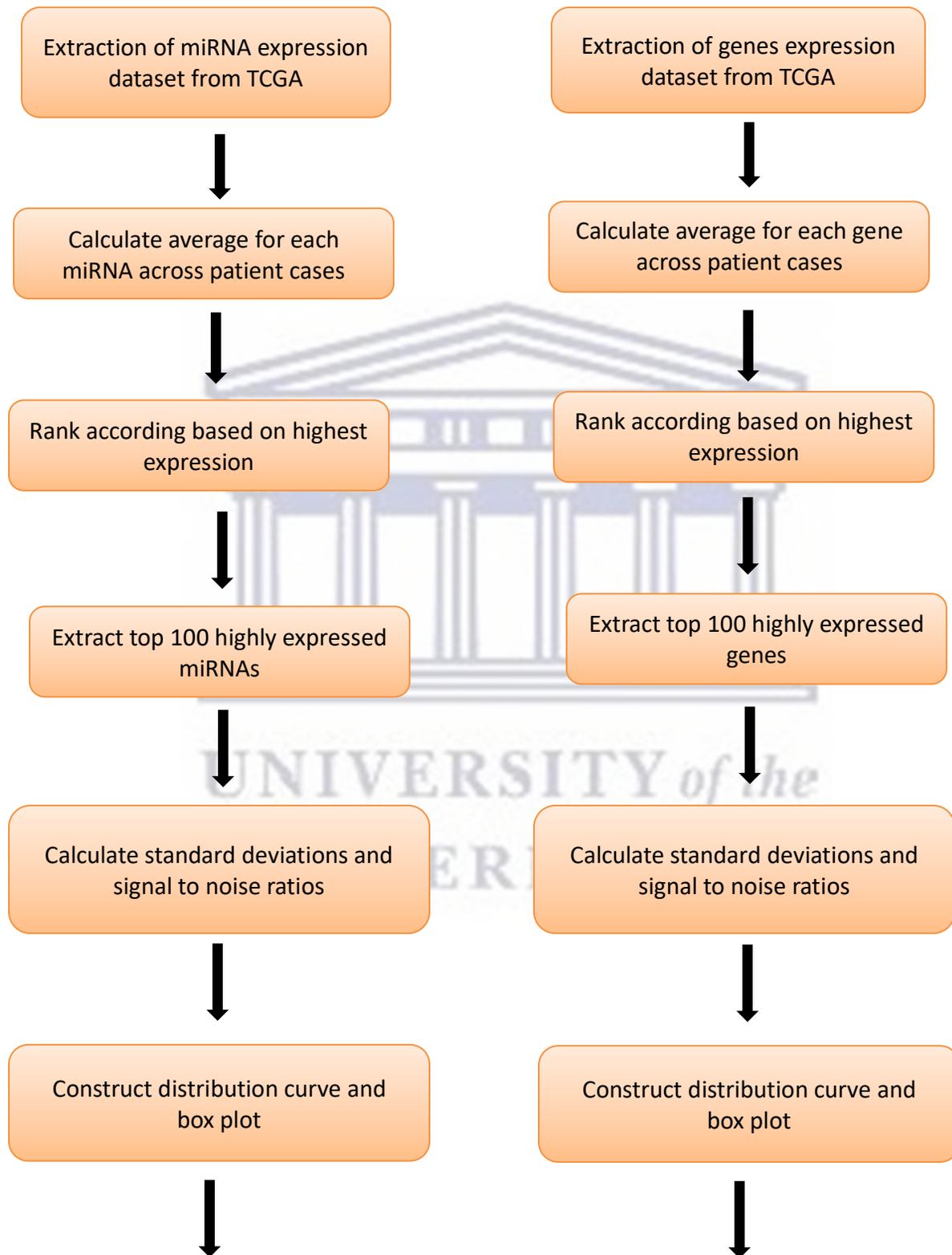
### 3.1.3 Objectives

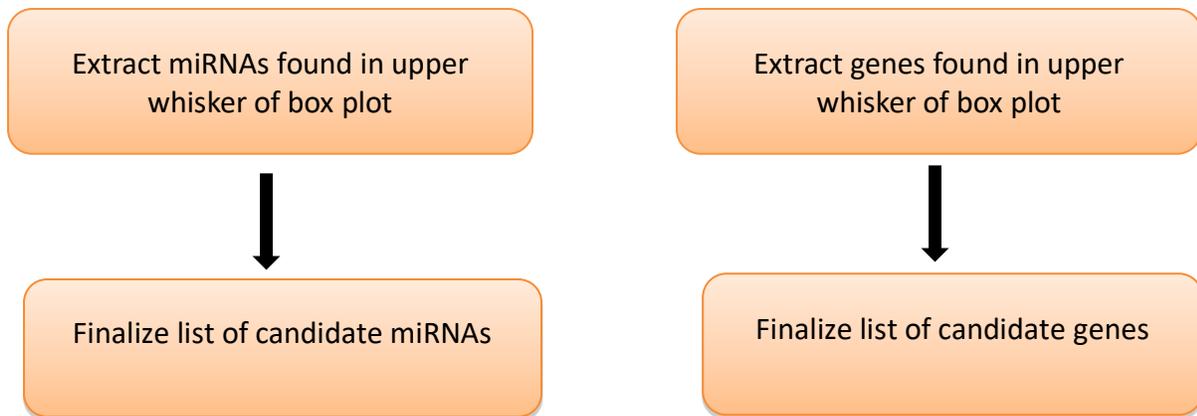
- Extract patient clinical OC cases of miRNA expression from TCGA
- Identify top 100 highly expressed miRNAs according to statistical measures
- Finalize a candidate miRNA list
- Extract patient clinical cases of gene expression from TCGA
- Identify top 100 highly expressed gene according to statistical measures
- Finalize candidate gene list
- Observe any targeting interactions between candidate miRNAs and genes



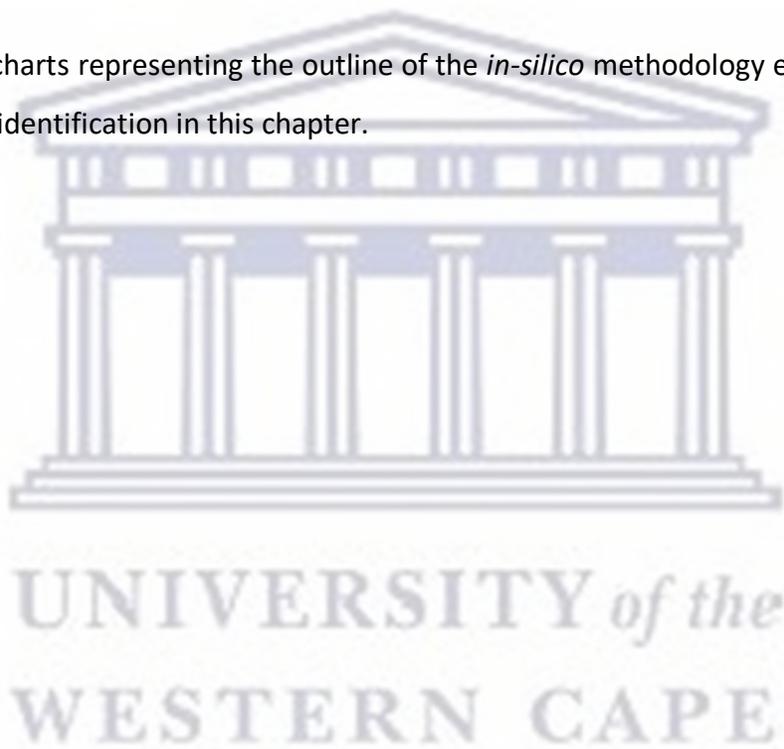
UNIVERSITY of the  
WESTERN CAPE

### 3.2 Methodology





**Figure 3.1:** Flow charts representing the outline of the *in-silico* methodology employed for both miRNA and gene identification in this chapter.



### **3.2.1. MiRNA data extraction**

To extract miRNAs expressed in OC cases, the TCGA data portal (available at <https://cancergenome.nih.gov/>) was launched and the repository tab was selected. “Ovary” was selected as the primary site, with “TCGA-OV” being the project of focus. “Transcriptome profiling” was selected under data category and “miRNA quantification” was specified as the data type. The files were extracted and analysed in Microsoft Excel.

#### **3.2.1.1. Statistical parameters to isolate candidates**

To identify and isolate the candidate miRNAs, Reads Per Million Mapped was used as the expression level of each miRNA and averages across all patient samples were calculated. The miRNA averages were ranked from highest to lowest expression. The top 100 highly expressed miRNAs were selected and their standard deviations and signal to noise (S/N) ratios were calculated. To ensure that the signal to noise values are in fact stable and accurate, a distribution curve was constructed. Using the signal to noise values, a box plot was created and miRNAs falling in the upper whisker were selected as the candidate miRNAs.

### **3.2.2. Gene data extraction**

To extract genes expressed in OC, the TCGA data portal (available at <https://cancergenome.nih.gov/>) was launched and the repository tab was selected. “Ovary” was selected as the primary site, with “TCGA-OV” being the project of focus. “Transcriptome profiling” was selected under data category and “Gene expression quantification” was specified as the data type. The files were extracted and analysed in Microsoft Excel.

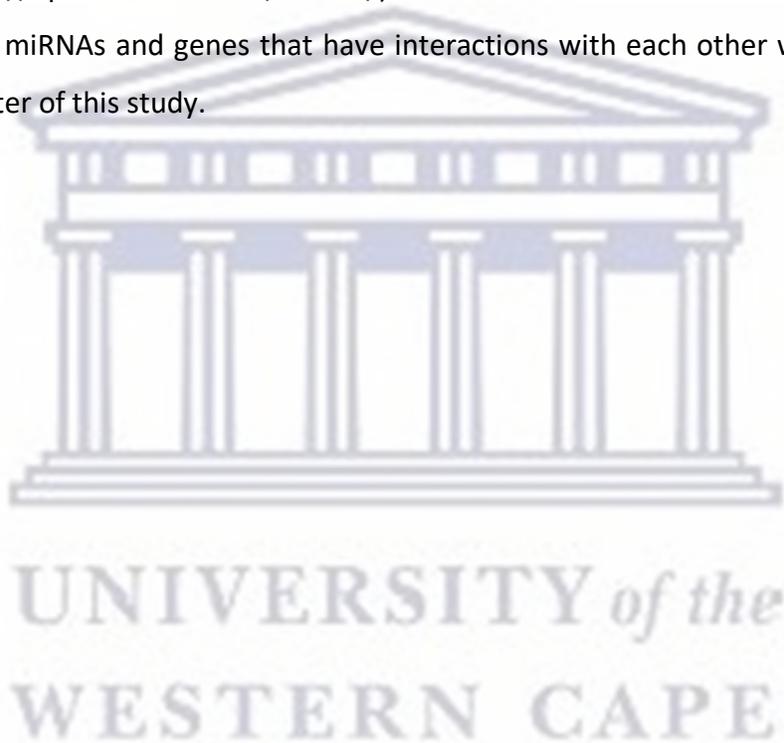
#### **3.2.2.1. Statistical parameters to isolate candidates**

To identify and isolate the candidate genes, Reads Per Million Mapped was used as the expression level of each gene and averages were calculated across all patient cases. The gene averages were ranked from highest to lowest expression. The top 100 highly expressed genes

were selected and their standard deviations and signal to noise ratios were calculated. To ensure that the signal to noise values are in fact stable and accurate, a distribution curve was constructed. Using the signal to noise values, a box plot was created and genes falling in the upper whisker were selected as the candidate genes.

### **3.2.3 MiRNA-gene targeting interactions**

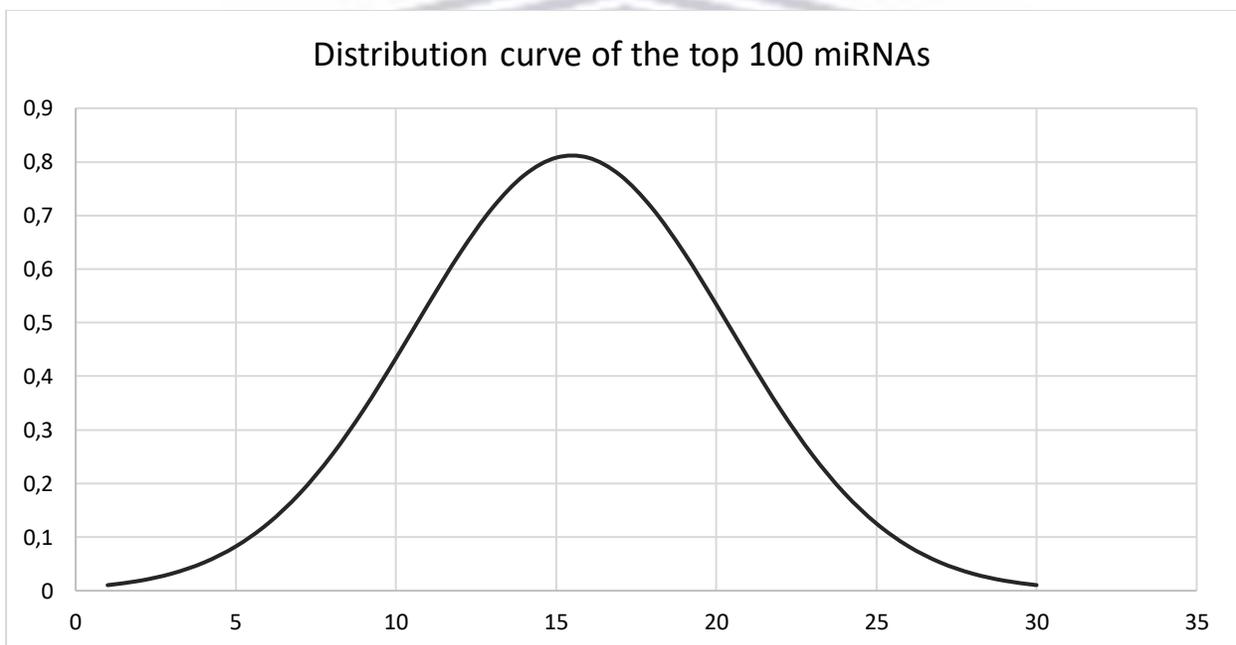
To determine if any of the candidate miRNAs target the candidate genes identified, miRDip (available at <http://ophid.utoronto.ca/mirDIP/>) was launched and the bidirectional search tool was utilized. The miRNAs and genes that have interactions with each other were used for the subsequent chapter of this study.



### 3.3 Results and discussion

#### 3.3.1 MiRNA data extraction and candidate identification

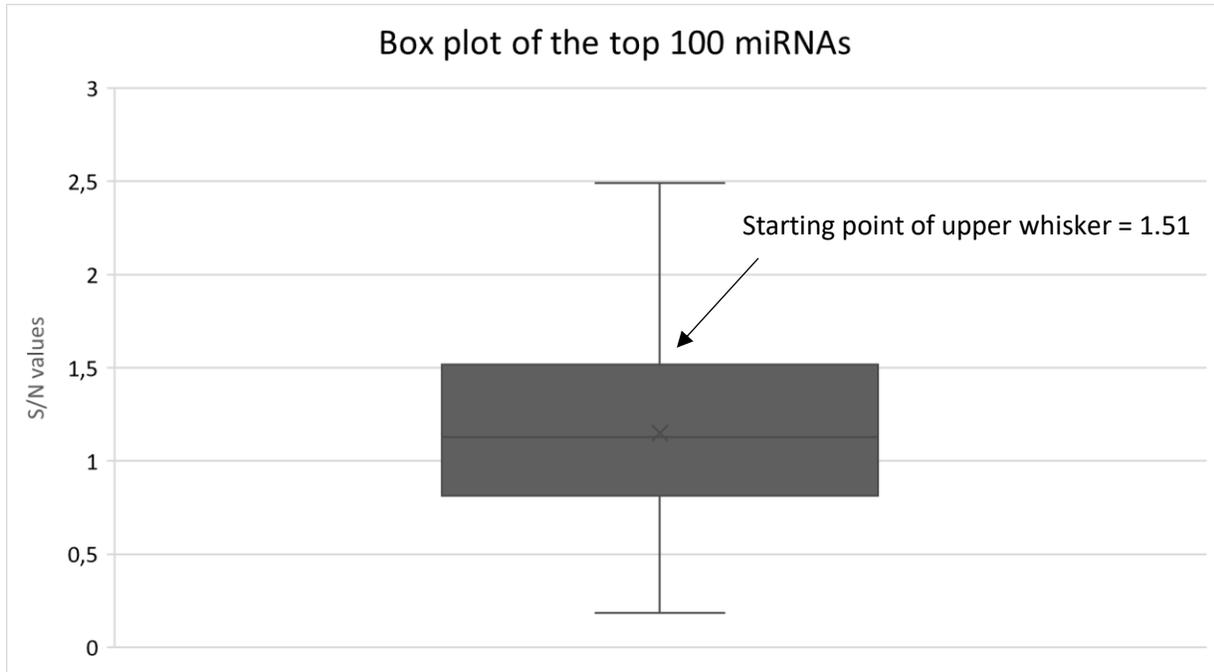
A total of 499 patient cases with expression level records for 1881 miRNAs were extracted from TCGA and analysed. The top 100 highly expressed miRNAs had average expression levels ranging from 274270,4552 to 277,3424168. This is notably a huge gap and as can be seen in Figure 3.2, the distribution curve of the top 100 miRNA's S/N ratios prove that the data is stable and accurate, based on the bell-shaped curve produced.



**Figure 3.2:** Distribution curve of the top 100 highly expressed miRNAs, based on their S/N values.

Many statistical tools assume that random variation in data follows normal distribution. “Normal” is the term given to continuous data distributed on either side of the mean as given by the standard deviation, that follows a bell-shape (Krithikadatta, 2014; Maltenfort, 2015). S/N values for each miRNA was calculated according to the formula “ $S/N = \bar{X}/s$ ”, where “ $\bar{X}$ ” is the mean and “ $s$ ” is the standard deviation (Busch and Busch, 2018).

Because a higher S/N value correlates to a higher level of accuracy (Mazziotta, 2002), Figure 3.2.1 depicts the box plot constructed to identify and isolate miRNAs with S/N values residing in the upper whisker.



**Figure 3.2.1:** Box plot of the top 100 miRNAs, based on their S/N values.

A total of 26 miRNAs had S/N values in the upper whisker of the box plot, with the range identified to be 1.51 to 2.5. These miRNAs are the candidates deemed to have accurate and reproducible expression levels in OC, which can be found in Table 3.1 along with their average expression and S/N values.

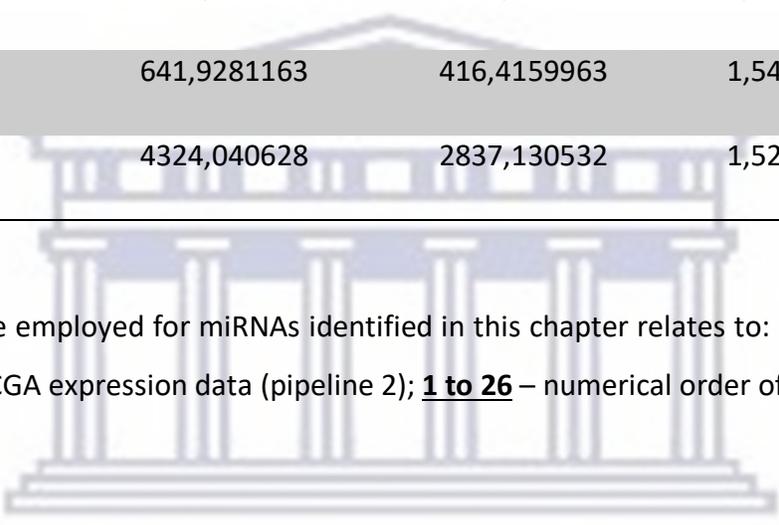
Box plots provide a graphical summary of data (Schlattmann and Dirnagl, 2010), based on the minimum value, the median of the first half of the data (Q1), the median (Q2), the median of the second half of the data (Q3) and the maximum value (Marmolejo-Ramos and Siva Tian, 2010). The upper whisker was chosen to identify candidates with highest levels of expression in OC as it contains the top 25% of data.

**Table 3.1:** Candidate miRNAs identified in this study, with their relative expression data.

<b>miRNA</b>	<b>Average expression</b>	<b>Standard deviation</b>	<b>S/N</b>
<b>miR-2-1</b>	4458,889373	2938,323243	1,517495
<b>miR-2-2</b>	4634,98713	1860,490687	2,491271
<b>miR-2-3</b>	1809,124095	1190,141792	1,520091
<b>miR-2-4</b>	1892,036496	1110,02132	1,704505
<b>miR-2-5</b>	2341,187127	1385,781261	1,689435
<b>miR-2-6</b>	94559,1116	52486,06122	1,801604
<b>miR-2-7</b>	1611,565932	1060,816006	1,519176
<b>miR-2-8</b>	18095,54756	10045,29476	1,801395
<b>miR-2-9</b>	1060,270432	691,26301	1,533816
<b>miR-2-10</b>	5543,778875	2559,821627	2,16569
<b>miR-2-11</b>	15755,83122	10153,83577	1,551712
<b>miR-2-12</b>	965,5993532	482,0611749	2,003064
<b>miR-2-13</b>	795,9024507	392,31845	2,028715
<b>miR-2-14</b>	6741,021503	4190,369678	1,608694
<b>miR-2-15</b>	274270,4552	130406,8453	2,103191
<b>miR-2-16</b>	54006,50543	28130,87368	1,91983
<b>miR-2-17</b>	59196,81886	30860,19656	1,918226
<b>miR-2-18</b>	26174,26004	13998,73795	1,869759

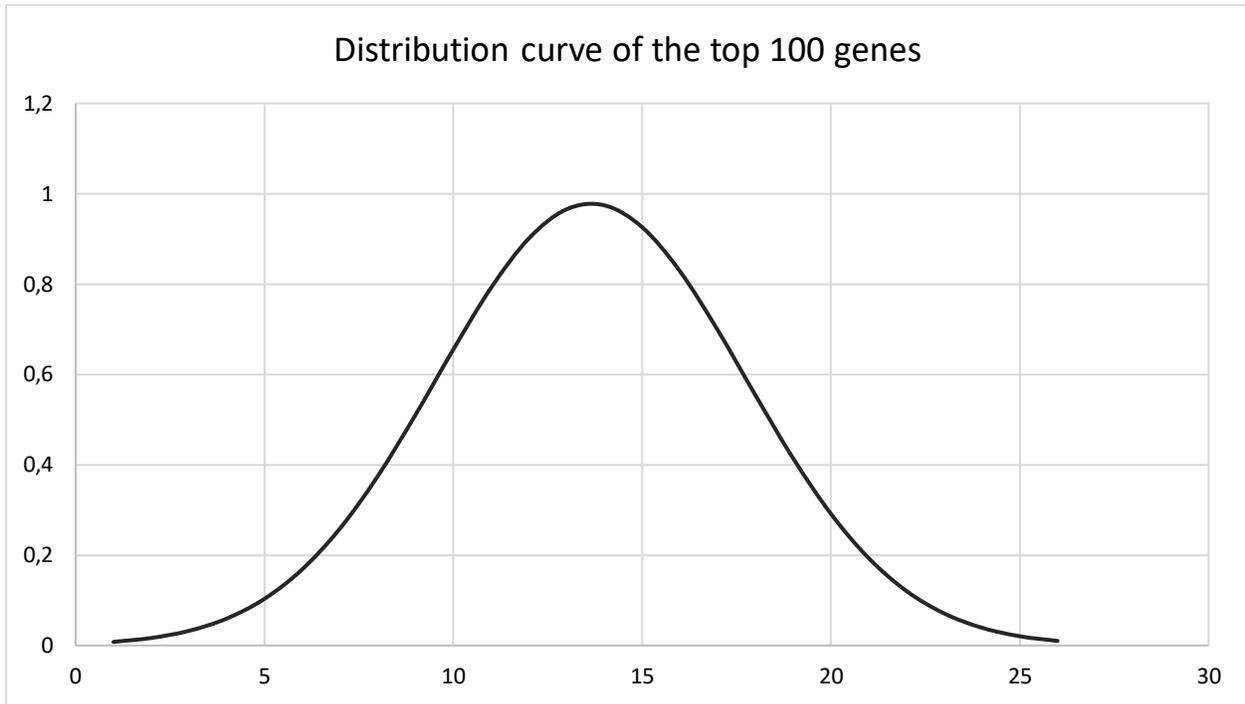
<b>miR-2-19</b>	26033,19381	13937,59924	1,867839
<b>miR-2-20</b>	26043,70234	13959,39537	1,865676
<b>miR-2-21</b>	771,4991879	487,3704161	1,582983
<b>miR-2-22</b>	9433,434034	5966,506702	1,581065
<b>miR-2-23</b>	9419,006612	5964,843608	1,579087
<b>miR-2-24</b>	2519,998189	1621,584784	1,554034
<b>miR-2-25</b>	641,9281163	416,4159963	1,541555
<b>miR-2-26</b>	4324,040628	2837,130532	1,524089

The nomenclature employed for miRNAs identified in this chapter relates to: **miR** – miRNA; **2** – identified from TCGA expression data (pipeline 2); **1 to 26** – numerical order of identification.



UNIVERSITY *of the*  
WESTERN CAPE

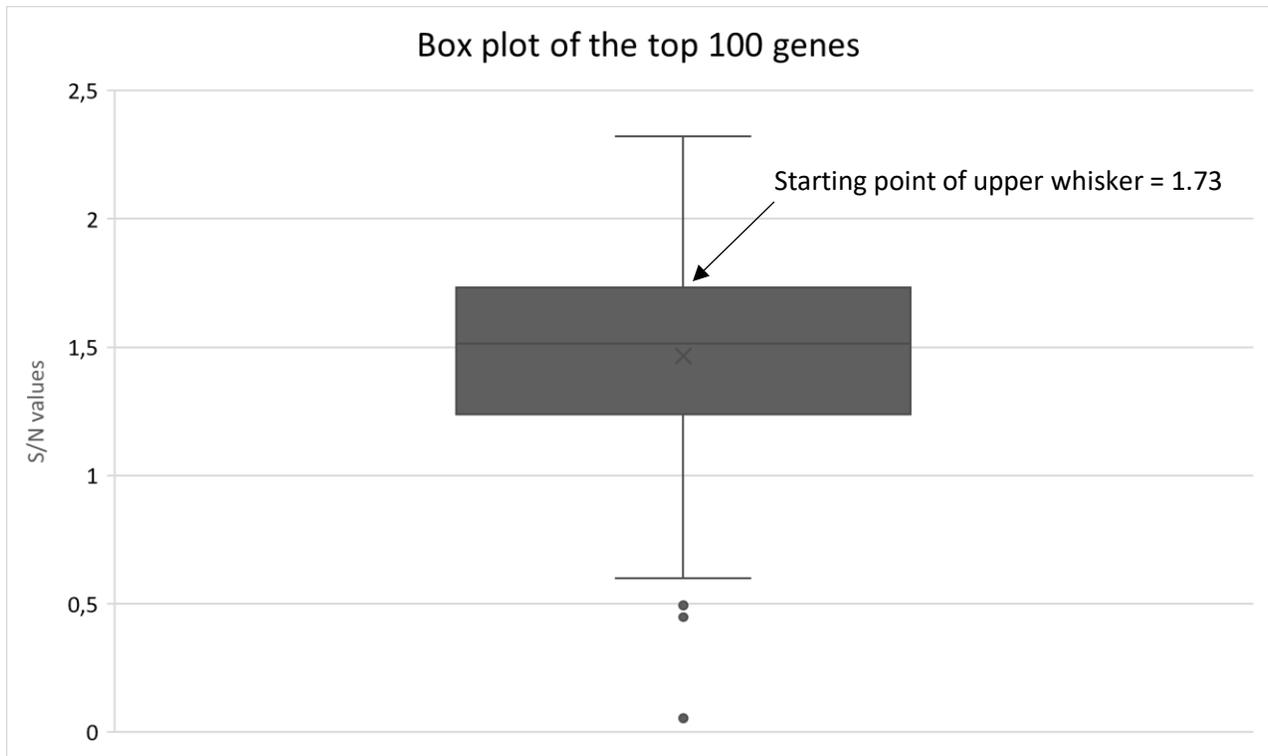
### 3.3.2 Gene data extraction and candidate identification



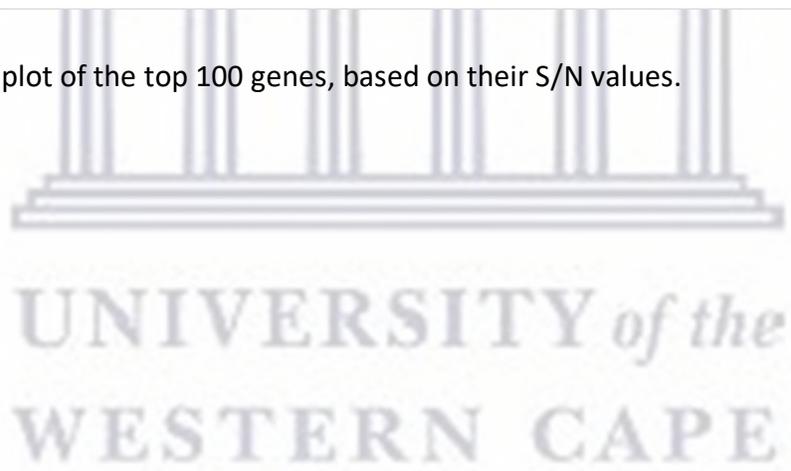
**Figure 3.3:** Distribution curve of the top 100 highly expressed genes, based on their S/N values.

The distribution curve in Figure 3.3 proves that the S/N values for the top 100 genes are stable and accurate, therefore allowing for them to be taken further in this study. In Figure 3.3.1 below, the box plot has an upper whisker starting point of 1.73. This indicates that the top 25% of the genes with high expression have a S/N value of  $\geq 1.73$ .

A total of 25 genes with S/N values of 1.73 and higher were identified and finalized as the candidate genes of this chapter. These genes are tabulated in Table 3.2 along with their average expression, standard deviation and S/N values.



**Figure 3.3.1:** Box plot of the top 100 genes, based on their S/N values.



**Table 3.2:** Candidate genes identified in this chapter, as well as their expression information.

<b>Gene</b>	<b>Average expression</b>	<b>Standard deviation</b>	<b>S/N</b>
<b>PTMA</b>	96127,20899	41419,54247	2,320818
<b>ACTB</b>	344376,4339	154014,9449	2,235994
<b>CALR</b>	74623,56614	34730,766	2,14863
<b>CFL1</b>	75154,06085	35857,30681	2,09592
<b>MT-CO1</b>	941874,9048	454346,8453	2,073031
<b>HSP90AB1</b>	97081,5	48063,50267	2,019859
<b>MT-ND4</b>	732800,5397	369430,8529	1,983593
<b>TMBIM6</b>	70141,12169	35738,85403	1,962601
<b>MT-ATP6</b>	236967,9841	121668,1398	1,947658
<b>ACTG1</b>	233980,4709	121570,2148	1,924653
<b>MT-ND2</b>	342401,9577	179211,1074	1,910607
<b>FTL</b>	156460,1455	82380,60824	1,899235
<b>GNAS</b>	113802,4021	60190,01033	1,890719
<b>TUBB</b>	76076,05556	40456,71676	1,880431
<b>MT-CO3</b>	440762,4339	240515,669	1,832573
<b>RPL15</b>	111229,2884	61468,23281	1,809541
<b>MT-RNR2</b>	832386,6693	460253,4606	1,80854
<b>MT-ND5</b>	169919,2884	93959,32637	1,808435

<b>RPL3</b>	114091,5767	63119,55505	1,807547
<b>MT-CO2</b>	405573,963	226899,4167	1,787461
<b>PSAP</b>	79962,5291	45078,27248	1,77386
<b>ALDOA</b>	109488,7989	62281,82887	1,757957
<b>RPL7A</b>	72758,38095	41587,67802	1,749518
<b>MT-ND1</b>	379320,0979	217706,7267	1,742344
<b>RPL4</b>	120127,7354	69255,61162	1,734556



### 3.3.3 MiRNA-gene targeting interactions

**Table 3.3:** Candidate miRNAs that target the candidate genes identified.

Candidate miRNAs	Candidate target genes															
	ACTB	ACTG1	ALDOA	CALR	CFL1	FTL	GNAS	HSP90AB1	PSAP	PTMA	RPL15	RPL3	RPL4	RPL7A	TMBIM6	TUBB
miR-2-1	✓	✓		✓	✓		✓		✓		✓				✓	✓
miR-2-2		✓	✓	✓	✓		✓		✓						✓	
miR-2-3	✓	✓		✓	✓							✓	✓		✓	
miR-2-4	✓	✓	✓				✓		✓	✓	✓				✓	✓
miR-2-5		✓	✓	✓	✓		✓		✓		✓				✓	✓
miR-2-6		✓					✓									✓
miR-2-7	✓	✓		✓			✓		✓		✓		✓	✓	✓	
miR-2-8	✓						✓		✓		✓				✓	
miR-2-9			✓	✓	✓		✓		✓						✓	
miR-2-10	✓			✓		✓	✓		✓		✓				✓	✓

miR-2-11	✓						✓			✓				
miR-2-12	✓					✓	✓			✓		✓		✓
miR-2-13	✓				✓		✓	✓	✓	✓	✓			✓
miR-2-14	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓
miR-2-15	✓				✓				✓					✓



Out of the 26 miRNAs and 25 genes identified in this chapter, there are targeting interactions between 15 miRNAs and 16 genes.

It was to be expected that not all miRNAs and genes would interact with each other, since the gene identification was not based on using the miRNAs as input, as in chapter 2. Solely for the purpose of the following chapter, was the targeting interactions between the candidate miRNAs and genes scrutinized. Both miRNA and gene lists were identified as candidates for diagnostic purposes based on their specific expression levels in OC clinical cases.

Cofilin 1 (CFL1) is a small ubiquitous protein involved in cytokinesis, endocytosis, apoptosis, cell migration and cell mobility (Mousavi *et al.*, 2018). The activity status of CFL1 is directly associated with invasion, intravasation and metastasis of tumours. It was reported that increased CFL1 expression results in the progression of OC, and that targeting the activities of this gene is sufficient to significantly inhibit tumour invasiveness. Additional studies are still required to further elucidate the clinical outcomes of CFL1 in OC (Zhou *et al.*, 2012).

Ferritin light chain (FTL) is a key protein involved in iron metabolism (Wu *et al.*, 2016). The regulation of iron metabolism in OC plays a crucial role in promoting cell proliferation and a study published in 2019 by Wang *et al.* revealed that high expression levels of FTL correlates to a poor prognostic outcome in patients with OC.

Transmembrane BAX Inhibitor Motif Containing 6 (TMBIM6) is an endoplasmic reticulum (ER) protein that regulates apoptosis in response to ER-stress triggers. Overexpression of TMBIM6 has been implicated in OC (Liu, 2017).

### **3.4 Conclusion**

TCGA is a huge repository containing patient clinical data for various cancer types, resulting in studies that have significantly advanced the understanding of cancer biology (Mounir *et al.*, 2019). Using expression data extracted from TCGA as well as distribution curves and box plots as statistical measures, 26 miRNAs and 25 genes were identified as candidate biomarkers for OC diagnosis.

Out of the 25 genes, 16 are involved in protein binding which is known to be a vital cellular process.

While all 26 miRNAs and 25 genes are still considered the candidate biomarkers identified in this chapter, only the 15 miRNAs and 16 genes that interact with each other, will be carried forward to the next chapter to assess their triplex-forming abilities.



## Chapter 4

### **Prioritization of candidate miRNAs and genes based on the triplex-forming potential between interacting miRNAs and genes.**

#### **4.1 Introduction**

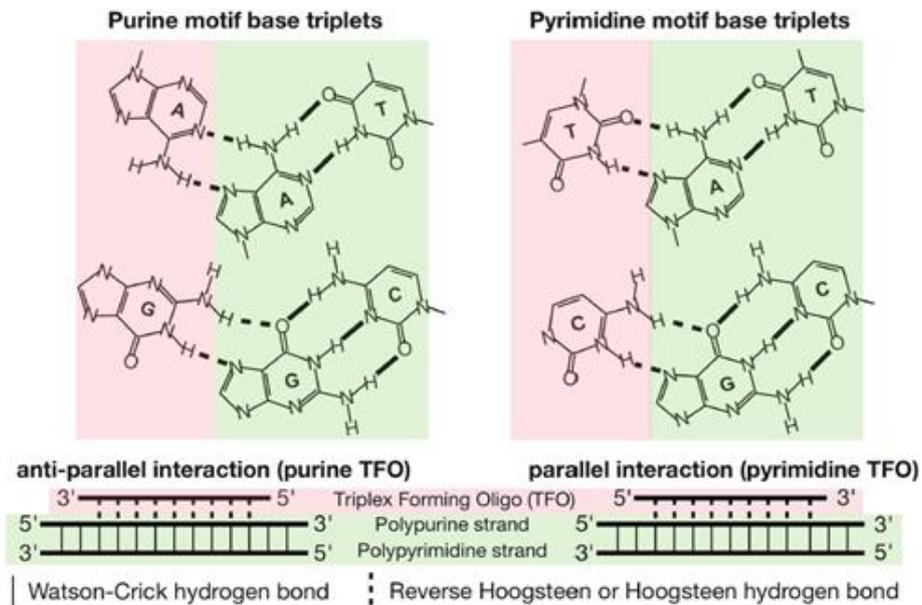
MiRNA-gene interactions are dynamic and depend on various factors, including subcellular location of miRNA, abundance of miRNA and target genes, and the affinity of their interactions. MiRNAs commonly interact with the 3'-UTR of target mRNAs, inducing mRNA degradation and translational inhibition. However, miRNAs are also capable of interacting with other sites in the target mRNA, such as the 5'-UTR, coding sequence as well as promoter regions. They have also been shown to upregulate gene expression in response to certain cellular conditions, sequences and co-factors. Binding of miRNA to the target gene's 3'-UTR, 5'-UTR and coding regions induces silencing effects on gene expression, whereas binding to the promoter region has been reported to induce transcription, thus stimulating gene expression (Valinezhad *et al.*, 2014; O'Brien *et al.*, 2018).

Eukaryotic promoters have a complex structure and several different sequence motifs. Transcription factors are required to bind to specific DNA sequences before RNA polymerase II can bind to the promoter and initiate transcription (Bhagavan and Ha, 2011). Commonly found in the promoter region of many genes are CpG islands which are an important feature of promoters as their methylation leads to silenced gene expression (Lim and Maher, 2010; Tollefsbol, 2011). With that being said, it is also known that approximately 45% of human gene promoters, particularly those of tissue-specific genes, do not contain CpG islands however, these genes are still transcriptionally silenced via methylation (Han *et al.*, 2011). Since promoter regions are significant regulatory sites, their sequences are of importance as well. Such sequences can be obtained from genome browsers which allow users to search, browse, extract and analyse genomic sequences (Wang *et al.*, 2012).

Two popular systems annotating and displaying genomic information are Ensembl Genome Browser and UCSC Genome Browser (Birney, 2004). Both browsers display gene annotations from various sources regarding sequence variation, conserved regions, CpG islands, as well as

regulatory and promoter sequences (Furey, 2006; Spudich *et al.*, 2010). Ensembl imports manually curated datasets, however if such evidence is not available, it annotates the gene set using a gene prediction pathway (Spudich *et al.*, 2007). UCSC produces several annotations based on mRNA alignments. mRNA and EST sequences are extracted from GenBank and aligned against the genome using a fast sequence alignment tool, BLAT (BLAST-like Alignment Tool). Data is filtered based on identity percentage as well as alignments that best match the sequence (Karolchik, 2003). BLAT is commonly used to locate sequences in a reference genome, identify homologous sequences from the genomes of closely related species, recognize exon-intron boundaries, determine gene structures, as well as aid in assembling and annotating genomic and transcriptomic sequences (Wang and Kong, 2019).

It has been postulated that miRNAs are capable of forming triplexes with the major groove of duplex DNA through either Hoogsteen or reverse Hoogsteen hydrogen bonds. Triplex formation involves a run of purines on one strand of the duplex. Purine bases contain more than one face from which they can form hydrogen bonds, which allows them to simultaneously participate in Watson-Crick pairing and either Hoogsteen or reverse Hoogsteen pairings (Paugh *et al.*, 2016). Bioinformatic analyses have uncovered enrichment of potential triplex targeting sites (TTS) in regulatory regions, primarily in promoters and enhancers. This direct interaction results in an altered gene function. As shown in Figure 4.1 below, triplex-forming oligonucleotides (TFOs) rich in pyrimidines interact with polypurine sequences via Hoogsteen base-pairing forming a parallel alignment whereas purine rich TFOs interact with polypurine sequences via reverse Hoogsteen base-pairing in an anti-parallel manner (Maldonado *et al.*, 2017).



**Figure 4.1:** Triplex forming interactions between TFOs and DNA polypurine strand. Solid lines indicate Watson-Crick hydrogen bonding whereas dotted lines represent either Hoogsteen or reverse Hoogsteen hydrogen bonding (Maldonado *et al.*, 2017).

Trident is a computational algorithm that assesses the landscape of potential miRNA triplex-binding sites in genomic DNA. It searches for triplex-forming units and Hoogsteen interactions according to the format of “XY:Z”, with “Z” representing the miRNA nucleotide. For example, TA:U and CG:C represents Hoogsteen interactions, whereas TA:A and CG:G indicates reverse Hoogsteen bonds. For each triplex binding site identified, Trident determines a thermodynamic binding energy as well as a heuristic score, with a higher heuristic score and lower thermodynamic energy relating to a stronger interaction (Paugh *et al.*, 2016).

The candidate miRNAs and genes from chapter 2 and 3 were identified through two distinct methodologies. Each chapter yielded a great number of candidates and since there was no overlap between the results, the candidates need to be ranked in order of priority. From there, candidates to be carried forward for future investigation can be distinguished. The Trident tool was implemented to prioritize the candidates for molecular validation, based on favourable thermodynamic interactions between a miRNA and its target genes. The motivation behind utilizing Trident to prioritize the candidates stem from the fact that: (a) it provides an additional line of evidence for an interaction between the miRNA and target

genes, and (b) triplex structures have been implicated in cancer due to its regulation of gene expression levels (Wang *et al.*, 2018).

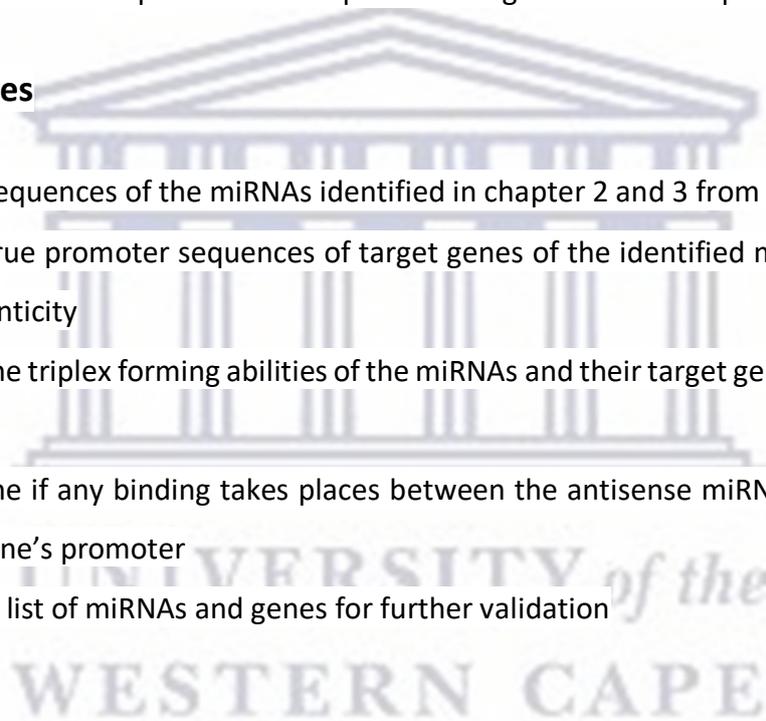
Since the purpose of this chapter was to rank and prioritize the candidates, the miRNAs and genes not involved in targeting interactions were still considered to part of the finalized candidate list. These miRNAs and genes were deemed to be of low priority.

#### **4.1.2 Aims**

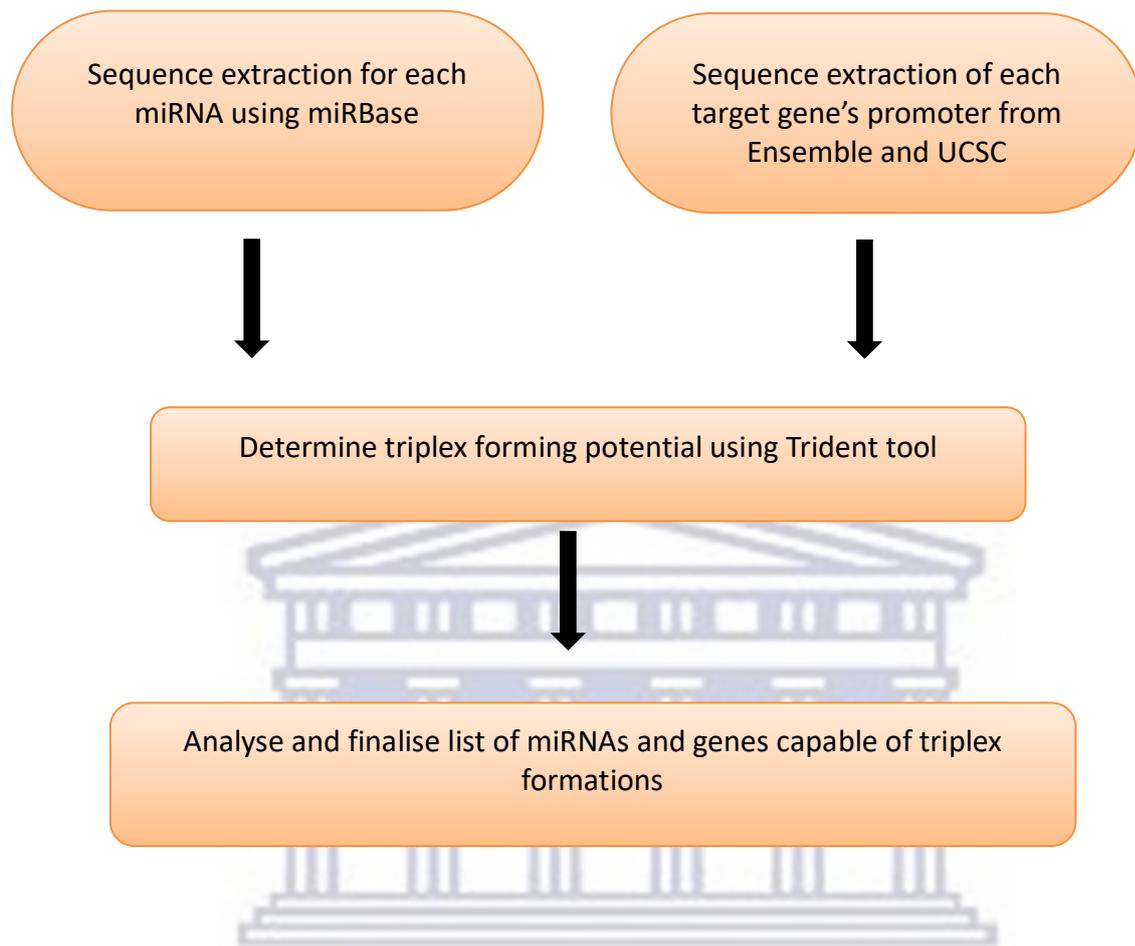
The aim of this chapter is to determine the potential of the candidate miRNAs identified in chapter 2 and 3, to form triplexes with the promoter regions of their respective target genes.

#### **4.1.3 Objectives**

- Extract sequences of the miRNAs identified in chapter 2 and 3 from miRBase
- Extract true promoter sequences of target genes of the identified miRNAs and verify its authenticity
- Determine triplex forming abilities of the miRNAs and their target genes, using Trident tool
- Determine if any binding takes places between the antisense miRNA strand and the target gene's promoter
- Highlight list of miRNAs and genes for further validation



## 4.2 Methodology



**Figure 4.2:** Flow chart depicting the outline of the *in-silico* methodology employed for identifying the candidates capable of triplex-formations.

### **4.2.1 MiRNA sequence extraction**

The candidate miRNAs and their target genes identified in both chapter 2 and chapter 3 were kept separate so their trident results could be compared.

MiRBase was employed to extract the mature sequence of each candidate miRNA by pasting the miRNA name into the search box. The mature sequences were saved in FASTA format and used for the Trident tool.

### **4.2.2 Promoter sequence extraction of miRNA target genes**

To ensure and verify that each gene's promoter sequence was authentic, both Ensembl Genome Browser and UCSC genome browser were used. Ensembl Genome Browser (available at <https://www.ensembl.org/index.html>) was launched and "Human" was specified as the species, with the gene name being pasted in the search box below. The query was submitted and the returned result with the correct gene name and species was selected. On the left tab of the gene result, "sequence" was selected and all configurations were left on default with the upstream flanking region, containing the promoter, being 600bp. This 600bp before the first exon was selected as the promoter region.

To verify that this sequence was in fact the true promoter, UCSC Genome Browser (available at <https://genome.ucsc.edu/>) was launched. The "BLAT" tool was utilized and the sequence extracted from Ensembl was pasted in the search box, with "Human" being specified and "Dec 2013" as the selected assembly. All other parameters were left on default and the query was submitted. The result with "100%" identity was analysed by selecting its "browser" link. Under the "Regulation" setting, CpG islands were selected to identify if it was present within the promoter region, however if no CpG islands were present, the promoter sequence was still deemed to be true based on the 100% identity between two different genome browsers.

### **4.2.3 Trident tool**

To predict triplex binding sites, Trident (available at <http://trident.stjude.org/>) was used. Under the "tools" tab, the miRNA sequence was pasted in the first search box with the specific target gene's promoter sequence in the second search box. "MiRanda Rules" was ticked and

"*Homo sapiens*" was specified. The query was submitted and if any results were returned, it was recorded and saved.

To serve as a negative control for this chapter, the reverse-complement of each miRNA sequence was generated using DNA Reverse Complement Sequence Generator Tool (available at [http://www.bugaco.com/calculators/dna\\_reverse\\_complement.php](http://www.bugaco.com/calculators/dna_reverse_complement.php)). RNA was specified as the sequence mode and each miRNA sequence was submitted as the input sequence. The reverse complementary sequence for each candidate miRNA was retained and used in the Trident tool along with its target gene's promoter sequence, to determine if anti-sense binding can occur.



### 4.3 Results and discussion

Gene promoter sequences are enriched with potential miRNA target sites in both sense and antisense directions (Catalanotto *et al.*, 2016), however the antisense miRNA has less potential to form favourable Hoogsteen bonds (Paugh *et al.*, 2016). Because of this, the antisense strand was used as a negative control to prioritize the miRNAs with triplex formations involving their sense strand only, thus resulting in a more enriched and validated interaction.

When searching for binding site pairs, Trident not only shares some similarities with MiRanda's miRNA-mRNA binding site algorithm but also incorporates its own rule adaptations. Trident binding rules assign a thermodynamic energy when searching for Hoogsteen (C:G and U:A) and Reverse Hoogsteen (G:G and A:A) binding (Paugh *et al.*, 2016).

Table 4.1 indicates the triplex-forming abilities of the candidate miRNAs and target genes identified in chapter 2. As can be seen there are a total of nine triplex interactions, involving seven sense-strand miRNAs and the promoters of five genes. For the antisense results, a total of five triplexes were returned involving four miRNAs and three genes. There is no overlap between the sense and antisense results except for miR-1-3a which binds to AMHR2 in both a sense and antisense manner. These interactions are tabulated in Table 4.1.1 along with their energy and heuristic scores, genomic position, and hit structure.

As mentioned previously, a higher heuristic score and lower thermodynamic energy correlates to a stronger interaction. The six miRNAs with only sense stand interactions have similar scores and are therefore, based on their favoured triplex formation abilities, equally prioritized. Due to the fact that miR-1-3a is capable of both sense and antisense bindings to AMHR2, it was deemed to be of low priority.

The five genes involved in sense triplexes (WISP1, WNT5A, AMHR2, EGR2 and PGR) are, as depicted in section 2.3.7.2 of chapter 2, linked to positive regulation of transcription from RNA polymerase II promoter, signal transduction, Wnt signalling pathway, and cell-cell signalling. Implication of the Wnt signalling pathway in cancer has been discussed in chapter 2, and it should be noted that when dysregulated, all other pathways mentioned here

contribute to tumorigenesis as well (Martin, 2003; Villicaña *et al.*, 2014; Sever and Brugge, 2015).



**Table 4.1:** Candidate miRNAs and target genes identified in chapter 2 that have triplex forming abilities.

	Binding type	WISP1	WNT5A	AMHR2	EGR2	ESR1	LHCGR	PGR
miR-1-2a	Sense		✓					✓
	Anti-sense						☑	
miR-1-2b	Sense				✓			
	Anti-sense							
miR-1-2c	Sense				✓			
	Anti-sense					☑		
miR-1-3a	Sense			✓				
	Anti-sense			☑				
miR-1-3b	Sense	✓	✓					
	Anti-sense							
miR-1-3c	Sense			✓				
	Anti-sense							

---

<b>miR-1-4a</b>	Sense	✓		
	Anti-sense		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

---



**Table 4.1.1:** Triplex interactions identified between each candidate miRNA and its target genes for both sense and antisense strands using Trident. Table indicates the Energy score (E), Heuristic score (H), Genomic position (P) and binding structure.

miR-1-2a					
Sense	WNT5A	E: -11.77	H: 144	P: 39-60	accguuuuggCGUUAUGAAGa  : :      cattcgtgtgGTTGTTACTTct gtaagcacacCAACAATGAAGa
	PGR	E: -8.98	H: 144	P: 31-50	accGUUUUGGCGUUAUGAAGa :    :   :       taaTAAATTAGT--TTACTTct attATTTAATCA--AATGAAGa
Antisense	LHCGR	E: -10.01	H: 142	P: 24-44	ucuucaUUAACGCCAAAACGGu             : aatctgAATTG-GATTTTGCTc ttagacTTAAC-CTAAAACGag
miR-1-2b					
Sense	EGR2	E: -14.3	H: 148	P: 195-216	ccGUUUUUGUCGUUAUGAAAg   :  : :      :    ttCTAGACTGTCAATTATTTTc aaGATCTGACAGTTAATAAAg

miR-1-2c

Sense	EGR2	E: -11.49	H: 152	P: 195-216	ccGUUUUGGGCGUUAUGAAAa    : :          :    ttCTAGACTGTCAATTATTTTc aaGATCTGACAGTTAATAAAAa
-------	------	-----------	--------	------------	---

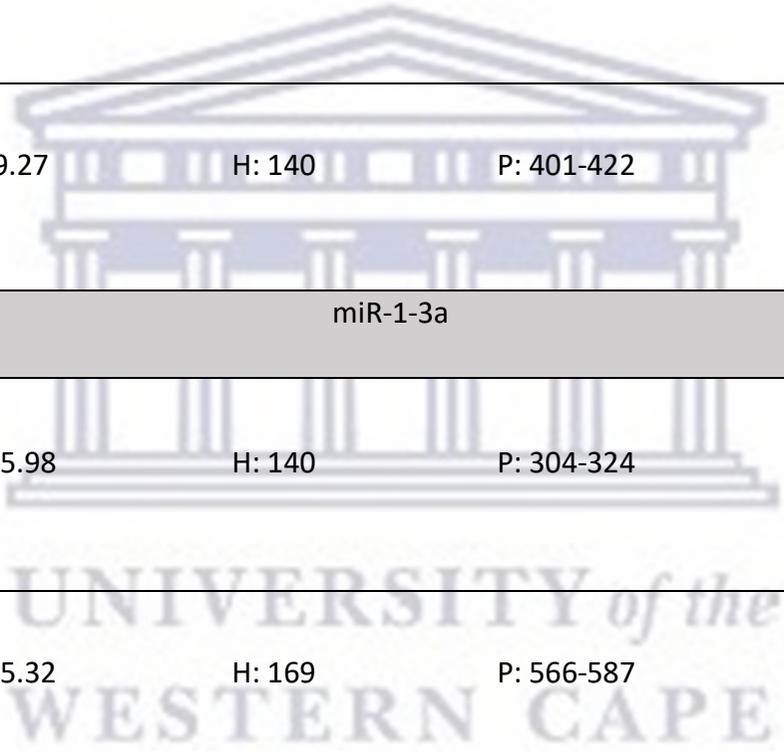
Antisense	ESR1	E: -9.27	H: 140	P: 401-422	uuUUCAUUAACGCCCAAAACgg                             aaAAATTACTGACGGTTTTGag ttTTTAATGACTGCCAAAACtc
-----------	------	----------	--------	------------	---

miR-1-3a

Sense	AMHR2	E: -15.98	H: 140	P: 304-324	ggaagacugaggaUCAGGUca                             gaagagtcaacagAGTCCAGc cttctcagttgtcTCAGGTCg
-------	-------	-----------	--------	------------	--

Antisense	AMHR2	E: -25.32	H: 169	P: 566-587	ugaCCUGA-UCCUCAGUCUUCc                             tggGGACTGAGGGATCAGAAGc accCCTGACTCCCTAGTCTTCg
-----------	-------	-----------	--------	------------	---

miR-1-3b





Antisense

AMHR2

E: -14.25

H: 144

P: 530-550

aaGU-UCAUUAAGUCCUAUCCa  
|| || |: ::||||||  
ccCACAG-AGGCTGGGATAGGa  
ggGTGTC-TCCGACCCTATCCt

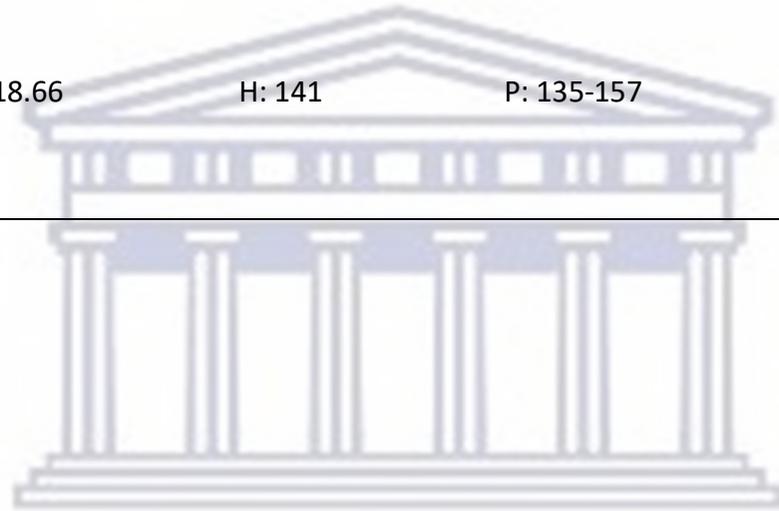
LHCGR

E: -18.66

H: 141

P: 135-157

aagUUCAUUA--GUCCUAUCCa  
::|||:| | ||||:|  
ctgGGGTAGTTGGGAGGATGGGc  
gacCCCATCAACCCTCCTACCCg



UNIVERSITY *of the*  
WESTERN CAPE

Table 4.2 indicates triplex-forming interactions between candidate miRNAs and target genes identified in chapter 3. A total of eleven sense-strand triplexes were formed between seven miRNAs and five genes. For the antisense results, a total of six triplexes were produced between four miRNAs and five genes. Once more, there is only one overlapping result namely miR-2-14 that binds to TMBIM6 in both a sense and antisense manner. The information regarding these interactions can be found in Table 4.2.1.

All sense strand interaction scores are relatively similar and equally prioritized, with miR-2-14 and ACTB having the strongest interaction and therefore being of highest priority.

The five genes (ACTB, CALR, GNAS, PTMA and TMBIM6) that are involved in sense-binding triplexes have roles in protein binding, with CALR and TMBIM6 also having roles in ubiquitin protein ligase binding. As previously discussed, protein binding is a process that when altered, contributes greatly to cancer onset and progression (Pereira *et al.*, 2017). Ubiquitin ligases are vital components of the ubiquitin proteasome system (UPS) which governs crucial processes regulating cellular homeostasis, cell cycle and metabolism in response to DNA damage and stress signals. Dysregulation of ubiquitin ligases results in alterations of substrate availability and activity, thereby promoting cellular transformation and tumorigenesis (Qi and Ronai, 2015).

Actin beta (ACTB) is a cytoskeleton structural protein that plays pivotal roles in cell development and migration, embryonic development, and gene expression. It is generally regarded as a housekeeping gene with an assumption that its expression is unaffected by experimental or physiological conditions, however it was found to be differentially expressed in a variety of cancers, including OC. ACTB expression is reported to be upregulated in OC samples when compared to healthy ovarian tissues (Guo *et al.*, 2013). This is especially interesting since triplex formation between a miRNA and its target gene is said to increase gene expression (Paugh *et al.*, 2016). This allows for the assumption that since miR-2-14 forms triplexes with ACTB and hence upregulates its expression, it directly leads to OC onset and progression.

**Table 4.2:** Candidate miRNAs and target genes identified in chapter 3 that have triplex forming abilities.

	Binding type	ACTB	ACTG1	CALR	GNAS	PSAP	PTMA	RPL15	TMBIM6
miR-2-1	Sense	✓			✓				✓
	Anti-sense								
miR-2-2	Sense				✓				
	Anti-sense								
miR-2-3	Sense			✓					
	Anti-sense								
miR-2-4	Sense								
	Anti-sense								☑
miR-2-5	Sense								✓
	Anti-sense								
miR-2-7	Sense								

	Anti-sense		<input checked="" type="checkbox"/>				
miR-2-8	Sense						✓
	Anti-sense						
miR-2-11	Sense					✓	
	Anti-sense						
miR-2-13	Sense						
	Anti-sense		<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>
miR-2-14	Sense	✓				✓	✓
	Anti-sense						<input checked="" type="checkbox"/>

UNIVERSITY of the  
WESTERN CAPE

**Table 4.2.1:** Triplex interactions identified between each candidate miRNA and its target genes for both sense and antisense strands using trident. Table indicates the Energy score (E), Heuristic score (H), Genomic position (P) and binding structure.

miR-2-1						
Sense	ACTB	E: -25.12	H: 149	P: 34-55		uugauAUGUUGGAGGAUGGAGu      :    :     ttcacTCCATTCTCCTGCCTCa aagtgAGGTAAGAGGACGGAGt
	GNAS	E: -21.9	H: 166	P: 392-412		uuGAUAUGUUGGAGGAUGGAGu     :           gaCT-TGCAAACTACTACCTCa ctGA-ACGTTTGATGATGGAGt
	TMBIM6	E: -25.4	H: 149	P: 134-156		uugaUAUGUU-GGAGGAUGGAGu   :   :    :     ggtcAAGCAATTCTCCTGCCTCa ccagTTCGTTAAGAGGACGGAGt
miR-2-2						
Sense	GNAS	E: -17.54	H: 157	P: 348-371		aguGUCCAUUUCCAGAGUCCCu                 tctCAGCTCAATCGAGCTCAGGGc agaGTCGAGTTAGCTCGAGTCCCg

miR-2-3					
Sense	CALR	E: -26.84	H: 142	P: 137-160	gucgGAUACCUUAAG-UCAAGAGu   :   ::    :      gcagCAGTGGGGTGCTGGTTCTCa cgtcGTCACCCCACGACCAAGAGt
miR-2-4					
Antisense	TMBIM6	E: -13.13	H: 141	P: 356-377	aucguCGUGUAGUACCAAAUGu    :        :  taattGCATACCTGGGTTTGCCc attaaCGTATGGACCCAAACGg
miR-2-5					
Sense	TMBIM6	E: -18.79	H: 153	P: 799-819	ugaCCUGAAACAUCCGGUCAa    :            cgaGGCGGTTGGT-GGCCAGTa gctCCGCCAACCA-CCGGTCat
miR-2-7					
Antisense	GNAS	E: -14.54	H: 140	P: 538-560	uaguguaACGGUCCUAAUGGUg        :      taatggcTGCCCGAAGTTACCAc attaccgACGGGCTTCAATGGTa

miR-2-8

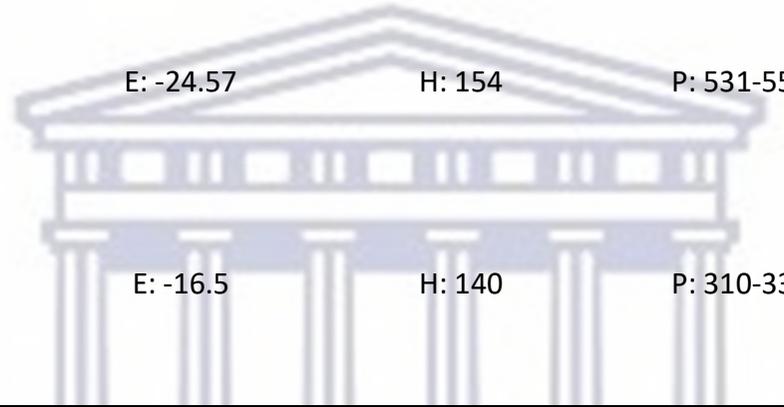
Sense	TMBIM6	E: -25.48	H: 140	P: 75-98	<pre> agucuGGCUCUG--UUCACGUUAc        :  :   :  tgtctCCCAGGCTGGAGTGCAGTg acagaGGGTCCGACCTCACGTAc           </pre>
-------	--------	-----------	--------	----------	---

miR-2-11

Sense	PTMA	E: -13.72	H: 140	P: 21-42	<pre> gaaggucagcccccACAAUGu                       gcgctctgtcccccTGTTTACg cgcgagacagggggACAAATGc           </pre>
-------	------	-----------	--------	----------	--

miR-2-13

Antisense	ACTG1	E: -26.37	H: 148	P: 218-241	<pre> ucGAGCCAGA-CUCCGGGAGUCa               :     gcCTCGGCCTCCATCCCTCTCAGt cgGAGCCGGAGGTAGGGAGAGTCa           </pre>
		E: -24.55	H: 143	P: 300-326	<pre> ucGAG-CCAGA---CUCCGGGAGUCa         :         :   ccCTCAGGTTTCCCAGAGCCCTTAGg ggGAGTCCAAGGGCTCTCGGGAATCc           </pre>



E: -24.31                      H: 142                      P: 282-307

ucgagccaGACUCCG---GGGAGUCa  
      |||: |: |||  
accggcacCTGGTGTCCACCCTCAGg  
tggccgtgGACCACAGGTGGGAGTCc

PSAP                      E: -24.57                      H: 154                      P: 531-553

ucGAGCCAGACUCCG-GGGAGUCa  
      ||: | | ||| |||  
ccCTTCG-CGAGGGCCCTCAGa  
ggGAAGC-GCGTCCGCGGGAGTCt

RPL15                      E: -16.5                      H: 140                      P: 310-332

ucgagccagacuccgGGGAGUCa  
                              |||  
cattccaaattgcaaCCCTCAGc  
gtaaggtttaacgttGGGAGTCg

miR-2-14

ACTB                      E: -28.42                      H: 165                      P: 281-303

gaUGGACGUGCUUGUCGUGAAAc  
      :||||:|    |||  
acGCCTGTAATCTCAGCACTTTg  
tgCGGACATTAGAGTCGTGAAAc

Sense

PTMA                      E: -26.98                      H: 142                      P: 210-231

gauGGACGUGCUUGUCGUGAAAc  
      ||| |: ||| ||| |  
agtCCT-CGGGAACAGCACTGTg  
tcaGGA-GCCCTTGTCGTGACAc

TMBIM6

E: -16.17

H: 140

P: 568-588

gauGGACGUGCUUGUCGUGAAAc  
||| :|||: :||:||||  
gccCCT-TGCGAG-GGCGCTTTt  
cggGGA-ACGCTC-CCGCGAAAa

Antisense

TMBIM6

E: -16.92

H: 141

P: 853-875

guuucacGACAAGCACGUCCAuc  
|||||: |||| |||  
gcagccGCTGTTTATGCACGTAc  
cgtcggCGACAAATACGTGCATg



UNIVERSITY *of the*  
WESTERN CAPE

## 4.4 Conclusion

Through regulating their target genes, miRNAs influence a broad spectrum of biological processes. Triplex structures involving miRNAs and gene promoter regions, formed in the major groove of duplex DNA, is a potential mechanism whereby miRNAs directly modify gene transcription. The ability of miRNAs to form triplexes depends on their length and sequence (Paugh *et al.*, 2016).

All triplex interactions identified in this chapter involve genes that govern crucial pathways and processes, which further prioritizes the miRNAs involved. This also indicates the strength that both pipelines offer in identifying miRNAs with significant triplex interactions. The candidate miRNAs with no triplex bindings are not disregarded in this study as it is possible that their triplex forming potential is yet to be uncovered and characterized through further studies.

Out of the set of miRNAs identified from the first pipeline and second pipeline, six and seven miRNAs respectively are enriched for strong triplex formations and additional investigations concerning them should be conducted. Due to the strong interaction between miR-2-14 and ACTB, based on them having the lowest energy binding and one of the highest heuristic scores, it is prioritized before all other candidates. All other miRNA-gene triplex interactions have similar binding scores and are equally prioritized after miR-2-14. Additional support for the great potential that ACTB offers, is its role in fundamental cellular pathways (Guo *et al.*, 2013).

Uncovering Hoogsteen and reverse Hoogsteen bonds is a task that can be performed by molecular simulations in the future, as the purpose of this chapter was to determine enriched triplex interactions between the candidate miRNAs and their target gene's promoter. Based on proven gene interactions in the form of triplexes, the prioritized miRNAs are deemed to be suitable biomarkers for OC diagnosis.

## Chapter 5

### General discussion and future prospects

With a 5-year survival rate of around 40%, OC is the most common gynaecologic malignancy. Accounting for this low survival rate, is the fact that it is often diagnosed in the advanced stages due to the lack of early symptoms. It is anticipated that if OC is diagnosed in the early stages, survival rates will increase to 80% (Burgess and Schmalfeldt, 2011). This stresses the urgency for identifying sensitive and specific biomarkers that can aid in the early diagnosis of this disease.

MiRNAs are appealing as biomarkers due to their roles in fundamental cellular processes as well as their proven dysregulation in cancer (Wang *et al.*, 2015). Since miRNAs are non-coding, they exert their function through the regulation of their target genes. For this reason, their target genes are equally appealing (Wong and Wang, 2014; Hammond, 2015). The discovery of biomarkers can be achieved through various techniques, with bioinformatics playing a pivotal role. Bioinformatics allows one to generate, store, annotate and analyse biological data, while remaining cost effective (Chowdhary *et al.*, 2016).

This study has relied on two bioinformatic techniques to separately identify miRNAs and genes with strong potential as diagnostic biomarkers for early stage OC.

The first pipeline incorporated a sequence similarity approach whereby novel miRNAs with no links to OC were identified as candidate biomarkers based on their similarity in sequence to miRNAs with a validated mechanism in OC. This is based on the notion that two sequences sharing a high degree of similarity, will most likely share a similar function (Lu *et al.*, 2008). In this study, a similarity range of 90-99% was set when using two similarity programs, namely BLAST and CD-HIT. In common to both programs, were 28 miRNAs with similarities to 19 miRNAs that have a proven dysregulation in OC. Following textmining, 9 miRNAs had no mechanism in OC to date and were finalized as the candidates for further study. Gene targets of the candidate miRNAs were extracted and those expressed in the ovary were identified and retained for further analysis. Functional annotation from DAVID returned 12 genes that have direct links in OC, with them being involved in processes related to cancer when dysregulated. Since the 9 candidate miRNAs target 12 genes, including FOS, WISP1, WNT5a,

EGR2 and PGR, with clear links to OC, it strengthens their potential in being suitable diagnostic biomarkers.

The second pipeline dealt with patient clinical data extracted from TCGA, a comprehensive atlas of genomic cancer profiles (Chu *et al.*, 2015). Distribution curves for both the miRNAs and genes proved that the extracted data was accurate and stable. Using box plots, the top 25% highly expressed miRNAs and genes were identified, resulting in 26 miRNAs and 25 genes shortlisted as the candidates of chapter 3. Out of this, it was noted that targeting interactions occur between 15 miRNAs and 16 genes, which were carried over to chapter 4 of this study.

The candidate miRNAs and genes with targeting interactions identified in each pipeline were subjected to the trident tool to further prioritize them based on their triplex-forming abilities. Triplexes, which directly alter gene function, can be formed between miRNAs and the major groove of duplex DNA (Paugh *et al.*, 2016). Out of the 9 miRNAs and 12 genes identified from the first pipeline, 6 miRNAs and the promoter regions of 5 genes were predicted to form sense triplexes. With a higher heuristic score and lower thermodynamic energy indicating a stronger interaction, all triplexes had similar scores and were thus equally prioritized. Out of the 15 miRNAs and 16 genes with targeting interactions identified in the second pipeline, 7 miRNAs and 5 genes are capable of forming sense triplexes. The triplex interaction between miR-2-14 and ACTB had the strongest interaction and was deemed as a top priority candidate, with the rest being equally prioritized. This interaction is especially interesting since ACTB is said to be upregulated in OC and it has been proposed that triplex formations increase a genes expression level (Guo *et al.*, 2013; Paugh *et al.*, 2016). This allows for the assumption that miR-2-14 directly causes the upregulation and dysregulation of ACTB which leads to OC onset and progression. The candidate miRNAs and genes not involved in any targeting interactions and triplex formations were still noted as candidates, just of low priority. Ranking of the candidates according to their priority can be found in Table 5.1 below.

**Table 5.1:** Prioritization of all candidate miRNAs based on their ability to form triplexes with their target genes.

	<b>MiRNAs</b>	<b>Genes</b>
<b>Top priority</b>	miR-2-14	ACTB
	miR-1-2a	WISP1
<b>Equally prioritized</b>	miR-1-2b	WNT5A
	miR-1-2c	AMHR2
	miR-1-3b	EGR2
	miR-1-3c	PGR
	miR-1-4a	CALR
	miR-2-1	GNAS
	miR-2-2	PTMA
	miR-2-3	TMBIM6
	miR-2-5	
	miR-2-8	
	miR-2-11	
<b>Low priority</b>	miR-1-1a	FOS
	miR-1-3a	BNC2
	miR-1-3d	CYP19A1
	miR-2-4	ESR1
	miR-2-6	ESR2
	miR-2-7	LHCGR

---

miR-2-9	PDGFB
miR-2-10	CFL1
miR-2-12	MT-CO1
miR-2-13	HSP90AB1
miR-2-15	MT-ND4
miR-2-16	MT-ATP6
miR-2-17	ACTG1
miR-2-18	MT-ND2
miR-2-19	FTL
miR-2-20	TUBB
miR-2-21	MT-CO3
miR-2-22	RPL15
miR-2-23	MT-RNR2
miR-2-24	MT-ND5
miR-2-25	RPL3
miR-2-26	MT-CO2
	PSAP
	ALDOA
	RPL7A
	MT-ND1
	RPL4

---

From the two approaches employed in this study, one relying on sequence similarity and the other based on patient clinical data, no overlapping results were obtained. The reason for each pipeline yielding its own distinct set of results could be attributed to the fact that the two pipelines share no underlying resemblance. The sequence similarity approach is a prediction methodology based purely on biological knowledge, whereas the TCGA approach is based on patient clinical data that depicts the confirmed miRNA/gene expression levels in OC patients.

This study has identified candidate miRNAs and genes that offer promising potential in being suitable diagnostic biomarkers for early stage OC. Future work regarding these candidates would entail molecular validation, followed by the development of a lateral flow device for diagnostic purposes.

In order to prove their dysregulated expression in OC, quantitative PCR (qPCR) should be performed on the candidate miRNAs and genes. qPCR involves monitoring DNA amplification in real time by tracking fluorescence. After each cycle, fluorescence is measured with the intensity of the fluorescent signal indicating the amount of DNA amplicons in the sample at that specific time. In terms of gene expression studies, qPCR is an extensively applied technique (Kralik and Ricchi, 2017). This would confirm the candidate genes dysregulated expression levels in OC samples when compared to healthy ovarian tissue. Possible OC cell lines that could be utilized are CaoV-3, Ovc3 and OAW28.

To determine the effect of the candidate miRNAs on their target genes, luciferase assays would need to be performed. A luciferase reporter gene assay is a common application used to examine the regulation of transcriptional activities by promoters and transcription factors. Adaptations to this assay allows for exploring the post-transcriptional regulation of miRNAs on their target genes. This can be attained by engineering a luciferase gene construct containing the predicted miRNA targeting sequence from the target gene (Jin *et al.*, 2013). Cells are co-transfected with the miRNA as well as a plasmid containing a luciferase coding sequence upstream of the mRNA gene of interest. If the miRNA targets the mRNA, the luminescence variation will be altered, thus reflecting the changes in the transcript's stability and/or translation efficiency (Campos-Melo *et al.*, 2014). This assay will offer confirmation of the predicted targeting interaction between the miRNAs and genes identified in this study.

Once all the predicted results from this *in silico* study has been validated, the end goal would be to construct a Point of Care (POC) device, such as a lateral flow assay, using nanotechnology. This device offers great potential due to its ease of use, and being robust and inexpensive (de Puig *et al.*, 2017). Since current OC biomarkers are failing, there is a great demand to identify ones that can offer an accurate diagnosis. If the identified candidate miRNAs and genes are validated to be sensitive and specific biomarkers for OC diagnosis, using them to construct a lateral flow device provides a promising and cost-effective means for early diagnosis of OC.



## Appendix A

Chapter 2 supplementary information

**Table A.1:** List of 198 OC validated miRNAs, after duplication removal.

hsa-miR-199a-3p	hsa-miR-200a-3p	hsa-miR-548a-5p
hsa-miR-499-3p	hsa-miR-301a-3p	hsa-miR-548b-5p
hsa-miR-371-3p	hsa-miR-513a-3p	hsa-miR-1224-3p
hsa-miR-129-2-3p	hsa-miR-548d-5p	hsa-miR-1225-3p
hsa-miR-509-3-5p	hsa-miR-181c-5p	hsa-miR-520d-5p
hsa-miR-1245b-5p	hsa-miR-146b-5p	hsa-miR-518a-5p
hsa-miR-1255b-5p	hsa-miR-135b-5p	hsa-miR-520a-3p
hsa-miR-219-1-3p	hsa-miR-302c-3p	hsa-miR-519b-3p
hsa-miR-193a-3p	hsa-miR-516a-3p	hsa-miR-516a-5p
hsa-miR-146b-3p	hsa-miR-1915-3p	hsa-miR-518d-3p
hsa-miR-125a-3p	hsa-miR-378a-3p	hsa-miR-548a-3p
hsa-miR-193a-5p	hsa-miR-200c-3p	hsa-miR-885-5p
hsa-miR-548c-3p	hsa-miR-193b-3p	hsa-miR-362-5p
hsa-miR-514b-5p	hsa-miR-1185-5p	hsa-miR-502-3p
hsa-miR-513a-5p	hsa-miR-374b-5p	hsa-miR-532-5p
hsa-miR-199a-5p	hsa-miR-302a-3p	hsa-miR-574-3p
hsa-miR-548d-3p	hsa-miR-130b-3p	hsa-miR-532-3p
hsa-miR-1225-5p	hsa-miR-135a-5p	hsa-miR-501-3p
hsa-miR-548c-5p	hsa-miR-1307-3p	hsa-miR-542-5p
hsa-miR-1207-3p	hsa-miR-519c-3p	hsa-miR-188-5p
hsa-miR-106a-5p	hsa-miR-302b-3p	hsa-miR-362-3p
hsa-miR-106b-5p	hsa-miR-199b-5p	hsa-miR-509-5p
hsa-miR-450b-3p	hsa-miR-181b-5p	hsa-miR-486-5p
hsa-miR-518a-3p	hsa-miR-1277-3p	hsa-miR-342-5p
hsa-miR-151a-3p	hsa-miR-548b-3p	hsa-miR-34c-3p
hsa-miR-550a-5p	hsa-miR-517c-3p	hsa-miR-34c-5p

hsa-miR-324-3p
hsa-miR-501-5p
hsa-miR-508-3p
hsa-miR-508-5p
hsa-miR-296-3p
hsa-miR-127-5p
hsa-miR-127-3p
hsa-miR-140-3p
hsa-miR-485-5p
hsa-miR-299-3p
hsa-miR-337-5p
hsa-miR-654-3p
hsa-miR-140-5p
hsa-miR-509-3p
hsa-miR-371-5p
hsa-miR-769-5p
hsa-miR-340-5p
hsa-miR-221-3p
hsa-miR-425-5p
hsa-miR-30d-5p
hsa-miR-27a-3p
hsa-miR-183-5p
hsa-miR-30b-5p
hsa-miR-92a-3p
hsa-miR-342-3p
hsa-miR-30c-5p
hsa-miR-18a-5p
hsa-miR-411-5p
hsa-miR-223-3p
hsa-miR-205-5p
hsa-miR-142-3p

hsa-miR-26a-5p
hsa-miR-367-3p
hsa-miR-486-3p
hsa-miR-30e-5p
hsa-miR-141-3p
hsa-miR-660-5p
hsa-miR-625-5p
hsa-miR-664-3p
hsa-miR-767-5p
hsa-miR-542-3p
hsa-miR-454-3p
hsa-miR-339-3p
hsa-miR-576-3p
hsa-miR-20a-5p
hsa-miR-192-5p
hsa-miR-10b-5p
hsa-miR-502-5p
hsa-miR-369-3p
hsa-miR-150-5p
hsa-miR-770-5p
hsa-miR-483-3p
hsa-miR-338-3p
hsa-miR-576-5p
hsa-miR-23a-3p
hsa-miR-337-3p
hsa-miR-659-3p
hsa-miR-409-5p
hsa-miR-34a-5p
hsa-miR-182-5p
hsa-miR-382-5p
hsa-miR-493-3p

hsa-miR-100-5p
hsa-miR-485-3p
hsa-miR-766-3p
hsa-miR-29a-3p
hsa-miR-361-5p
hsa-miR-455-3p
hsa-miR-541-3p
hsa-miR-144-3p
hsa-miR-23b-3p
hsa-miR-330-5p
hsa-miR-187-3p
hsa-miR-29c-3p
hsa-miR-185-5p
hsa-miR-34b-3p
hsa-miR-33b-5p
hsa-miR-331-3p
hsa-miR-154-5p
hsa-miR-101-3p
hsa-miR-214-3p
hsa-miR-143-3p
hsa-miR-139-3p
hsa-miR-126-3p
hsa-miR-483-5p
hsa-miR-888-5p
hsa-miR-188-3p
hsa-miR-615-5p
hsa-miR-525-5p
hsa-miR-875-5p
hsa-miR-139-5p
hsa-miR-628-3p
hsa-miR-769-3p

hsa-miR-490-5p
hsa-miR-409-3p
hsa-miR-490-3p
hsa-miR-142-5p
hsa-miR-574-5p
hsa-miR-219a-5p
hsa-miR-628-5p
hsa-miR-875-3p
hsa-miR-876-5p

hsa-miR-590-5p
hsa-miR-129-5p
hsa-miR-876-3p
hsa-miR-339-5p
hsa-miR-491-5p
hsa-miR-93-5p
hsa-miR-96-5p
hsa-let-7d-5p
hsa-miR-21-5p

hsa-let-7i-5p
hsa-miR-17-5p
hsa-miR-516-5p
hsa-miR-30a-5p
hsa-mir-135a-3p
hsa-mir-200b-3p
hsa-mir-551b-3p
hsa-mir-224-5p
hsa-mir-28-5p



UNIVERSITY *of the*  
WESTERN CAPE

**Table A.2:** Ovary-specific genes extracted, following duplication removal.

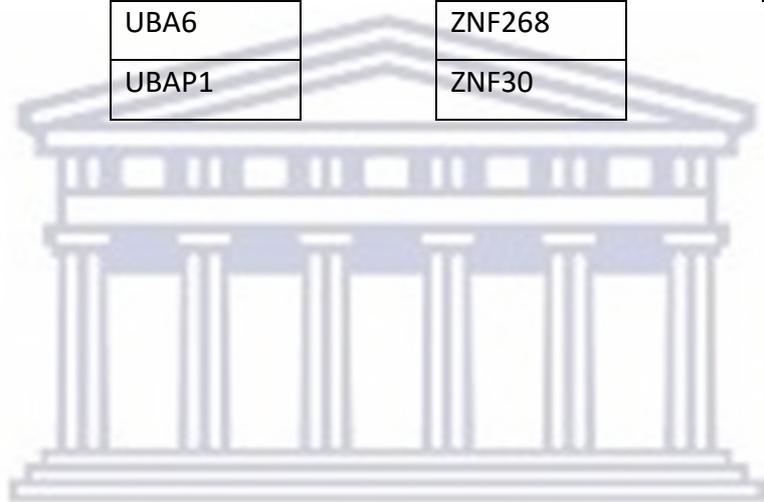
ACSM1	COL6A2	GOT1L1	MUM1L1
ADAMTS16	COL6A3	GSTM2	NPC1L1
ADRB3	CPZ	GSTM5	NR2F2
AMHR2	CRYGD	GTF2A1L	NR4A2
ANGPTL2	CSF2	HAS1	NRK
ANGPTL5	CYP11A1	HS3ST1	OGN
AQR	CYP19A1	HSD11B1	OR4D10
ARMS2	CYP2W1	HSD3B2	OR5K1
ARX	DDX17	IQCH	OTOR
ASMT	DISC1	KCNK7	OVGP1
ATP4B	ECEL1	KERA	P2RX2
BAIAP3	EGR1	KLHDC8A	PAEP
BCL2L2	EGR2	KRT32	PDGFB
BMP6	EGR3	KRTAP10-12	PDLIM4
BNC2	EML5	LCN1	PDZK1
C12orf71	EPHB2	LDHAL6A	PEG3
C6	EPYC	LEFTY2	PGR
CCBE1	ESR1	LHCGR	PHACTR3
CCDC17	ESR2	LHX9	PKNOX2
CCDC94	ETV7	LPAR4	POU1F1
CDC42BPA	FAM153A	LRRC17	PRPS2
CDH11	FAM19A3	LTBP4	PRSS35
CDH3	FOS	MAMLD1	PUS7L
CDON	FOSB	MAP4K1	RBMS3
CELF5	FOXL2	MDFI	RBP1
CHRNA4	FOXO1	MDK	REN
CHST9	GABRB2	MEX3A	RHBG
CLSTN2	GLI3	MGA	RPP25
CLUL1	GNGT1	MMP11	RRS1

SCGB1D4
SCGB2A2
SCML1
SEMA6D
SIGLEC11
SLC22A1
SLC24A2
SLC25A41
SLC7A8
SNCAIP
SOX4

SPATA5L1
SPRR2F
SSTR3
STIP1
TAB3
TCF23
TMEM190
TRPC1
TSHZ3
UBA6
UBAP1

UBXN8
USH1G
VWCE
WFIKKN2
WISP1
WNT4
WNT5A
XCR1
ZFPM2
ZNF268
ZNF30

ZNF469
ZNF540
ZNF549
ZNF556
ZNF660
ZNF714
ZNF83
ZNF837
ZP4



UNIVERSITY *of the*  
WESTERN CAPE

## References

- Allegra, A., Alonci, A., Campo, S., Penna, G., Petrunaro, A., Gerace, D. and Musolino, C. (2012). Circulating microRNAs: New biomarkers in diagnosis, prognosis and treatment of cancer (Review). *International Journal of Oncology*. 41(6):897-1912.
- Antunović, B., Barlow, S., Chesson, A., Flynn, A., Hardy, A., Jeger, M., Knaap, A., Kuiper, H., Lovell, D., Nørrung, B., Pratt, I., Rietjens, I., Schlatter, J., Silano, V., Smulders, F. and Vannier, P. (2011). Statistical Significance and Biological Relevance. *EFSA Journal*. 9(9):2372
- Azuaje, F. (2013). *Bioinformatics and biomarker discovery*. Hoboken, New Jersey.: Wiley, 115-117.
- Babu, M. (1997). "Biological Databases and Protein Sequence Analysis", Center for Biotechnology, Anna University, Chennai-25, India.
- Ballman, K. (2015). Biomarker: Predictive or Prognostic? *Journal of Clinical Oncology*. 33(33):3968-3971.
- Barca-Mayo, O. and Lu, Q. (2012). Fine-Tuning Oligodendrocyte Development by microRNAs. *Frontiers in Neuroscience*. 6:1-7.
- Bayat, A. (2002). Bioinformatics. *British Medical Journal*. 324(7344):1018–1022.
- Bhagavan, N. and Ha, C. (2011). *Essentials of medical biochemistry*. Amsterdam: Elsevier/Academic Press.
- Bhajun, R., Guyon, L., Pitaval, A., Sulpice, E., Combe, S., Obeid, P., Haguët, V., Ghorbel, I., Lajaunie, C. and Gidrol, X. (2015). A statistically inferred microRNA network identifies breast cancer target miR-940 as an actin cytoskeleton regulator. *Scientific Reports*. 5(1):8336.
- Birney, E. (2004). An Overview of Ensembl. *Genome Research*. 14(5):925-928.
- Brain, K., Smits, S., Simon, A., Forbes, L., Roberts, C., Robbé, I., Steward, J., White, C., Neal, R. and Hanson, J. (2014). Ovarian cancer symptom awareness and anticipated delayed presentation in a population sample. *BMC Cancer*. 14(1).

- Burges, A. and Schmalfeldt, B. (2011). Ovarian Cancer: Diagnosis and Treatment. *Deutsches Ärzteblatt International*. 108(38):635–641.
- Busch, K. and Busch, M. (2018). Light Polarization and Signal Processing in Chiroptical Instrumentation. *Chiral Analysis*. 73-151.
- Cai, X., Yang, X., Jin, C., Li, L., Cui, Q., Guo, Y., Dong, Y., Yang, X., Guo, L. and Zhang, M. (2018). Identification and verification of differentially expressed microRNAs and their target genes for the diagnosis of esophageal cancer. *Oncology Letters*.
- Calin, G., Dumitru, C., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., Rassenti, L., Kipps, T., Negrini, M., Bullrich, F. and Croce, C. (2002). Nonlinear partial differential equations and applications: Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences*. 99(24):15524-15529.
- Campos-Melo, D., Droppelmann, C., Volkening, K. and Strong, M. (2014). Comprehensive Luciferase-Based Reporter Gene Assay Reveals Previously Masked Up-Regulatory Effects of miRNAs. *International Journal of Molecular Sciences*. 15(9):15592-15602.
- Cancer Association of South Africa (CANSAs) (2016). Fact Sheet on Ovarian Cancer. [online] Available at: <http://www.cansa.org.za/files/2016/12/Fact-Sheet-Ovarian-Cancer-NCR-2011-web-Dec-2016> [Accessed 16 May. 2018].
- Catalanotto, C., Cogoni, C. and Zardo, G. (2016). MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions. *International journal of molecular sciences*. 17(10):1712.
- Chen, L. and Berek, J. (2017). Ovarian cancer diagnosis and staging. [online] Uptodate.com. Available at: <https://www.uptodate.com/contents/ovarian-cancer-diagnosis-and-staging-beyond-the-basics> [Accessed 17 May. 2018].
- Chen, Q., Zobel, J. and Verspoor, K. (2017). Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. *Database*. 2017:p.baw163.
- Chen, V., Ruiz, B., Killeen, J., Cote, T., Wu, X., Correa, C. and Howe, H. (2003). Pathology and classification of ovarian tumors. *Cancer*. 97(10 Suppl):2631-2642.

- Cheng, G. (2015). Circulating miRNAs: Roles in cancer diagnosis, prognosis and therapy. *Advanced Drug Delivery Reviews*. 81:75-93.
- Chornokur, G., Amankwah, E. K., Schildkraut, J. M. and Phelan, C. M. (2013). *Global ovarian cancer health disparities*. *Gynecologic Oncology*. 129(1):258–264.
- Chow, A. Y. (2010) Cell Cycle Control by Oncogenes and Tumor Suppressors: Driving the Transformation of Normal Cells into Cancerous Cells. *Nature Education*. 3(9):7
- Chu, A., Robertson, G., Brooks, D., Mungall, A., Birol, I., Coope, R., Ma, Y., Jones, S. and Marra, M. (2015). Large-scale profiling of microRNAs for The Cancer Genome Atlas. *Nucleic Acids Research*. 44(1):e3-e3.
- CNS reporter (2016). New study reveals three deadliest cancers in South Africa. [online] The Citizen. Available at: <http://citizen.co.za/news/south-africa/1367371/new-study-reveals-three-deadliest-cancers-in-south-africa/> [Accessed 30 May. 2018].
- Cooper GM. The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates; 2000. The Eukaryotic Cell Cycle. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK9876/>
- Coticchia, C. M., Yang, J. and Moses, M. A. (2008). Ovarian Cancer Biomarkers: Current Options and Future Promise. *Journal of the National Comprehensive Cancer Network*. 6(8):795–802.
- Daniilidis, A. and Karagiannis, V. (2007). Epithelial ovarian cancer. Risk factors, screening and the role of prophylactic oophorectomy. *Hippokratia*. 11(2):63–66.
- Davidson, B., Trope, C. and Reich, R. (2014). The Role of the Tumor Stroma in Ovarian Cancer. *Frontiers in Oncology*. 4.
- de Puig, H., Bosch, I., Gehrke, L. and Hamad-Schifferli, K. (2017). Challenges of the Nano-Bio Interface in Lateral Flow and Dipstick Immunoassays. *Trends in Biotechnology*. 35(12):1169–1180.
- Deb, B., Uddin, A. and Chakraborty, S. (2017). miRNAs and ovarian cancer: An overview. *Journal of Cellular Physiology*. 233(5):3846-3854.

- Delfino, K. and Rodriguez-Zas, S. (2013). Transcription Factor-MicroRNA-Target Gene Networks Associated with Ovarian Cancer Survival and Recurrence. *PLoS ONE*. 8(3):e58608.
- Doubeni, C., Doubeni, A. and Myers, A. (2016). Diagnosis and Management of Ovarian Cancer. *American Academy of Family Physicians*. 93(11):937-944.
- Du, B., Wu, D., Yang, X., Wang, T., Shi, X., Lv, Y., Zhou, Z., Liu, Q. and Zhang, W. (2018). The expression and significance of microRNA in different stages of colorectal cancer. *Medicine*. 97(5):e9635.
- Du, L., Qian, X., Dai, C., Wang, L., Huang, D., Wang, S. and Shen, X. (2015). Screening the molecular targets of ovarian cancer based on bioinformatics analysis. *Tumori Journal*. 101(4):384-389.
- Duchartre, Y., Kim, Y. and Kahn, M. (2016). The Wnt signaling pathway in cancer. *Critical Reviews in Oncology/Hematology*. 99:141-149.
- Eser, E., Can, T. and Ferhatosmanoğlu, H. (2014). Div-BLAST: Diversification of Sequence Search Results. *PLoS ONE*. 9(12), p.e115445.
- Fabbri, M. (2014). *Non-coding RNAs and Cancer*. New York, NY: Springer New York.
- Fagotti, A. (2010). Peritoneal carcinosis of ovarian origin. *World Journal of Gastrointestinal Oncology*. 2(2):102.
- Fathalla, M. (1971). Incessant ovulation—a factor in ovarian neoplasia? *The Lancet*. 298(7716):163.
- Fathi, E., Mesbah-Namin, S. and Farahzadi, R. (2013). Biomarkers in Medicine: An Overview. *British Journal of Medicine and Medical Research*. 4(8):1701-1718.
- Feng, Y. (2015). The association between obesity and gynecological cancer. *Gynecology and Minimally Invasive Therapy*. 4(4):102-105.
- Ford, C., Punnia-Moorthy, G., Henry, C., Llamosas, E., Nixdorf, S., Olivier, J., Caduff, R., Ward, R. and Heinzelmann-Schwarz, V. (2014). The non-canonical Wnt ligand, Wnt5a, is

- upregulated and associated with epithelial to mesenchymal transition in epithelial OC. *Gynecologic Oncology*. 134(2):338-345.
- Fouad. Y.A. and Aanei, C. (2017). Revisiting the hallmarks of cancer. *Am J Cancer Res*. 7(5):1016-1036.
- Furey T. S. (2006). Comparison of human (and other) genome browsers. *Human genomics*. 2(4):266–270.
- Garces de los Fayos Alonso, I., Liang, H., Turner, S., Lagger, S., Merkel, O. and Kenner, L. (2018). The Role of Activator Protein-1 (AP-1) Family Members in CD30-Positive Lymphomas. *Cancers*. 10(4):93.
- Giannopoulou, L., Mastoraki, S., Buderath, P., Strati, A., Pavlakis, K., Kasimir-Bauer, S. and Lianidou, E. (2018). ESR1 methylation in primary tumors and paired circulating tumor DNA of patients with high-grade serous OC. *Gynecologic Oncology*. 150(2):355-360.
- Goodman, M., Lurie, G., Thompson, P., McDuffie, K. and Carney, M. (2008). Association of two common single-nucleotide polymorphisms in the CYP19A1 locus and ovarian cancer risk. *Endocrine Related Cancer*. 15(4):1055-1060.
- Goossens, N., Nakagawa, S., Sun, X. and Hoshida, Y. (2015). Cancer biomarker discovery and validation. *Translational Cancer Research*. 4(3):256–269.
- Govindarajan, R., Duraiyan, J., Kaliyappan, K. and Palanisamy, M. (2012). Microarray and its applications. *Journal of Pharmacy & Bioallied Sciences*. 4 (Suppl 2): S310–S312.
- Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*. 34 (Database issue):D140–D144.
- Grimm, D., Bauer, J., Pietsch, J., Infanger, M., Eucker, J., Eilles, C. and Schoenberger, J. (2011). Diagnostic and Therapeutic Use of Membrane Proteins in Cancer Cells. *Current Medicinal Chemistry*. 18(2):176-190.

- Grossmann, C. (2010). *Clinical Data As the Basic Staple of Health Learning: Creating and Protecting a Public Good: Workshop Summary (Learning Health System Series)*. Washington (DC): National Academies Press.
- Guo, C., Liu, S., Wang, J., Sun, M. and Greenaway, F. (2013). ACTB in cancer. *Clinica Chimica Acta*. 417:39-44.
- Hammond, S. (2015). An overview of microRNAs. *Advanced Drug Delivery Reviews*. 87:3-14.
- Han, H., Cortez, C. C., Yang, X., Nichols, P. W., Jones, P. A. and Liang, G. (2011). DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter. *Human molecular genetics*. 20(22):4299–4310.
- Hanahan, D. and Weinberg, R. (2000). The Hallmarks of Cancer. *Cell*. 100:57-70.
- Hanahan, D. and Weinberg, R. (2011). Hallmarks of Cancer: The Next Generation. *Cell*. 144(5):646-674.
- Hein, S., Mahner, S., Kanowski, C., Löning, T., Jänicke, F. and Milde-Langosch, K. (2009). Expression of Jun and Fos proteins in ovarian tumors of different malignant potential and in OC cell lines. *Oncology Reports*. 22:177-183.
- Hejmadi, M. (2010) *Introduction to Cancer Biology*. 2nd ed. Frederiksberg, Denmark: BoonBooks.com, 1-48.
- Henry, N. and Hayes, D. (2012). Cancer biomarkers. *Molecular Oncology*. 6(2):40-146.
- Hennebold, J., Tanaka, M., Saito, J., Hanson, B. and Adashi, E. (2000). Ovary-Selective Genes I: The Generation and Characterization of an Ovary-Selective Complementary Deoxyribonucleic Acid Library\*. *Endocrinology*. 141(8):2725-2734.
- Hess, J. (2004). AP-1 subunits: quarrel and harmony among siblings. *Journal of Cell Science*. 117(25):5965-5973.
- Hu, Z.-Z., Huang, H., Wu, C. H., Jung, M., Ditschilo, A., Riegel, A. T. and Wellstein, A. (2011). Omics-Based Molecular Target and Biomarker Identification. *Methods in Molecular Biology (Clifton, N.J.)*. 719:547–571.

- Huang, Y., Niu, B., Gao, Y., Fu, L. and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*. 26(5):680-682.
- Hudler, P., Kocevar, N. and Komel, R. (2014). Proteomic Approaches in Biomarker Discovery: New Perspectives in Cancer Diagnostics. *The Scientific World Journal*. 2014:1-18.
- Hudson, W. and Ortlund, E. (2014). The structure, function and evolution of proteins that bind DNA and RNA. *Nature Reviews Molecular Cell Biology*. 15(11):749-760.
- Hunn, J. and Rodriguez, G. (2012). Ovarian Cancer. *Clinical Obstetrics and Gynecology*. 55(1):3-23.
- Ilyin, S. E., Belkowski, S. M. and Plata-Salamán, C. R. (2004). Biomarker discovery and validation: Technologies and integrative approaches. *Trends in Biotechnology*. 22(8):411-416.
- Jain, K. (2010). *The handbook of biomarkers*. New York: Springer.
- Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G. and Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Research*. 37(Database issue).
- Jin, H., Won, M., Shin, E., Kim, H., Lee, K. and Bae, J. (2017). EGR2 is a gonadotropin-induced survival factor that controls the expression of IER3 in ovarian granulosa cells. *Biochemical and Biophysical Research Communications*. 482(4):877-882.
- Jin, Y., Chen, Z., Liu, X. and Zhou, X. (2013). Evaluating the microRNA targeting sites by luciferase reporter gene assay. *Methods in Molecular Biology (Clifton, N.J.)*. 936:117–127.
- Kamanu, T., Radovanovic, A., Archer, J. and Bajic, V. (2013). Exploration of miRNA families for hypotheses generation. *Scientific Reports*. 3(1).
- Kampen, K. (2011). Membrane Proteins: The Key Players of a Cancer Cell. *The Journal of Membrane Biology*. 242(2):69-74.
- Karolchik, D. (2003). The UCSC Genome Browser Database. *Nucleic Acids Research*. 31(1):51-54.

- Kastrinos, F. (2009). Risk of Pancreatic Cancer in Families with Lynch Syndrome. *Journal of the American Medical Association*. 302(16):1790.
- Katz, B., Tropé, C., Reich, R. and Davidson, B. (2015). MicroRNAs in Ovarian Cancer. *Human Pathology*. 46(9):1245-1256.
- Kehl, T., Backes, C., Kern, F., Fehlmann, T., Ludwig, N., Meese, E., Lenhof, H. and Keller, A. (2017). About miRNAs, miRNA seeds, target genes and target pathways. *Oncotarget*. 8(63).
- Kerr, J.F.; Wyllie, A.H. and Currie, A.R. (1972) Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *British Journal of Cancer*. 26(4):239-257.
- Khayeka-Wandabwa, C., Ma, X., Cao, X., Nunna, V., Pathak, J., Bernhardt, R., Cai, P. and Bureik, M. (2019). Plasma membrane localization of CYP4Z1 and CYP19A1 and the detection of anti-CYP19A1 autoantibodies in humans. *International Immunopharmacology*. 73:64-71.
- Klasberg, S., Bitard-Feildel, T. and Mallet, L. (2016). Computational Identification of Novel Genes: Current and Future Perspectives. *Bioinformatics and Biology Insights*. 10:BBI.S39950.
- Kozomara, A. and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*. 39(suppl1):D152–D157.
- Kozomara, A., Birgaoanu, M. and Griffiths-Jones, S. (2018). miRBase: from microRNA sequences to function. *Nucleic Acids Research*. 47(D1):D155-D162.
- Kralik, P. and Ricchi, M. (2017). A Basic Guide to Real Time PCR in Microbial Diagnostics: Definitions, Parameters, and Everything. *Frontiers in Microbiology*. 8.
- Krithikadatta J. (2014). Normal distribution. *Journal of Conservative Dentistry: JCD*. 17(1):96–97.
- Lacey, Jr, J. (2002). Menopausal Hormone Replacement Therapy and Risk of Ovarian Cancer. *Journal of the American Medical Association*. 288(3):334.
- Lan, H., Lu, H., Wang, X. and Jin, H. (2015). MicroRNAs as Potential Biomarkers in Cancer: Opportunities and Challenges. *BioMed Research International*. 2015:125094.

- Larrea, E., Sole, C., Manterola, L., Goicoechea, I., Armesto, M., Arestin, M., Caffarel, M., Araujo, A., Araiz, M., Fernandez-Mercado, M. and Lawrie, C. (2016). New Concepts in Cancer Biomarkers: Circulating miRNAs in Liquid Biopsies. *International Journal of Molecular Sciences*. 17(5):627.
- Leber, M.F. and Efferth, T. (2009) Molecular principles of cancer invasion and metastasis (Review). *International Journal of Oncology*. 34(4):881-895
- Leitzmann, M. F., Koebnick, C., Danforth, K. N., Brinton, L. A., Moore, S. C., Hollenbeck, A. R., Schatzkin, A. and Lacey, J. V. (2009). Body mass index and risk of ovarian cancer. *Cancer*. 115(4): 812–822.
- Lengyel, E. (2010) Ovarian Cancer Development and Metastasis. *The American Journal of Pathology*. 177(3):1053-1064
- Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 22(13):1658–1659.
- Li, W., Jaroszewski, L. and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*. 17(3):282-283.
- Lim, D. and Maher, E. (2010). DNA methylation: a form of epigenetic control of gene expression. *The Obstetrician and Gynaecologist*. 12(1):37-42.
- Liu Q. (2017). TMBIM-mediated Ca<sup>2+</sup> homeostasis and cell death. *Biochimica et biophysica acta. Molecular Cell Research*. 1864(6):850–857.
- Liu, B., Li, J. and Cairns, M. J. (2014). Identifying miRNAs, targets and functions. *Briefings in Bioinformatics*. 15(1), 1–19.
- Liu, J., Lichtenberg, T., Hoadley, K., Poisson, L., Lazar, A., Cherniack, A., Kovatich, A., Benz, C., Levine, D., Lee, A., Omberg, L., Wolf, D., Shriver, C. and Thorsson, V. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*. 173(2):400-416.e11.

- Liu, Y., Asch, H. and Kulesz-Martin, M.F. (2001). Functional Quantification of DNA-binding Proteins p53 and Estrogen Receptor in Cells and Tumor Tissues by DNA Affinity Immunoblotting. *American Association for Cancer Research*. 61(14):5402-5406.
- Lu, K. and Daniels, M. (2013). Endometrial and ovarian cancer in women with Lynch syndrome: update in screening and prevention. *Familial Cancer*. 12(2):273-277.
- Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W. and Cui, Q. (2008). An Analysis of Human MicroRNA and Disease Associations. *PLoS ONE*. 3(10):p.e3420.
- Ludwig, J. and Weinstein, J. (2005). Biomarkers in Cancer Staging, Prognosis and Treatment Selection. *Nature Reviews Cancer*. 5(11):845-856.
- Luscombe, N., Greenbaum, D. and Gerstein, M. (2001). What is Bioinformatics? A Proposed Definition and Overview of the Field. *Methods of Information in Medicine*. 40(04):346-358.
- Mäbert, K., Cojoc, M., Peitzsch, C., Kurth, I., Souchelnytskyi, S. and Dubrovskaya, A. (2014). Cancer biomarker discovery: Current status and future perspectives. *International Journal of Radiation Biology*. 90(8):659-677.
- MacFarlane, L. and R. Murphy, P. (2010). MicroRNA: Biogenesis, Function and Role in Cancer. *Current Genomics*. 11(7):537-561.
- Majewska, M., Wysokińska, H., Kuźma, Ł. and Szymczyk, P. (2018). Eukaryotic and prokaryotic promoter databases as valuable tools in exploring the regulation of gene transcription: a comprehensive overview. *Gene*. 644:38-48.
- Makondi, P., Wei, P., Huang, C. and Chang, Y. (2019). Development of novel predictive miRNA/target gene pathways for colorectal cancer distant metastasis to the liver using a bioinformatic approach. *PLOS ONE*. 14(2):e0211968.
- Maldonado, R., Filarsky, M., Grummt, I. and Längst, G. (2017). Purine- and pyrimidine-triple-helix-forming oligonucleotides recognize qualitatively different target sites at the ribosomal DNA locus. *RNA*, 24(3):371-380.
- Maltenfort, M. (2015). Understanding a Normal Distribution of Data. *Journal of Spinal Disorders and Techniques*. 28(10):377-378.

- Manconi, A., Moscatelli, M., Armano, G., Gnocchi, M., Orro, A. and Milanesi, L. (2016). Removing duplicate reads using graphics processing units. *BMC Bioinformatics*. 17(Suppl 12):59–71.
- Marco, A., Ninova, M., Ronshaugen, M. and Griffiths-Jones, S. (2013). Clusters of microRNAs emerge by new hairpins in existing transcripts. *Nucleic Acids Research*. 41(16):7745-7752.
- Marmolejo-Ramos, F. and Siva Tian, T. (2010). The shifting boxplot. A boxplot based on essential summary statistics around the mean. *International Journal of Psychological Research*. 3(1):37.
- Martin, G. (2003). Cell signalling and cancer. *Cancer Cell*. 4(3):167-174.
- Mazziotta, J. (2002). Time and Space. *Brain Mapping: The Methods*. 33-46.
- McGinnis, S. and Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*. 32 (Web Server issue):W20–W25.
- McLemore, M., Miaskowski, C., Aouizerat, B., Chen, L. and Dodd, M. (2009). Epidemiological and Genetic Factors Associated With Ovarian Cancer. *Cancer Nursing*. 32(4):281-288.
- Mishra, P. (2016). Application & Advantages of Bioinformatics. [online] prezi.com. Available at: <https://prezi.com/86tn7qvwe7vo/application-advantages-of-bioinformatics/> [Accessed 5 Jul. 2018].
- Moore, R., MacLaughlan, S. and Bast, R. (2010). Current state of biomarker development for clinical application in epithelial ovarian cancer. *Gynecologic Oncology*. 116(2):240-245.
- Mounir, M., Lucchetta, M., Silva, T., Olsen, C., Bontempi, G., Chen, X., Noushmehr, H., Colaprico, A. and Papaleo, E. (2019). New functionalities in the TCGAblinks package for the study and integration of cancer data from GDC and GTEx. *PLOS Computational Biology*. 15(3), p.e1006701.
- Mousavi, S., Safaralizadeh, R., Hosseinpour-Feizi, M., Azimzadeh-Isfanjani, A. and Hashemzadeh, S. (2018). Study of *cofilin 1* gene expression in colorectal cancer. *Journal of Gastrointestinal Oncology*. 9(5):791–796.

- Mungenast, F. and Thalhammer, T. (2014). Estrogen Biosynthesis and Action in Ovarian Cancer. *Frontiers in Endocrinology*. 5.
- Nakamura, K., Sawada, K., Yoshimura, A., Kinose, Y., Nakatsuka, E. and Kimura, T. (2016). Clinical relevance of circulating cell-free microRNAs in ovarian cancer. *Molecular Cancer*. 15(1).
- National Cancer Institute (no date). Cancer of the Ovary - Cancer Stat Facts. [online] Seer.cancer.gov. Available at: <https://seer.cancer.gov/statfacts/html/ovary.html> [Accessed 30 May. 2018].
- National Institutes of Health (US); Biological Sciences Curriculum Study. NIH Curriculum Supplement Series [Internet]. Bethesda (MD): National Institutes of Health (US); 2007. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK20362/>
- Neal, C. and Berry, D. (2006) Basic principles of the molecular biology of cancer II: angiogenesis, invasion and metastasis. *Surgery (Oxford)*. 24(4):120-125
- Negm, R., Verma, M. and Srivastava, S. (2002). The promise of biomarkers in cancer screening and detection. *Trends in Molecular Medicine*. 8(6):288-293.
- Negri, E., Pelucchi, C., Franceschi, S., Montella, M., Conti, E., Dal Maso, L., Parazzini, F., Tavani, A., Carbone, A. and La Vecchia, C. (2002). Family history of cancer and risk of ovarian cancer. *European Journal of Cancer*. 39:505–510.
- O'Brien, J., Hayder, H., Zayed, Y. and Peng, C. (2018). Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Frontiers in Endocrinology*. 9:402
- Ohyagi-Hara, C., Sawada, K., Kamiura, S., Tomita, Y., Isobe, A., Hashimoto, K., Kinose, Y., Mabuchi, S., Hisamatsu, T., Takahashi, T., Kumasawa, K., Nagata, S., Morishige, K., Lengyel, E., Kurachi, H. and Kimura, T. (2013). miR-92a Inhibits Peritoneal Dissemination of Ovarian Cancer Cells by Inhibiting Integrin  $\alpha 5$  Expression. *The American Journal of Pathology*. 182(5):1876-1889.
- Olsen, C., Nagle, C., Whiteman, D., Ness, R., Pearce, C., Pike, M., Rossing, M., Terry, K., Wu, A., Risch, H., Yu, H., Doherty, J., Chang-Claude, J., Hein, R., Nickels, S., Wang-Gohrke, S.,

- Goodman, M., Carney, M., Matsuno, R., Lurie, G., Moysich, K., Kjaer, S., Jensen, A., Hogdall, E., Goode, E., Fridley, B., Vierkant, R., Larson, M., Schildkraut, J., Hoyo, C., Moorman, P., Weber, R., Cramer, D., Vitonis, A., Bandera, E., Olson, S., Rodriguez-Rodriguez, L., King, M., Brinton, L., Yang, H., Garcia-Closas, M., Lissowska, J., Anton-Culver, H., Ziogas, A., Gayther, S., Ramus, S., Menon, U., Gentry-Maharaj, A. and Webb, P. (2013). Obesity and risk of ovarian cancer subtypes: evidence from the Ovarian Cancer Association Consortium. *Endocrine Related Cancer*. 20(2):251-262.
- Origoni, m., Bergamini, A., Pella, F., Ottolina, J., Giorgione, V., Del Prato, D., Almirante, G. and Candiani, M. (2013). Small Cell Carcinoma of the Ovary: Report of Three Cases of a Poor Prognosis Disease. *Journal of Medical Cases*. 4(3):189-192.
- Pan, J. B., Hu, S. C., Shi, D., Cai, M. C., Li, Y. B., Zou, Q. and Ji, Z. L. (2013). PaGenBase: a pattern gene database for the global and dynamic understanding of gene function. *PLoS ONE*. 8(12):e80747.
- Paranjape, T., Slack, F. J. and Weidhaas, J. B. (2009). MicroRNAs: tools for cancer diagnostics. *Gut*. 58(11):1546–1554.
- Paugh, S., Coss, D., Bao, J., Laudermilk, L., Grace, C., Ferreira, A., Waddell, M., Ridout, G., Naeve, D., Leuze, M., LoCascio, P., Panetta, J., Wilkinson, M., Pui, C., Naeve, C., Uberbacher, E., Bonten, E. and Evans, W. (2016). MicroRNAs Form Triplexes with Double Stranded DNA at Sequence-Specific Binding Sites; a Eukaryotic Mechanism via which microRNAs Could Directly Alter Gene Expression. *PLOS Computational Biology*. 12(2):e1004744.
- Pearson, W. (2013). An Introduction to Sequence Similarity (“Homology”) Searching. *Current Protocols in Bioinformatics*. 42(1):3.1.1-3.1.8.
- Peled, S., Leiderman, O., Charar, R., Efroni, G., Shav-Tal, Y. and Ofran, Y. (2016). De-novo protein function prediction using DNA binding and RNA binding proteins as a test case. *Nature Communications*. 7(1).
- Peng, Y., Zhang, X., Feng, X., Fan, X. and Jin, Z. (2016). The crosstalk between microRNAs and the Wnt/ $\beta$ -catenin signaling pathway in cancer. *Oncotarget*. 8(8).

- Pereira, B., Billaud, M. and Almeida, R. (2017). RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends in Cancer*. 3(7):506-528.
- Peter, M. (2010). Targeting of mRNAs by multiple miRNAs: the next step. *Oncogene*. 29(15):2161-2164.
- Pruthi, S., Gostout, B. and Lindor, N. (2010). Identification and Management of Women With BRCA Mutations or Hereditary Predisposition for Breast and Ovarian Cancer. *Mayo Clinic Proceedings*. 85(12):1111-1120.
- Qi, J. and Ronai, Z. A. (2015). Dysregulation of ubiquitin ligases in cancer. *Drug resistance updates: reviews and commentaries in antimicrobial and anticancer chemotherapy*. 23:1–11.
- Rastogi, M., Gupta, S. and Sachan, M. (2016). Biomarkers towards Ovarian Cancer Diagnostics: Present and Future Prospects. *Brazilian Archives of Biology and Technology*. 59(0).
- Rauh-Hain, J.A., Krivak, T.C., del Carmen, M.G. and Olawaite, A.B. (2011) Ovarian Cancer Screening and Early Detection in the General Population. *Reviews in Obstetrics & Gynecology*. 4(1):15-21
- Raza, K. (2012). Application of data mining in bioinformatics. *Indian Journal of Computer Science and Engineering*. 1(2): 114-118.
- Razi, S., Ghoncheh, M., Mohammadian-Hafshejani, A., Aziznejhad, H., Mohammadian, M. and Salehiniya, H. (2016). The incidence and mortality of ovarian cancer and their relationship with the Human Development Index in Asia. *Ecancermedicalscience*. 10:628.
- Redaniel, M., Laudico, A., Mirasol-Lumague, M., Gondos, A., Pulte, D., Mapua, C. and Brenner, H. (2009). Cancer survival discrepancies in developed and developing countries: comparisons between the Philippines and the United States. *British Journal of Cancer*. 100(5):858-862.
- Richards, J. and Pangas, S. (2010). The ovary: basic biology and clinical implications. *Journal of Clinical Investigation*. 120(4):963-972.

- Riman, T. (2002). Hormone Replacement Therapy and the Risk of Invasive Epithelial Ovarian Cancer in Swedish Women. *CancerSpectrum Knowledge Environment*, 94(7):497-504.
- Robertson, D. C. (2005). Diagnosing Cancer Earlier With Blood Markers. *Biotechnology Healthcare*. 2(1):14–16.
- Rolle, K., Piwecka, M., Belter, A., Wawrzyniak, D., Jeleniewicz, J., Barciszewska, M. and Barciszewski, J. (2016). The Sequence and Structure Determine the Function of Mature Human miRNAs. *PLoS ONE*. 11(3):e0151246.
- Rostgaard, K., Wohlfahrt, J., Andersen, P., Hjalgrim, H., Frisch, M., Westergaard, T. and Melbye, M. (2003). Does Pregnancy Induce the Shedding of Premalignant Ovarian Cells? *Epidemiology*. 14(2):168-173.
- Russo, A., Calo, V., Bruno, L., Rizzo, S., Bazan, V. and Di Fede, G. (2009) Hereditary ovarian cancer. *Critical Reviews in Oncology/Hematology*. 69(1):28-44
- Russo, F., Di Bella, S., Nigita, G., Macca, V., Laganà, A., Giugno, R., Pulvirenti, A. and Ferro, A. (2012). miRandola: extracellular circulating microRNAs database. *EMBnet.journal*. 18(A):135.
- Sahdev, A. (2016). CT in ovarian cancer staging: how to review and report with emphasis on abdominal and pelvic disease for surgical planning. *Cancer Imaging*. 16(1):19.
- Salehi, F., Dunfield, L., Phillips, K., Krewski, D. and Vanderhyden, B. (2008). Risk Factors for Ovarian Cancer: An Overview with Emphasis on Hormonal Factors. *Journal of Toxicology and Environmental Health, Part B*. 11(3-4):301-321.
- Sathya, R. and Abraham, A. (2013). Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *International Journal of Advanced Research in Artificial Intelligence*. 2(2):34-38.
- Schlattmann, P. and Dirnagl, U. (2010). Statistics in experimental cerebrovascular research-comparison of two groups with a continuous outcome variable. *Journal of cerebral blood flow and metabolism: official journal of the International Society of Cerebral Blood Flow and Metabolism*. 30(3):474–479.

- Sever, R. and Brugge, J. S. (2015). Signal transduction in cancer. *Cold Spring Harbor perspectives in medicine*. 5(4):a006098.
- Shaaban, A., Rezvani, M., Elsayes, K., Baskin, H., Mourad, A., Foster, B., Jarboe, E. and Menias, C. (2014). Ovarian Malignant Germ Cell Tumors: Cellular Classification and Clinical and Imaging Features. *RadioGraphics*. 34(3):77-801.
- Smith, T. and Guidozi, F. (2009). Epithelial ovarian cancer in Southern Africa. *Southern African Journal of Gynaecological Oncology*. 1(1):23-27.
- South African Medical Research Council (no date). FAQ - cancer in SA. [online] Mrc.ac.za. Available at: <http://www.mrc.ac.za/bod/faqcancer.htm> [Accessed 30 May. 2018].
- Spudich, G. and Fernández-Suárez, X. (2010). Touring Ensembl: A practical guide to genome browsing. *BMC Genomics*. 11(1):295.
- Spudich, G., Fernandez-Suarez, X. and Birney, E. (2007). Genome browsing with Ensembl: a practical overview. *Briefings in Functional Genomics and Proteomics*. 6(3):202-219.
- Stephen, S., Sarojini, S. and Milinovic, N. (2013). Ovarian Cancer Biomarkers: Current Trends in Translational Research for Early Detection. *Translational Medicine*. 3(1).
- Stępień, E., Costa, M. and Enguita, F. (2018). miRNAtools: Advanced Training Using the miRNA Web of Knowledge. *Non-Coding RNA*. 4(1):5.
- Sundar, S., Neal, R. and Kehoe, S. (2015). Diagnosis of ovarian cancer. *British Medical Journal*. 351.
- Svoronos, A. A., Engelman, D. M. and Slack, F. J. (2016). OncomiR or Tumor Suppressor? The Duplicity of MicroRNAs in Cancer. *Cancer Research*. 76(13):3666–3670.
- Thébault, P., Bourqui, R., Benchimol, W., Gaspin, C., Sirand-Pugnet, P., Uricaru, R. and Dutour, I. (2015). Advantages of mixing bioinformatics and visualization approaches for analyzing sRNA-mediated regulatory bacterial networks. *Briefings in Bioinformatics*. 16(5):795–805.
- Thind, A. and Wilson, C. (2016). Exosomal miRNAs as cancer biomarkers and therapeutic targets. *Journal of Extracellular Vesicles*. 5(1):31292.

- Tian, F., Shen, Y., Chen, Z., Li, R. and Ge, Q. (2017). No Significant Difference between Plasma miRNAs and Plasma-Derived Exosomal miRNAs from Healthy People. *BioMed Research International*. 2017:1-5.
- Tokar, T., Pastrello, C., Rossos, A., Abovsky, M., Hauschild, A., Tsay, M., Lu, R. and Jurisica, I. (2017). mirDIP 4.1—integrative database of human microRNA target predictions. *Nucleic Acids Research*. 46(D1):D360-D370.
- Tollefsbol, T. (2011). *Handbook of epigenetics*. Amsterdam: Elsevier.
- Tomczak, K., Czerwińska, P. and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology (Poznan, Poland)*. 19(1A):A68–A77.
- Torpy, J. (2011) Ovarian Cancer. *Journal of the American Medical Association*. 305(23):2484
- Toss, A., Tomasello, C., Razzaboni, E., Contu, G., Grandi, G., Cagnacci, A., Schilder, R. and Cortesi, L. (2015). Hereditary Ovarian Cancer: Not Only BRCA1 and 2 Genes. *BioMed Research International*. 2015:1-11.
- Ueland F. R. (2017). A Perspective on Ovarian Cancer Biomarkers: Past, Present and Yet-To-Come. *Diagnostics (Basel, Switzerland)*. 7(1):14.
- Valinezhad Orang, A., Safaralizadeh, R. and Kazemzadeh-Bavili, M. (2014). Mechanisms of miRNA-Mediated Gene Regulation from Common Downregulation to mRNA-Specific Upregulation. *International Journal of Genomics*. 2014:1-15.
- Van Nagell, J. R. and Hoff, J. T. (2014). Transvaginal ultrasonography in ovarian cancer screening: current perspectives. *International Journal of Women's Health*. 6:25–33.
- Villicaña, C., Cruz, G. and Zurita, M. (2014). The basal transcription machinery as a target for cancer therapy. *Cancer cell international*. 14(1):18.
- Wang, J., Chen, J. and Sen, S. (2015). MicroRNA as Biomarkers and Diagnostics. *Journal of Cellular Physiology*. 231(1):25-30.
- Wang, J., Kong, L., Gao, G. and Luo, J. (2012). A brief introduction to web-based genome browsers. *Briefings in Bioinformatics*. 14(2):131-143.

- Wang, J., Zuo, Y., Man, Y., Avital, I., Stojadinovic, A., Liu, M., Yang, X., Varghese, R., Tadesse, M. and Resson, H. (2015). Pathway and Network Approaches for Identification of Cancer Signature Markers from Omics Data. *Journal of Cancer*. 6(1):54-65.
- Wang, K., Li, L., Fu, L., Yuan, Y., Dai, H., Zhu, T., Zhou, Y. and Yuan, F. (2019). Integrated Bioinformatics Analysis the Function of RNA Binding Proteins (RBPs) and Their Prognostic Value in Breast Cancer. *Frontiers in Pharmacology*. 10.
- Wang, L., Li, X., Mu, Y., Lu, C., Tang, S., Lu, K., Qiu, X., Wei, A., Cheng, Y. and Wei, W. (2019). The iron chelator desferrioxamine synergizes with chemotherapy for cancer treatment. *Journal of Trace Elements in Medicine and Biology*.
- Wang, M. and Kong, L. (2019). pblat: a multithread blat algorithm speeding up aligning sequences to genomes. *BMC Bioinformatics*. 20(1).
- Wang, S., Ke, H., Zhang, H., Ma, Y., Ao, L., Zou, L., Yang, Q., Zhu, H., Nie, J., Wu, C. and Jiao, B. (2018). LncRNA MIR100HG promotes cell proliferation in triple-negative breast cancer through triplex formation with p27 loci. *Cell Death & Disease*. 9(8).
- Wang, Z., Jensen, M. and Zenklusen, J. (2016). A Practical Guide to The Cancer Genome Atlas (TCGA). *Methods in Molecular Biology*. 1418:111-141.
- Weinberg, R.A. (2014) *The Biology of Cancer*. 2nd ed. New York: Garland Science, 45-50
- Wheeler, D. and Bhagwat, M. (2007). *BLAST QuickStart: example-driven web-based BLAST Tutorial*. Totowa (NJ): Humana Press.
- Whittemore, A., Harris, R. and Itnyre, J. (1992) Characteristics relating to ovarian cancer risk: Collaborative Analysis of 12US case-control studies. *American Journal of Epidemiology*. 136(10):1212-1220
- Williams, G.H. and Stoeber, K. (2011) The cell cycle and cancer. *Journal of Pathology*. 226: 352-364
- Winter, J., Jung, S., Keller, S., Gregory, R. and Diederichs, S. (2009). Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nature Cell Biology*. 11(3):228-234.

- Wong, N. and Wang, X. (2014). miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Research*. 43(D1):D146-D152.
- Wong, R. (2011) Apoptosis in cancer: from pathogenesis to treatment. *Journal of Experimental & Clinical Cancer Research*. 30(1):87
- Wu, D., Rice, C. and Wang, X. (2012). Cancer bioinformatics: A new approach to systems clinical medicine. *BMC Bioinformatics*. 13(1):71.
- Wu, T., Li, Y., Liu, B., Zhang, S., Wu, L., Zhu, X. and Chen, Q. (2016). Expression of Ferritin Light Chain (FTL) Is Elevated in Glioblastoma, and FTL Silencing Inhibits Glioblastoma Cell Proliferation via the GADD45/JNK Pathway. *PLoS ONE*. 11(2):e0149361.
- Xie, B., Ding, Q., Han, H. and Wu, D. (2013). miRCancer: a microRNA-cancer association database constructed by text mining on literature. *Bioinformatics*. 29(5):638-644.
- Xue, J., Yang, G., Ding, H., Wang, P. and Wang, C. (2015). Role of NSC319726 in ovarian cancer based on the bioinformatics analyses. *Oncotargets and Therapy*. 8:3757–3765.
- Yang, Z., Ren, F., Liu, C., He, S., Sun, G., Gao, Q., Yao, L., Zhang, Y., Miao, R., Cao, Y., Zhao, Y., Zhong, Y. and Zhao, H. (2010). dbDEMC: a database of differentially expressed miRNAs in human cancers. *BMC Genomics*. 11(Suppl 4):S5.
- Yang, Z., Wu, L., Wang, A., Tang, W., Zhao, Y., Zhao, H. and Teschendorff, A. (2016). dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. *Nucleic Acids Research*. 45(D1):D812-D818.
- Zaki, M., Karypis, G. and Yang, J. (2007). Data Mining in Bioinformatics (BIOKDD). *Algorithms for Molecular Biology*. 2(1).
- Zhang, B., Cai, F. F. and Zhong, X. Y. (2011) An overview for biomarkers for the ovarian cancer diagnosis. *European Journal of Obstetrics & Gynecology and Reproductive Biology*. 152(2):119-123
- Zhang, J., Li, S., Li, L., Li, M., Guo, C., Yao, J. and Mi, S. (2015). Exosome and Exosomal MicroRNA: Trafficking, Sorting, and Function. *Genomics, Proteomics & Bioinformatics*. 13(1):17-24.

Zhang, Y., Sui, J., Shen, X., Li, C., Yao, W., Hong, W., Peng, H., Pu, Y., Yin, L. and Liang, G. (2017). Differential expression profiles of microRNAs as potential biomarkers for the early diagnosis of lung cancer. *Oncology Reports*. 37(6):3543-3553.

Zhong, Y., Wang, Y., Huang, J., Xu, X., Pan, W., Gao, S., Zhang, Y. and Su, M. (2019). Association of hCG and LHCGR expression patterns with clinicopathological parameters in ovarian cancer. *Pathology - Research and Practice*. 215(4):748-754.

Zhou, J., Wang, Y., Fei, J. and Zhang, W. (2012). Expression of cofilin 1 is positively correlated with the differentiation of human epithelial ovarian cancer. *Oncology letters*. 4(6):1187–1190.

