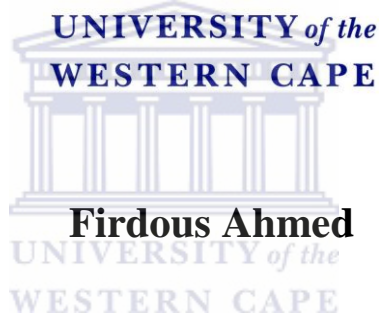


**Identification of potential biomarkers in lung cancer as
possible diagnostic agents using bioinformatics and
molecular approaches**



Thesis presented in the fulfilment of the requirements for the
Magister Scientae

Department of Biotechnology, University of the Western Cape

Supervisor: Dr. Ashley Pretorius

Co-supervisors: Dr. Kareemah Gamieldien

Dr. Junaid Gamieldien

DECLARATION

I declare that the work presented here, *Identification of potential biomarkers in lung cancer as possible diagnostic agents using bioinformatics and molecular approaches* 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.

Firdous Ahmed

December 2015

.....

Signature



ACKNOWLEDGEMENTS

I would like to thank my Supervisor Dr. Ashley Pretorius for his time and support in allowing me this opportunity.

I would like to express my sincere gratitude to my co-Supervisors, Dr. Kareemah Gamieldien and Dr. Junaid Gamieldien, for their guidance and support throughout my M.Sc. My sincere thanks also go to Dr. Junaid Gamieldien for his time and vast knowledge of bioinformatics, and to Dr. Kareemah Gamieldien for her continuous patience and motivation.

To my parents, for their unwavering support, love and sacrifices. None of this would have been possible without them.

Thank you to my friends for supporting me throughout the course of my Masters. A special thank you, to Junaid Rawoot, for believing in me, even when I doubted myself, and for being my support, through the darkest of days.

Last but not least, I would like to thank NRF, for funding this research project and making it a reality.

ABSTRACT

Lung cancer remains the leading cause of cancer deaths worldwide, with the majority of cases attributed to non-small cell lung carcinomas. At the time of diagnosis, a large percentage of patients present with advanced stage of disease, ultimately resulting in a poor prognosis. The identification circulatory markers, overexpressed by the tumour tissue, could facilitate the discovery of an early, specific, non-invasive diagnostic tool as well as improving prognosis and treatment protocols. The aim was to analyse gene expression data from both microarray and RNA sequencing platforms, using bioinformatics and statistical analysis tools. Enrichment analysis sought to identify genes, which were differentially expressed ($p < 0.05$, $FC > 2$) and had the potential to be secreted into the extracellular circulation, by using Gene Ontology terms of the Cellular Component. Results identified 1 657 statically significant genes between normal and early lung cancer tissue, with only 1 gene differentially expressed (DE) between the early and late stage disease. Following statistical analysis, 171 DE genes selected as potential early stage biomarkers. The overall sensitivity of RNAseq, in comparison to arrays enabled the identification of 57 potential serum markers. These genes of interest were all downregulated in the tumour tissue, and while they did not facilitate the discovery of an ideal diagnostic marker based on the set criteria in this study, their roles in disease initiation and progression require further analysis.

Key Words: lung cancer, early diagnosis, bioinformatics, gene enrichment analysis, microarray, RNAseq

TABLE OF CONTENTS

DECLARATION	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
TABLE OF CONTENTS	v
LIST OF ABBREVIATIONS	xii
LIST OF FIGURES	xviii
LIST OF TABLES	xxii
LIST OF APPENDICES	xxiv
Chapter 1: Literature Review	
1.1. Cancer Overview	1
1.2. Carcinogenesis	2
1.2.1. Sustaining Proliferative Signaling	4
1.2.2. Evading Growth Suppressors	4
1.2.3. Resisting Cell Death	5
1.2.4. Enabling Replicative Immortality	6



1.2.5. Inducing Angiogenesis	7
1.2.6. Activating Invasion and Metastasis	8
1.3. Lung Cancer	8
1.4. Classification of Lung Cancer	9
1.5. Genetic Alterations in Lung Carcinomas	11
1.6. Epigenetics and Lung Cancer	13
1.7. Causes and Risk Factors Associated with Lung Cancer	15
1.8. Staging and Grading	15
1.9. Diagnosis of Lung Cancer	17
1.10. Treatment and Prognosis	18
1.11. The Burden of Disease of Lung Cancer	20
1.12. Biomarker Application in Cancer	22
1.13. Lung Cancer Biomarkers	23
1.14. Protein Biomarkers	24
1.15. Gene Biomarkers	25
1.16. Sources of Biomarkers	27




1.17. Applications of Bioinformatics into Biomarker Discovery	29
1.18. Biomarker Validation	31
1.19. Aims and Objectives	33
1.20. References	34
Chapter 2: Identification of Potential Circulatory Biomarkers using Microarray	
Data	
2.1. Introduction	43
2.2. Data Mining	43
2.2.1. Microarray Data Mining	44
2.2.2. Digital Expression Profiling using EST and SAGE	46
2.3. Biological Databases	47
2.3.1. Oncomine	47
2.3.2. Gene Expression Atlas	48
2.3.3. Intergrative OncoGenomics	48
2.3.4. C-It	48
2.3.5. Tissue-specific Gene Expression and Regulation	49



2.3.6. VeryGene	49
2.3.7. Database for Annotation, Visualisation and Integrated Discovery	50
2.4. Text Mining	50
2.4.1. Text Mining Databases	51
2.4.1.1. The Universal Protein Knowledgebase	51
2.4.1.2. PolySearch	51
2.4.1.3. Human Genome Epidemiology Network	52
2.4.1.4. Google Scholar	52
2.5. Materials and Methods	54
2.5.1. Extraction of Candidate Gene Biomarkers	54
2.5.1.1. Oncomine Database	55
2.5.1.2. Gene Expression Atlas Database	56
2.5.1.3. IntOGen Database	56
2.5.1.4. C-It Database	56
2.5.1.5. TiGER Database	57
2.5.1.6. VeryGene Database	57



2.5.1.7. Excluded Databases	57
2.5.2. Analysis of Gene Lists	57
2.5.2.1. Functional Characterisation using DAVID	57
2.5.3. Literature Mining of Candidate Entities	58
2.6. Results and Discussion	59
2.6.1. Identification of Eligible Cancer Biomarkers	59
2.6.2. Gene Enrichment Analysis	62
2.6.3. Literature Mining of Candidate Genes	66
2.7. Discussion and Conclusion	67
2.8. References	72
 UNIVERSITY of the WESTERN CAPE	
Chapter 3: Identification of Potential Circulatory Biomarkers using RNAseq	
Data	
3.1. Introduction	79
3.2. Next Generation Sequencing	79
3.2.1. RNA Sequencing	80
3.2.1.1. RNAseq Version 2	82

3.3. The Cancer Genome Atlas	83
3.4. Bioinformatics Analysis Tools	84
3.4.1. Bioinformatics Enrichment Tools	84
3.4.1.1. MultiExperiment Viewer	86
3.4.1.2. Enrichr	87
3.4.2. Databases and Platforms	88
3.4.2.1.1. Molecular Signatures Database	88
3.4.2.2. Gene Expression Atlas	89
3.5. Materials and Methods	90
3.5.1. Data Retrieval from TCGA	90
3.5.2. Analysis of Data using Bioinformatics Tools	91
3.5.2.1. Analysis using MultiExperiment Viewer	91
3.5.2.2. Enrichment Analysis using Enrichr	93
3.5.2.2.1. Ontology Annotation Sources	93
3.5.3. Gene Expression Analysis	94
3.5.3.1. Molecular Signatures Database	94



3.5.3.2. Gene Expression Atlas	95
3.6. Results and Discussion	95
3.6.1. Data Collection from TCGA	95
3.6.2. Statistical Analysis using MultiExperiment Viewer	96
3.6.3. Enrichment Analysis using Enrichr Feature and Annotation Tool	100
3.6.4. Expression Analysis	112
3.7. Discussion and Conclusion	114
3.8. References	117
Chapter 4: Future Perspectives	129
4.1. References	132
Appendices	135



LIST OF ABBREVIATIONS

IARC	:	International Agency for Research on Cancer
WHO	:	World Health Organization
RB	:	Retinoblastoma
P53	:	Tumour protein P53
G ₀	:	Rest phase of cell cycle
DNA	:	Deoxyribonucleic acid
ECM	:	Extracellular matrix
NSCLC	:	Non-small cell lung carcinomas
SCLC	:	Small-cell lung cancer
EGFR	:	Epidermal growth factor receptor
KRAS	:	Kirsten rat sarcoma viral oncogene homolog
p16	:	Cyclin-dependent kinase inhibitor 2A
RNA	:	Ribonucleic acid
miRNA	:	Micro ribonucleic acid
CpG	:	Cytosine phosphate guanosine
mRNA	:	Messenger RNA
SNP	:	Single-nucleotide polymorphism
TNM	:	Tumour node metastasis
NCI	:	National Cancer Institute
AJCC	:	American Joint Committee on Cancer
IUCC	:	International Union for Cancer Control

CT	:	Computerized tomography
MRI	:	Magnetic resonance imaging
PET	:	Positron emission tomography
SEER	:	Surveillance, Epidemiology, and End Results
RNase	:	Ribonuclease
<i>H. influenza</i>	:	<i>Haemophilus influenza</i>
cDNA	:	Complementary DNA
EST	:	Expressed Sequence Tags
SAGE	:	Serial Analysis Gene Expression
MPPS	:	Massively parallel signature sequencing
HT	:	High-throughput
MS	:	Mass spectrometry
DE	:	Differentially expressed
RT-PCR	:	Real-time polymerase chain reaction
ELISA	:	Enzyme-linked, immunosorbent assays
GSEA	:	Gene set enrichment analysis
RNAseq	:	RNA sequencing
DEG	:	Differentially expressed genes
KDD	:	Knowledge Discovery in Databases
GO	:	Gene Ontology
GEA	:	Gene Expression Atlas
EBI	:	European Bioinformatics Institute
IntOGen	:	Integrative OncoGenomics

TiGER	:	Tissue-specific Gene Expression and Regulation
TSG	:	tissue-specific gene
CRM	:	cis-regulatory module
KEGG	:	Kyoto Encyclopedia of Genes and Genomes
MGI	:	Mouse Genome Informatics
DAVID	:	Database for Annotation, Visualization and Integrated Discovery
UniProtKB	:	The Universal Protein Knowledgebase
HuGENet	:	Human Genome Epidemiology Network
MeSH	:	Medical Subject Headings
NCBI	:	National Center for Bioinformatics
CC	:	Cellular components
MF	:	Molecular function
BP	:	Biological process
CGAP	:	Cancer Genome Characterization Initiative
COPZ1	:	Coatomer Protein Complex, Subunit Zeta 1
SEC23B	:	<i>S. Cerevisiae</i> Homolog B
SEC24A	:	(<i>S. Cerevisiae</i> Family Member A)
SEC24D	:	(<i>S. Cerevisiae</i> Family Member D)
NGS	:	Next Generation Sequencing
CDS	:	Coding DNA sequence
RNAseq V2	:	RNASeq Version 2

RPKM	:	Reads Per Kilobase of exon model per Million
TCGA	:	The Cancer Genome Atlas
RSEM	:	RNAseq by Expectation-Maximization
NHGRI	:	National Human Genome Research Institute
LUAD	:	Lung adenocarcinoma
LUSC	:	Lung squamous cell carcinoma
SEA	:	Singular enrichment analysis
MEA	:	Modular enrichment analysis
MeV	:	MultiExperiment Viewer
KEA	:	Kinase enrichment analysis
GeneSigDB	:	Gene Signatures Database
MSigDB	:	Molecular Signatures Database
OMIM	:	Online Mendelian Inheritance in Man
N	:	Normal lung tissue
E	:	Early stage lung cancer (Stage I & II)
L	:	Late stage lung cancer (Stage III & IV)
TDMS	:	Tab delimited, Multiple Sample Files
FDC	:	False discovery corrections
FDR	:	False discovery rate
FC	:	Fold change
HCL	:	Hierarchical clustering
CS	:	Combined score
GPRC	:	G-protein coupled receptors

MGI	:	Midrand Graduate Institute
MP	:	Mammalian Phenotype
Ca ²⁺	:	Calcium
GEO	:	Gene Expression Omnibus
PPI	:	Protein-protein interactions
ANGPTL7	:	Angiopoietin-like 7
EDN3	:	Endothelin 3
RETN	:	Resistin
NRG3	:	Neuregulin 3
CMTM2	:	CKLF-like MARVEL transmembrane domain containing 2
CAMP	:	Cathelicidin antimicrobial peptide
FGF10	:	Fibroblast growth factor 10
AGRP	:	Agouti related neuropeptide
ANGPTL5	:	Angiopoietin-like
CMTM5	:	CKLF-like MARVEL transmembrane domain containing 5)
FGF	:	Fibroblast growth factor
IRX1	:	Iroquois homeobox 1
ITLN2	:	Intelectin 2
CD5L	:	CD5 molecule-like
FIGF	:	c-fos induced growth factor
VEGFD	:	Vascular endothelial growth factor D

WNT	:	Wingless type proteins
WNT7A	:	Wingless type protein family member
JNK	:	cJun N-terminal kinase
GRIA1	:	Glutamate receptor, ionotropic, AMPA 1
CHRM1	:	Cholinergic receptor, muscarinic 1
CHRM2	:	Cholinergic receptor, muscarinic 2
HOP92	:	Lung adenocarcinoma cell line
HOP62	:	Lung adenocarcinoma cell line
A549	:	Lung adenocarcinoma cell line
NCI H23	:	Lung adenocarcinoma cell line
EKVX	:	Lung adenocarcinoma cell line
NCI 460	:	Large cell lung carcinoma cell line
NCI H322	:	Unspecified lung carcinoma cell line
NCI H226	:	Squamous cell lung carcinoma cell line
IM	:	Immunofluorescence
2D-PAGE	:	Two- dimensional polyacrylamide gel electrophoresis
SELDI-ToF	:	Surface enhanced laser desorption/ionisation time of flight
ICAT	:	Isotope coded affinity tags
MudPIT	:	Multidimensional protein identification technology

LIST OF FIGURES

Figure 1.1: Cancer as the consequence of combined genetic and epigenetic alterations (Herceg & Hainaut 2007).....	3
Figure 1.2: The six biological hallmarks of cancer (Hanahan & Weinberg 2011).....	3
Figure 1.3: Molecular evolution of lung cancer. Showing the interaction of environmental factors, genetic susceptibility and unknown factors to influence carcinogenesis and resulting in genetic and epigenetic alterations, which influence the process of angiogenesis and metastases (Esteller 2008).....	9
Figure 1.4: Genetic mutations specific to SCLCs and NSCLCs (Esteller 2008).....	13
Figure 1.5: Lung cancer deaths and 5 year median survival rate in relation to (A) clinical stages and (B) pathologic stage (Detterbeck 2009).....	19
Figure 1.6: Estimated new cancer diagnoses and deaths of most common types of cancer in the U.S. in 2014. With lung cancer displaying the 3rd most common cancer type, representing 13.5 % of all new cancer cases (National Cancer Institute (NCI), Surveillance, Epidemiology, and End Results (SEER) 2015).....	21

Figure 1.7: Mortality and incidence rates of lung cancer based on geographical location in females and males (Altintas & Tothill 2013).....	21
Figure 1.8: Biocomputing tools for discovery and validation of biomarkers (Phan et al. 2009).....	32
Figure 2.1: Steps involved in knowledge discovery (Fayyed et al. 1996).....	44
Figure 2.2: Outline of the methodology for biomarker discovery.....	54
Figure 2.3: Functional characterisation of genes in DAVID based on their biological process using GO analysis.....	63
Figure 2.4: Functional characterisation of genes in DAVID based on their cellular component using GO analysis.....	64
Figure 2.5: Functional characterisation of genes in DAVID based on their molecular function using GO analysis.....	65
Figure 3.1: The typical infrastructure of enrichment tools with three distinct layers: backend annotation database, data mining, and result presentation (Huang et al. 2009).....	85
Figure 3.2: Expression matrix displaying rows of genes with high (red) and low (green) expression in relation to samples in each column generated by MeV v4.9 (http://tm4.org/mev.html).....	87

Figure 3.3: Methodological approach for the retrieval and analysis of lung adenocarcinoma RNAseq V2 Level 3 data from the Cancer Genome Atlas.....90

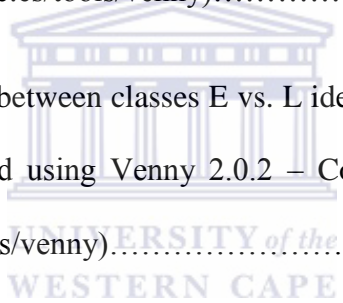
Figure 3.4: Gene expression data of LUAD between classes (N, E and L) generated from TCGA Level 3 RNAseq V2 data depicted using Venny 2.0.2 - Computational Genomics Service (bioinfogp.cnb.csic.es/tools/venny).....96

Figure 3.5: DEG identified in classes E vs. N using MeV parametric t-Tests and multiple FDC (t-Test 1: $p \leq 0.01$, no correction, t-Test 2: $p \leq 0.05$ and Bonferroni, t-Test 3: $p \leq 0.05$ and maxT) depicted using Venny 2.0.2 – Computational Genomics Service (bioinfogp.cnb.csic.es/tools/venny).....99

Figure 3.6: Unique DEG between classes E vs. L identified using MeV parametric t-Test 1, $p \leq 0.01$, depicted using Venny 2.0.2 – Computational Genomics Service (bioinfogp.cnb.csic.es/tools/venny).....99

Figure 3.7: DEG identified between classes N vs. L using parametric t-Tests and multiple FDC (t-Test 1: $p \leq 0.01$, no correction, t-Test 2: $p \leq 0.05$ and Bonferroni, t-Test 3: $p \leq 0.05$ and maxT) in MeV, depicted using Venny 2.0.2 Computational Genomics Service (bioinfogp.cnb.csic.es/tools/venny).....100

Figure 3.8: Unique DEG in classes E vs. N and E vs. L determined to be statistically significant following statistical analysis using MeV. Depicted using Venny 2.0.2 –



Computational	Genomics	Service
(bioinfoqp.cnb.csic.es/tools/venny).....100		
Figure 3.9: Histogram of enriched GO terms of BP generated from annotation analysis in Enrichr.....102		
Figure 3.10: Histogram of enriched GO terms of CC generated from annotation analysis in Enrichr.....103		
Figure 3.11: Histogram of enriched GO terms of MF generated from annotation analysis in Enrichr.....104		



LIST OF TABLES

Table 1.1: Summary of lung tumour types (Herceg & Hainaut 2007; Herbst et al. 2008; Patel et al. 2008).....	11
Table 1.2: Cancer staging based on TNM criteria (Detterbeck 2009).....	16
Table 1.3: Lung cancer protein biomarkers currently available (Sung & Cho 2008; Altintas & Tothill 2013).....	25
Table 1.4: Genes and associated mutation types reported in lung cancer (Sung & Cho 2008; Altintas & Tothill 2013).....	27
Table 2.1: Summary of genes extracted from Oncomine.....	60
Table 2.2: Summary of genes extracted from GEA based on GO terms.....	60
Table 2.3: Summary of genes extracted from databases.....	61
Table 3.1: Statistical parameters implemented to identify DEG between LUAD N, E and L samples using MeV.....	92
Table 3.2: Ontology enrichment terms extracted from Enrichr (p < 0.01, (CS) > 2).....	105

Table 3.3: Pathway enrichment terms extracted from Enrichr ($p < 0.01$, (CS) > 2).....108

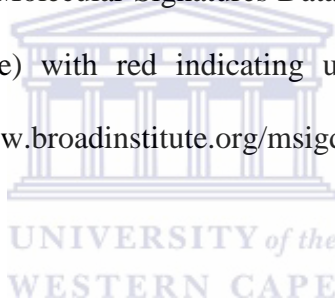
Table 3.4: Candidate genes identified as located in the extracellular cellular component using GO annotations and having the potential to be serum markers.....110

Table 3.5: Genes of interest identified using annotation, statistical analysis and enrichment analysis.....111



LIST OF APPENDICES

- Appendix A:** Average sample means of groups early versus normal lung cancer used to identify genes as downregulated (FC > 2).....137
- Appendix B:** Heat map visulisation of oncogenic signatures of candidate genes of interest generated by the Molecular Signatures Database (MSigDB) NCI-60 cell line (National Cancer Institute) with red indicating upregulation and blue depicting downregulation (<http://www.broadinstitute.org/msigdb>).....142
- Appendix C:** Heat map visulisation of oncogenic signatures of candidate serum markers generated by the Molecular Signatures Database (MSigDB) NCI-60 cell line (National Cancer Institute) with red indicating upregulation and blue depicting downregulation (<http://www.broadinstitute.org/msigdb>).....143



Chapter 1

Literature Review

1.1. Cancer Overview

Cancer arises as the result of abnormal cell growth and can be identified as a hyper-proliferative disorder, characterized by deregulation of apoptosis, increased cell proliferation, cell invasion, angiogenesis as well as metastasis (Cooper & Hausman 2007; Herceg & Hainaut 2007). A tumour, an abnormal mass of cells is defined as either malignant or benign. A benign tumour does not invade surrounding tissue or spread to distant body sites and remains confined to its location of origin. A malignant tumour is capable of both invasion of surrounding tissue as well as metastasis via the lymphatic or circulatory systems (Cooper & Hausman 2007).

Tumours are classed according to the type of cell from which they arise, and commonly fall into three main groups: carcinomas, sarcomas and leukemias or lymphomas (Cooper & Hausman 2007; Herceg & Hainaut 2007). Approximately 90 % of all malignancies are carcinomas, which are malignancies of epithelial tissue. Less common in humans is sarcomas, which are solid tumours of connective tissues such as; bone, muscle, cartilage and fibrous tissue. Leukemias and lymphomas originate from blood forming cells and cells of the immune system, respectively and account for approximately 8 % of human cancers (Cooper & Hausman 2007).

According to the International Agency for Research on Cancer (IARC) and World Health Organization (WHO), cancer is a leading cause of death worldwide, resulting in 8.2 million deaths in 2012. The IARC estimates that annual cancer cases will rise from 14 million in 2012 to 22 million in the next two decades (De Martel et al. 2012; Stewart & Wild 2014).

The most common human cancers, accounting for more than half of all neoplasias are; breast, prostate, lung and colon cancers, with lung cancer being by far the most lethal and resulting in approximately 30 % of all cancer deaths (Cooper & Hausman 2007; Stewart & Wild 2014).

1.2. Carcinogenesis

Carcinogenesis occurs due to the accumulation of genetic as well as epigenetic alterations, which alter the structure and/or function of the genome (Figure 1.1). These changes can be induced by dietary and/or environmental factors which trigger inappropriate activation or inactivation of specific genes which result in neoplastic transformation (Herceg & Hainaut 2007). Studies conducted by Hanahan and colleagues (2000) demonstrated that the transformation of a primary cell into a malignant one, involves alterations in mechanisms governing cell growth, homeostatic balance, cell differentiation and cell death (Figure 1.2) (Herceg & Hainaut 2007). More specifically there are six biological hallmarks of cancer.

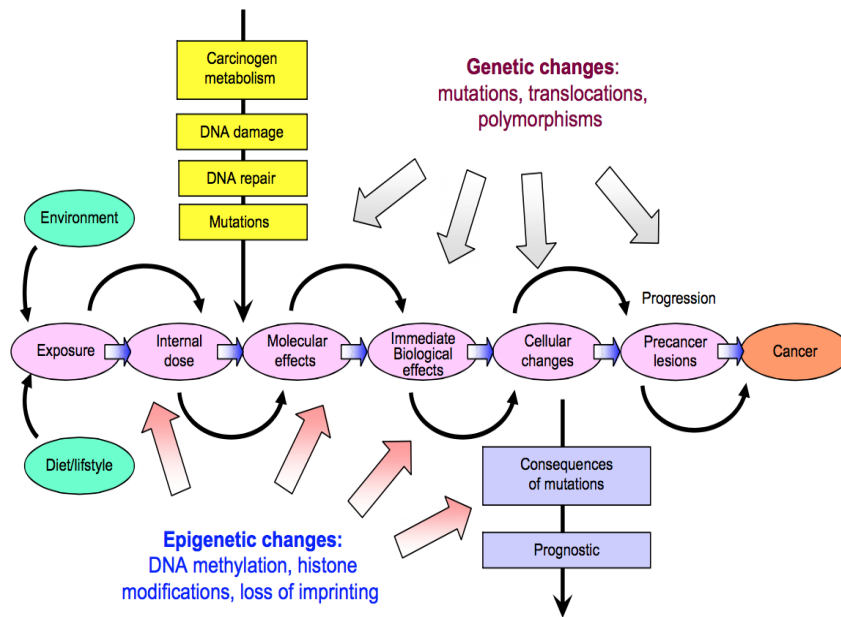


Figure 1.1: Cancer as the consequence of combined genetic and epigenetic alterations (Herceg & Hainaut 2007).

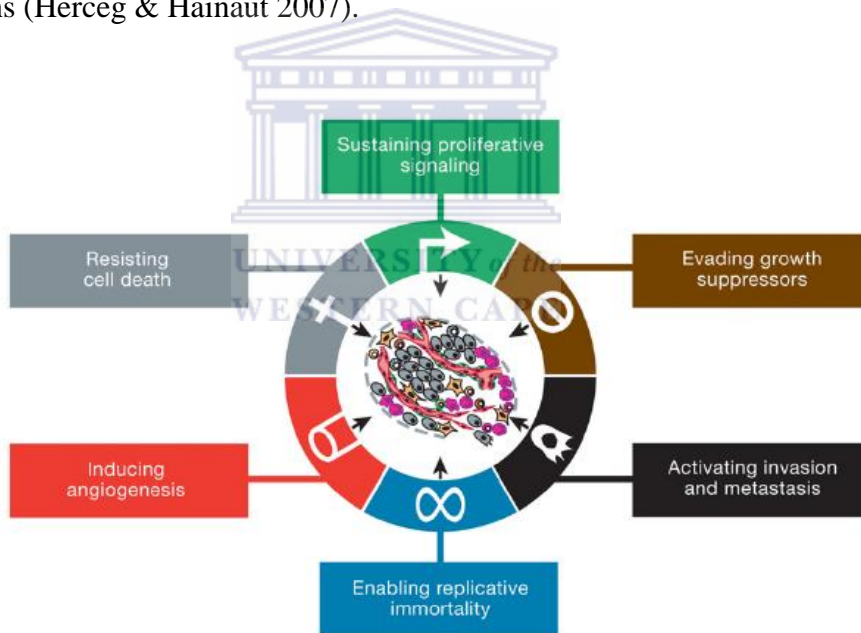


Figure 1.2: The six biological hallmarks of cancer (Hanahan & Weinberg 2011).

1.2.1. Sustaining Proliferative Signaling

One of the most important characteristics of tumour cells is their ability to sustain cell proliferation. The uncontrolled growth of malignancies distinguishes them from their normal counterparts (Cooper & Hausman 2007; Stratton et al. 2009). In normal cells, production and release of growth-promoting signals control homeostasis through cell growth and division cycles. These signals are communicated by growth factors, which bind cell-surface receptors, commonly containing tyrosine kinase domains. The receptors proceed to emit signals through branched intracellular pathways to regulate cell cycle progression (Hanahan & Weinberg 2011).

Cancer cells however, are able to deregulate these mechanisms resulting in uncontrolled proliferation. The ability of a tumour to sustain proliferation may be brought about in several ways. In some cases, cancer cells may produce growth factor ligands themselves, resulting in autocrine proliferation stimulation. Alternatively tumour cells may signal normal cells within the neoplasia to supply the cancer cells with growth signals. In other instances the reduced growth factor dependence of the tumour may result from elevated levels of receptors at the cancer cell surface (Cooper & Hausman 2007; Hanahan & Weinberg 2011; Chaffer & Weinberg 2011).

1.2.2. Evading Growth Suppressors

Within normal tissues, various anti-proliferative signals maintain cellular quiescence and homeostasis. These signals include soluble growth inhibitors and inhibitors embedded in the extracellular matrix and on surfaces of nearby cells (Hanahan &

Weinberg 2011; Stratton et al. 2009). The ability to bypass anti-proliferative signals is another fundamental trait of cancer cells. These signals are most typically regulated by tumour suppressor genes. Numerous tumour suppressors act in various ways to limit cell growth as well as proliferation. The two quintessential tumour suppressor genes encode the retinoblastoma (RB) and tumour protein P53 (P53) proteins which operate as central controls within two complementary cell regulatory circuits of cell proliferation (Chaffer & Weinberg 2011; Hanahan & Weinberg 2011). The RB protein integrates signals from both extracellular as well as intracellular sources and determines whether a cell would proceed through its growth and division cycle. While RB responds largely to extracellular signaling; P53 receives input from stress intracellularly, acting to halt proliferation or cause the cell to undergo apoptosis. Tumour cells with defects in the RB or P53 pathway are therefore, lacking important gatekeepers of cell cycle proliferation which may cause cells to cease proliferation and enter the G₀ (rest) phase of the cell cycle (Hanahan & Weinberg 2000; Hanahan & Weinberg 2011).

1.2.3. Resisting Cell Death

Apoptosis, programmed cell death, is a mechanism that enables multi-cellular organisms to tightly regulate or control cell growth in order to prevent pathological processes such as cancer (Simon et al. 2000). Apoptosis is triggered in response to various physiological stresses such as deoxyribonucleic acid (DNA) damage associated with hyper-proliferation and signaling imbalances, due to elevated levels of oncogenes (Brodie & Blumberg 2003; Hanahan & Weinberg 2011). Oncogenes are

mutated genes, which initially acted in cell cycle regulation. The failure of cancer cells to undergo apoptosis ultimately contributes substantially to tumour development (Cooper & Hausman 2007; Simon et al. 2000). Tumor cells utilize a variety of strategies to resist apoptosis, the most common being the loss of P53 tumor suppressor function. Tumors may also evade apoptosis by elevating expression of anti-apoptotic regulators or by down-regulating pro-apoptotic factors of survival signals (Hanahan & Weinberg 2011; Gibbons et al. 2014).

1.2.4. Enabling Replicative Immortality

Cancer cells require an infinite replicative potential in order to produce macroscopic tumours. This is in direct contrast to the tumours' normal cell counterparts, which only pass through a limited number of successive cell growth and division cycles. This limitation has been associated with two barriers to proliferation, namely; senescence, the irreversible entrance into a non-proliferative but viable state, and crisis, which results in cell death. Rarely do the cells emerge from crisis, this transition is called immortalization (Hanahan & Weinberg 2011).

Telomeres, which protect the ends of chromosomes, are implicated in being intricately involved in unlimited proliferation. In non-immortalized cells, telomeres shorten progressively, eventually losing the ability to protect the chromosomal DNA (Hanahan & Weinberg 2000; Hanahan & Weinberg 2011). The length of telomeric DNA also dictates the number of successive cell generations it may pass through before entering into crisis. Telomerase, a specialized DNA polymerase, adds telomere

repeat segments to the ends of telomeric DNA and is generally absent in normal cells but may be highly expressed in immortalized cells such as human cancer cells (Hanahan & Weinberg 2000; Chaffer & Weinberg 2011). By extending the telomeric DNA, telomerase counters the normal erosion that should occur. The presence of telomerase activity is directly correlated with resistance of both senescence and apoptosis (Simon et al. 2000; Hanahan & Weinberg 2011).

1.2.5. Inducing Angiogenesis

Tumours secrete growth factors that promote the formation of new blood vessels (angiogenesis). Angiogenesis is necessary to support growth beyond the size of an estimated million cells, which at this point require new blood vessels to supply oxygen and nutrients to the proliferating cancer cells. Blood vessels are formed in response to growth factors, secreted by the tumor cells that stimulate proliferation of endothelial cells within the walls of capillaries in surrounding tissue. This results in the outgrowth of new capillaries into the tumor (Sabine et al. 2002). The formation of such new blood vessels is critical not only in supporting tumor growth, but also in metastasis. The actively growing new capillaries formed in response to angiogenic stimulation are easily penetrated by the tumor cells, providing a ready opportunity for cancer cells to enter the circulatory system and begin the metastatic process (Cooper & Hausman 2007; Garraway & Lander 2013).

1.2.6. Activating Invasion and Metastasis

Most cancer cells are less adhesive than normal cells, often as a result of reduced expression of cell surface adhesion molecules. The reduced adhesiveness also results in morphological and cytoskeletal alterations in which many tumor cells are resultantly rounder than normal, in part because of the reduced attachment to either the extracellular matrix (ECM) or neighboring cells (Cooper 2000; Hunter et al. 2008; Hanahan & Weinberg 2011). Tumor cells are able to migrate and continue moving after contact with their neighbors, migrating over adjacent cells, and growing in disordered, multilayered patterns. The multistep process of invasion and metastasis is a distinct sequence of events often termed the invasion-metastasis cascade. This begins with local invasion, followed by intravasation of the cancer cells into the blood and lymphatic vessels and then the escape of the cancer cells into distant tissues and the formation of cancer nodules or micrometastasis resulting in the formation of tumours (Kenific et al. 2010; Hanahan & Weinberg 2011).

1.3. Lung Cancer

The molecular origins of lung cancer are the consequence of complex interactions between the environment and combined genetic and epigenetic host susceptibility (Figure 1.3) (Herceg & Hainaut 2007; Herbst et al. 2008). Certain environmental factors and genetic susceptibility may influence the initiation or promotion of carcinogenesis. The former may result in tissue injury, which initially can be seen in

the form of genetic and epigenetic alterations (Panov 2005; Herceg & Hainaut 2007; Herbst et al. 2008)

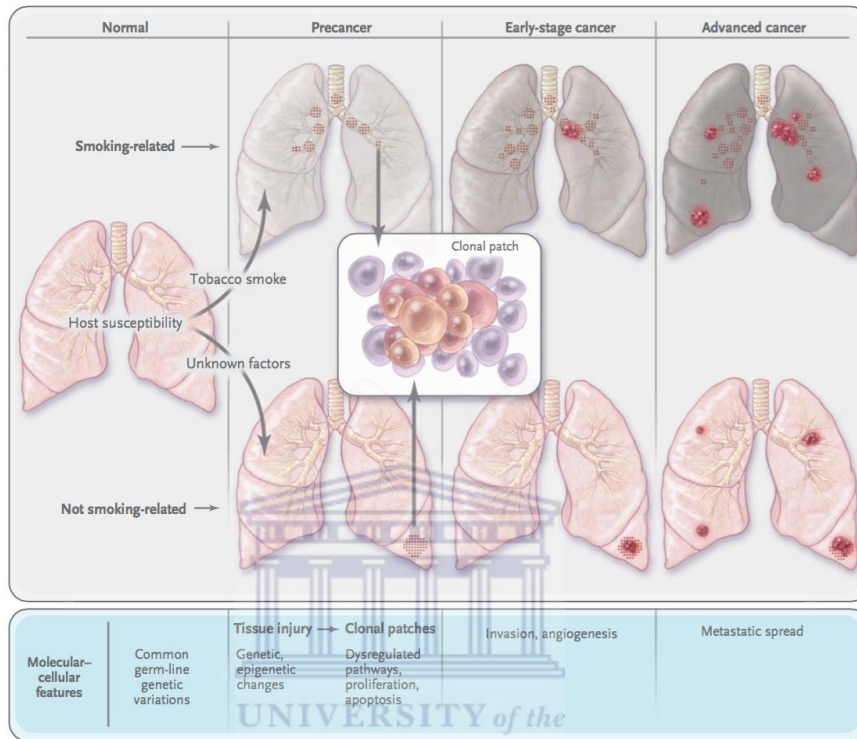


Figure 1.3: Molecular evolution of lung cancer. Showing the interaction of environmental factors, genetic susceptibility and unknown factors to influence carcinogenesis and resulting in genetic and epigenetic alterations, which influence the process of angiogenesis and metastases (Esteller 2008).

1.4. Classification of Lung Cancer

Lung cancer can be classified based on the size and appearance of the malignant cells (Table 1.1) as either; non-small cell lung carcinomas (NSCLC) or small-cell lung cancer (SCLC) (Travis et al. 2004). NSCLC can be further histologically categorized as; squamous cell carcinoma, adenocarcinoma and large cell lung carcinoma. 85 % of

all lung cancers are attributed to NSCLC, of which squamous cell carcinoma is most common in males and adenocarcinoma in females and non-smokers (Wynder & Muscat 1995; Johnson 1998; Brescia 2001).

Lung tumours present with heterogeneous patterns of genetic and epigenetic changes as well as gene expression, even in homogenous histological groups (Herceg & Hainaut 2007). Each class of tumour can be seen to progress via a different mechanism of carcinogenesis in association with specific genetic lesions (Wakamatsu et al. 2007). Lung carcinomas related to smoking display a very different molecular profile when compared to lung cancers unrelated to tobacco products (Herceg & Hainaut 2007). Studies report that epidermal growth factor receptor (EGFR) kinase mutations are observed in early adenocarcinoma development in never smokers, whilst mutations in Kirsten rat sarcoma viral oncogene homolog (KRAS) are seen in smokers (Herceg & Hainaut 2007; Herbst et al. 2008). Squamous cell carcinoma and SCLC are most commonly related to tobacco smoke and generally develop in the central airway. Tumours usually unrelated to smoking such as adenocarcinomas, tend to develop in the peripheral airways (Esteller 2008).

Table 1.5: Summary of lung tumour types (Herceg & Hainaut 2007; Herbst et al. 2008; Patel et al. 2008)

	Non-small Cell Carcinomas	Small Cell
--	----------------------------------	-------------------

	Squamous Cell Carcinomas	Adenocarcinomas	Large Cell Carcinomas	Carcinomas
Incidence (%)	55	15	5	25
Gender Incidence	M>F	F>M	M>F	M>F
Location	Hilar	Peripheral	Peripheral/Central	Hilar
Histological Stain	keratin	mucin	-	-
Relationship to Smoking	High	Low	High	High
Growth Rate	Slow	Medium	Rapid	Very rapid
Metastasis	Late	Intermediate	Early	Very early
Prognosis	2 year survival rate = 50 %			1 year if treated

1.5. Genetic Alterations in Lung Carcinomas

Overall, genetic alterations disrupt normal patterns of gene expression, which can result in abnormal expression of proteins (Herceg & Hainaut 2007; Lilloglou et al. 2014). DNA damage in the lung may fail to be repaired, resulting in incorrect nucleotide incorporations and ultimately mutations (Massion & Carbone 2003). Studies have revealed that in a single tumour, approximately 11 genes, including oncogenes and tumour suppressor genes, were mutated at a significant level (Herceg & Hainaut 2007; Risch & Plass 2008). Mutations envelop a variety of structural changes in DNA and include; changes in chromosome copy numbers, chromosomal

alterations such as translocations, amplifications, deletions and changes in nucleotide sequences (Massion & Carbone 2003; Herceg & Hainaut 2007; El-Zein et al. 2012).

EGFR is an example of a protein often over expressed as a result of a mutation and in the protein kinase domain. EGFR regulates important carcinogenic processes such as proliferation, apoptosis, angiogenesis and metastasis (Esteller 2008). Other often mutated functional domains involve DNA binding and transcriptional regulation domains (Herceg & Hainaut 2007).

Alteration in the P53 tumour suppressor gene is a typical example of a DNA binding and transcriptional regulation domain mutation, and is present in two thirds of lung cancers (Massion & Carbone 2003). Other common genetic mutations include that of KRAS, an oncogene mutated in approximately 30 % of lung carcinomas and cyclin-dependent kinase inhibitor 2A (p16), a tumour suppressor gene mutated in approximately 40 % of NSCLCs (Figure 1.4) (Wakamatsu et al. 2007; Estela et al. 2010; Fang et al. 2013). Chromosomal translocation is the most common type of mutation, while the protein kinase domain is functionally most frequently encoded by cancer genes (Herceg & Hainaut 2007; Kandoth et al. 2013).

Abnormality	Non-Small-Cell Lung Cancer		Small-Cell Lung Cancer
	Squamous-Cell Carcinoma	Adenocarcinoma	
Precursor			
Lesion	Known (dysplasia)	Probable (atypical adenomatous hyperplasia)	Possible (neuroendocrine field) †
Genetic change	<i>p53</i> mutation	<i>KRAS</i> mutation (atypical adenomatous hyperplasia in smokers), <i>EGFR</i> kinase domain mutation (in nonsmokers)	Overexpression of c-MET
Cancer			
<i>KRAS</i> mutation	Very rare	10 to 30% ‡	Very rare
<i>BRAF</i> mutation	3%	2%	Very rare
<i>EGFR</i>			
Kinase domain mutation	Very rare	10 to 40% ‡	Very rare
Amplification §	30%	15%	Very rare
Variant III mutation	5% ¶	Very rare	Very rare
<i>HER2</i>			
Kinase domain mutation	Very rare	4%	Very rare
Amplification	2%	6%	Not known
<i>ALK</i> fusion §	Very rare	7%	Not known
<i>MET</i>			
Mutation	12%	14%	13%
Amplification	21%	20%	Not known
<i>TTF-1</i> amplification	15%	15%	Very rare
<i>p53</i> mutation	60 to 70%	50 to 70% ‡	75%
<i>LKB1</i> mutation	19%	34%	Very rare
<i>PIK3CA</i>			
Mutation	2%	2%	Very rare
Amplification	33%	6%	4%

Figure 1.4: Genetic mutations specific to SCLCs and NSCLCs (Esteller 2008).

1.6. Epigenetics and Lung Cancer

Epigenetics refers to all heritable alterations in genetic expression and chromatin structure which is not directly coded in the DNA sequence (Herceg & Hainaut 2007). Epigenetic mechanisms which include; DNA methylation, histone modifications, micro ribonucleic acid (microRNAs) (miRNAs) and nucleosome remodeling, work together to regulate gene expression (Risch & Plass 2008; Liloglou et al. 2014). Epigenetic changes deregulate important mechanisms such as transcriptional control leading to inappropriate gene activation or silencing (Kanwal & Gupta 2012; Liloglou et al. 2014).

DNA methylation is an early event in the process of lung cancer (Risch & Plass 2008). Two types of DNA methylation are found in lung tumorigenesis; global hypomethylation, (the overall loss of 5-methyl-cytosine) contributing to genomic instability and gene promoter-associated hypomethylation (cytosine phosphate guanosine (CpG) island specific) (Herceg & Hainaut 2007; Liloglou et al. 2014). DNA promoter sequence methylation in association with histone tail modifications acts as the silencing mechanism of tumour suppressor genes. P16 is a tumour suppressor gene well studied in promoter-associated hypomethylation (Risch & Plass 2008).

MiRNAs, short, (20-22 nucleotides) non-coding RNAs capable of acting as either oncogenes or tumour suppressor genes, regulate gene expression post-transcriptionally (Herceg & Hainaut 2007). They may therefore affect messenger RNA (mRNA) stability and translational rate (Laird 2003). To date two miRNAs, miR-23 and miR-225 have been found to be specific to NSCLCs (Liloglou et al. 2014).

In addition, recent genome association research found a correlation between single-nucleotide polymorphism (SNP) variation at 15q 24-15q25.1 and lung cancer susceptibility. This SNP region includes two nicotinic acetylcholine alpha receptor genes encoding subunits which are regulated by nicotine exposure (Esteller 2008).

1.7. Causes and Risk Factors Associated with Lung Cancer

Tobacco smoke is attributed to approximately 75 % of all lung cancer cases worldwide, with remaining cases being linked to other environmental factors, heritable conditions and chronic inflammatory diseases (Johnson 1998; Coté et al. 2012). Despite the major correlation between cigarette smoke and lung cancer, lung carcinomas in never smokers represents the seventh leading cause of cancer related deaths globally (Pallis & Syrigos 2013).

Epidemiological studies have shown an association between an increased risk of lung cancer development and family history (Esteller 2008). Risk of susceptibility to this form of cancer is also increased in patients with inherited cancer syndromes resulting from rare germ-line mutations in P53, RB and EGFR (Gibbons et al. 2014).

1.8. Staging and Grading

At the time of diagnosis, the progression of the cancer is an important factor used to determine a treatment protocol as well as prognosis. The TNM system is the most commonly used cancer staging system (Edge & Compton 2010; McLoud & Swenson 1999). This system is accepted and maintained by the NCI (National Cancer Institute), American Joint Committee on Cancer (AJCC) and the International Union for Cancer Control (IUCC). It codifies cancers (Table 1.2) based on the size and extent of the primary tumour (T), the degree of spread to regional lymph nodes (N), and the presence of metastasis (M) or the formation of secondary tumours. A numerical index is added to each letter to indicate the extent of the primary tumour

and degree of cancer. These TNM combinations correspond to specific stages of cancer (Edge & Compton 2010; Edge et al. 2010).

Using the TNM system, early stage lung carcinomas are classified to be that of Stage I and II while late stage lung carcinomas are deemed Stage III and IV. Stage I form of cancer is located solely in the lungs with no spread to any lymph nodes; Stage II represents a tumour in the lungs with nearby lymph node spread. Stage III is termed a locally advanced disease with cancer spreading to lymph nodes in the middle of the chest, which are considered to be outside the lung. The most advanced stage of lung cancer, Stage IV is when the disease has spread to both lungs, the fluid surrounding the lungs or any other organ of the body (Edge & Compton 2010; Maldonado & Jett 2014).

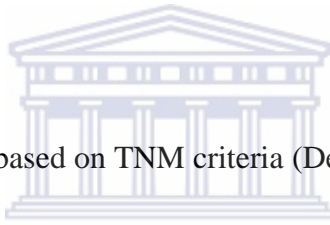


Table 1.6: Cancer staging based on TNM criteria (Detterbeck 2009).

T/M	Subgroup	N0	N1	N2	N3
T1	T1a	Ia	IIa	IIIa	IIIb
	T1b	Ia	IIa	IIIa	IIIb
T2	T2a	Ib	IIa	IIIa	IIIb
	T2b	IIa	IIb	IIIa	IIIb
T3	T3 _{>7}	IIb	IIIa	IIIa	IIIb
	T3 _{Inv}	IIb	IIIa	IIIa	IIIb
	T3 _{Satell}	IIb	IIIa	IIIa	IIIb
T4	T4 _{Inv}	IIIa	IIIa	IIIb	IIIb
	T4 _{Ipsi Nod}	IIIa	IIIa	IIIb	IIIb
M1	M1a _{Contra Nod}	IV	IV	IV	IV
	M1a _{Pl Disem}	IV	IV	IV	IV
	M1b	IV	IV	IV	IV

1.9. Diagnosis of Lung Cancer

Current diagnostic tools include; chest X-ray, computerized tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), sputum cytology and biopsy (Altintas & Tothill 2013). Chest radiography requires good view of both the posteroanterior and lateral lung regions. Chest X-rays demonstrate over 90 % of carcinomas. However, the mass is required to be between 1-2 cm in size for a reliable diagnosis. Pleural effusions as well as lobar collapse may be present (Patel et al. 2008). CT scans allow for a more accurate visualization of the mediastinum and is, therefore, better at identifying smaller lesions. This form of diagnosis is used to assess the extent of the tumour metastases as well as the operability of the mass (Patel et al. 2008).

In transthoracic fine-needle aspiration biopsy, a needle is guided by means of X-ray or CT. Direct aspiration of peripheral lung lesions occurs through the chest wall. Although implantation metastases does not occur, 25 % of patients may suffer a pneumothorax during this procedure (Patel et al. 2008; Altintas & Tothill 2013). These tools are expensive, tedious and may not be suitable for all cases, as other pathologies consider the needs of the patient. Individually it may also result in pain or complications for the patient (Altintas & Tothill 2013). At the time of diagnosis only 20 % of all lung cancer cases are localized, with the remainder being distant metastatic carcinomas which are non-resectable (Patz et al. 2007; Tu 2010). Current diagnostic techniques do not allow for early stage diagnosis, or detection of lung

cancer in asymptomatic patients, ultimately resulting in disease progression and a poor prognosis (Tu 2010).

1.10. Treatment and Prognosis

Surgery is currently the gold standard of NSCLC treatment, however, only 15 % of cases are operable at the time of diagnosis. Surgery is performed only upon confirmation of lung function tests displaying sufficient respiratory reserve, with no evidence of metastases on CT scans (Patel et al. 2008). Radiation treatment is often used for inoperable tumours and is effective particularly with slow growing squamous carcinomas. Radiation pneumonitis is a complication found in approximately 10-15 % of patients, while radiation fibrosis may occur in varying degrees in all cases (Patel et al. 2008; Flynn et al. 2013). Chemotherapy is the only effective treatment for SCLC, and is undertaken as a treatment therapy only and not a cure (Patel et al. 2008; Planque et al. 2009). Endoscopic therapy and transbronchial stenting may be used to provide symptomatic relief in patients. Daily administration of prednisone is used to improve appetite, while opioid analgesics are used to control pain (Brescia 2001; Patel et al. 2008).

Within one year of diagnosis, 45 % of lung cancer patients die despite receiving treatment (Brescia 2001). The average 5 year survival rate (Figure 1.5) is only 10-15 %, even patients presenting with clinical stage I lung cancer have a 60 % median survival rate of 5 years, indicating that a large portion of these patients possibly have undetectable metastatic lung cancer at the time of presentation of the disease (Patz et

al. 2007; Rose-James & TT 2012). The majority of NSCLC patients present with stage III and IV of the disease, those with stage IV cancer dying within 6-10 months of diagnosis (Brescia 2001).

In patients who survive surgical resection of a NSCLC, the risk of developing a second lung carcinoma is approximately 1 % to 2 % and 6 % for SCLC. The median survival rate of a secondary lung cancer diagnosis in these patients is 1-2 years, with less than 20 % of these cancers being resectable (Nicholson et al. 2001; Zeng et al. 2015).

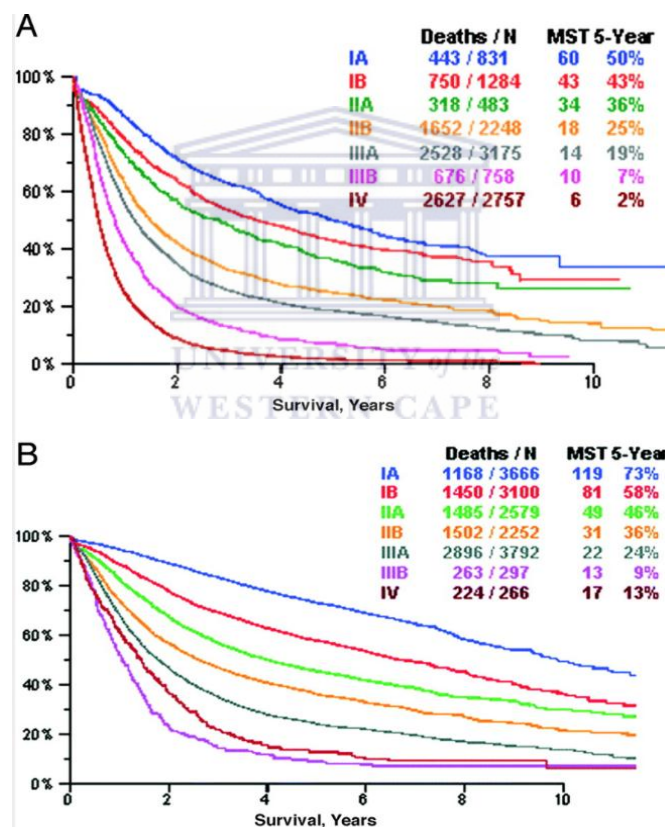


Figure 1.5: Lung cancer deaths and 5 year median survival rate in relation to (A) clinical stages and (B) pathologic stage (Detterbeck 2009).

1.11. The Burden of Disease of Lung Cancer

Lung cancer remains a significant public health issue, resulting in the most cancer-related deaths globally (Kim et al. 2007; Altintas & Tothill 2013). In 2012 lung cancer was responsible for 1.59 million deaths (De Martel et al. 2012). Variations in incidence rates of lung cancer can be seen based on age, gender and global geographical location, with the greatest incidence of disease being observed in men in eastern and central Europe and northern America (Altintas & Tothill 2013). Surveillance, Epidemiology, and End Results (SEER) statistics indicate that in the USA lung cancer represented 13.5 % of all new cancer cases (Figure 1.6) with an estimated 159 260 deaths predicted in 2014 (Howlader et al. 2015). This neoplasia presents with similar mortality and incidence rates in contrast to other common cancers such as breast, colon and prostate carcinomas with relatively low mortality rates. With approximately 22 % of all cancers stemming from this disease, it is the second most common cancer in men and third in women (Figure 1.7) (Altintas & Tothill 2013).

With one in six males and one in eight females at risk of developing cancer, South Africa has one of the highest incidence rates of cancer in Africa (Albrecht 2006; Nema & Khare 2011). Data obtained from the National Cancer Registry (2004) showed lung cancer to be one of the leading cancers to affect South African males. In South Africa, lung cancer is the second most common cancer in men and the sixth leading cancer in women in terms of diagnosis (CANSAs 2008). Approximately 60 % of all lung cancer deaths in South Africa are due to tobacco smoking with over 8 % of

all deaths attributed to smoking. Over 42 000 South Africans a year die of tobacco-related diseases, which includes lung cancer (Mayosi et al. 2009).

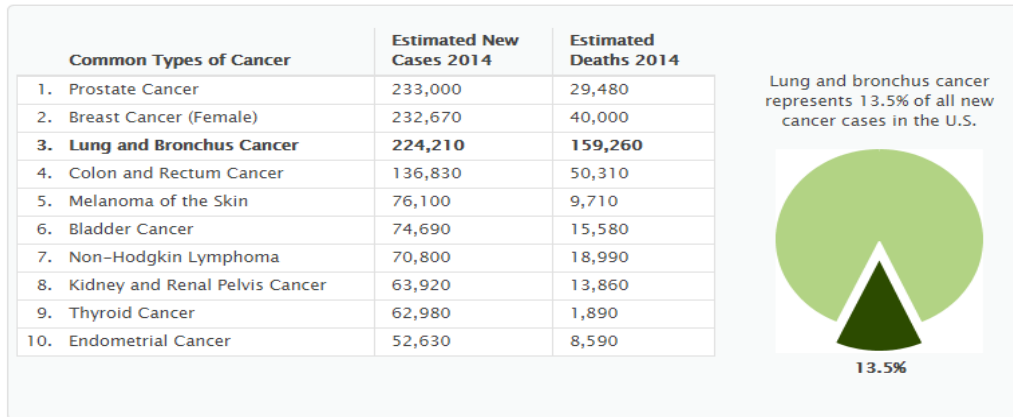


Figure 1.6: Estimated new cancer diagnoses and deaths of most common types of cancer in the U.S. in 2014. With lung cancer displaying the 3rd most common cancer type, representing 13.5 % of all new cancer cases (National Cancer Institute (NCI), Surveillance, Epidemiology, and End Results (SEER) 2015).

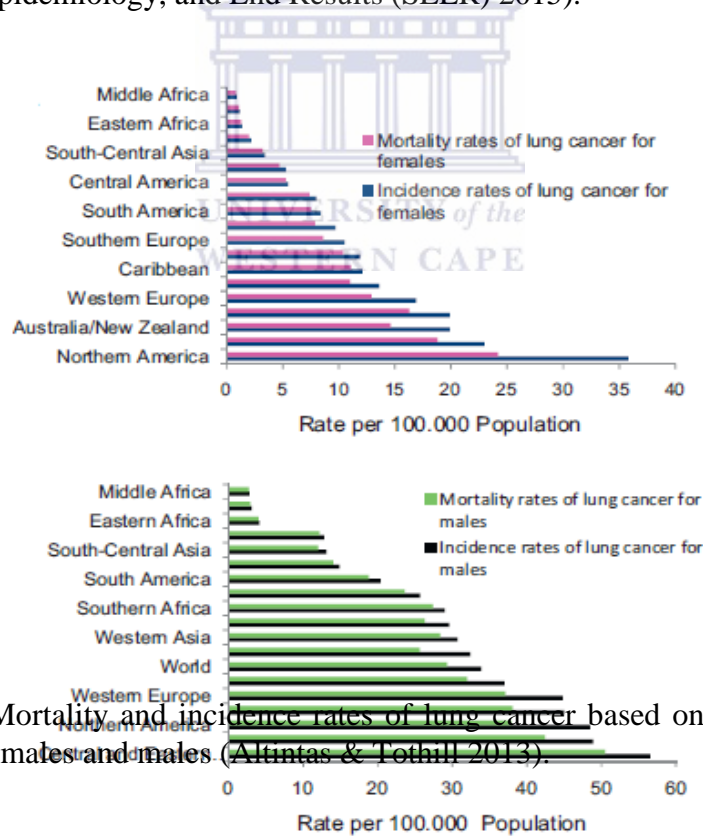


Figure 1.7: Mortality and incidence rates of lung cancer based on geographical location in females and males (Altintas & Tothill 2013).

1.12. Biomarker Application in Cancer

The main reason associated with poor prognosis is due to the difficulty in making an early stage diagnosis, unresectability as well as the high rate of recurrence after treatment (Brescia 2001; Kim et al. 2007). Currently no diagnostic tools exist that are able to detect lung cancer in asymptomatic patients. Indeed, a rapid, sensitive, easily applicable, non-invasive screening mechanism is needed to detect lung cancer at a stage in which intervention could alter the natural progression of the disease (Altintas & Tothill 2013; Vannini et al. 2013).

Sophisticated molecular techniques have made it possible to detect genetic alterations in tumours; with research highlighting the fact that certain of these changes are specific to homogenous malignant diseases (Fleischhacker et al. 1999; Kim et al. 2007).

Tumour biomarkers are molecules used as indicators of biological homeostasis and are produced by cancer cells as a direct response by the body to the tumour (Altintas & Tothill 2013). Cancer markers can be differentiated into several distinct groups based on; genetics, epigenetics and proteomics (Sung & Cho 2008).

Post-translational and translational expression analysis in single cells have to date identified many biomarkers as screening tools for cancer research. These research areas include; cancer diagnosis, prognosis and therapy development techniques, to predict the response to specific therapy types such as chemotherapy or evaluate the risk of future relapse (Schwarzenbach et al. 2011; Altintas & Tothill 2013).

Correct diagnosis of cancer using biomarkers is expected to significantly benefit in molecular based cancer patient care, with the potential to predict possible cancer progression and possibly prevent cancer development in individuals identified as high risk (Hassanein et al. 2012). These biomarkers are expected to not just predict predisposition but to also diagnose patients at an early stage of the disease. This would greatly increase the patients' prognosis and ultimately decrease mortality. In addition, certain tumour markers, referred to as secondary biomarkers, change in expression levels in response to therapy and treatment and could be used as a guide to the most effective therapy required (Sung & Cho 2008).

Thus molecular markers may potentially be used to signify risk in individuals without the disease, and in prognosis of those affected. It could also determine sensitivity to treatment, spanning the course of a disease through its various stages (Esteller 2008). It is unlikely that one single biomarker will meet all these conditions, due to the heterogeneity reported among cancers. It is also unrealistic that a single biomarker will provide the specificity and sensitivity necessary throughout the various stages of tumour progression and development (Phan et al. 2009). Therefore, identifying a panel of tumour markers would improve the efficacy of diagnosis as well as prognosis (Planque et al. 2009; Travis et al. 2011).

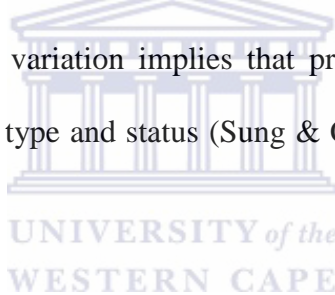
1.13. Lung Cancer Biomarkers

Unlike the specific parameters used in the TNM system, tumour markers are reported to be far more suited to the heterogenous nature of cancer (Sung & Cho 2008).

Biomarkers have greater potential for differential diagnosis and histological sub-typing, particularly in lung carcinomas of unknown origins (Chi-Shing Cho 2007). Although lung cancer is histologically categorized into SCLC and NSCLC, there may be various other criteria dividing the sub-types such as genetic mutations. These specific sub-categories cannot be determined without the use of invasive biopsies and screening of tumour specific biomarkers, and may prove more useful in an attempt to accurately diagnose the cancer (Sung & Cho 2008).

1.14. Protein Biomarkers

The human genome is currently known to contain 20 488 genes. Proteins result in far greater variety due to post-translation modifications, protease cleavages and splice variants. This increase in variation implies that protein biomarkers can contribute more specificity to cancer type and status (Sung & Cho 2008; Makridakis & Vlahou 2010).



Protein lung cancer biomarkers can be classified from the source of the proteins into three categories, namely: serum biomarkers, tissue biomarkers, and sputum biomarkers with a broad range associated with lung cancer (Table 1.3). The concentration and levels of these biomarkers however are quite complex. Both the specificity as well as the response ratio of the protein biomarkers show alterations depending on the histological subtype of the lung carcinoma (Altintas & Tothill 2013).

Table 1.7: Lung cancer protein biomarkers currently available (Sung & Cho 2008; Altintas & Tothill 2013).

	Diagnosis	Therapy monitoring	Prognosis monitoring	Ontology
Carcinoembryonic antigen (CEA)	Adenocarcinoma Large cell lung cancer	Adenocarcinoma NSCLC	Adenocarcinoma NSCLC	Cellular component
Cytokeratin fragment (CYFRA 21-1)	NSCLC, SCLC	NSCLC	NSCLC, SCLC	Structural constituent of cytoskeleton
Tissue polypeptide antigen (TPA)	NSCLC, SCLC	-	NSCLC	
Progastrin-releasing peptide (ProGRP)	SCLC	SCLC	-	Neuropeptide hormone activity
Neuron-specific endolase (NSE)	SCLC	SCLC	SCLC	Phosphoglycerate dehydrogenase activity, sub-cellular location
Tumour M2 pyruvate kinase	Adenocarcinoma	-	Adenocarcinoma	Pyruvate kinase activity, Glycolysis, Cytoplasm

1.15. Gene Biomarkers

A large variety of genes have been exposed to somatic mutations in human tumour cells or tissue. These mutated genes include oncogenes and tumour suppressor genes as well as genes which encode proteins that perform vital functions in cell cycle regulation, DNA repair, apoptosis and telomerase activity (Altintas & Tothill 2013).

A range of mutation types have also been identified, these include; missense, nonsense and splicing mutations, gene amplification, micro deletions, translocations and promoter hyper-methylation (Table 1.4). The roles of these somatic mutations play in lung carcinogenesis is understood in terms of their ability to promote cellular growth, interfere with DNA repair, evasion of host immunity, to confer resistance to apoptosis or to induce cellular transformation to name a few (Altintas & Tothill 2013; Fang et al. 2013).

Inactivation of tumour suppressor genes during cell division is a key factor driving clonal cancer cells into hyper-proliferation, migration and metastasis. In many cases the inactivation is initiated by loss of DNA chromosomal rearrangement occurring during cell division. The most frequently occurring abnormality is deletion of the short arm of chromosome 3 (3p) (Altintas & Tothill 2013; Fang et al. 2013). Loss of chromosomal material has been reported to be detected in metaplastic epithelium tissue of smokers (Sung & Cho 2008).

Altered hyper-methylation and methylation of CpG rich regions of several promoter regions is representative of epigenetic changes in the cell and may lead to gene silencing. Due to this, specific methylation status in genes can serve as biomarkers especially in tumour suppressor genes (Sung & Cho 2008). Activation of genes involves growth factors, their receptors, their messengers or cell cycle activators of mutations, which drives tumorigenesis (Estela et al. 2010).

Table 1.8: Genes and associated mutation types reported in lung cancer (Sung & Cho 2008; Altintas & Tothill 2013).

Groups	Types of genes	Prevalence in sample
Chromosomal changes	Deletion of the short arm of chromosome 3 (3p)	27-88 % in circulating DNA in lung cancer patients
Hypermethylation	Serine protease family member-trypsinogen IV (PRSS3) Tissue inhibitor of metalloproteinase (TIMP)-3 Death associated protein (DAP)-kinase P16 FHIT	Associated with increased risk of lung cancer recurrence after therapy
Genetic Changes	KRAS P53	20-30 % in circulating DNA of lung cancer patients 27 % in circulating DNA of lung cancer patients

1.16. Sources of Biomarkers

Tumour markers can be detected in the blood, urine or serum in higher than normal ranges and may include; hormones, specific antigens, oncogenes and proteins, etc. (Altintas & Tothill 2013). The increased entry of these molecules into serum circulation is facilitated by mechanisms such as secretion, angiogenesis, invasion and destruction of tissue architecture (Prassas et al., 2012).

Plasma is a target of interest for biomarker detection as it would contain small amounts of circulating DNA fragments shed by normal and tumour cells undergoing apoptosis or necrosis (Herceg & Hainaut 2007). Serum tests for oncogene mutations and hyper-methylation of promoter regions are used for cancer detection. Since recent advances in genomic and proteomic technologies, specific changes in tumour cells expression levels can distinctly be measured, even with the presence of DNA shed by normal cells (Wulfkühle et al. 2003).

Recent studies into identifying sources of potential cancer markers has focused largely into secretome, which focuses on studies monitoring molecules shed from the surfaces of living cells, including proteins. Secreted molecules, proteins and extracellular matrix components from tumour cells are therefore a promising source of potential tumour markers (Makridakis & Vlahou 2010).

In lung cancer specifically, serum, tissue and sputum serve as important sources of potential biomarkers. In sputum, cells from cancer sites are major protein sources. In biopsied lung tissue, cancer sites as well as cells involved in immune reactions such as, cytokines and derivatives from immune or inflammatory response can be found. In blood, however, biomarkers with potentially greater significance can be found (Saijo 2012). These include biomarkers found in biopsied cancer tissue as well as many circulating proteins and cells derived from the tumour tissue. Since the end goal of biomarker discovery is the specific, early and non-invasive diagnosis and post treatment monitoring of the disease, blood is thought of as an important biological

material. Resulting in many biomarker investigations carried out with blood based strategies (Sung & Cho 2008).

Since many tumour markers exist in more significant concentrations in tumour tissue than in body fluids, biopsies still tend to be both invasive and uncomfortable for patients. Some genetic markers for lung cancer may be obtained from sputum or studying pleural fluid, but blood still appears to be a more suitable source for biomarkers (Altintas & Tothill 2013).

When circulating RNAs are obtained from sputum, due to the high levels of ribonuclease (RNase) in the sputum, some of the total RNAs are degraded. Circulating nucleic acids is reported to be a crucial parameter for detecting the disease without invasiveness (Schwarzenbach et al. 2011). DNA and RNA molecules are present in the serum of both healthy and ill patients. The existence of circulating DNA in blood with malignant neoplasm has been known since the 1970s and since then researchers have attempted to develop methods using biological materials obtained from non-invasive procedures in order to locate potential biomarkers for early diagnosis (Altintas & Tothill 2013).

1.17. Applications of Bioinformatics into Biomarker Discovery

Bioinformatics is the application of computational techniques to analyse information associated with biological data on a large scale (Luscombe et al. 2001). There are three main aims in the field of bioinformatics. The first and most basic is to order data in such a way that users are able to access existing data as well as submit new

findings. The second, is to develop tools to assist in data analysis and the third, to use these tools to analyse and interpret results in such a way that they become biologically significant (Luscombe et al. 2001; Wu et al. 2012;).

Biological studies have traditionally intricately examined isolated systems and often compared them to a few related studies. Bioinformatics in contrast, allows for the global analyses of data, aiming to uncover novel features and highlighting principles which apply to various disciplines (Luscombe et al. 2001).

The major advancements achieved in unraveling the molecular mechanisms of human diseases, molecular diagnostics and therapy over the past two decades, is largely due to the substantial growth in the amount of genomic, proteomic and transcriptomic data being generated (Phan et al. 2009).

As of August 2000, a repository of 8 214 000 nucleic acid sequences and 88 166 protein sequences were publicly available, with datasets doubling in size every 15 months (Luscombe et al. 2001). Since the publication of the *Haemophilus influenzae* (*H. influenzae*) genome, complete sequences for over 40 organisms have been released, ranging from 450 to over 100 000 genes (Fleischmann et al. 1995). This surge in data availability, coupled with a myriad of related studies; into gene expression, protein structure and interactions between various biomolecules resulted in many of these previously known biological challenges becoming challenges of computing. Thus firmly establishing bioinformatics as a discipline in molecular biology (Luscombe et al. 2001).

Gene expression levels can be determined by measuring mRNA levels with various techniques such as microarrays, expressed complementary DNA (cDNA) sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signature sequencing (MPSS) etc. (Raza 2012). These techniques have allowed an unbiased overview of changes occurring at transcriptional levels and have revolutionised cancer research, resulting in numerous potential biomarkers being generated (Rhodes & Chinnaiyan 2004; Werner 2008).

Protein expression is one of the most accurate indicators of actual gene activity since proteins are usually final catalysts of cell activity. Protein microarrays and high-throughput (HT) mass spectrometry (MS) can provide an image of the proteins present in a biological sample. Bioinformatics is integral in making sense of protein microarray and high throughput data (Raza 2012).

1.18. Biomarker Validation

Before biomarkers can be utilized in clinical practice, each biomarker needs to be discovered and validated by means of a process involving several important steps (Figure 1.8). The first step of this process consists of experimental design and data acquisition, generally in the form of large amounts of genomic or proteomic expression data (Prassas et al. 2012). Once acquired, data needs to be organised and annotated, this can be done using various databases and web based tools. The next stage in data processing is identifying candidate biomarkers, which are differentially expressed (DE), using classification methods and feature extraction. Functional

relevance of candidate biomarkers needs to then be evaluated by determining their biological expression level (Phan et al. 2009). Validation of these markers for example by means of real-time polymerase chain reaction (RT-PCR) and enzyme-linked, immunosorbent assays (ELISAs) can be both labour and resource intensive, making validation of these markers of critical importance (Rhodes & Chinnaiyan 2004). Given the necessity for disease specific biomarkers along with the flood of genomic and proteomic data, it is therefore, up to biological computation systems to provide methods to evaluate, integrate and translate the data (Kim et al. 2007).

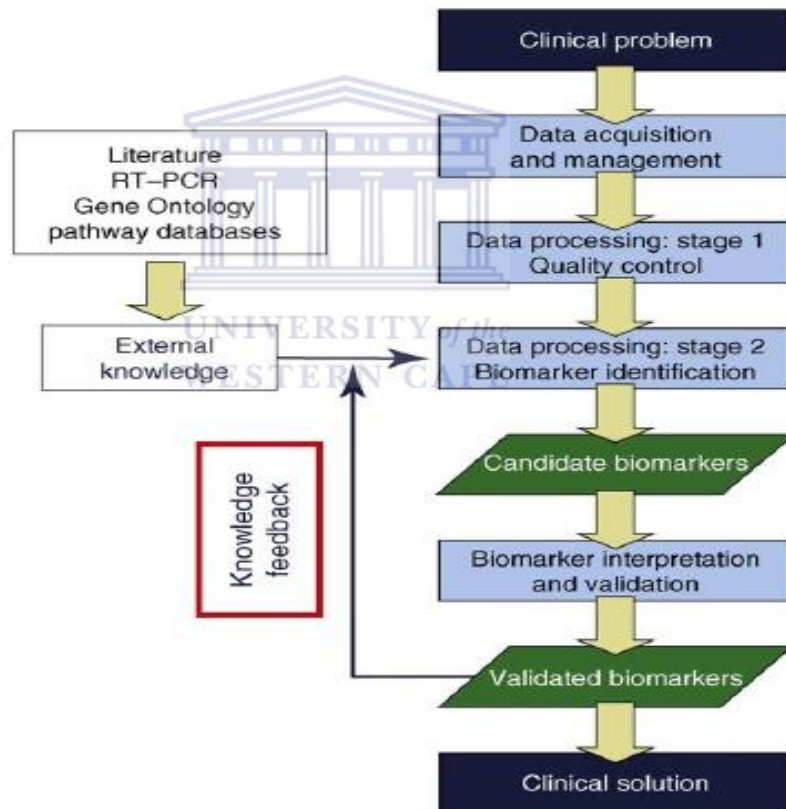


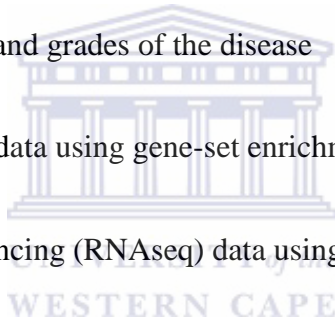
Figure 1.8: Biocomputing tools for discovery and validation of biomarkers (Phan et al. 2009).

1.19. Aims and Objectives

Lung cancer remains the leading cause of cancer deaths worldwide. The efficiency of current treatment depends strongly on the time of diagnosis, at the time of diagnosis, the progression of the cancer is an important factor used to determine a treatment protocol as well as prognosis (Risch & Plass 2008). Current diagnostic techniques do not allow for early stage diagnosis, or detection of lung cancer in asymptomatic patients, ultimately resulting in disease progression and a poor prognosis.

This project aims to identify potential circulatory biomarkers only in Stage I and II lung cancer as possible diagnostic agents by:

- Data mining public cancer databases for novel genes related to mechanisms involved in the stages and grades of the disease
- Analysing microarray data using gene-set enrichment analysis (GSEA) techniques
- Analysing RNA sequencing (RNAseq) data using bioinformatics enrichment tools
- Correlating differentially expressed genes (DEG) between samples of:
 - Early stage lung cancer vs. normal lung tissue and
 - Early stage disease vs. late stage lung carcinoma
- Identifying co-expression of genes involved in the pathogenic phenotype



1.20. References

- Albrecht, C. 2006. Overview of the South African cancer research environment as a basis for discussions concerning the activation of CARISA (Cancer Research Initiative of South Africa). Cape Town.
- Altintas, Z. & Tohill, I. 2013. Biomarkers and biosensors for the early diagnosis of lung cancer. *Sensors and Actuators B: Chemical*. 188:988–998.
- Brescia, F.J. 2001. Lung cancer--a philosophical, ethical, and personal perspective. *Critical Reviews in Oncology/Hematology*. 40(2):139–148.
- Brodie, C. & Blumberg, P.M. 2003. Regulation of cell apoptosis by protein kinase c. *Apoptosis*. 8(1):19–27.
- CANSA. 2008. *Fact Sheet on Lung Cancer*. Available: <http://www.cansa.org.za/files/2014/05/Fact-Sheet-Lung-Cancer> [2014, June 07].
- Chaffer, C.L. & Weinberg, R.A. 2011. A perspective on cancer cell metastasis. *Science*. 331(6024):1559–1564.
- Chi-Shing Cho, W. 2007. Potentially useful biomarkers for the diagnosis, treatment and prognosis of lung cancer. *Biomedicine & Pharmacotherapy*. 61(9):515–519.
- Cooper, G.M. 2000. The Cell. <http://www.ncbi.nlm.nih.gov/books/NBK9839/> [2013, November 07].
- Coté, M.L. et al. 2012. Increased risk of lung cancer in individuals with a family history of the disease: A pooled analysis from the International Lung Cancer

Consortium. *European Journal of Cancer*. 48(13):1957–1968.

Detterbeck, F. 2009. The new lung cancer staging system. *CHEST*. 136(1):6-8.

Edge, S.B. & Compton, C.C. 2010. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Annals of Surgical Oncology*. 17(6):1471–4.

Edge, S., Byrd, D., Compton, C., Greene, F. & Fritz, A. 2010. *AJCC Cancer Staging Manual*. 7th ed. New York: Springer-Verlag.

El-Zein, R. A., Young, R.P., Hopkins, R.J. & Etzel, C.J. 2012. Genetic predisposition to chronic obstructive pulmonary disease and/or lung cancer: important considerations when evaluating risk. *Cancer Prevention Research*. 5(4):522–527.

Estela, R., Cabral, C., Bispo, J., Neto, C. & Carvalho, C. 2010. Circulating DNA as a biomarker for early detection of cancer : A Brief Update with an Emphasis on Lung Cancer. *Lung Cancer*. 38–44.

Esteller, M. 2008. Molecular origins of cancer epigenetics in cancer. *N Engl J Med*. 358:1148–59.

Fang, X., Netzer, M., Baumgartner, C., Bai, C. & Wang, X. 2013. Genetic network and gene set enrichment analysis to identify biomarkers related to cigarette smoking and lung cancer. *Cancer Treatment Reviews*. 39(1):77–88.

Fleischhacker, M., Beinert, T. & Possinger, K. 1999. Molecular genetic characteristics of lung cancer—useful as “real” tumor markers? *Lung Cancer*. 25(1):7–24.

Fleischmann, R.D. et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science (New York, N.Y.)*. 269(5223):496–512.

Flynn, A.E., Peters, M.J. & Morgan, L.C. 2013. Attitudes towards lung cancer screening in an Australian high-risk population. *Lung Cancer International*. 2013:1–7.

Garraway, L. A. & Lander, E.S. 2013. Lessons from the cancer genome. *Cell*. 153(1):17–37.

Gibbons, D.L., Byers, L. A. & Kurie, J.M. 2014. Smoking, p53 mutation, and lung cancer. *Molecular Cancer Research*. 12(1):3–13.

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

Hanahan, D. & Weinberg, R.A. 2011. Hallmarks of cancer: the next generation. *Cell*. 144(5):646–74.

Hassanein, M., Callison, J.C., Callaway-Lane, C., Aldrich, M.C., Grogan, E.L. & Massion, P.P. 2012. The state of molecular biomarkers for the early detection of lung cancer. *Cancer Prevention Research*. 5(8):992–1006.

Herbst, R., Heymach, J. & Lippman, S. 2008. Molecular origins of cancer: lung cancer. *N Engl J Med*, 359: 1367–1380.

Herceg, Z. & Hainaut, P. 2007. Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis. *Molecular Oncology*. 1(1):26–41.

Howlader, N. et al. 2015. *Cancer Statistics Review, 1975-2012 - SEER Statistics*. Available: http://seer.cancer.gov/csr/1975_2012/ [2015, January 01].

Hunter, K.W., Crawford, N.P. & Alsarraj, J. 2008. Mechanisms of metastasis. *Breast Cancer Research*. 10(Suppl 1):S2.

Johnson, B.E. 1998. Tobacco and lung cancer. *Primary Care*. 25(2):279–291.

Kandoth, C. et al. 2013. Mutational landscape and significance across 12 major cancer types. *Nature*. 502(7471):333–9.

Kanwal, R. & Gupta, S. 2012. Epigenetic modification in cancer. *Clinical Genetics*. 81(4):303–311.

Kenific, C.M., Thorburn, A. & Debnath, J. 2010. Autophagy and metastasis: another double-edged sword. *Current Opinion in Cell Biology*. 22(2):241–245.

Kim, B. et al. 2007. Clinical validity of the lung cancer biomarkers identified by bioinformatics analysis of public expression data. *Cancer Research*. 67(15):7431–7438.

Laird, P.W. 2003. The power and the promise of DNA methylation markers. *Nature Reviews. Cancer*. 3(4):253–266.

Liloglou, T., Bediaga, N.G., Brown, B.R.B., Field, J.K. & Davies, M.P.A. 2014. Epigenetic biomarkers in lung cancer. *Cancer Letters*. 342(2):200–212.

Luscombe, N.M., Greenbaum, D. & Gerstein, M. 2001. What is bioinformatics? An introduction and overview. *Yearbook of Medical Informatics*. 83–100.

Makridakis, M. & Vlahou, A. 2010. Secretome proteomics for discovery of cancer biomarkers. *Journal of Proteomics*. 73(12):2291–2305.

Maldonado, F. & Jett, J.R. 2014. Invasive and noninvasive advances in the staging of lung cancer. *Seminars in Oncology*. 41(1):17–27.

De Martel, C., Ferlay, J., Franceschi, S., Vignat, J., Bray, F., Forman, D. & Plummer, M. 2012. Global burden of cancers attributable to infections in 2008: A review and synthetic analysis. *The Lancet Oncology*. 13(6):607–615.

Massion, P.P. & Carbone, D.P. 2003. The molecular basis of lung cancer: Molecular abnormalities and therapeutic implications. *Respiratory Research*. 4(12).

Mayosi, B.M., Flisher, A.J., Lalloo, U.G., Sitas, F., Tollman, S.M. & Bradshaw, D. 2009. The burden of non-communicable diseases in South Africa. *The Lancet*. 374(9693):934–947.

McLoud, T. & Swenson, S. 1999. Lung carcinoma. *Clinics in Chest Medicine*. Available: <http://www.sciencedirect.com/science/article/pii/S0272523105702495> [2014, October 26].

Nema, R. & Khare, S. 2011. A review on: breast, lung and ovary cancer. *International Journal of Research in Biological Sciences*. 2(2):73–76.

Nicholson, R.I., Gee, J.M. & Harper, M.E. 2001. EGFR and cancer prognosis. *European Journal of Cancer*. 37(4):9–15.

Pallis, A.G. & Syrigos, K.N. 2013. Lung cancer in never smokers: Disease characteristics and risk factors. *Critical Reviews in Oncology/Hematology*. 88(3):494–503.

Panov, S.Z. 2005. Molecular biology of the lung cancer. *Radiology and Oncology*.

39(3):197.

Patel, H., Gwilt, C. & McGowan, P. 2008. *Respiratory System*. London: Elsevier. 3744-3750.

Patz, E.F., Campa, M.J., Gottlin, E.B., Kusmartseva, I., Xiang, R.G. & Herndon, J.E. 2007. Panel of serum biomarkers for the diagnosis of lung cancer. *Journal of Clinical Oncology*. 25(35):5578–5583.

Phan, J.H., Moffitt, R. A., Stokes, T.H., Liu, J., Young, A.N., Nie, S. & Wang, M.D. 2009. Convergence of biomarkers, bioinformatics and nanotechnology for individualized cancer treatment. *Trends in Biotechnology*. 27(6):350–358.

Planque, C., Kulasingam, V., Smith, C.R., Reckamp, K., Goodglick, L. & Diamandis, E.P. 2009. Identification of five candidate lung cancer biomarkers by proteomics analysis of conditioned media of four lung cancer cell lines. *Molecular & Cellular Proteomics*. 8(12):2746–2758.

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

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

Rhodes, D.R. & Chinnaiyan, A.M. 2004. Bioinformatics strategies for translating genome-wide expression analyses into clinically useful cancer markers. *Annals of the New York Academy of Sciences*. 1020(1):32–40.

- Risch, A. & Plass, C. 2008. Lung cancer epigenetics and genetics. *International Journal of Cancer*. 123(1):1–7.
- Rose-James, A. & TT, S. 2012. Molecular markers with predictive and prognostic relevance in lung cancer. *Lung Cancer International*. 2012:1–12.
- Sabine, Z., Gazdar, A.F. & Minna, J.D. 2002. Molecular biology of lung cancer: clinical implications *Thorax*. 58(10):681–708.
- Saijo, N. 2012. Critical comments for roles of biomarkers in the diagnosis and treatment of cancer. *Cancer Treatment Reviews*. 38(1):63–67.
- Schwarzenbach, H., Hoon, D.S.B. & Pantel, K. 2011. Cell-free nucleic acids as biomarkers in cancer patients. *Nature Reviews Cancer*. 11(6):426–437.
- Simon, H.U., Haj-Yehia, A. & Levi-Schaffer, F. 2000. Role of reactive oxygen species (ROS) in apoptosis induction. *Apoptosis*. 5(5):415–418.
- Stewart, B.S. & Wild, C.P. 2014. World Cancer Report 2014. *WHO*. Available: <http://www.mendeley.com/research/world-cancer-report-2014-21/> [2014, October 26].
- Stratton, M., Campbell, P. & Futreal, P. 2009. The cancer genome. *Nature*. 458(7239):719-724.
- Sung, H.-J. & Cho, J.-Y. 2008. Biomarkers for the lung cancer diagnosis and their advances in proteomics. *BMB reports*. 41(9):615–625.
- Travis, W.D., Brambilla, E., Müller-Hermelink, H.K. & Harris, C.C. 2004. Pathology and genetics of tumours of the lung. *Bulletin of the World Health Organization*. 50(1-

2):9–19.

Travis, W.D. et al. 2011. International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society international multidisciplinary classification of lung adenocarcinoma. *J Thorac Oncol.* 6(2):244–285.

Tu, S.-M. 2010. Diagnosis and prognosis. *Cancer Treatment and Research.* 154:103–114.

Vannini, I., Fanini, F. & Fabbri, M. 2013. MicroRNAs as lung cancer biomarkers and key players in lung carcinogenesis. *Clinical Biochemistry.* 46(10-11):918–925.

Wakamatsu, N., Devereux, T.R., Hong, H.-H.L. & Sills, R.C. 2007. Overview of the molecular carcinogenesis of mouse lung tumor models of human lung cancer. *Toxicologic Pathology.* 35(1):75–80.

Werner, T. 2008. Bioinformatics applications for pathway analysis of microarray data. *Current Opinion in Biotechnology.* 19(1):50–4.

Wu, D., Rice, C.M. & Wang, X. 2012. Cancer bioinformatics: A new approach to systems clinical medicine. *BMC Bioinformatics.* 13(1):71.

Wulfkuhle, J.D., Liotta, L.A. & Petricoin, E.F. 2003. Early detection: Proteomic applications for the early detection of cancer. *Nature Reviews Cancer.* 3(4):267–275.

Wynder, E.L. & Muscat, J.E. 1995. CA - The changing epidemiology of smoking and lung cancer histology. *Environmental Health Perspectives.* 103:143–148.

Zeng, J., Zhan, P., Wu, G., Yang, W., Liang, W., Lv, T. & Song, Y. 2015. Prognostic value of twist in lung cancer : systematic review and meta-analysis. 4(6):236–241.



Chapter 2

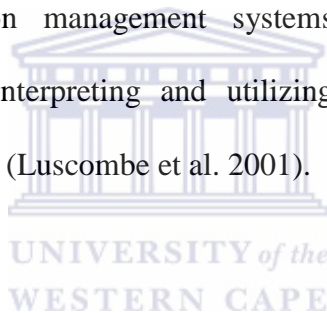
Identification of Potential Circulatory Biomarkers using Microarray Data

2.1. Introduction

The development of high-throughput technologies such as microarrays and Serial Analysis Gene Expression (SAGE) has led to a flood of cancer gene expression profiling data in the public domain (Gellert et al. 2010). Due to the large volume of data being generated, sifting through this data has become near impossible for the laboratory researcher. As a result, bioinformatics tools are used to provide methods to evaluate, integrate and translate the data (Rhodes & Chinnaiyan 2004). Thus, these information management systems aid in storing, extracting, organising, analyzing, interpreting and utilizing information from biological sequences and molecules (Luscombe et al. 2001).

2.2. Data Mining

The process of extracting or “mining” information from large amounts of data requires the application of specific algorithms for discovering novel correlations, trends and patterns from large amounts of data stored in computational warehouses (Fayyad et al. 1996; Raza 2012). Data mining is also often referred to as Knowledge Discovery in Databases (KDD) (Raza 2012). Although more accurately KDD (Figure 2.1) refers to the overall process of discovering useful knowledge from data, while data mining refers to a particular step in this process (Fayyad et al. 1996).



Data mining approaches are ideally suited for bioinformatics, due to data being collected and accumulated across a variety of fields at a rapid pace (Fayyad et al. 1996). Moreover, the mining of biological data assists in extracting useful knowledge from large datasets. Many applications of data mining include; gene finding, protein function domain detection, protein and gene interactions, disease diagnosis, disease prognosis and disease treatment optimization, to name a few (Raza 2012).

Post the evolution of microarrays and large scale databases of SAGE and expressed sequence tags (EST), bioinformatics tools can be used to integrate this public data in the search for potential biomarkers (Kim et al. 2007).

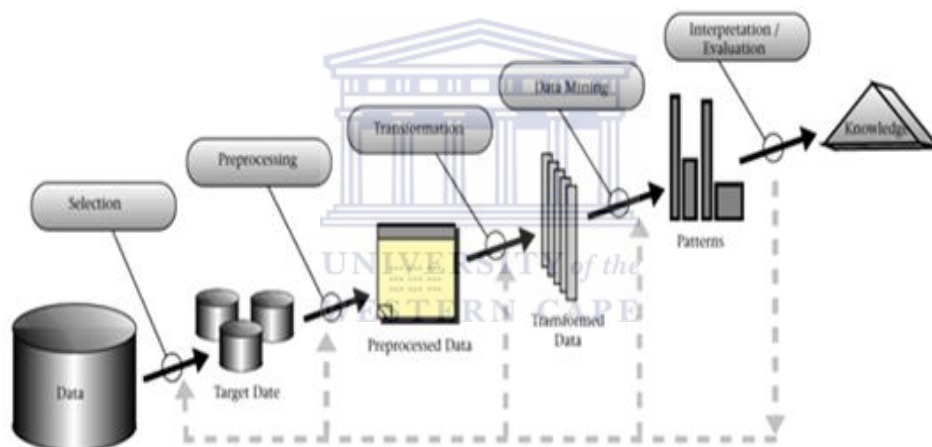


Figure 2.4: Steps involved in knowledge discovery (Fayyad et al. 1996).

2.2.1. Microarray Data Mining

Advances in DNA microarrays have revolutionized cancer research, resulting in numerous gene expression profiling studies and generating a number of potential biomarkers (Rhodes & Chinnaiyan 2004). Expression profiling allows for the simultaneous measurement of cellular concentration of different mRNAs (Guo et

al. 2013), in which these tissue and serum markers are reported to have the potential to aid in more accurate diagnosis, prognosis as well as the potential early diagnosis of the cancer and the effectiveness of therapy (Rhodes & Chinnaiyan 2004).

Microarrays, also known as gene chips, quantitatively measure relative expressed mRNA levels between different samples (Luscombe et al. 2001). Expression data measured using microarray technology arise from a single, large experiment in which a collection of gene standards are always included in order to normalize experimental data (Munoz et al. 2004).

At the beginning of the microarray era, bioinformatics tools were focused on unsupervised clustering, aiming to discover novel properties of data structure. More recently however, the interest of analysis of data has shifted to more supervised and guided analysis, focusing on differentially expressed genes (DEG) under various known conditions (Phan et al. 2009).

Lists of candidate biomarkers generated from microarray data analysis depend on both the availability of samples as well as selection algorithms. High-throughput (HT) assay platforms are typically comprised of thousands of genes, making the interpretation of their results a daunting task. The association of candidate genes with biological functions aids in the process of understanding underlying mechanisms of the relevant disease and the biological relevance of the feature selection algorithm. Subsequently, Gene Ontology (GO) is used to facilitate interpretations of gene functions on a large scale (Phan et al. 2009).

2.2.2. Digital Expression Profiling using EST and SAGE

The estimation of protein expression levels remains a significant area of interest in both genomics and proteomics. Protein expression levels indicate links between the genetics and an organism's functional property, with average levels of protein expression allowing environments within cells to be determined. Changes in these levels can provide information regarding; developmental biology, stress-response and progression of disease (Munoz et al. 2004).

Large scale sequencing of cDNAs provides a complementary approach to structural analysis of the human genome by generating ESTs (Okubuku, et al., 1992). These fragments of mRNA sequences are derived through single sequencing reactions performed on randomly selected clones from cDNA libraries. Currently over 45 million ESTs have been generated from over 1 400 different species of eukaryotes. EST projects are generally used to complement existing genome projects or serve as low-cost alternatives for gene discovery (Parkinson & Blaxter 2009).

Unlike microarray studies which pools data from one large experiment, ESTs arise from the entire database, which is constructed from various experiments performed under different conditions which often examine different subsets of genes of interest (Munoz et al. 2004).

Consequently, gene expression levels can be determined by measuring mRNA levels using expressed cDNA sequence tag sequencing and SAGE tag sequencing (Raza 2012).

SAGE is a method of obtaining quantitative absolute gene expression profiles from cells under selected physiological conditions (Margulies & Innis 2000; Luscombe et al. 2001). Unlike arrays, SAGE does not require any prior knowledge of genes to be analyzed (Hu & Polyak 2006). In the original EST approach, tags are 100 - 300 nucleotides in length. SAGE, however only requires 9 nucleotides, therefore, allowing for a larger throughput (Audic & Claverie 1997).

2.3. Biological Databases

Databases are ordered collections of data, generally stored in one or more associated files. The data is stored in tables, allowing cross referencing between them, with existing relationships among these tables producing a relational database (Niland & Rouse 2010).



2.3.1. Oncomine

Oncomine (<http://www.oncomine.org>) is a public cancer microarray database and web-based data-mining platform. Its primary aim is to facilitate discovery from genome-wide expression analyses (Rhodes et al. 2004). Oncomine incorporates 65 gene expression datasets consisting of approximately 50 million gene expression measurements from over 4 700 microarray studies (Rhodes & Chinnaiyan 2004). Differential analyses in Oncomine compares cancer tissues with their respective normal type (Rhodes et al. 2004). Genes most under- and overexpressed are defined by over 100 differential expression analyses in nearly every major cancer as well as various clinical and pathology based subtypes (Rhodes & Chinnaiyan 2004).

2.3.2. Gene Expression Atlas

The Gene Expression Atlas (GEA) (www.ebi.ac.uk/gxa/) launched by the European Bioinformatics Institute (EBI) is a public database which allows users to query gene expression under various biological conditions, including different cell types, developmental stages, physiological states, phenotypes and disease states. GEA content is derived from curation and statistical analysis of selected data from the ArrayExpress Archive of Functional Genomics Data (Kapushesky et al. 2009). To date, GEA contains data from over 200 000 genes of 9 different species and over 1 000 different independent studies, with the database being updated on a monthly basis (Kapushesky et al. 2009).

2.3.3. Integrative OncoGenomics

Integrative OncoGenomics (IntOGen)(<http://www.intogen.org/>) is a web platform, which provides support to researchers and aids in identifying tumour drivers in various cohorts. IntOGen identifies and visualizes cancer drivers, analyzing 4 623 exomes from 13 cancer sites. Somatic mutations, genes and pathways involved in cancer have been summarized and made available for public curation (Gundem & Perez-Llamas 2010; Gonzalez-Perez & Perez-Llamas 2013).

2.3.4. C-It

C-It (<http://c-It.mpi-bn.mpg.de>) is a knowledge database that focuses on genes previously uncharacterised. The database contains expression profiles of various tissues, including human, mouse, rat, chicken and zebrafish. C-It is designed to

provide a comprehensive coverage of gene expression patterns and tissue-enriched splicing isoforms (Gellert et al. 2010).

Included in the C-It database is literature information from PubMed, assisting in the identification of genes lacking publication records. Tissue expression data of ESTs are used to identify tissue-enriched genes and microarray and SAGE data provide comprehensive transcriptional profiles (Gellert et al. 2009; Gellert et al. 2010).

2.3.5. Tissue-specific Gene Expression and Regulation

Tissue-specific Gene Expression and Regulation (TiGER) (<http://bioinfo.wilmer.jhu.edu/>) is a publicly available database which provides large scale data sets for tissue-specific gene (TSG) expression and regulation in various human tissues. The database includes three types of data, namely; tissue-specific gene expression profiles, combinatorial gene regulations, and cis-regulatory module (CRM) detections. TiGER currently contains expression profiles for 19 526 UniGene genes, combinatorial regulations for 7 341 transcription factor pairs and 6 232 putative CRMs for 2 130 reference sequencing (RefSeq) genes (Liu et al. 2008).

2.3.6. VeryGene

VeryGene (<http://www.verygene.com/>) is a knowledge database of tissue-specific genes. The VeryGene web platform integrates TSGs from large-scale data analyses with respective information on subcellular localization, GO, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway, Mouse Genome

Informatics (MGI) Mammalian Phenotype, disease association, and targeting drugs. The database presently consists of 3 960 annotated TSGs derived from 127 normal human tissues and cell types, including 5 672 gene-disease and 2 171 drug-target relationships (Yang et al. 2011).

2.3.7. Database for Annotation, Visualization and Integrated Discovery

Database for Annotation, Visualization and Integrated Discovery (DAVID) (<http://david.abcc.ncifcrf.gov/>) is a HT knowledge platform which aims to provide a functional interpretation of large gene lists derived from genomic studies. The database integrates a multitude of public bioinformatics resources to combine tens of millions of diverse gene/protein identifiers and annotation terms from a variety of public bioinformatics databases (Huang et al. 2007). The grouping of identifiers improves cross-reference capability, enabling more than 40 publicly available functional annotation sources to be comprehensively integrated and utilized by the DAVID gene clusters (Sherman et al. 2007).

DAVID is able to provide GO analysis, as well as condense large gene lists into gene functional groups and convert between gene/protein identifiers. By mapping genes to GO terms and then statistically highlighting the most enriched, increases the likelihood that the biological process of interest will be identified (Huang et al. 2007).

2.4. Text Mining

Scientific literature represents a rich source for knowledge retrieval on associations between biomedical concepts such as genes, diseases and cellular

processes (Frijters et al. 2010). Biomedical text mining allows researchers to identify relevant information more accurately. Text mining facilitates establishment of relationships hidden within large amounts of available biomedical data currently in literature. It moves the burden of information overload from the researcher to the computer by the application of algorithmic, statistical and data analysis methods with (Cohen 2005) with various databases existing and facilitate text mining.

2.4.1. Text Mining Databases

2.4.1.1. The Universal Protein Knowledgebase

The Universal Protein Knowledgebase (UniProtKB) provides a comprehensive resource for protein sequences and functional information. UniProtKB (<http://www.uniprot.org>) consists of two sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. UniProtKB/Swiss-Prot contains manually annotated records with information extracted from literature and computational analysis. The annotation consists of; function, enzyme-specific information, biologically relevant domains and sites, post-translational modifications, sub-cellular location, tissue specificity, developmental specific expression, structure, interaction and associated diseases, deficiencies or abnormalities (The UniProt Consortium 2010).

2.4.1.2. PolySearch

PolySearch (<http://wishart.biology.ualberta.ca/polysearch>) is a public web tool, specifically designed for extracting and analyzing text-derived relationships between human diseases, genes/proteins, mutations, drugs, metabolites, pathways,

tissues, organs and sub-cellular localizations (Liu 2013). PolySearch extracts and analyses not only PubMed data, but also text data from multiple databases (DrugBank, SwissProt, HGMD, Entrez SNP, etc.). PolySearch utilizes various techniques in text mining and information retrieval to identify, highlight and rank informative abstracts, paragraphs or sentences (Cheng et al. 2008).

2.4.1.3. Human Genome Epidemiology Network

The Human Genome Epidemiology Network (HuGENet) (<http://www.hugenavigator.net/>) maintains a database of published, population-based epidemiologic studies of human genes extracted from PubMed. The introduction of machine learning search strategies have reduced the labour intense manual curation and increased both the sensitivity and specificity of the screening (Yu et al. 2008). The database is updated weekly with articles newly added to PubMed and has to date indexed more than 30 000 articles, referenced more than 3 000 genes and indexed nearly 2 000 disease terms article with Medical Subject Headings (MeSH) terms and gene information from the National Center for Bioinformatics (NCBI) Entrez Gene database (Yu et al. 2008; Yu et al. 2009).

2.4.1.4. Google Scholar

Google Scholar (<https://scholar.google.co.za/>) is a subset of the Google search engine, consisting of full-text journal articles, technical reports, preprints, theses, books, and other documents, including selected Web pages. Google Scholar encompasses a diverse range of topical areas, but is deemed to be strongest in the sciences. Google Scholar's index is obtained from a crawl of full-text journal content of both commercial and open source publishers. It retrieves document or

page matches based on the keywords searched and then organizes the results using a closely guarded relevance algorithm. Since much of Google Scholar's content comes from licensed commercial journal content, search results may reveal only an abstract and not full text articles (Vine 2006).

Aims:

1. Data mining of several cancer databases (as outlined above) for the extraction of potential circulating biomarkers for the early diagnosis of lung cancer
2. Refining the compiled genes, using literature mining tools to generate a candidate gene list



2.5. Materials and Methods

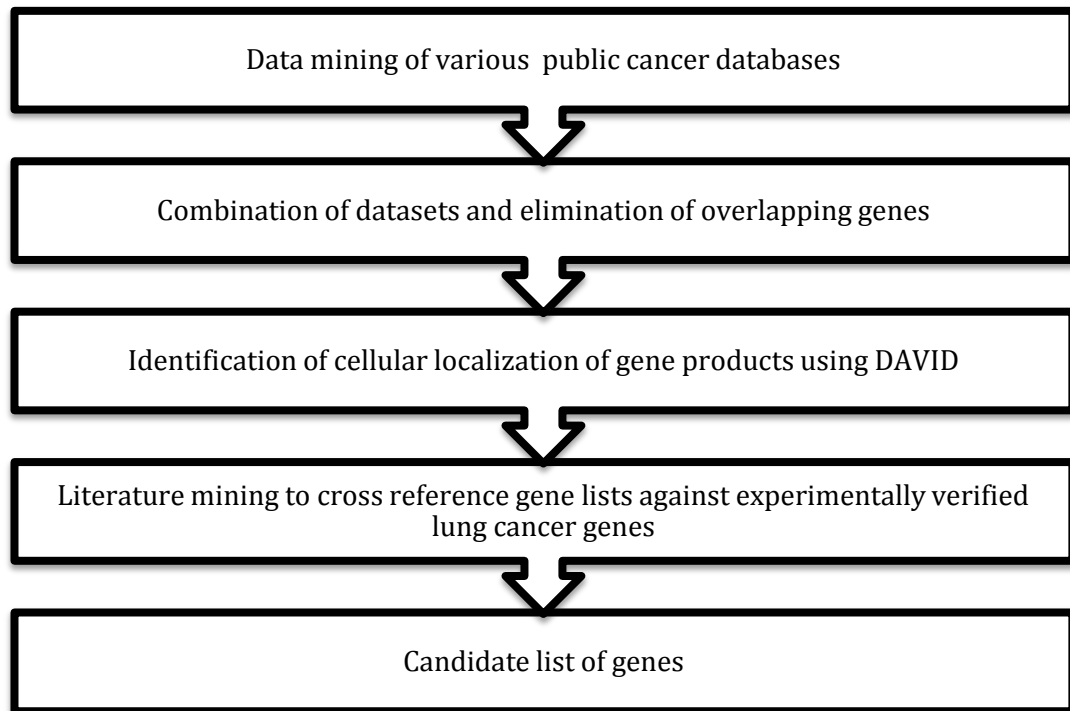


Figure 2.5: Outline of the methodology for biomarker discovery.

2.5.1. Extraction of Candidate Gene Biomarkers

The purpose of the cancer biomarker discovery pipeline analysis was to retrieve genes which were differentially expressed (DE) in lung cancer in comparison to normal lung tissue and to combine and filter these genes into a feasible gene list. In addition to identify genes specific for potential biomarkers found in the circulatory system. A bioinformatics approach was used to integrate public cancer databases.

In this study, six databases were mined to identify novel genes highly expressed in lung cancer. Querying multiple databases were used to help overcome the shortcomings which are associated with using only one methodology or a single data type (Prassas et al. 2012). Input parameters used were key words/phrases that

allowed for the extraction of genes highly specific to lung cancer tissue e.g. < lung cancer>, <homo sapiens>, <differential analysis> etc.

The bioinformatics pipeline was divided into two main sections:

- I. Data mining of publicly available databases (candidate and reference gene lists), and
- II. Literature mining of experimentally validated lung cancer-associated genes.

The gene extraction pipeline was followed according to the protocol described by Prassas et al. (2012), for the identification of tissue-specific serological biomarkers.

2.5.1.1. Oncomine Database

Oncomine was mined for differentially expressed genes in lung cancer with the following input query:

- *Analysis type*: Differential analysis, cancer vs. normal
- *Cancer type*: Lung cancer.
- *Data type*: mRNA
- *Pathology subtype*: Stage and grade type.

All datasets containing both up and down-regulated genes with respect to lung cancer were extracted.

2.5.1.2. Gene Expression Atlas Database

Differentially expressed genes were queried in GEA using the following input parameters:

- *All genes*
- Up/down in
- Homo sapiens
- Lung cancer

GO terms relating to cellular components (CC) (e.g. cytoplasm, integral to plasma membrane) were used to further refine the list of differentially expressed genes retrieved.



2.5.1.3. IntOGen Database

The Integrative OncoGenomics database was searched for genes, which were shown to be mutated in lung cancer. <Lung> was selected as the cancer site query and all experiments were selected for retrieval with <all> genes/modules.

2.5.1.4. C-It Database

The C-It database was searched for genes/proteins enriched in lung tissue. The query was specified for human data only. The following literature information search parameters were selected:

- Fewer than five publications in PubMed and

- Fewer than three publications with the MeSH.

2.5.1.5. TiGER Database

The TiGER database was mined for ESTs. <Lung> was selected under *Tissue View* for the acquisition of relevant genes.

2.5.1.6. VeryGene Database

VeryGene was searched using *Tissue View* for human lung tissue specific/enriched genes.

2.5.1.7. Excluded Databases

Datasets from BioGPS (<http://biogps.org/>) and the Cancer Genome Characterization Initiative (CGAP) (<http://cgap.nci.nih.gov/>) were excluded, as the databases contained no available data on lung cancer.

2.5.2. Analysis of Gene Lists

The candidate gene list consisting of 12 combined datasets was submitted to DAVID Version 6.7 (<http://david.abcc.ncifcrf.gov/>) for Gene Enrichment analysis, with all duplicated genes deleted prior to submission.

2.5.2.1. Functional Characterization using DAVID

All datasets were submitted to DAVID for clustering and functional annotation by means of a 3 step process:

Step 1. Uploading Gene List of Interest

"Start Analysis" was selected and subsequently the candidate gene list was uploaded. "OFFICIAL_GENE_SYMBOL " was chosen as the identifier of choice and "gene list" for viewing purposes.

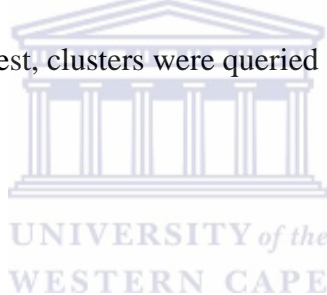
Step 2. Analysis of Candidate Genes

The uploaded gene list was analysed using, "Functional Annotation Clustering", selected from DAVID's functional annotation tools. Class stringency was set to medium, with the following selected: display, fold change and Bonferoni Analysis. The dataset was then rerun using the newly selected options.

Step 3. Annotational Clustering of Genes

GO terms were used to select annotation clusters. Since the Cellular Component was the GO term of interest, clusters were queried using the following terms:

- cell surface
- secreted,
- secretory granules
- extracellular matrix
- extracellular space
- extracellular membrane.



A sub-list of the newly acquired genes was created, exported and saved.

2.5.3. Literature Mining of the Candidate Entities

Literature mining was used to eliminate genes already experimentally linked to lung cancer, in an attempt to ensure that genes selected as potential biomarkers

would be novel. The databases used were: Uniprot, Polysearch, Google Scholar and HuGENavigator.

Genes were searched for by entering the gene name with the Boolean term “AND” and the cancer of interest e.g. <lung cancer> AND <gene name>. All relevant literature, abstracts and journal articles, were searched for information linking the genes as biomarkers for lung cancer. All genes found to be experimentally validated or suggested as lung cancer biomarkers were eliminated and a final candidate gene list was subsequently created.

2.6. Results and Discussion

2.6.1. Identification of Eligible Cancer Biomarkers

The methodology of mining several cancer databases was used to identify DE genes that encode proteins secreted into bodily fluids with potential application as biomarkers of early diagnosis in lung cancer. Genes are usually considered as potential targets or markers if they are highly over expressed in a particular type of cancer (Rhodes et al. 2004).

For each database utilized, specific criteria, tools and data mining steps were used to increase stringency and reduce the volume of data retrieved. All queries into respective databases followed the protocols as outlined in 2.5.1.1 to 2.5.1.7.

Mining of the Oncomine microarray database for genes differentially expressed in lung cancer in comparison with normal lung tissue resulted in the identification of 590 genes. Genes were ranked-ordered by the p-value and seed lists of the top 1 %, 5 % and 10 % relative expression were retrieved resulting in a total of 6 output

gene lists. Datasets were categorized based on the different levels of gene expression as compared to its normal counterpart (Table 2.1.). A total of 1 749 DE genes were retrieved from Oncomine, and subsequent combining of all seed lists and eliminating duplicates, 1 159 genes remained.

Table 2.9: Summary of genes extracted from Oncomine.

Relative expression compared to normal tissue counterpart	Number of genes extracted
1 %	122
5 %	547
10 %	1 080

GEA is a database consisting of high quality microarray experimental data. Querying this platform searched for genes up or down regulated in lung cancer tissue and included CC GO terms that generated 25 039 genes (Table 2.2.). The use of GO annotation provided a platform for the discovery of potential markers that were up-regulated in cancers and are used to further filter analysis. Following curation of the list a total of 10 512 DE genes were formatted.

Table 2.10: Summary of genes extracted from GEA based on GO terms.

GO Term	Number of Genes
Protein Binding	8 631
Cytoplasm	4 961
Nucleus	5 087
Cytosol	2 594
Integral to Membrane	3 766

GEA microarray data is sourced from ArrayExpress, a generic microarray database designed to hold data from all microarray platforms (Brazma et al. 2003). Biological relevance is ensured by comparing expression in healthy and the relevant diseased tissue, maintaining a minimum of 3 sample replicates and providing both p-values and t-statistics for all microarray analyses (Petryszak et al. 2014) IntOGen is a platform which displays somatic mutations identified in various cancers. Copy number changes and changes in gene expression were used to identify cancer drivers in the tissue of choice (Gonzalez-Perez et al. 2013). Mining the IntOGen database initially produced 53 466 genes, but elimination of duplicate genes resulted in 2 934 unique genes (Table 2.3).

C-It, VeryGene and TiGER are databases containing tissue specific enriched genes. C-It focused on uncharacterised tissue-enriched gene variants and TSGs, and generated 1 819 unique genes. While TiGER and VeryGene are both based on ESTs, each had a unique data output of 117 and 92 genes, respectively. Even though databases are based on similar sources of data, individual databases still identified unique outputs, further validating the initial approach of mining several databases (Table 2.3).

Table 2.11: Summary of genes extracted from databases.

Database	# Genes Identified	# Duplicated Genes	# Unique Genes
Oncomine	1 749	590	1 159
GEA	25 039	14 527	10 512
IntOGen	53 466	50 532	2 934
C-It	2 708	889	1 819
TiGER	156	39	117

VeryGene	94	2	92
Total Number of Combined Genes	83 212	66 579	16 633

2.6.2. Gene Enrichment Analysis

DAVID allows for the extraction of biological meaning from large gene or protein lists by using text and pathway mining tools (Huang et al. 2007). A total of 16 633 genes were uploaded to DAVID for gene enrichment analysis with a resultant output of 117 genes generated.

Enrichment analyses of GO terms: biological process (BP), cellular component (CC) and molecular function (MF) was performed on the 117 genes using the functional clustering annotation tools.

GO is a set vocabulary as stipulated by NCBI, applied to the functions of genes and proteins (Dennis et al. 2003).

Statistical significance of GO terms was analysed using the p-values (<0.05) and false discovery rate (FDR <0.05), which corresponds to a 95 % confidence of enrichment. Default options of medium/high classification stringency were used, and cluster names were extracted from the most biologically relevant GO term assigned to each cluster.

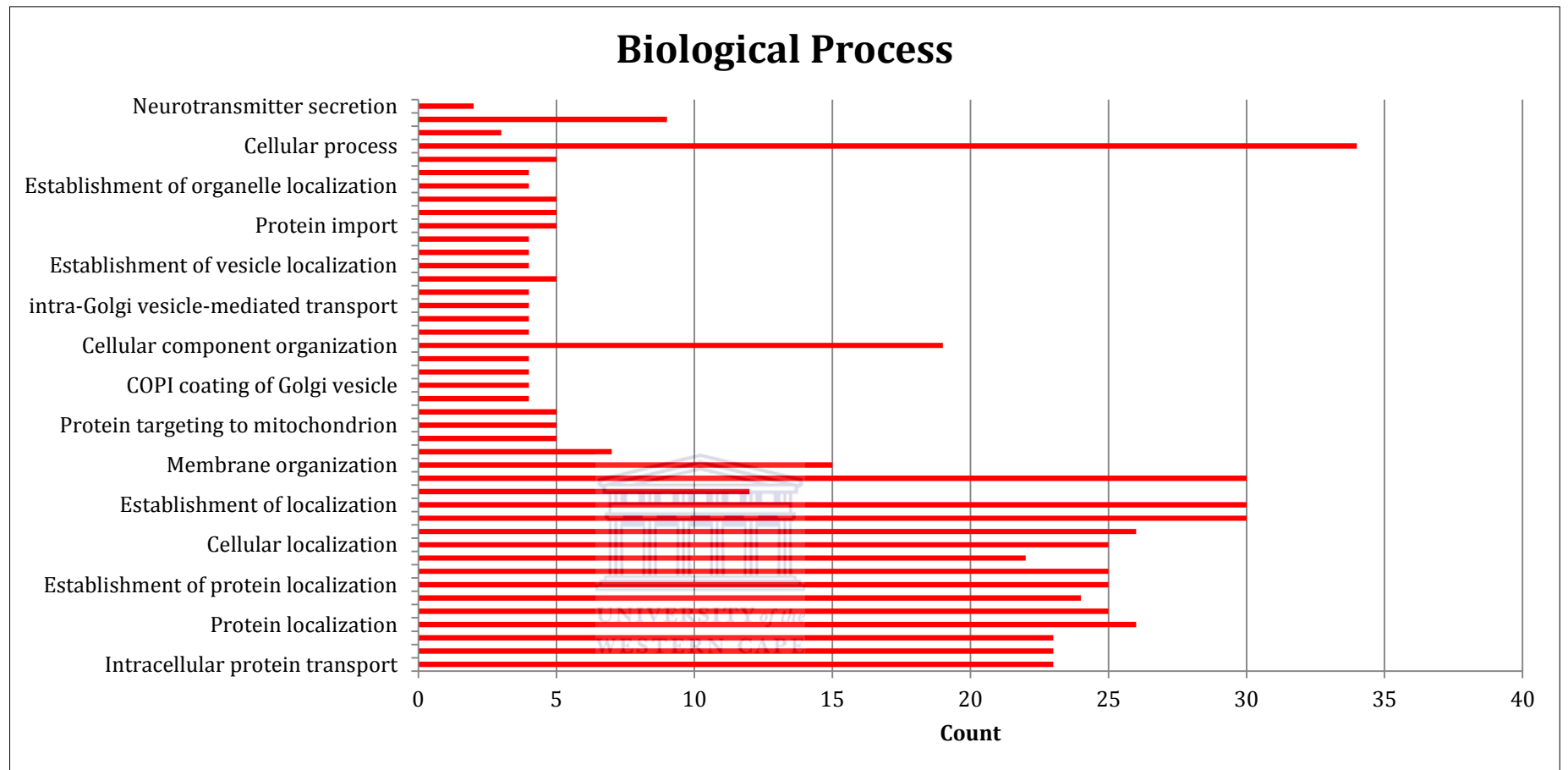


Figure 2.6: Functional characterisation of genes in DAVID based on their biological process using GO analysis.

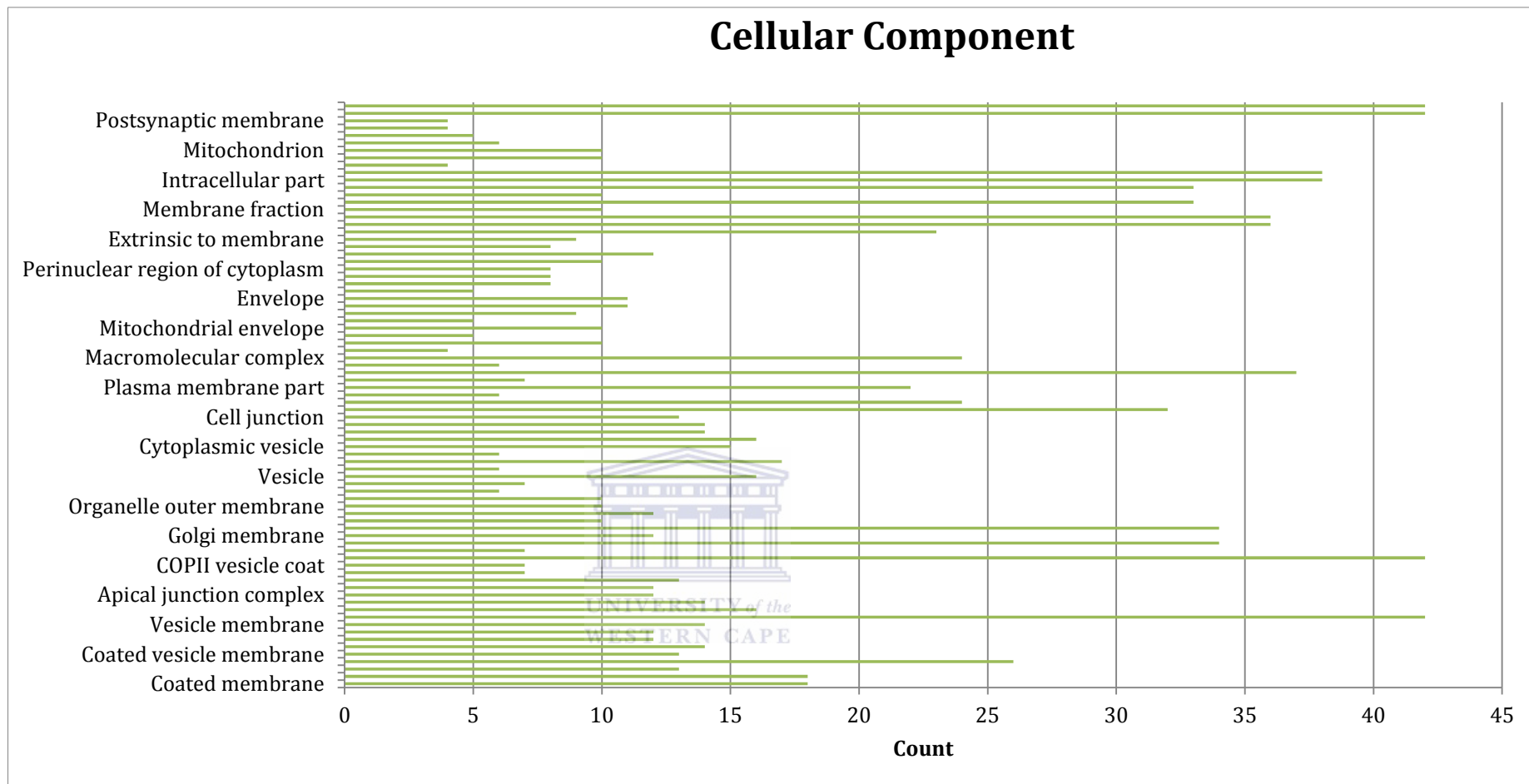


Figure 2.7: Functional characterisation of genes in DAVID based on their cellular component using GO analysis.

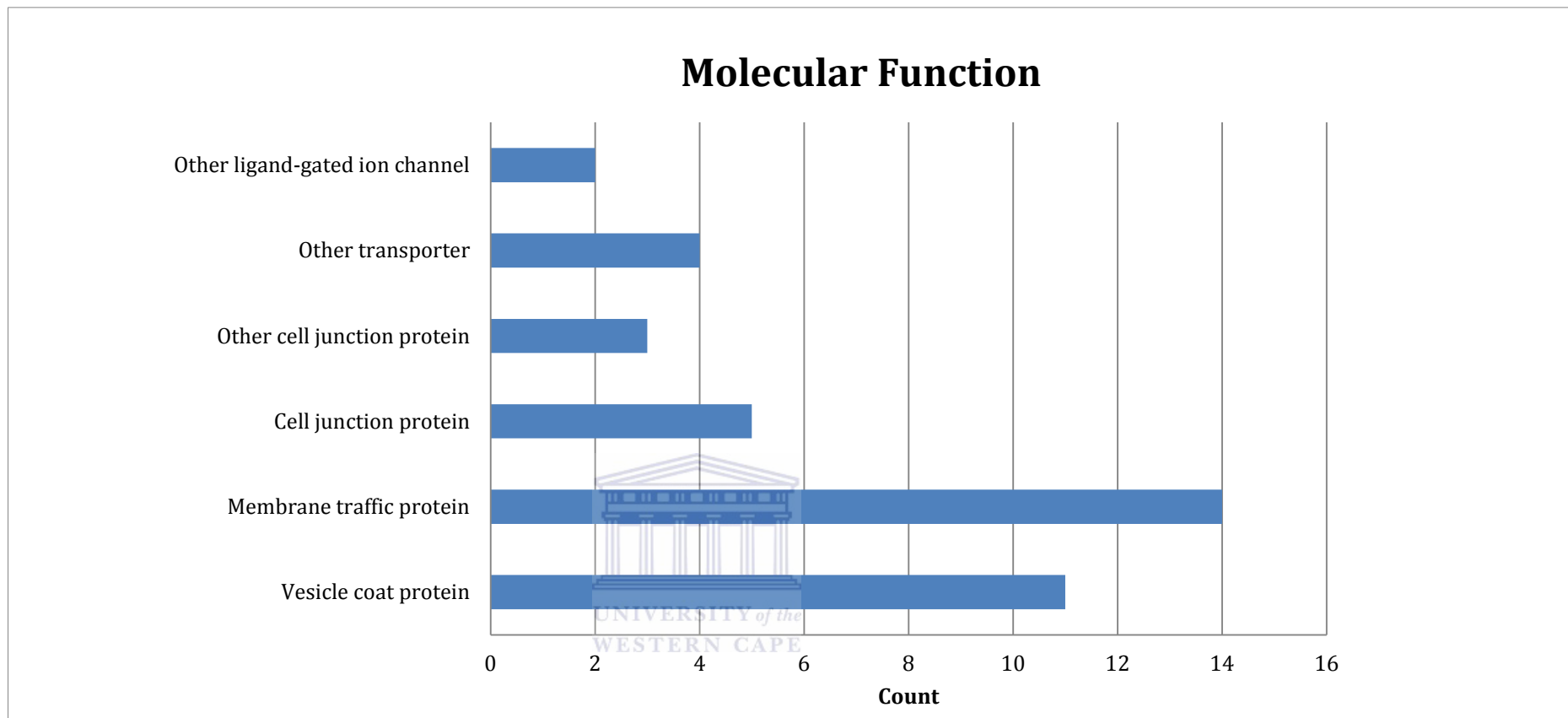


Figure 2.8: Functional characterisation of genes in DAVID based on their molecular function using GO analysis.

2.6.3. Literature Mining of Candidate Genes

Following functional clustering and GO annotation in DAVID, the list 117 genes were investigated using literature mining, in order to obtain a subset of genes of greater relevance to be validated as novel potential biomarkers for lung cancer. Text mining was performed using the following databases; Uniprot, PolySearch, Google Scholar and HuGENavigator. All cited literature; articles, abstracts and references linked to genes of interest were carefully scrutinized. Genes not yet experimentally validated as having any involvement in lung cancer were chosen for the new subset of candidate genes. Of the 16 633 genes queried in DAVID, only 117 genes met the criteria selected for this study. Literature mining further reduced this number to 20 candidate genes not experimentally linked to lung cancer. Further curation of these 20 candidate genes was performed by literature mining using the same databases. All genes found to be experimentally validated or linked to any other cancer was eliminated to further increase the stringency, resulting in a list of 4 candidate markers:

- COPZ1 (Coatomer Protein Complex, Subunit Zeta 1)
- SEC23B (*S. Cerevisiae* Homolog B)
- SEC24A (*S. Cerevisiae* Family Member A)
- SEC24D (*S. Cerevisiae* Family Member D)

2.7. Discussion and Conclusion

Lung cancer remains a serious health burden and the leading cause of cancer related deaths globally. A good prognosis as well as efficiency of treatment strongly depend on early stage diagnosis of the cancer (Risch & Plass 2008). However, most tumours are diagnosed at a late stage, presenting with distant metastasis. There is no validated screening method for lung cancer and the overall five-year survival rate of less than 10 % has not changed significantly in the last 20 years. There is therefore, a need for an early, rapid detection method which is both non-invasive and cost efficient (Brambilla et al. 2003).

Lung carcinogenesis is a multistep process involving the accumulation of genetic and epigenetic mutations. Much of what is known about the biological pathways and processes in tumorigenesis is derived from the investigation of genes and their functions (Stratton et al. 2009). Many genetic alterations which occur during tumour development disrupt paracrine signaling networks, resulting in the release of cancer cells from regular growth constraints. Cancer specific autocrine and paracrine signals is often accompanied by the inappropriate expression of secreted proteins or their receptors (Welsh et al. 2003).

Increasing evidence that the interaction and network between genes and proteins play a pivotal role in the understanding of the molecular mechanism of cancer has resulted in a systems approach to the study of this disease. The concept of systems biology into cancer research, integrates omics-based technology, clinical science, molecular biology, bioinformatics and computer science to aid in diagnosis, therapies and prognosis (Wu et al. 2012).

Cancer bioinformatics presents an essential tool to the process of early diagnoses and has the potential to play a critical role in the identification and validation of biomarkers, specific to clinical phenotypes (Chen et al. 2013).

Circulatory biomarkers would represent a non-invasive aid in the clinical management of cancer patients, particularly in areas of disease diagnosis, prognosis, monitoring and therapeutic stratification (Chi-Shing Cho 2007). The marker needs to be produced by the tumor or its microenvironment and enter the circulation, resulting in increased serum levels. For a serological biomarker to be ideal for early detection, its presence in serum must be low in healthy individuals. The mechanisms which facilitate the entry into circulation include secretion or shedding, angiogenesis, invasion and destruction of tissue architecture (Altintas & Tothill 2013). The biomarker would also need to be tissue specific such that a change in serum level can be directly attributed to lung cancer.

In this study, several *in silico* approaches were used to investigate high-throughput databases. Microarray platforms were queried and various bioinformatics tools were used to identify genes encoding secreted proteins in human lung cancer. Lists of candidate biomarkers generated from microarray data analysis depend on sample availability as well as the respective selection algorithm. These lists may often vary from sample to sample or be highly unstable (Phan et al. 2009). When investigating gene databases for this analysis, stringency was set to high for all platforms, so as to filter the number of genes generated to a more specific group of interest.

Data generated from these platforms typically consists of tens of thousands of genes, making interpretation of their results a daunting task. The association of candidate genes with biological functions, assists in understanding underlying mechanisms of the associated disease as well as relevance of feature selection algorithms (Harris et al. 2004). The gene-annotation enrichment analysis (HT strategy) increased the likelihood of identification of biological processes most relevant to the specific study. Bioinformatics methods, using GO allowed for the systematic dissection of large gene lists in an attempt to assemble a summary of the most enriched and pertinent candidates (Huang et al. 2009).

GO collected biological knowledge in a gene-to-annotation format, suitable for HT bioinformatics scanning for enrichment analysis. The tools allowed for systematic mapping of large lists of genes of interest, associated with biological annotation terms (GO Terms), which then statistically examined the enrichment of gene members for each of the annotation terms by comparing the outcome to the control (Smith et al. 2003; Huang et al. 2009).

The GO covers 3 domains: (1) MF, which describes activities, such as catalytic or binding activities, and represent activities rather than the molecules or complexes performing the actions. It does not specify where, when or in what context the action takes place. (2) BP describes the biological goals achieved by one or more ordered assemblies of molecular functions and can be used to describe processes such as apoptosis or chromatin condensation. (3) CC describes the sub-cellular structures and macromolecular complexes locations (Harris et al. 2004)

GO cellular component annotations (Figure 2.4.) were of most relevance to this analysis as it allowed for grouping according to mechanisms that facilitate biomarker entry into the circulation (Huang et al. 2007). Significance of these terms is given by their assigned p-values, which denotes the probability of a term occurring in the set by chance or not. This GO enrichment analysis with very high significance ($p < 0.05$), represents a set of genes highly similar in its properties. The methodological approaches previously described resulted in the identification of 16 633 genes found to be highly expressed in lung cancer tissue. Further enrichment analysis in DAVID through annotation and sequence analysis-based approaches, generated 117 candidate genes.

GO analysis of these genes showed the majority were enriched for CC (intrinsic and integral to membrane and cell surface) (Figure 2.4.). These results proved promising since the targeted biomarkers for this subsection of the study were those, which would be easily detectable in bodily fluids. Results of MF (Figure 2.5.) were consistent with BP (Figure 2.3.) showing a large majority of the genes to be involved in membrane trafficking, transport and localization.

The strategy of mining gene and protein databases was described by Prassas and colleagues (2012). Mining of databases for proteins highly specific to or strongly expressed in a specific tissue, selects proteins which are secreted or shed to prioritize candidates for further validation (Prassas et al. 2012). Even though proteomics provides a more functional approach than genomics, proteomics alone might be insufficient as post-translational modifications, such as phosphorylation, control the stability and function of many proteins (Welsh et al. 2003). Mining

both gene and protein platforms with different data sources (ESTs and microarrays) aimed to minimize the limitations of each resulting in the identification of more specific markers (Prassas et al. 2012).

Literature mining was used to further reduce the generated list of candidate markers. These genes/protein have been well studied but not as potential cancer biomarkers and thus represent potential candidates. The emphasis of this study was to identify novel candidates, which have not been experimentally linked or validated in lung cancer or any other mechanisms of carcinogenesis.

This *in silico* analysis identified four genes; SEC 23B, SEC 24A, SEC 24D and COPZ1, pending validation, as early diagnostic biomarkers for lung cancer. Laboratory based validation of data would provide independent, experimental validation of gene expression levels. There are several molecular methodologies available to validate these results such as, RT-PCR, northern blot, RNase protection assay, and *in situ* hybridization or immunohistochemistry using tissue microarrays (Chuaqui et al. 2002). More specifically, real-time RT-PCR, which quantitatively measures specific mRNAs could be used to validate expression patterns. These genes identified *in silico* will be linked to specific cell function in Chapter 4 and compared to the genes filtered for Next Generation Sequencing (NGS) (Chapter 3).

2.8. References

Altintas, Z. & Tothill, I. 2013. Biomarkers and biosensors for the early diagnosis of lung cancer. *Sensors and Actuators B: Chemical*. 188:988–998.

Audic, S. & Claverie, J.M. 1997. The significance of digital gene expression profiles. *Genome Research*. 7(10):986–995.

Brambilla, C. et al. 2003. Early detection of lung cancer: role of biomarkers. *European Respiratory Journal*. 21(39):36–44.

Brazma, A. et al. 2003. ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*. 31(1):68–71.

c-IT. Available: <http://c-IT.mpi-bn.mpg.de> [2014, June, 23].

Chen, J., Zhang, D., Yan, W., Yang, D. & Shen, B. 2013. Translational bioinformatics for diagnostic and prognostic prediction of prostate cancer in the next-generation sequencing era. *BioMed Research International*. (8):90-157.

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

Chi-Shing Cho, W. 2007. Potentially useful biomarkers for the diagnosis, treatment and prognosis of lung cancer. *Biomedicine & Pharmacotherapy*. 61(9):515–519.

Chuaqui, R.F. et al. 2002. Post-analysis follow-up and validation of microarray experiments. *Nature Genetics*. 32:509–514.

Cohen, A.M. 2005. A survey of current work in biomedical text mining. *Briefings*

in *Bioinformatics*. 6(1):57–71.

Database Database for Annotation, Visualization and Integrative Discovery (DAVID). Available: <http://david.abcc.ncifcrf.gov/> [2014, August,12].

Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H. & Lempicki, R. 2003. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biology*. 4(9).

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. 1996. From data mining to knowledge discovery in databases. *AI Magazine*. 37–54.

Frijters, R., van Vugt, M., Smeets, R., van Schaik, R., de Vlieg, J. & Alkema, W. 2010. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Computational Biology*. 6(9).

Gene Expression Atlas (GEA). Available: <http://ebi.ac.uk/gxa> {2014, June, 21}.

Gellert, P., Uchida, S. & Braun, T. 2009. Exon Array Analyzer: a web interface for Affymetrix exon array analysis. *Bioinformatics*. Available: <http://bioinformatics.oxfordjournals.org/content/25/24/3323>. [2014, October 10].

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

Gonzalez-Perez, A. & Perez-Llamas, C. 2013. IntOGen-mutations identifies cancer drivers across tumor types. *Nature*. Available: <http://www.nature.com/articles/nmeth.2642> [2014, October 11].

Gonzalez-Perez, A. et al. 2013. IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods*. 10(11):1081–2.

Gundem, G. & Perez-Llamas, C. 2010. IntOGen: integration and data mining of multidimensional oncogenomic data. *Nature*. Available: <http://www.nature.com/articles/nmeth0210-92> [2014, October 21].

Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D. & Shyr, Y. 2013. Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PloS One*. 8(8).

Harris, M. A. et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*. 32:258–261.

Hu, M. & Polyak, K. 2006. Serial analysis of gene expression. *Nature Protocols*. 1(4):1743–1760.

Huang, D.W. et al. 2007. DAVID Bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*. 35:169–175.

Huang, D.W., Sherman, B.T. & Lempicki, R. a. 2009. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*. 37(1):1–13.

Human Genome Epidemiology Network (HuGeNet). Available: www.hugenavigator.net [2014, July, 17].

Integrative OncoGenomics (IntOGen). Available: <http://intogen.org/> [2014, July, 16].

Kapushesky, M. et al. 2009. Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Research*. 38:690–698.

- Kim, B. et al. 2007. Clinical validity of the lung cancer biomarkers identified by bioinformatics analysis of public expression data. *Cancer Research*. 67(15):7431–7438.
- Liu, Y. 2013. The University of Alberta participation in the BioASQ challenge: The Wishart system. *Proceedings of the first Workshop on Bio-Medical*. Available: http://bioasq.org/sites/default/files/wishart_system.pdf [2014, October 21].
- Liu, X., Yu, X., Zack, D.J., Zhu, H. & Qian, J. 2008. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*. 9(1):271.
- Luscombe, N.M., Greenbaum, D. & Gerstein, M. 2001. What is bioinformatics? An introduction and overview. *Yearbook of Medical Informatics*. 83–100.
- Margulies, E.H. & Innis, J.W. 2000. Gene expression with Serial Analysis of Gene Expression (SAGE). *Bioinformatics*. 16(7):650–651.
- Munoz, E.T., Bogarad, L.D. & Deem, M.W. 2004. Microarray and EST database estimates of mRNA expression levels differ: The protein length versus expression curve for *C. elegans*. *Nature Biotechnology*. 6:1–6.
- Niland, J.C. & Rouse, L. 2010. *Biomedical Informatics for Cancer Research*. Available: <http://link.springer.com/10.1007/978-1-4419-5714-6> [2015, October 21].
- Oncomine. Available: <http://www.oncomine.org/> [2014, June, 17].
- Parkinson, J. & Blaxter, M. 2009. Expressed sequence tags: an overview. *Methods in Molecular Biology (Clifton, N.J.)*. 533:1–12.

Petryszak, R. et al. 2014. Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Research*. 42:926–32.

Phan, J.H., Moffitt, R. A., Stokes, T.H., Liu, J., Young, A.N., Nie, S. & Wang, M.D. 2009. Convergence of biomarkers, bioinformatics and nanotechnology for individualized cancer treatment. *Trends in Biotechnology*. 27(6):350–358.

Polysearch. Available: <http://wishart.biolog.ualberta.ca/polysearch> [2014, August, 20].

Prassas, I., Chrystoja, C.C., Makawita, S. & Diamandis, E.P. 2012. Bioinformatic identification of proteins with tissue-specific expression for biomarker discovery. *BMC Medicine*. 10:39.

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

Rhodes, D.R. & Chinnaiyan, A.M. 2004. Bioinformatics strategies for translating genome-wide expression analyses into clinically useful cancer markers. *Annals of the New York Academy of Sciences*. 1020(1):32–40.

Rhodes, D.R. et al. 2004. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression . *Proceedings of the National Academy of Sciences of the United States of America*. 101(25):9309–9314.

Risch, A. & Plass, C. 2008. Lung cancer epigenetics and genetics. *International Journal of Cancer*. 123(1):1–7.

Sherman, B.T. et al. 2007. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics*. 8(1):426.

Smith, B., Williams, J. & Schulze-Kremer, S. 2003. The ontology of the gene ontology. *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*. 609–613.

Stratton, M.R., Campbell, P.J. & Futreal, P.A. 2009. The cancer genome. *Nature*. 458(7239):719–724.

The UniProt Consortium. 2010. The universal protein resource (UniProt) in 2010. *Nucleic acids research*. 38:142–8.

Tissue -specific Gene Expression and Regulation (TiGER). Available: <http://bioinfo.wilmer.jhu.edu/> [2015, October, 18].

Universal Protein Knowledge Base (UniProt KB). Available: <http://uniprot.org> [2014, November, 10].

VeryGene. Available: www.verygene.com [2014, July, 19].

Vine, R. 2006. Google scholar. *J Med Libr Assoc*. 94(1):97-99.

Welsh, J.B. et al. 2003. Large-scale delineation of secreted protein biomarkers overexpressed in cancer tissue and serum. *Proceedings of the National Academy of Sciences of the United States of America*. 100(6):3410–5.

Wu, D., Rice, C.M. & Wang, X. 2012. Cancer bioinformatics: A new approach to systems clinical medicine. *BMC Bioinformatics*. 13(1):71.

Yang, X., Ye, Y., Wang, G. & Huang, H. 2011. VeryGene: linking tissue-specific

genes to diseases, drugs, and beyond for knowledge discovery. *Physiological*. Available: <http://physiolgenomics.physiology.org/content/43/8/457.short> [2014, October 11].

Yu, W., Gwinn, M., Clyne, M., Yesupriya, A. & Khoury, M. 2008. A navigator for human genome epidemiology. *Nature Genetics*. Available: <http://www.nature.com/ng/journal/v40/n2/full/ng0208-124.html> [2015, October 21].

Yu, W., Clyne, M., Khoury, M.J. & Gwinn, M. 2009. Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics*. 26(1):145–146.



Chapter 3

Identification of Potential Circulatory Biomarkers using RNAseq Data

3.1. Introduction

Cancer presents in various forms depending on; the location, cell of origin as well as the range of genomic alterations, which promote oncogenesis. Many genomic events with direct phenotypic correlation have been identified, yet the complex molecular landscapes remain uncharted for many cancer lineages (Chang et al. 2013). The inception of gene expression microarrays led to the possibility of genomic phenotype classification. Fundamentally however, two major problems have hindered this endeavour: (1) the inaccuracy of microarray measurements and (2) small sample sizes (Knight et al. 2014). Improvements in sequencing and array-based profiling have resulted in an influx of diverse genome related data, including whole genome sequencing and exome based data, with expression profiling of both coding and non-coding RNAs and SNPs (Robinson et al. 2011; Guo et al. 2013). Analysis of these large, diverse datasets holds the potential for a comprehensive understanding of the human genome and its relation to disease (Robinson et al. 2011).

3.2. Next Generation Sequencing

Next Generation Sequencing (NGS), also known as massively parallel signature sequencing (MPSS) has revolutionised the characterisation of cancer at the genomic, transcriptomic and epigenetic levels. This technology has allowed for

cataloguing of mutations, copy number aberrations and somatic rearrangements in an entire cancer genome at base pair resolution (Reis-Filho 2009). NGS can provide unbiased transcriptomic analysis of mRNAs, small RNAs and non-coding RNAs, genome-wide methylation assays and high throughput (HT) chromatin immuno-precipitation assays (Reis-Filho 2009). Whole genome sequencing allows a deeper understanding into the full spectrum of genetic variation as well as phenotypic variation and pathogenesis (Mamanova et al. 2010).

3.2.1. RNA Sequencing

The introduction of RNA Sequencing (RNAseq) has revolutionised expression research. It refers to the use of NGS technologies to sequence cDNA to obtain information about the respective sample's RNA content (Guo et al. 2013). RNAseq allows for complete sequencing of transcriptomes in almost any tissue or population and is often used to measure gene expression (Davey et al. 2011).

RNAseq technologies sequence small mRNA fragments to measure gene expression and is viewed as the transcriptome analog to whole genome shotgun sequencing (Li et al. 2010; Knight et al. 2014). However, RNAseq is primarily used to estimate the copy number of transcripts in a sample (Li et al. 2010). During a standard RNAseq experiment, an RNA sample is converted to cDNA fragments and sequenced by a commercially available HT platform. Raw data consists of large amounts of sequences of DNA fragments, called reads, that undergo a series of steps of analysis including; mapping the reads, summarizing each genes map counts, normalization and detection of differentially expressed

genes (DEG) (Oshlack et al. 2010; Kvam et al. 2012). Subsequently, gene expression is determined by measuring the number of reads mapped to a gene (Knight et al. 2014). RNAseq therefore provides a discrete measurement for gene expression, unlike fluorescence intensity measurements from microarray platforms. Consequently, new statistical methods are needed to appropriately handle the large volume of RNAseq data being generated (Kvam et al. 2012).

Detection of DEG is often the end goal of statistical analysis of RNAseq data and aids in elucidating gene function (Robinson & Oshlack 2010). They can also serve as an initial step in gene expression clustering profiling or gene set enrichment (Kvam et al. 2012). Since the recent advent of RNAseq technologies and its continuous development, no standard methods have yet been determined to detect DEG based on the data (Oshlack et al. 2010; Kvam et al. 2012).

In comparison to microarrays, the RNAseq method offers several advantages. The detection range of RNAseq is not limited to a set of predetermined probes as with microarray methods. RNAseq can detect expression at the gene, exon, transcript and coding DNA sequence (CDS) level while microarrays are limited to the gene level or the exon level for specially designed exon arrays. RNAseq is also able to detect structural variants such as alternative splicing and gene fusion (Guo et al. 2013).

3.2.1.1. RNAseq Version 2

RNASeq Version 2 (RNAseq V2) similarly to RNAseq uses mapped counts to determine gene expression levels, however a different set of algorithms are used to determine the expression levels (Li et al. 2010). Since the number of reads from a gene is a function of the length of the mRNA as well as its molar concentration, it is necessary to normalize the read count while preserving molarity (Pepke et al. 2009). Two analysis pipelines are used to create and normalize Level 3 expression data from this data.

The first approach relies on the Reads Per Kilobase of exon model per Million mapped reads (RPKM) method and is utilized in various databases such as The Cancer Genome Atlas (TCGA) (Li et al. 2010). RPKM quantifies gene expression from RNA sequencing data by normalizing for total read length and the number of sequencing reads, making them directly comparable within the sample (Mortazavi et al. 2008). The second method uses RNAseq by Expectation-Maximization (RSEM) for quantitation (Li et al. 2010). RPKM is most commonly used and is calculated using the formula: $RPKM = 10^9(C/NL)$, where C is the number of reads mapped to the gene, N is the total number of reads mapped to all genes and L represents the length of the gene (Guo et al. 2013). The key difference between RPKM and RSEM is that the normalization factor of RPKM is proportional to the mean length of a transcript unlike RSEM which is independent of the mean expressed transcript length (Guo et al. 2013; Li et al. 2010).

3.3. The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) (<http://cancergenome.nih.gov/>) is a public systematic cancer genomics project which applies emerging technologies to the analysis of specific tumour types (Chang et al. 2013). TCGA was launched by the NCI and the National Human Genome Research Institute (NHGRI) with the goal of improving the ability to diagnose, treat and prevent cancer. TCGA provides molecular profiles at DNA, RNA, protein and epigenetic levels for various cancers and their subtypes as well as hundreds of clinical tumour samples (Chang et al. 2013).

Samples are characterized using technologies that evaluate the sequence of the exome; copy number variation (measured by SNP arrays), DNA methylation, mRNA expression and sequence, miRNA expression and transcript splice variation. Whole-genome sequencing may also be applied to a subset of the tumors (Kandoth et al. 2013). As of July 2013, TCGA had molecularly mapped patterns across 7992 cases, which represented 27 different tumour types. TCGA aims to have analyzed the genomic, epigenomic and gene expression profiles of more than 10 000 specimens from over 25 various tumour types by the end of 2015 (Chang et al. 2013). TCGA has currently archived 497 and 470 specimens from lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) biopsies, respectively. The datasets are composed of level 3 RNAseq V2 data which are mapped read counts against 20 531 known human RNA transcripts (Knight et al. 2014).

3.4. Bioinformatics Analysis Tools

3.4.1. Bioinformatics Enrichment Tools

Current enrichment tools are categorized into three classes; singular enrichment analysis (SEA), gene set enrichment analysis (GSEA), and modular enrichment analysis (MEA) based on their respective algorithms (Huang et al. 2009). SEA measures expression levels in each gene individually; MEA adopts this same enrichment calculation, while also incorporating a network discovery algorithm (Huang et al. 2009; Laukens et al. 2015). GSEA evaluates gene set data, considering expression profiles from samples belonging to the two aforementioned classes (Subramanian et al. 2005) and is commonly used in the analysis of differential expression data, providing greater statistical power compared to SEA methodologies (Draghici et al. 2007).

Changes and regulation of genome-wide genes can be measured simultaneously using HT technologies. These approaches typically generate large gene or protein lists as their final output (Berriz et al. 2009; Huang et al. 2009). The challenge lies in translating these results within the context of their underlying biological mechanisms (Laukens et al. 2015). Bioinformatics enrichment approaches may facilitate identification of these pertinent processes and pathways (Subramanian et al. 2005). Biological processes involve several genes as opposed to a single gene alone, forming the foundation of enrichment analysis. If these mechanisms are altered or abnormal, co-functioning genes should have a greater potential to be

selected as a relevant group when analyzed (Huang et al. 2009). Enrichment tools query lists of DEG against prior knowledge gene-set libraries. Gene-set libraries organize and store functional knowledge, such as pathways and transcription factors of each gene in the group (Chen et al. 2013). Most of these analysis tools focus on mapping genes to associated biological annotation terms (e.g. GO or pathway) and then statistically compute enrichment (Draghici et al. 2007; Alaimo et al. 2015). Regardless of their specific features, three main layers can characterize all of these tools, namely; backend annotation database, or data support, data mining which includes algorithms and statistics, and result resenatation (Figure 3.1) (Huang et al. 2009).

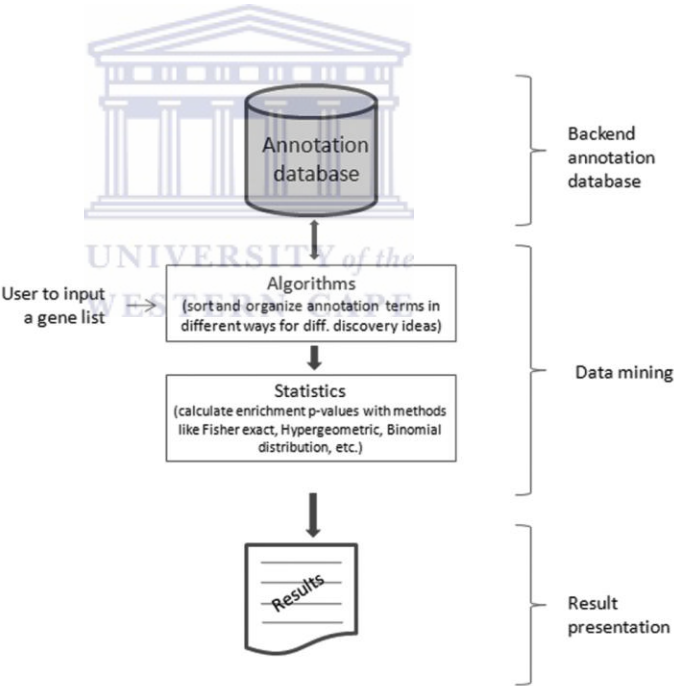


Figure 3.1: The typical infrastructure of enrichment tools with three distinct layers: backend annotation database, data mining, and result presentation (Huang et al. 2009).

3.4.1.1. MultiExperiment Viewer

MultiExperiment Viewer (MeV) is a java based free software application, which utilizes modern bioinformatics tools for integrative data analysis. MeV v4.9 (<http://tm4.org/mev.html>) is a component of the TM4 Microarray Software Suite which incorporates sophisticated algorithms for clustering, visualization, classification, statistical analysis, and biological theme discovery from single or multiple experiments (Howe et al. 2010).

Robust statistical methods and data analysis tools are imperative to users of RNAseq data. MeV allows users to load raw or normalized data and supplies a variety of algorithms for clustering, classification and statistical analysis (Saeed et al. 2006). Currently 24 analysis techniques are available in MeV. These algorithms are broadly categorized into three types based on the objectives they aim to accomplish, namely; exploratory techniques, hypothesis testing and classification (Howe et al. 2011).

Once data is loaded MeV generates an expression matrix, which is a two-dimensional array of expression elements from genes (Figure 3.2). Each row is an expression vector from a specific gene and each column corresponds to a given experiments expression vector. Typically low expression is indicated in green and high expression in red (Saeed et al. 2006).

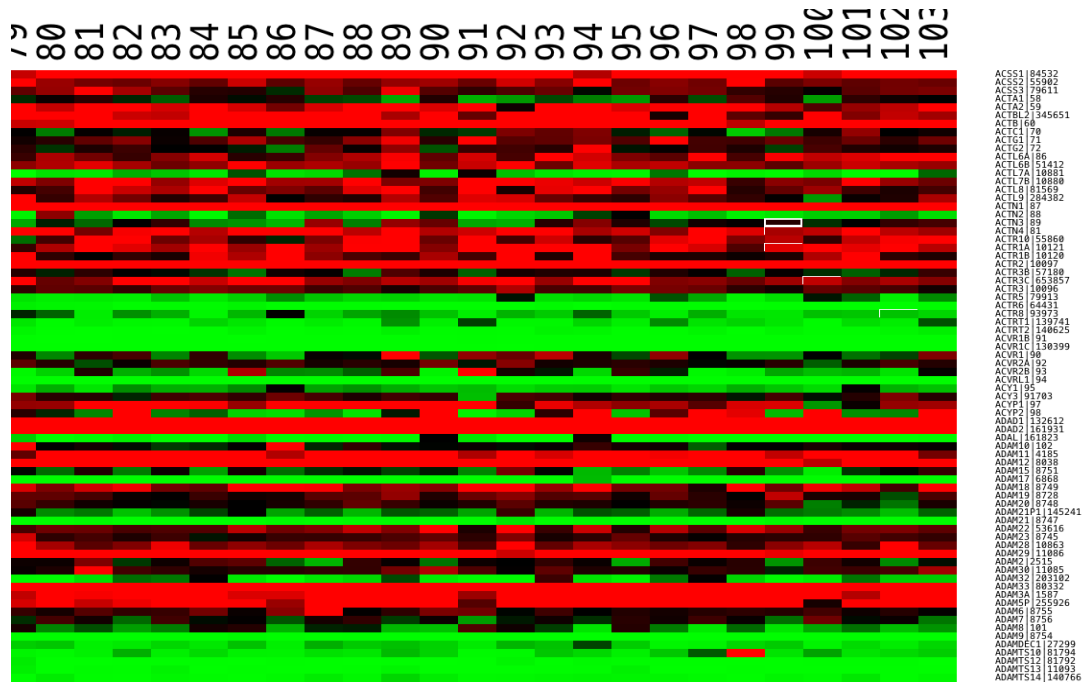


Figure 3.2: Expression matrix displaying rows of genes with high (red) and low (green) expression in relation to samples in each column generated by MeV v4.9 (<http://tm4.org/mev.html>).

3.4.1.2. Enrichr

Enrichr, (<http://amp.pharm.mssm.edu/Enrichr/>) is a HTML5 web-based, bioinformatics enrichment analysis application. This tool provides a novel approach to rank enriched terms and displays innovative, interactive visualizations of results. A total of 35 gene-set libraries are available, totaling 31 026 gene-sets, which encompass both the human and mouse genome and proteome. Each gene-set consists of approximately 350 genes, with more than six million connections between genes and terms. Some libraries are incorporated from other tools or databases while many remain exclusive to Enrichr. Libraries are divided into six categories: transcription, pathways, ontologies, diseases/ drugs, cell types and miscellaneous (Chen et al. 2013).

The three ontology trees, namely biological process (BP), cellular component (CC) and molecular function (MF), supply gene-set libraries to the ontology category, while well-known pathway databases provide knowledge to Enrichr's pathway category, which include, BioCarta, Kyoto Encyclopedia of Genes and Genomes (KEGG), WikiPathways and Reactome. Gene-set libraries created from kinase enrichment analysis (KEA), unique to this platform are also included in this category (Chen et al. 2013). The category of diseases/drugs incorporates libraries from the Connectivity Map database, the Gene Signatures Database (GeneSigDB), the Molecular Signatures Database (MSigDB), Online Mendelian Inheritance in Man (OMIM), and VirusMINT (Chen et al. 2013).

Enrichment scores assess significance of overlap between the input genes and the tool's available knowledge contained in the gene-set libraries. These evaluations include; p-values, z-score test statistics; and a combined score incorporating both of the former. With these features, Enrichr can be used to obtain a global view of cell regulation in cancer by comparing highly expressed genes in cancer tissues with their normal counterpart (Chen et al. 2013).

3.4.2. Databases and Platforms

3.4.2.1. Molecular Signatures Database

The Molecular Signatures Database (MSigDB) v5.0 (<http://www.broadinstitute.org/msigdb>) is a knowledge based repository containing over 6700 annotated gene sets (Liberzon et al. 2011). Developed and maintained by the Broad Institute to facilitate GSEA, incorporated tools allow for,

the computing of gene set overlaps, categorizing of gene families, and heat map visualization of expression profiles from referenced compendia (Liberzon 2014). MSigDB gene families include oncogenes, tumour suppressors, translocated cancer genes, transcription factors, protein kinases, homeodomain proteins, cell differentiation markers and cytokines/growth factors (Liberzon et al. 2011).

3.4.2.2. Gene Expression Atlas

Gene Expression Atlas (GEA) (www.ebi.ac.uk/gxa/) is an annotated database which provides gene and protein expression profiles from over 200 000 genes in various cell types, biological conditions, phenotypes and disease states (Kapushesky et al. 2009). The platform consist of high quality baseline and differential expression data obtained from microarray and RNAseq experiments (Petryszak et al. 2014).



Aims:

1. Collection of clinical LUAD samples and generation of RNAseq V2 Level 3 data for each stage of the disease as well as its normal tissue counterpart
2. Analysis of this data using bioinformatics feature selection and enrichment tools to identify relevant biological phenomena pertinent to the disease to aid in the development of early stage LUAD biomarkers

This chapter focused only on clinical data of adenocarcinomas of the lung. Since although lung cancer in general remains a global burden, LUAD specifically has surpassed all other lung carcinoma types to become the most common histologic subtype of this pathology. While most incidences of the disease are related to

smoking, LUAD still develops more frequently than any other lung cancer type, particularly in females who have never smoked (Travis et al. 2004).

3.5. Materials and Methods

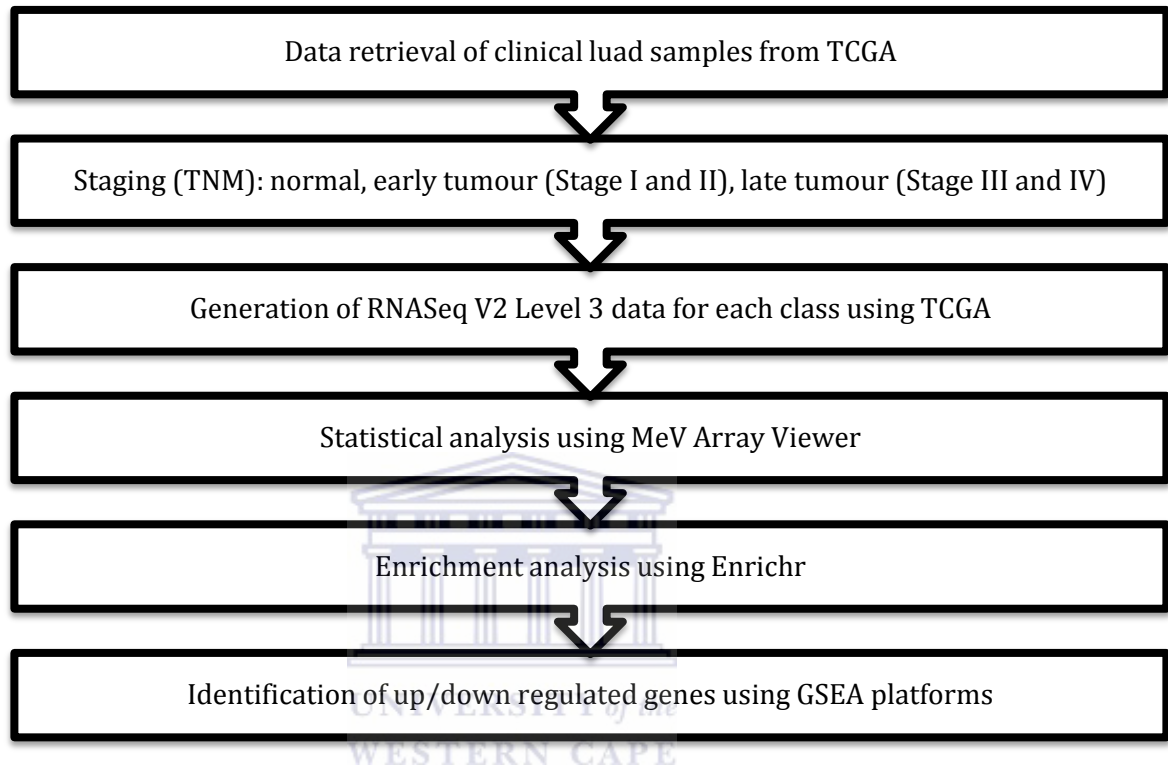


Figure 3.3: Methodological approach for the retrieval and analysis of lung adenocarcinoma RNAseq V2 Level 3 data from the Cancer Genome Atlas

3.5.1. Data Retrieval from TCGA

Clinical LUAD samples were collected from TCGAs Data Portal, via the Open-Access HTTP Directory using the following pipeline; *Lung adenocarcinoma (luad), bcr/biotab/clin/nationwidechildrens.org_clinical_patient_luad.txt*.

All data retrieved was filtered using TNM staging into three classes, normal (N), early (E) (Stage I & II) and late (L) (Stage III & IV) and uploaded to the Data Matrix using the following he following filters:

- *Disease:* Lung adenocarcinoma
- *Data type:* RNASeq V2
- *Data level:* 3
- *Access Tier:* Public
- *Availability:* Available data
- *Frozen preservation*
- *Tumour/normal filters* for the respective samples

RNASeqV2 archives were built as the output of choice, and gene based normalised results saved for each respective sample.

3.5.2. Analysis of Data using selected Bioinformatics Tools

3.5.2.1. Analysis using MultiExperiment Viewer

Data was uploaded to MeV Multiple Array Viewer from the *File* menu. The Tab delimited, Multiple Sample Files (TDMS) option was selected and the files of interest were uploaded. The upper-leftmost cell, containing an expression value was selected and the data was imported. Log 2 transform was selected under the Adjust data tab. The Cluster Manager tab was used to cluster samples as E, L and N, using previously outlined criteria to allow statistical analysis between samples. Changes in expression between samples, was determined by performing parametric t-Tests between unpaired samples. Three variations of this test were implemented, using different false discovery corrections (FDC) and p-value parameters, (Table 3.1) to determine p-values for each class (E vs. N, E vs. L, N

vs. E). Fold change (FC) was used as a complementary step to evaluate significance, with $FC > 2$ used to indicate relevance. Data sets of interest were selected as group 1 and 2 and significant genes were determined for each class. For each class, DEG were deemed significant to this study only if they were identified using all three FDC methods implemented (Table 3.1). Since this study aims to identify markers useful in early stage diagnosis, enriched genes in classes, E vs. N and E vs. L were selected for further investigation.

Table 3.12: Statistical parameters implemented to identify DEG between LUAD N, E and L samples using MeV

	t-Test 1	t-Test 2	t-Test 3
p-value parameters	p = 0.01 based on t-distribution	p = 0.05 based on t-distribution	p = 0.05 based on permutation, randomly group: 100x
p-value/ FDC	Just alpha (α) no corrections	Standard Bonferroni correction	maxT, proportion of false genes not exceed 0.05
Hierarchical clustering (HCL)	Hierarchical trees for significant genes only	Hierarchical trees for significant genes only	Hierarchical trees for significant genes only
HCLTree selection	Gene tree Sample tree	Gene tree Sample tree	Gene tree Sample tree
Ordering Optimization	Gene leaf order, Sample leaf order	Gene leaf order, Sample leaf order	Gene leaf order, Sample leaf order

3.5.2.2. Enrichment Analysis using Enrichr

DEG identified in MeV, as statistically significant in classes E vs. N and E vs. L were further evaluated to determine whether they contained any commonality before conducting enrichment analysis. Any genes identified in both groups, were eliminated due to lack of specificity to early stage diagnoses. These genes would

be involved in both developmental stages of the cancer and not present the ideal early stage biomarker. The final, curated E vs. N gene list was then uploaded to Enrichr. Data generated from the gene-set libraries of; Pathways, Ontologies, and Diseases/Drugs were investigated using computed p-value, z-score and combined test statistics generated. Results were viewed as network visualizations and histograms to better understand interactions and putative mechanisms involved in each analysis.

3.5.2.2.1. Ontology Annotation Sources

The Gene Ontology (GO) (<http://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz>) provides functional descriptions of gene products in relation to their cellular location and involvement in biological processes and molecular functions.

To identify genes likely to be secreted into the extracellular exome, and therefore serve as potential serum markers, CC GO terms of, extracellular space, extracellular region and extracellular exome were used.

The Pathway Ontology (<http://www.ncbi.nlm.nih.gov/pubmed/24499703>) characterises biological pathways, including altered and disease pathways in relation to gene expression.

The Disease Ontology (<http://www.ncbi.nlm.nih.gov/pubmed/26093607>) is a standardized vocabulary aimed at providing descriptions of human disease phenotype characteristics.

The Human Phenotype Ontology

(<http://www.ncbi.nlm.nih.gov/pubmed/22961259>) aims to provide correlations between biological data in relation to disease and depicts human disease phenotypic features.

The Mammalian Phenotype Ontology (<http://www.ncbi.nlm.nih.gov/pubmed/24217912>) allows for annotation of mammalian phenotypes in the context of genotypic variations and gene knockout models, and are used as models of human disease.

3.5.3. Gene Expression Analysis

3.5.3.1. Molecular Signatures Database

Candidate gene lists (Table 3.4) (Table 3.5) were queried in MSigDB to evaluate genes both under and over expressed across a multitude of cancer cell lines. Genes were referenced using the NCI-60 cell lines National Cancer Institute referendium and expression data was generated from the categories of cancer gene neighbourhoods and oncogenic signatures, derived directly from gene expression cancer profiles.

3.5.3.2. Gene Expression Atlas

Genes of interest and those identified as potential serum markers, (Table 3.4) (Table 3.5) were uploaded to GEA and searched against *homo sapiens* normal and matched experimental data of *lung adenocarcinoma*, and *all cancer cites*

respectively. Differential expression results generated up- and downregulated genes with log₂ FC values (FC > 2).

3.6. Results and Discussion

3.6.1. Data Collection from TCGA

The primary goal of gene expression profiling is to identify DEG. TCGA collected 497 clinical specimens from lung adenocarcinoma biopsies and expression data of 58 healthy lung specimens. TNM staging allowed tumour samples to be classified as either early stage or late stage lung adenocarcinoma, with stage I and II representing early stage, and III and IV late stage tumour, respectively (Detterbeck 2009). Categorizing specimens resulted in 394 early stage and 103 late stage specimens.

Level 3 RNA RNAseq V2 data generated 20 531 genes expressed in healthy lung tissue, 5371 in early stage carcinogenesis and 20 351 in late stage LUAD, respectively (Figure 3.4). Comparison of expressed data showed that 15 160 of the genes are common to both normal lung tissue and late stage tumour, while 5371 genes were found in all stages of disease and healthy samples.

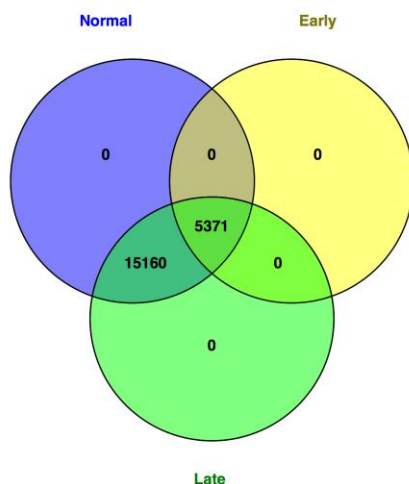


Figure 3.4: Gene expression data of LUAD between classes (N, E and L) generated from TCGA Level 3 RNAseq V2 data depicted using Venny 2.0.2 - Computational Genomics Service (bioinfo.gp.cnb.csic.es/tools/venny).

3.6.2. Statistical Analysis using MultiExperiment Viewer

DEG of paired LUAD RNAseq V2 Level 3 data from TCGA were identified using MeV and normalized using RPKM. Parametric, unpaired t-Tests with multiple FDC were used to evaluate statistical significance between samples. The null hypothesis of this analysis was that there was no difference between the means of the two groups, and was rejected at p-values selected for each test performed (Table 3.1).

Statistical significance of the DEG was evaluated by p-values generated for each sample (Subramanian et al. 2005). Hypothesis testing techniques such as t-Tests use information about the experimental design to identify a subset of genes that show statistical differences in patterns of expression across groups of samples (Saeed et al. 2006). Unpaired t-Tests compare the means of 2 independent samples. The null hypothesis, (presumed to be true), is that the 2 groups have the same average value with p-values indicating the accuracy of the null hypothesis (Cui et al. 2005). Small p-values denote a high probability of true difference in expression and a low probability that the observed difference occurred by chance (Morozova et al. 2008).

Multiple hypothesis testing, such as GSEA analyses large lists of DEG and may result in an increased false discovery rate (FDR) since multiple annotations are tested (Draghici et al. 2007). FDR is a multiple testing error, controlled using

multiple FDCs, which is an important step in the analysis of RNAseq data (Kvam et al. 2012). Corrections such as Bonferroni, adjust p-values derived from multiple statistical tests to correct for false positives (Li et al. 2010). Permutation tests re-samples n times the total number of observations, in a population sample, to build an estimate of the null distribution from which the test statistic has been drawn (Ge et al. 2003; Camargo et al. 2008). Fold change (FC) is a valuable complement to p-values and provides a way to assess differences between groups where p-values may be significant due to large sample numbers and low sample variability within groups, but the actual difference in the magnitude, or FC, between groups is low (Dudoit et al. 2002; Morozova et al. 2008).

The incorporation of multiple FDC and FC, into this study aimed to increase the statistical relevance of output data generated. For a gene to be significantly differentially expressed between the tumor and normal samples, it has to satisfy two conditions: FDR adjusted p-value and $FC > 2$ (Fonseca et al. 2014). The criteria were implemented in this study, to aid in identifying DEG with statistical relevance.

Analysis of samples E vs. N identified both common and unique DEG for all parametric tests performed (Figure 3.5). From this output, 171 genes were statistically significant to this study, as they were enriched for all FDC parameters implemented. Genes found to be unique to a single test, or only commonly expressed in two were excluded from further analysis. Evaluation of statistical analysis performed on samples E vs. L identified only one significant output,

which was generated using a FDC of only α , no corrections and a p-value threshold of 0.01 (Figure 3.6).

Analysis of N vs. L samples generated many outputs of interest, which may be pertinent to many biological systems (Figure 3.7). They were, however excluded from further investigation in this study, as a primary aim of this research was to identify DEG for the diagnosis of early stage LUAD. Statistically relevant genes from the tests of classes E vs. N and E vs. L were evaluated and compared to identify any overlap in data (Figure 3.8). No common genes were found between samples and therefore zero eliminated, resulting in a final DEG list of 171 candidates identified from the group E vs. N.



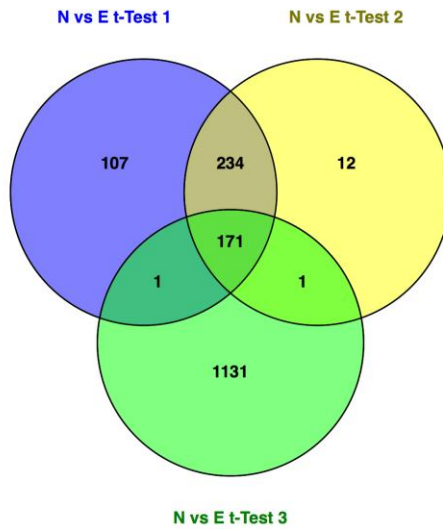


Figure 3.5: DEG identified in classes E vs. N using MeV parametric t-Tests and multiple FDC (t-Test 1: $p \leq 0.01$, no correction, t-Test 2: $p \leq 0.05$ and Bonferroni, t-Test 3: $p \leq 0.05$ and maxT) depicted using Venny 2.0.2 – Computational Genomics Service (bioinfogp.cnb.csic.es/tools/venny).

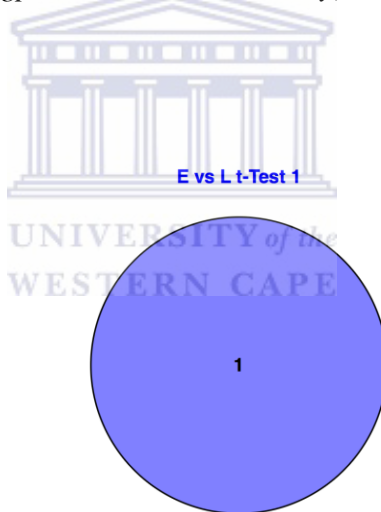


Figure 3.6: Unique DEG between classes E vs. L identified using MeV parametric t-Test 1, $p \leq 0.01$, depicted using Venny 2.0.2 – Computational Genomics Service (bioinfogp.cnb.csic.es/tools/venny).

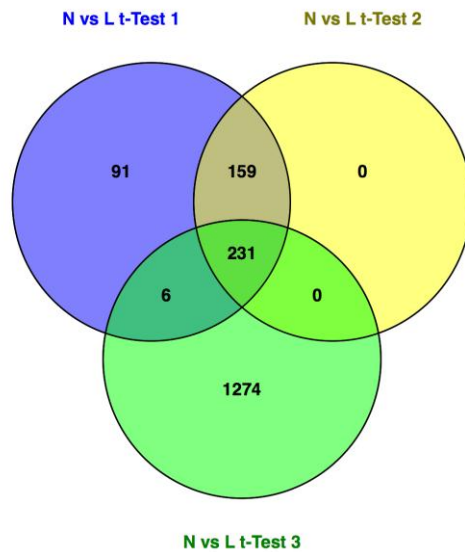


Figure 3.7: DEG identified between classes N vs. L using parametric t-Tests and multiple FDC (t-Test 1: $p \leq 0.01$, no correction, t-Test 2: $p \leq 0.05$ and Bonferroni, t-Test 3: $p \leq 0.05$ and maxT) in MeV, depicted using Venny 2.0.2 Computational Genomics Service (bioinfogp.cnb.csic.es/tools/venny).

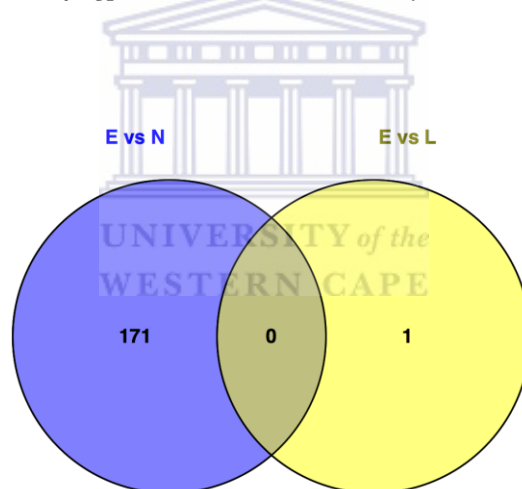


Figure 3.8: Unique DEG in classes E vs. N and E vs. L determined to be statistically significant following statistical analysis using MeV. Depicted using Venny 2.0.2 – Computational Genomics Service (bioinfogp.cnb.csic.es/tools/venny).

3.6.3. Enrichment Analysis using Enrichr Feature and Annotation Tool

The final candidate list of 171 DEG was uploaded to Enrichr for feature and annotation analysis. This tool computed enrichment for the input list against its

myriad of gene-set libraries (Huang et al. 2009). Comparison of enrichment signature patterns, between LUAD and its matched normal tissue counterpart, may aid in the identification of changes in critical biological regulatory mechanisms (Alexa et al. 2006). This knowledge may allow for the elucidation of disease specific genes involved in the process of carcinogenesis to be identified, enabling more accurate diagnosis (Huang et al. 2011). Enrichment scores were computed using three tests to rank relevance of queried genes against the gene set libraries. (1) The Fisher exact test, generated p-values, (2) a z-score test statistic, of the deviation from the expected rank of the Fisher exact test, and (3) a combined score, which multiplied the log of the given p-value by the computed z-score (Chen et al. 2013). The Fisher exact test, commonly used in GSEA, makes no assumption about sample size. This assumption may affect the ranking of terms, based only on the length of the gene set, the z-score statistic is computed as a correction for this possible bias (Bullard et al. 2010). The incorporation of these scores as well as the computation of a combined test statistic into each category, provided increased statistical power (Azuaje 2014). To determine relevance of enriched results for the purpose of this analysis, $p < 0.01$, and combined scores (CS) > 2 were denoted as significant, and gene sets, incorporating less than 5 overlaps were excluded. Characterisation of genetic expression patterns using ontologies, remains the most commonly used resource in uncovering molecular mechanisms associated with tumour initiation and progression (Young et al. 2010). Gene annotation provides a platform to facilitate the interpretation of gene signatures, in relation to its role in phenotypic diseases (Rhodes et al. 2007).

The Enrichr ontology annotation tool incorporates six gene-set libraries; GO Biological Process, GO Cellular Component, GO Molecular Function, MGI Mammalian Phenotype Level 3, MGI Mammalian Phenotype Level 4 and Human Phenotype Ontology. Querying of data against these libraries generated enriched terms from all but the latter (Table 3.2).

In the GO category of BP, terms of regulation of system process, secretion by cell, regulation of blood circulation and negative regulation of developmental growth were identified as enriched (Figure 3.9) (Table 3.2). Regulation and secretion are crucial to the cell-cycle, abnormal expression of factors involved in either of these processes have been identified in tumour formation (Lægreid et al. 2003).

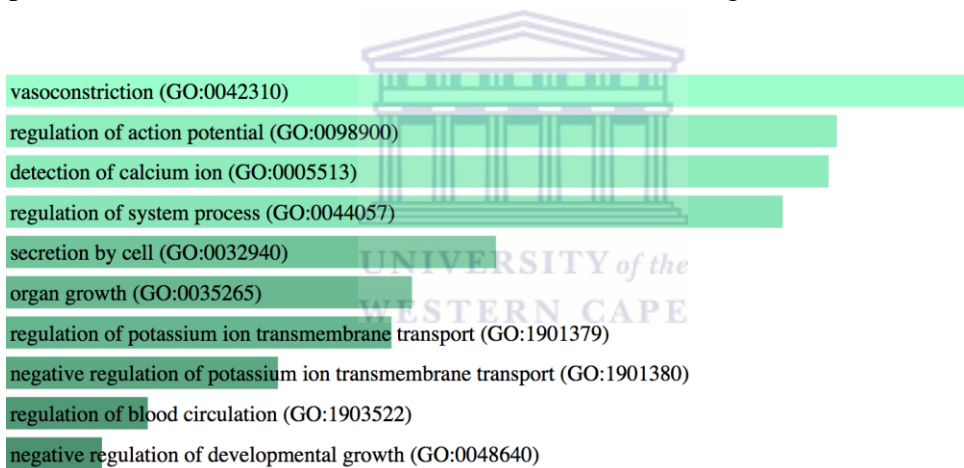


Figure 3.9: Histogram of enriched GO terms of BP generated from annotation analysis in Enrichr

Investigation of CC terms identified this category to be most highly expressed of all queried gene set ontologies (Figure 3.10) (Table 3.2). Terms, extracellular region and extracellular space were identified as the most highly ranked and statistically significant in terms of p-values and CS. Cell surface and extracellular

annotations, presented as terms of interest in this study, as serum proteins are excreted extracellularly (Lai et al. 2009). These terms allowed for correlation of gene expression patterns, according to mechanisms which facilitated entry into circulation (Huang et al. 2007). Elucidation of genes signatures relating to these sub-categories would facilitate the identification of potential circulatory LUAD markers (Nogales-Cadenas et al. 2009). Molecular Function described the tasks performed by individual genes (Sherman et al. 2007). Evaluation of the data identified receptor activator activity and receptor regulator activity to be highly enriched (Figure 3.11) (Table 3.2). Exopeptidase activity, was also identified in MF terms, and is known to increase as a tumour transforms and proliferates (Villanueva et al. 2006).

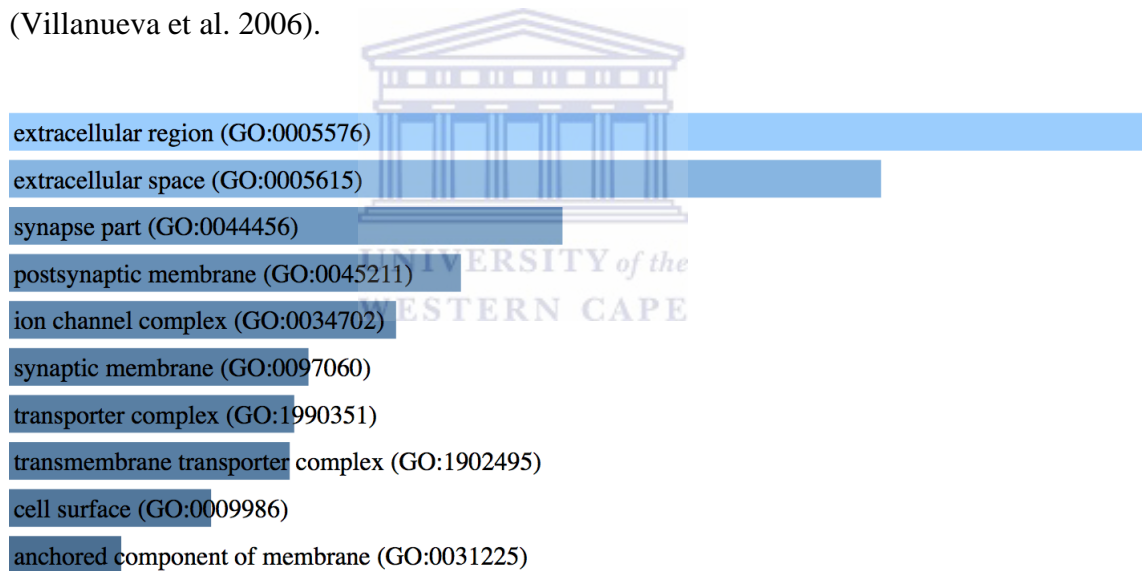


Figure 3.10: Histogram of enriched GO terms of CC generated from annotation analysis in Enrichr

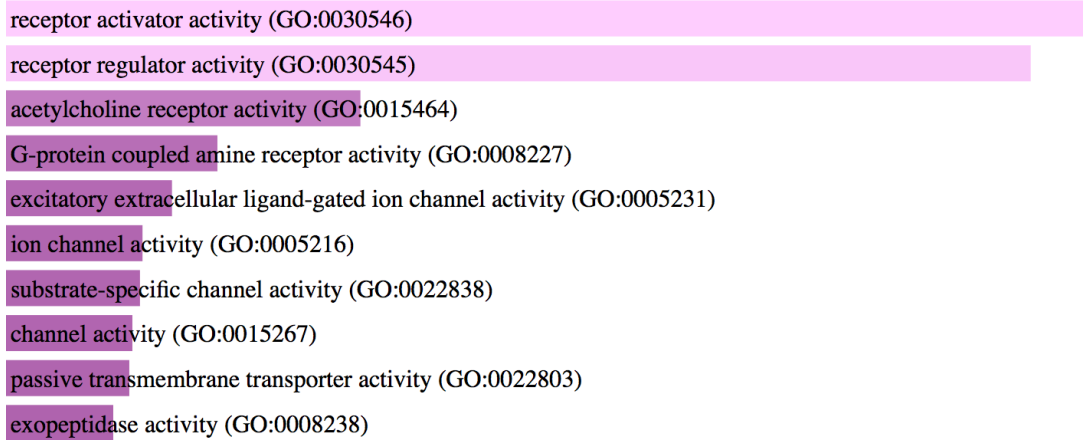


Figure 3.11: Histogram of enriched GO terms of MF generated from annotation analysis in Enrichr

Querying of the Midrand Graduate Institute (MGI) Mammalian Phenotype (MP) data enabled mammalian phenotypes to be annotated in the context of mutations and abnormal traits, used to model human disease biology. Different levels of phenotypic knowledge is supported and expressed, according to gene annotations (Smith et al. 2005). Investigation of Level 3 and 4 MP data revealed a trend in enriched abnormal neuronal terms (Table 3.2). Abnormal synaptic activity and abnormal neuron morphology was enriched in Level 4, while abnormal nervous system, presented with multiple terms in both levels. A pattern of abnormalities relating to the nervous system was seen in all output lists, and provided the analysis with an interesting theme (Table 3.2).

Table 3.2: Ontology enrichment terms extracted from Enrichr ($p < 0.01$, $(CS) > 2$)

Term	P-value	Combined Score (CS)	Source
Regulation of system process (GO:0044057)	0.000066301456224	8.43	GO Biological Process (BP)
Secretion by cell (GO:0032940)	0.000184527278327	8.15	
Regulation of blood circulation (GO:1903522)	0.000336913928736	7.81	
Negative regulation of developmental growth (GO:0048640)	0.000056995289897	7.77	
Extracellular region (GO:0005576)	0.000003314146713	21.77	GO Cellular Component (CC)
Extracellular space (GO:0005615)	0.000001231578227	18.93	
Synapse part (GO:0044456)	0.000027485228831	15.48	
Postsynaptic membrane (GO:0045211)	0.000043541600901	14.39	
Ion channel complex (GO:0034702)	0.000067021933900	13.68	
Synaptic membrane (GO:0097060)	0.000137750591689	12.74	
Transporter complex (GO:1990351)	0.000174426121335	12.58	
Transmembrane transporter complex (GO:1902495)	0.000152213467642	12.53	
Cell surface (GO:0009986)	0.000296352725425	11.68	
Anchored component of membrane (GO:0031225)	0.000298003137136	10.71	
Receptor activator activity (GO:0030546)	0.000006295227776	17.80	
Receptor regulator activity (GO:0030545)	0.000001903659722	17.10	
Ion channel activity (GO:0005216)	0.001313360572904	6.72	
Substrate-specific channel activity (GO:0022838)	0.001574493746172044	6.69	
Passive transmembrane transporter activity (GO:0022803)	0.002261191016990231	6.60	
Channel activity (GO:0015267)	0.002261191016990231	6.57	
Exopeptidase activity (GO:0008238)	0.002183918074263835	6.38	
Abnormal nervous system (MP0003633)	0.000006020503366089	10.37	MGI Mammalian Phenotype Level 3
Abnormal behavior (MP0004924)	0.000091394502025639	10.09	
Abnormal touch/nociception (MP0001968)	0.000422545186545143	9.33	
Abnormal muscle physiology (MP0002106)	0.001017379291353520	6.33	
Abnormal physiological response (MP0008872)	0.002716882239327337	5.57	
Abnormal nervous system (MP0003632)	0.000312940184740873	5.45	
Abnormal induced morbidity (MP0001657)	0.003231540392301920	5.07	

Abnormal synaptic transmission (MP0003635)	0.0000048882978824 83	10.40	MGI Mammalia n Phenotype Level 4
Abnormal nervous system (MP0002272)	0.0006146689696155 86	5.78	
Decreased physiological sensing (MP0008874)	0.0017821265753240 83	4.72	
Abnormal muscle fiber (MP0004087)	0.0015456066378186 70	4.44	
Abnormal neuron morphology (MP0002882)	0.0004290949439187 47	4.17	
Abnormal muscle contractility (MP0005620)	0.0031912080243396 20	3.69	
Abnormal pain threshold (MP0001970)	0.0052571379470811 96	3.49	

Carcinogenesis is a multifaceted phenomenon, involving changes in a multitude of cellular signaling mechanisms and pathways (Segal et al. 2004). Computational interrogation of specific regulatory networks of tumour cells, have revealed cryptic master regulator proteins, whose loss or gain affect the initiation and/or progression of carcinogenesis.

These proteins have proven to be powerful 'integrators' of various genetic and epigenetic alterations, which contribute to malignant phenotypes and therefore hold great promise in biomarker discovery (Schreiber et al. 2010). To date, a multitude of cancer pathways have been elucidated, and identified aberrations in regulation of key proliferation and survival pathways, common to all cancers (Segal et al. 2004; Subramanian et al. 2005).

Pathway analysis in Enrichr cross-referenced the candidate gene list against its 15 incorporated libraries. KEGG pathways, presented neuroactive ligand receptor interaction, as the only term, which met the set parameters of this section of analysis, while a multitude of terms, were generated from WikiPathways and Reactome (Table 3.3). These terms included amongst others, neurotransmitter

receptor binding, GPCRs (G-protein coupled receptors), calcium regulation in the cardiac cell and transmission across chemical synapses.

Lung cancers have been described to possess several neuropeptide receptors, as well as being able to synthesize and secrete various of these neuropeptides, thereby establishing autocrine-stimulated growth (Beekman et al. 1998). These peptides are part of a membrane receptor family, which initiates a cascade of intracellular signal transduction by interacting with G proteins and activating several kinase pathways and the mobilization of intracellular calcium (Ca^{2+}), ultimately resulting in cell proliferation (Prassas et al. 2012). Kinase perturbations from Gene Expression Omnibus (GEO) identified BRD4 and AKT1 as proteins relevant to the gene set of interest (Table 3.3).

Several hub protein-protein interactions (hub PPI) were expressed in this pathway analysis (Table 3.3). PPI hub, a small number of highly connected protein nodes, attempt to infer function to networks (Beekman et al. 1998). Gene encoded proteins can be expressed in increased quantities as a result of gene amplification, or through increased transcriptional activity, resulting in imbalances between gene repressors and gene activators (Kulasingam & Diamandis 2008).

Although identification of these regulatory proteins revealed relationships to cancer pathways, their identification alone lacks specificity to disease diagnosis (Wing et al. 2011). Alterations in the modulators of signaling networks, specific to lung adenocarcinoma may facilitate the identification of diagnostic markers

(Makridakis & Vlahou 2010).

Evaluation of disease/drugs libraries provided no meaningful data. While platforms of OMIM and Achilles, aimed at identifying and cataloging genetic vulnerabilities across diseases, generated several outputs, results were found to incorporate gene lists below the set threshold and/ or a significantly low CS < 2.

Table 3.3: Pathway enrichment terms extracted from Enrichr ($p < 0.01$, $(CS) > 2$)

Term	P-value	Combined Score	Source
Neuroactive ligand receptor interaction	0.00443048485907919 2	3.06	KEGG
Non-odorant GPCRs (Mus musculus)	0.00075073319118259 64	7.28	WikiPathways
Calcium regulation in the cardiac cell (Homo sapiens)	0.00322462435875148 1	5.16	
Calcium regulation in the cardiac cell (Mus musculus)	0.00264699333251841 7	5.13	
GPCRs, other (Homo sapiens)	0.00230020299762629 7	4.99	
Adipogenesis genes (Mus musculus)	0.00836209088503532 2	3.89	
Neuronal system	0.00075010074740886 4	4.53	
Transmission across chemical synapses	0.00195791471706012 7	4.25	Reactome
Class A/1 (Rhodopsin-like receptors)	0.00597171128439237 8	2.89	
Neurotransmitter receptor binding & downstream transmission - postsynaptic cell	0.00804394544789450 3	2.60	
NGF signaling via TRKA from the plasma membrane	0.03492034654718689	2.10	

DLG4	0.00000028478289991 5	14.61	PPI Hub Protein
CAMK2A	0.00011556154483942 18	10.14	
YWHAB	0.00001897636077605 2	7.84	
PRKACA	0.00179856717598491 6	6.39	
CALM1	0.00405231320914435 9	4.05	
FYN	0.00874983512168222 8	3.38	
BRD4	0.00000000050167503 26	26.32	Kinase Pertubations from GEO
AKT1	0.00475176034717670 4	5.35	

In order to facilitate the discovery of potential circulatory biomarkers, candidate genes, most likely to be found in the extracellular exome were investigated using ontology annotation sources (as described in 3.5.2.2.1). Of the 171 DEG, 57 were identified as most likely to be secreted into circulation (Table 3.4). Of these genes, a large number were identified as cytokines and growth factors, including, ANGPTL7 (angiopoietin-like 7), EDN3 (endothelin 3), RETN (resistin), NRG3 (neuregulin 3), CMTM2 (CKLF-like MARVEL transmembrane domain containing 2), CAMP (cathelicidin antimicrobial peptide), FGF10 (fibroblast growth factor 10), AGRP (agouti related neuropeptide), ANGPTL5 (angiopoietin-like 5) and CMTM5 (CKLF-like MARVEL transmembrane domain containing 5).

Cytokines and growth factors are signaling molecules, which regulate an array of biological processes such as cell proliferation, activation and differentiation by binding to specific receptors on the surface of their target cells and inducing intracellular signaling pathways (Welsh et al. 2003). The fibroblast growth factor

(FGF) family has been implicated in several disorders of bone growth, as well as in tumor formation and progression FGF10, a member of this family, has been proposed to play unique roles in the brain, in lung development, and wound healing (Beer et al. 2005).

Table 3.4: Candidate genes identified as located in the extracellular cellular component using GO annotations and having the potential to be serum markers

Potential Candidate Serum Markers						
ACR	ACTN2	AGTR2	AGRP	ANGPT 4	ANGPT L5	ANGPT L7
ASPA	C2orf40	C8B	CA4	CAMP	CD300L G	CD5L
CMTM2	CMTM5	CNKSR 2	CPB2	CST5	CRHBP	DPP6
EDN3	ENPP6	FABP4	F11	FAM150 B	FGF10	FGFBP2
FIGF	GKN2	GPA33	GPC5	GPM6A	ITLN2	KRT27
LIN7A	NRG3	ODAM	OVCH1	OVCH2	PCDH15	PLA2G1 B
RS1	RSPO1	RSPO2	RETN	SCUBE1	SFRP5	SH3GL2
SIRPD	SLC6A1 3	TNR	TRHDE	UPK3B	VWC2	WNT3A
WNT7A						

Investigation of statistical, enrichment and annotation data identified 15 genes of interest (Table 3.5). Candidates displaying the largest FC and determined to be the most highly differentially expressed were selected from MeV based analysis. Enrichr results were curated based on the number of times genes were found expressed in terms of ontologies and/or pathways. The incorporation of enhanced annotation, together with the most highly ranked statistical and enrichment findings, assisted to more accurately identify outputs of interest. Of these genes of interest, IRX1 (iroquois homeobox 1), a homeodomain transcription factor known to play a critical role in cellular processes, presented as the most significantly

differentially expressed (Guo et al. 2010). ITLN2 (intelectin 2) and was also identified as expressed in several instances using enrichment analysis. DEG identified as having the potential to enter circulation, such as FGF10 and CD5L (CD5 molecule-like) were also expressed in several categories. FIGF (c-fos induced growth factor (vascular endothelial growth factor D)), also identified as VEGFD is a growth factor actively involved in the P13K-akt pathway, known to be mutated in tumours (Ding et al. 2008).

Wingless type proteins (WNT) have a firmly established role in carcinogenesis. WNT7A, identified as a gene of interest, is a member of this family, identified as overexpressed in NSLC (Table3.4) (Kirikoshi & Katoh 2002). WNT signaling involves several other pathways including, Ca^{2+} flux, protein kinase A, cJun N-terminal kinase (JNK), and G protein, which have all been implicated in cancer (Brodie & Blumberg 2003; Stewart 2014). In normal tissue, WNT7A is associated with neuronal differentiation but has been identified as downregulated in almost all lung cancer types (Stewart 2014).

The identification of, neuron signaling, receptor and membrane terms were frequently identified in this analysis. GRIA1 (glutamate receptor, ionotropic, AMPA 1) is often involved in synaptic transmission, Ca^{2+} and kinase activity (Lisman et al. 2012). While cholinergic receptors CHRM1 (cholinergic receptor, muscarinic 1) and CHRM2 are known to be involved in G protein receptor activity and neurodegenerative disorders (Lai et al. 2001).

Table 3.5: Genes of interest identified using annotation, statistical analysis and enrichment analysis

MeV (Statistical Analysis)*	Enrichr (Enrichment Analysis)**	GO CC Extracellular Exosome***
IRX1	GRIA1	FGF10
SLC6A	CHRM2	WNT3A
C13orf36	GRIK4	WNT7A
ITLN2	AGTR2	FIGF
CD300LG	CHRM1	CD5L

* DEG representing largest FC > 2

** DEG expressed most often in terms of ontologies and/or pathways

*** DEG identified as having the potential to be secreted into the extracellular exosome as well as presenting with multiple expressions in terms of ontologies and/or pathways

Correlation of the candidate genes to lung cancer could possibly be identified following further investigation into their functioning and networks involvement (Ooi et al. 2009).



3.6.4. Expression Analysis

Enrichment analysis allowed for the identification of DEG, signifying statistically relevant biological differences between two test samples (Subramanian et al. 2007). While DEG indicated changes in comparative expression levels, it was necessary to determine whether genes were up- or downregulated, to better evaluate biological mechanisms and functions (Nam & Kim 2008). Regulation of gene expression was evaluated, between matched samples in groups, normal and early LUAD, by comparison of their respective means. All of the 171 genes identified as differentially expressed were found to be downregulated in the

cancer tissue ($FC > 2$) (Appendix A). To further assess these findings, genes were queried using GEA and MSigDB platforms.

GEA differential expression data for LUAD provided validation to the observations of downregulation, of genes of interest (Table 3.5), SCL6A4, C13orf36, CD300LG, GRIA1, GRIK4, AGTR2, FGF10, FIGF, CD5L, WNT3A and WNT7A. No experimental data was available in GEA for remaining input genes of interest in relation to LUAD.

Querying of the candidate genes (Table 3.5) against NCI-60 cell lines (National Cancer Institute) oncogenic signatures in MSigDB generated expression profiles showing both up and downregulation of genes (Appendix B). Of the eight lung cancer profiles available, five were identified as LUAD, namely, HOP92, HOP62, A549, NCI H23 and EKVX. Comparison of expression signatures yielded conflicting results. GRIA1 was seen to be downregulated in all LUAD except NCI H23, while FIGF presented with down regulation in LUAD and over expression in large cell lung cancer, NCI 460, and unspecified lung tumour cell line NCI H322. GRIK4 was identified as upregulated most lung cancers, showing a particularly high expression profile for squamous cell lung cancer line, NCI H226. CD5L was also seen to present with moderate upregulation in most lung cancers except EKVX. Analysis of these profiles yielded no definitive upregulation of any of the genes of interest queried. Variability across studies could arise due to biological differences amongst samples and populations or technological differences between the platforms. Gene expression profiling patterns could

facilitate in distinguishing the major morphological classes of lung tumours as well as enable subgroups of adenocarcinomas to be defined (Parmigiani et al. 2004).

Evaluation of the 57 potential serum markers, identified as downregulated, using MsigDB NCI-60 oncogenic signatures lead to no conclusive upregulation of candidate genes (Appendix C). Confirmation of these findings would need to be assessed using wet lab techniques such as RT PCR.

3.7. Discussion and Conclusion

The identification of candidate serum markers and genes of interest (Table 3.4) (Table 3.5) may hold relevant implications into understanding mechanisms involved in lung adenocarcinoma initiation and progression. However, they did not meet the criteria in order to facilitate identifying early stage circulatory markers, specific to the disease phenotype.

Genetic alterations in tumour tissue often involves growth-stimulatory autocrine and paracrine signaling (Hanahan & Weinberg 2011). Two distinct genetic alterations are involved in tumour development, the activation of oncogenes and the inactivation of tumor suppressor genes. While oncogenes drive abnormal cell proliferation as a consequence of genetic alterations that increase gene expression, tumor suppressor genes act to inhibit cell proliferation and tumor development. In

many tumors, these genes are lost or downregulated, removing negative regulators of cell proliferation and contributing to the abnormal proliferation of tumor cells (Cooper & Sunderland 2000).

Despite, several of the identified candidate markers offering links to these networks, for a protein to be useful in diagnosis it is necessary that it be highly expressed in comparison to its normal counterpart (Welsh et al. 2003). The principle behind the discovery of serum biomarkers require that the tumour secrete these product at an elevated level into bodily fluids (Diamandis 2004). As a tumour develops, proteins required for growth and metastasis are secreted and sheds cells into the circulation (Patz et al. 2007). The upregulation of DEG, is therefore of critical importance in the identification of tissue specific circulatory markers (Hassanein et al. 2012).

Lung cancer is currently classified according to morphologically as squamous cell carcinoma, adenocarcinoma, small cell carcinoma, and large cell carcinoma. This classification, however is often ineffective in predicting the biological behavior of these cancers. Gene expression profiles report distinct molecular profiles which has lead to refinement of classification (Parmigiani et al. 2004). Global gene expression profiling has routinely been used to uncover the underlying differences between normal and cancer cells, and these signatures are commonly used in facilitating diagnosis and prognosis of lung cancer (Ben-hamo et al. 2013). However to a large extent, expression profiling has failed to uncover genes that are upregulated and involved in cancer initiation, as the overlap at the gene level

between profiles is often poor, resulting in questioning of their robustness (Rapin et al. 2014). As most studies compare cancer with cancer, many of the detected transcriptional changes between different cancer samples may be attributed to differences in cell type and developmental stage and, consequently, will not identify gene expression signatures that underlie the malignant phenotype of interest (Rapin et al. 2014).

Overexpressed genes provide relevance not only for diagnosis, but because they constitute potential targets for therapeutic intervention. Expressed genes are a major determinant of cellular phenotype and function and are also responsible for variation of cellular responses to environmental stimuli (Chengalvala et al. 2007).

Gene expression offers assistance to guide drug discovery by illustrating involvement of the desired cellular pathways, as well as avoidance of acting on the toxicological pathways (Bai et al. 2013).

Investigation of DEG aimed to identify candidates, which were seen to be overexpressed in LUAD in order to facilitate biomarker discovery. However, all genes presented as being downregulated, resulting in no classic biomarker being identified in this analysis.

Studies by Danielsson and colleagues (2013) identified the majority of genes involved in malignant transformation to be downregulated, with only 20 % of genes evaluated being over expressed. While upregulated genes were seen as

being involved in cellular proliferation control, downregulated genes consisted of proteins exposed or secreted from the cell surface (Danielsson et al. 2013).

Altogether, the RNAseq data showed that in early stage LUAD, the enriched group of 171 genes presented as downregulated and related to a diverse set of functions, such as receptor binding and signaling, as well as consisting of a large proportion of cytokines and growth factors. To fully understand lung cancer dysregulation, as well as the potential of these genes being tumour suppressors, further evaluation of protein expression pattern and function of the proteins *in vitro* and *in vivo* is needed (Volinia et al. 2006).



3.8. References

Alaimo, S., Giugno, R. & Acunzo, M. 2015. Post-transcriptional knowledge in pathway analysis increases the accuracy of phenotypes classification. *Bioinformatics*. 29:2004-2008.

Alexa, A., Rahnenfuhrer, J. & Lengauer, T. 2006. Improved scoring of functional

groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*. 22(13):1600–1607.

Azuaje, F.J., 2014. Selecting biologically informative genes in co-expression networks with a centrality score. *Biology Direct*. 9(1):12.

Bai, J.P.F., Alekseyenko, A.V., Statnikov, A., Wang, I.M., & Wong, P.H. 2013. Strategic applications of gene expression: from drug discovery/development to bedside. *The AAPS Journal*. 15(2):427–37.

Beekman, A., Bunn, P.A., & Heasley, L.E. 1998. Expression of catalytically inactive phospholipase c-beta disrupts phospholipase c-beta and mitogen-activated protein kinase signaling and inhibits small cell lung cancer growth. *Cancer Research*. 58:910–913.

Beer, H.-D., Bittner, M., Niklaus, G., Munding, C., Max, N., Goppelt, A., & Werner, S. 2005. The fibroblast growth factor binding protein is a novel interaction partner of FGF-7, FGF-10 and FGF-22 and regulates FGF activity: implications for epithelial repair. *Oncogene*. 24(34):5269–77.

Ben-hamo, R., Boue, S., Martin, F., Talikka, M., & Efroni, S. 2013. Classification of lung adenocarcinoma and squamous cell carcinoma samples based on their gene expression profile in the sbv IMPROVER Diagnostic Signature Challenge. *Systems Biomedicine*. 8130:83–92.

Berriz, G.F., Cenik, C., Tasan, M., & Roth, F.P. 2009. Next generation software for functional trend analysis. *Bioinformatics*. 25(22):3043–3044.

Brodie, C. & Blumberg, P.M., 2003. Regulation of cell apoptosis by protein kinase c. *Apoptosis*. 8(1):19–27.

Bullard, J.H., Purdom, E., Hansen, K.D., & Dudoit, S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*. 11(1):94.

Camargo, A., Azuaje, F., Wang, H., & Zheng, H. 2008. Source Code for Biology and Medicine Permutation – based statistical tests for multiple hypotheses. *Biol Med*. 8:1–8.

Chang, K. et al. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*. 45(10):1113–1120.

Chen, E.Y. et al. 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*. 14:128.

Chengalvala, M.V., Chennathukuzhi, V.M., Johnston, D.S., Stevis, P.E., & Kopf, G.S. 2007. Gene expression profiling and its practice in drug development. *Current Genomics*. 8(4):262–70.

Cooper, G. & Sunderland, M. 2000. *The cell: a molecular approach: tumour suppressor genes* 2nd ed. Sinauer Associates.

Cui, X., Hwang, J., Blades, N.J., & Churchill, G.A. 2005. Improved statistical tests for differential gene expression by shrinking variance components estimates. 6(1):59–75.

Danielsson, F. et al. 2013. Majority of differentially expressed genes are down-

regulated during malignant transformation in a four-stage model. *Proceedings of the National Academy of Sciences of the United States of America*. 110(17):6853–8.

Davey, J.W., Hoenlohe, P.A., Etter, P. Boone, J.Q., Catchen, J.M., & Blaxter, M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature genetics*. 12:499–510.

Detterbeck, F. 2009. The new lung cancer staging system. *CHEST*. 136(1):6-8.

Diamandis, E.P. 2004. Analysis of serum proteomic patterns for early cancer diagnosis: Drawing attention to potential problems. *Journal of the National Cancer Institute*. 96(5):353–356.

Ding, L. et al. 2008. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*. 455(7216):1069–75.

Draghici, S. et al. 2007. A systems biology approach for pathway level analysis. *Genome*. 17:1537–1545.

Dudoit, S., Yang, Y., Callow, M., & Speed, T.P. 2002. Statistical methods for identifying differentially expressed genes in replicated c{DNA} microarray experiments. *Stat. Sinica*. 12(1):111–139.

Enrichr. Available: <http://amp.pharm.mssm.edu/Enrichr/> [2015, November, 11].

Fonseca, N.A., Marioni, J. & Brazma, A., 2014. RNA-Seq gene profiling--a systematic empirical comparison. *PloS one*, 9(9), p.e107026.

Ge, Y., Dudoit, S. & Speed, T.P., 2003. Resampling-based multiple testing for microarray data analysis. *Stat Sinica*. (510):1–41.

- Gene Expression Atlas (GEA). Available: www.ebi.ac.uk/gxa/ [2015, November, 20].
- Guo, X. et al. 2010. Homeobox gene IRX1 is a tumor suppressor gene in gastric carcinoma. *Oncogene*. 29(27):3908–20.
- Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D.C., & Shyr, Y. 2013. Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PloS One*. 8(8):71462.
- Hanahan, D. & Weinberg, R.A. 2011. Hallmarks of cancer: the next generation. *Cell*. 144(5):646–74.
- Hassanein, M., Callison, J., Callaway-Lane, C., Aldrich, M., Grogan, E.L., & Massion, P. 2012. The state of molecular biomarkers for the early detection of lung cancer. *Cancer Prevention Research*. 5(8):992–1006.
- Howe, E. et al. 2010. *Biomedical Informatics for Cancer Research* M. F. Ochs, J. T. Casagrande, & R. V. Davuluri, eds., Boston, MA: Springer US.
- Howe, E., Holton, K., Nair, S., Schluch, D., Sinha, R., & Quackenbush, J. 2011. RNA-Seq analysis in MeV. *Bioinformatics*. 27(22):3209–3210.
- Huang, D.W. et al. 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*. 35:169–75.
- Huang, D.W., Sherman, B.T. & Lempicki, R. A. 2009. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*. 37(1):1–13.

Huang, Q. et al. 2011. RNA-Seq analyses generate comprehensive transcriptomic landscape and reveal complex transcript patterns in hepatocellular carcinoma. *PLoS One*. 6(10):26168.

Kandoth, C. et al., 2013. Mutational landscape and significance across 12 major cancer types. *Nature*. 502(7471).

Kapushesky, M. et al. 2009. Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Research*. 38:690–698.

Kirikoshi, H. & Katoh, M. 2002. Expression of WNT7A in human normal tissues and cancer, and regulation of WNT7A and WNT7B in human cancer. *International Journal of Oncology*. 21(4):895–900.

Knight, J., Ivanov, I. & Dougherty, E. 2014. MCMC implementation of the optimal Bayesian classifier for non-Gaussian models: model-based RNA-Seq classification. *BMC bioinformatics*. 15(40).

Kulasingam, V. & Diamandis, E.P. 2008. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nature Clinical Practice Oncology*. 5(10):588–99.

Kvam, V.M., Liu, P. & Si, Y. 2012. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American Journal of Botany*. 99(2):248–256.

Lægreid, A., Hvidsten, T., Midelfart, H., Komorowski, J., & Sandvik, A. 2003. Predicting gene ontology biological process from temporal gene expression patterns.

Genome Research. 13(5):965–979.

Lai, M.K. et al. 2001. Psychosis of Alzheimer's disease is associated with elevated muscarinic M2 binding in the cortex. *Neurology*. 57(5):805–11.

Lai, X., Liangpunsakul, S., Crabb, D., Ringham, H.N., & Witzmann, F.A. 2009. A proteomic workflow for discovery of serum carrier protein-bound biomarker candidates of alcohol abuse using LC-MS/MS. *Electrophoresis*. 30(12):2207–2214.

Laukens, K., Naulaerts, S., & Berghe, W. 2015. Bioinformatics approaches for the functional interpretation of protein lists: from ontology term enrichment to network analysis. *Proteomics*. 15(5-6):981–96.

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., & Dewey, C.N. 2010. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*. 26(4):493–500.

Liberzon, A. 2014. A description of the Molecular Signatures Database (MSigDB) Web site. *Methods in Molecular Biology (Clifton, N.J.)*. 1150:153–60.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., & Meisirov, J.P. 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England)*. 27(12):1739–40.

Lisman, J., Yasuda, R., & Raghavachari, S. 2012. Mechanisms of CaMKII action in long-term potentiation. *Nature reviews. Neuroscience*. 13(3):169–82.

Makridakis, M., & Vlahou, A. 2010. Secretome proteomics for discovery of cancer biomarkers. *Journal of Proteomics*. 73(12):2291–2305.

Mamanova, L. et al. 2010. Target-enrichment strategies for next-generation sequencing. *Nature Methods*. 7(2):111–118.

MultiExperiment Viewer (MeV). Available: <http://tm4.org/mev.html> [2015, January, 20].

Molecular Signatures Database (MSigDB) NCI-60 cell line (National Cancer Institute) Subramanian, Tamayo, et al. (2005, *PNAS* 102:15545-15550) Available: <http://www.broadinstitute.org/msigdb> [2015, November, 20].

Morozova, O. et al. 2008. Rank-based procedures for mixed paired and two-sample designs. *Genomics*. 8(1):3209–3210.

Mortazavi, A., Williams, B.A, McCue, K., Schaeffer, L., & Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*. 5(7):621–8.

Nam, D., & Kim, S.-Y. 2008. Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*. 9(3):189–197.

Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Yang, X., Tirado, F., Carazo, J.M., & Pascual-Montano, A. 2009. GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Research*. 37:317–322.

Ooi, C.H. et al. 2009. Oncogenic pathway combinations predict clinical prognosis in gastric cancer. *PLoS Genetics*. 5(10):1000676.

Oshlack, A., Robinson, M.D., & Young, M.D. 2010. From RNA-seq reads to

differential expression results. *Genome Biology*. 11(12):.220.

Parmigiani, G., Garrett-mayer, E.S., & Anbazhagan, R. 2004. A cross-study comparison of gene expression studies for the molecular classification of lung cancer a cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clinical Cancer Research*. 10(410):2922–2927.

Patz, E.F., Campa, M.J., Gottlin, E.B., Kumartseva, I., Xiang, R.G., & Herndon, J.E. 2007. Panel of serum biomarkers for the diagnosis of lung cancer. *Journal of Clinical Oncology*. 25(35):5578–5583.

Pepke, S., Wold, B., & Mortazavi, A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nature Methods*. 6(11):22–32.

Petryszak, R. et al. 2014. Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Research*. 42(1):926–932.

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

Rapin, N. et al. 2014. Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in AML patients. *Blood*. 123(6):894–904.

Reis-Filho, J.S. 2009. Next-generation sequencing. *Breast Cancer Research*. 11(3):12.

Rhodes, D.R. et al. 2007. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia (New York, N.Y.)*. 9(2):166–80.

Robinson, J., Thorvaldsdóttir, H., & Winckler, W., 2011. Integrative genomics viewer. *Nature Biotechnology*. 29(1):24-26.

Robinson, M.D. & Oshlack, A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*. 11(3):25.

Saeed, A.I. et al. 2006. [9] TM4 Microarray Software Suite. *Methods in Enzymology*. 411:134–193.

Schreiber, S.L. et al. 2010. Towards patient-based cancer therapeutics. *Nature Biotechnology*. 28(9):904–6.

Segal, E., Friedman, N., Koller, D., & Regev, A. 2004. A module map showing conditional activity of expression modules in cancer. *Nature Genetics*. 36(10):1090–1098.

Sherman, B.T. et al. 2007. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics*. 8(1):426.

Smith, C.L., Goldsmith, C.-A.W., & Eppig, J.T. 2005. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*. 6(1):7.

Stewart, D.J. 2014. Wnt signaling pathway in non-small cell lung cancer. *Journal of*

the National Cancer Institute. 106(1):1–11.

Subramanian, A. et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*. 102(43):15545–50.

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., & Mesirov, J.P. 2007. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics (Oxford, England)*. 23(23):3251–3.

The Cancer Genome Atlas (TCGA). Available: <http://cancergenome.nih.gov/> [2015, January, 11].

The Disease Ontology. Available: <http://www.ncbi.nlm.nih.gov/pubmed/26093607/> [2015, November, 20].

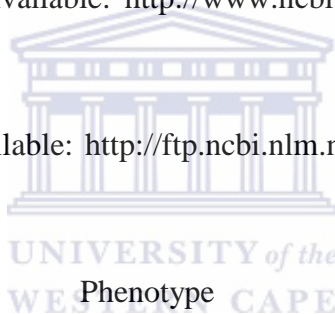
The Gene Ontology. Available: <http://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz/> [2015, November, 20].

The Human Phenotype Ontology. Available: <http://www.ncbi.nlm.nih.gov/pubmed/22961259/> [2015, November, 20].

The Mammalian Ontology. Available: <http://www.ncbi.nlm.nih.gov/pubmed/24217912/> [2015, November, 20].

The Pathway Ontology. Available: <http://www.ncbi.nlm.nih.gov/pubmed/24499703/> [2015, November, 20].

Travis, W.D., Brambilia, E., Muller-Hermelink, H.K., & Harris, C.C. 2004. Pathology and Genetics of Tumours of the lung. *Bulletin of the World Health Organization*. 50(1-2):9–19.



Venny 2.0.2 – Computational Genomics Service. Available: <http://www.bioinfogp.cnb.csic.es/tools/venny> [2015, August, 11].

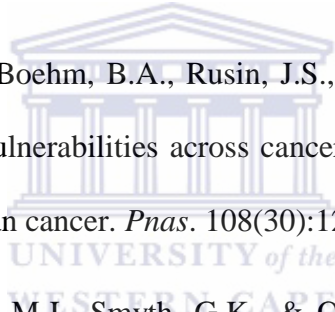
Villanueva, J. et al. 2006. Differential exoprotease activities confer tumor-specific serum peptidome patterns. *The Journal of Clinical Investigation*. 116(1):271–84.

Volinia, S. et al. 2006. A microRNA expression signature of human solid tumors defines cancer gene targets. *Proceedings of the National Academy of Sciences of the United States of America*. 103(7):2257–2261.

Welsh, J.B. et al. 2003. Large-scale delineation of secreted protein biomarkers overexpressed in cancer tissue and serum. *Proceedings of the National Academy of Sciences of the United States of America*. 100(6):3410–3415.

Wing, H., Cowley, G.S., Boehm, B.A., Rusin, J.S., & Scott, J.A. 2011. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Pnas*. 108(30):12372–12377.

Young, M.D., Wakefield, M.J., Smyth, G.K., & Oshlak, A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*. 11(2):14.



Chapter 4

Future Perspectives

Diagnostic biomarker development is based on the biological properties of cancer as a systemic disease, and entails the search for plasma proteins identified as over expressed by the tumour tissue when compared to its normal counterpart (Hanash et al. 2008). Classification of lung cancer has traditionally been based on tumour morphology. However, tumours identified as histologically similar can exhibit different responses to therapy, denoting variations at the molecular level (Cuperlovic-Culf et al. 2005).

Therefore, identification of circulatory tumour specific markers, could provide an early, rapid and non-invasive diagnostic technique which would most certainly improve the prognosis of patients with lung carcinomas (Altintas & Tothill 2013). Gene expression data, obtained using high throughput (HT) technologies offers an ideal platform to facilitate biomarker discovery (Prassas et al. 2012).

RNAseq allows for whole transcriptome profiling using deep sequencing. This technique compares favorably to previously used methods for gene expression measurement, such as DNA microarrays. This is due to be its higher sensitivity, lower background and ability to detect previously unknown transcripts (Howe et al. 2011).

Microarray analysis has become a widely used tool in the interrogation of gene expression data in many biological settings. The ability of these arrays to

simultaneously interrogate thousands of transcripts, has led to several insights into developmental processes and differences in gene expression between samples (Choi et al. 2014). Microarray data, however, presents with several limitations, which have hindered the development and acceptance of markers based on their profiling. These include the use of multiple platforms and protocols in determining differential expression profiles as well as the lack of assay reproducibility on multiple samples of the same tissue specimen (Simon 2003). Classes used in comparison of expression profiling may represent different tumour types and therefore not yield tissue specific gene expression (Szczurek et al. 2010). Accurate identification of DEG, requires high-quality specimens with well-matched controls, and an efficient process to confirm discoveries through independent validation studies (Hanash et al. 2008).

RNAseq technologies, unlike array sequencing, allow for the mapping of previously unknown organisms and do not require the use of probes (Oshlack et al. 2010). While cross-hybridization of microarray probes affect expression measures non-uniformly (Petryszak et al. 2014).

RNAseq also presents with the advantage of analyzing expression at exon levels, and provides detail about transcriptional features that arrays are not able to. Novel transcribed regions, splicing variants and allele-specific expression can be identified using this technology, reflecting the high overall sensitivity of RNAseq compared with other whole-transcriptome expression quantification platforms (Huang et al. 2011; Trapnell et al. 2012).

The overall advantages provided by RNAseq, allow for the more accurate assessment of DEG. Genes identified as potential candidate markers (Chapter 2), COPZ1, SEC23B, SEC24A and SEC24D, did not present as differentially expressed using this NGS platform, and were also found to be located in GO cellular components, which would not facilitate secretion into the extracellular exosome. In addition, none of the 4 above mentioned potential candidate genes were identified using RNAseq (Chapter 3). This in Combination with their intracellular predisposition eliminated them as potential circulatory markers for early lung cancer diagnosis.

The 57 potential serum markers and 10 genes of interest identified as downregulated (Chapter 3) would require validation at a molecular level. Quantitative real-time PCR, (RT PCR) is a tool commonly used when validating HT gene expression results (Morey et al. 2006). Changes in gene expression at the RNA level would then need to be evaluated at the protein level, with several techniques such as, immunofluorescence (IM) microscopy, two- dimensional polyacrylamide gel electrophoresis (2D-PAGE), surface enhanced laser desorption/ionisation time of flight (SELDI-ToF), protein arrays, isotope coded affinity tags (ICAT), and multidimensional protein identification technology (MudPIT) currently used (Berriz et al. 2009; Danielsson et al. 2013).

Genes identified as downregulated in cancer may reveal themselves to be tumour suppressor genes, which encode proteins that normally inhibit the formation of

tumours (Kumar et al. 2005). Inactivation of both copies of a tumor suppressor gene is required before their function can be eliminated, and thus further investigation of these genes is necessary to accurately assess their roles in tumorigenesis. While the genes identified may not be targets as diagnostic tools for early stage lung cancer, they might however reveal novel pathways implicated in tumorigenesis (Westbrook et al. 2005).

4.1. References

Altintas, Z., & Tothill, I. 2013. Biomarkers and biosensors for the early diagnosis of lung cancer. *Sensors and Actuators B: Chemical*, 188:988–998.

Berriz, G.F., Cenik, C., Tasan, M., & Roth, F.P. 2009. Next generation software for functional trend analysis. *Bioinformatics*. 25(22):3043–3044.

Choi, N. et al. 2014. Comparison of combination of feature selection methods and classification methods for multiclass cancer classification from RNA-seq gene expression data. In *The 7th International Conference FITAT/ISPM*. South Korea: Database/Bioinformatics Laboratory, Department of Computer Science, Chungbuk National University. 1–3.

Cuperlovic-Culf, M., Belacel, N., & Ouellette, R.J. 2005. Determination of tumour marker genes from gene expression data. *Drug Discovery Today*. 10(6):429–37.

Danielsson, F. et al. 2013. Majority of differentially expressed genes are down-regulated during malignant transformation in a four-stage model. *Proceedings of the*

National Academy of Sciences of the United States of America. 110(17):6853–6858.

Hanash, S.M., Pitteri, S.J., & Faca, V.M., 2008. Mining the plasma proteome for cancer biomarkers. *Nature*. 452(7187):571–9.

Howe, E., Holton, K., Nair, S., Schluch, D., Sinha, R., & Quackenbush, J. 2011. RNA-Seq analysis in MeV. *Bioinformatics*. 27(22):3209–3210.

Huang, Q. et al. 2011. RNA-Seq analyses generate comprehensive transcriptomic landscape and reveal complex transcript patterns in hepatocellular carcinoma. *PLoS One*. 6(10):26168.

Kumar, V., Abbas, A., & Fausto, N. 2005. Chapter 7: Tumor suppressor genes and oncogenes. In *Robbins and Cotran: Pathologic basis of disease*. Elsevier/ Saunders. 292–306.

Morey, J.S., Ryan, J.C., & Van Dolah, F.M. 2006. Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biological Procedures Online*. 8:175–93.

Oshlack, A., Robinson, M.D., & Young, M.D. 2010. From RNA-seq reads to differential expression results. *Genome Biology*. 11(12):220.

Petryszak, R. et al. 2014. Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Research*. 42:926–932.

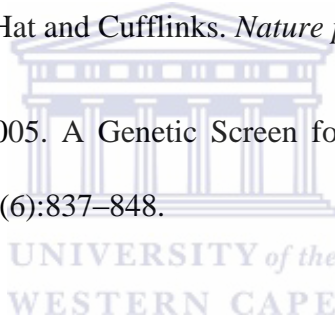
Prassas, I., Chrystoja, C.C., Makawita, S., & Diamandis, E.P. 2012. Bioinformatic identification of proteins with tissue-specific expression for biomarker discovery. *BMC Medicine*. 10(1):39.

Simon, R. 2003. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *British Journal of Cancer*. 89(9):1599–604.

Szczurek, E., Biecek, P., Tiuryrn, J., & Vingron, M. 2010. Introducing knowledge into differential expression analysis. *Journal of computational biology : A Journal of Computational Molecular Cell Biology*, 17(8):953–67.

Trapnell, C. et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*. 7(3):62–78.

Westbrook, T.F. et al., 2005. A Genetic Screen for Candidate Tumor Suppressors Identifies REST. *Cell*. 121(6):837–848.



LIST OF APPENDICES



UNIVERSITY *of the*
WESTERN CAPE

Appendix A: Average sample means of groups early versus normal lung cancer used to identify genes as downregulated (FC > 2)

GENE ID	GroupA (normal) Mean Average	GroupB (early) Mean Average	FC
ABCC13 150000	4,2816224	1,4065871	3,043979573
ACADL 33	9,009417	4,443536	2,027533253
ACR 49	2,6779847	1,0424821	2,568854372
ADCY8 114	5,6271825	0,9243555	6,087682174
ADCYAP1R1 117	1,7422048	0,4905618	3,551448156
ADH1A 124	6,1923056	2,2539406	2,74732422
ADRA1A 148	4,755768	0,9787929	4,858809254
ADRA1D 146	5,065104	2,3533425	2,152302098
AGBL1 123624	3,967218	0,9221507	4,302136299
AGRP 181	5,0293627	1,5410669	3,263558967
AGTR2 186	9,102922	4,2310033	2,151480714
ANGPT4 51378	4,962274	1,6346674	3,035647496
ANGPTL5 253935	3,3710234	0,8047033	4,189150709
ANGPTL7 10218	4,5971813	0,8931059	5,147408947
ANKRD1 27063	9,901326	4,0223947	2,461550081
ANO2 57101	5,1373405	2,4721	2,07812811
ART4 420	5,445078	2,58589	2,105688177
ASPA 443	5,4383144	2,2289684	2,439834679
BAI3 577	5,414573	2,5566201	2,117863737
BET3L 100128327	1,3542717	0,29541817	4,58425323
BTBD18 643376	1,1017184	0,41362333	2,663578962
BTNL3 10917	2,1138666	0,8886269	2,378801047
C10orf67 256815	6,4371624	1,3119464	4,906574232
C13orf36 400120	8,287082	0,7133158	11,61769023
C15orf51 196968	2,8569746	1,3251755	2,155921687
C19orf69 10017076 5	1,4497871	0,25641677	5,654026061
C1orf150 148823	4,702532	2,337855	2,01147291
C20orf202 400831	5,0562468	2,2932894	2,204801016
C21orf71 282566	1,2679096	0,35393432	3,582330191
C2orf40 84417	7,753121	3,732006	2,077467453
C8B 732	7,6111403	2,8394043	2,680541232
CA4 762	9,431905	3,4145806	2,762244066
CAMP 820	5,6810355	1,7828605	3,186472245
CASP12 120329	3,6950889	0,8951441	4,127926331
CASP5 838	3,9735572	1,9769943	2,009898157
CASQ2 845	6,5151634	3,0123134	2,162843813
GENE ID	GroupA (normal)	GroupB (early)	FC

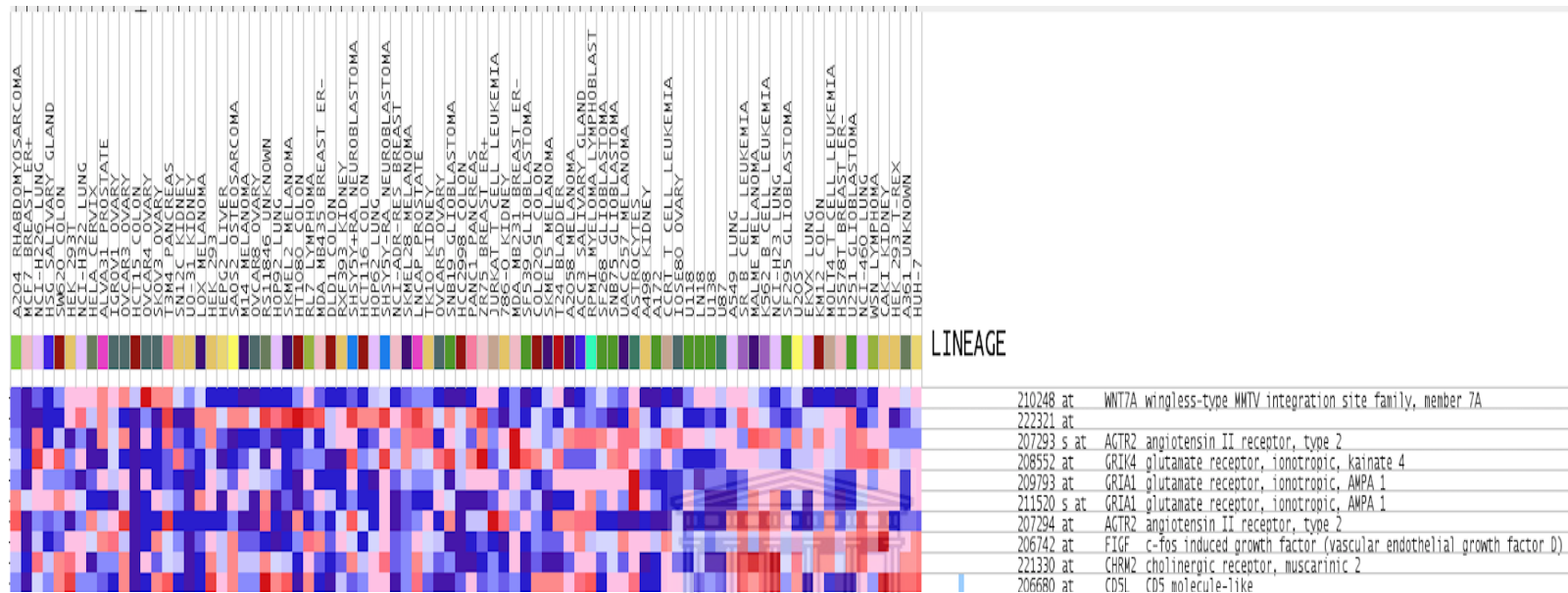
	Mean Average	Mean Average	
CCDC141 285025	7,720403	3,4250622	2,254091327
CCDC54 84692	1,389033	0,1639819	8,470648285
CD300LG 146894	9,059065	0,7597923	11,92308082
CD5L 922	5,2576313	0,98691463	5,327341535
CDH19 28513	4,6428785	1,3723037	3,383273324
CHRM1 1128	6,2181664	1,3174869	4,71971782
CHRNA2 1135	3,9502168	0,5350508	7,382881775
CMTM2 146225	4,3510895	1,782931	2,440413847
CMTM5 116173	2,6354039	0,41136432	6,406496071
CNKS2 22866	5,233461	2,3568158	2,220564288
CNTFR 1271	5,1314034	1,6777856	3,058438098
CNTN6 27255	8,261649	2,7290616	3,027285643
CPB2 1361	9,513501	3,9313664	2,419896807
CRHBP 1393	2,8420057	1,353906	2,099115965
CST5 1473	4,5923557	1,912013	2,401843345
CYP3A7 1551	4,765544	0,8792027	5,420301826
DCC 1630	5,146027	1,9704973	2,611537199
DPP6 1804	5,502035	1,8839203	2,920524292
EDN3 1908	4,3012714	1,0384432	4,142038197
ELMOD1 55531	3,9549534	1,6000018	2,471843094
EMR3 84658	6,095803	2,4613633	2,476596202
ENDOU 8909	3,0747228	0,92509377	3,323687717
ENPP6 133121	4,228245	2,0811806	2,031656935
ERVFRDE1 405754	4,5414157	1,5628172	2,905916124
F11 2160	7,680035	2,7858522	2,756799158
FABP4 2167	11,316286	5,073104	2,230643409
FAM150B 285016	6,6600966	2,5018952	2,662020615
FAM189A1 23359	6,1947474	2,9185467	2,122545238
FGF10 2255	2,8450553	0,5674932	5,013373376
FGFBP2 83888	6,5299954	2,6398761	2,473599197
FIGF 2277	10,009997	4,9667354	2,015407747
FLJ37543 285668	1,3223774	0,2869659	4,608134277
G6PC2 57818	1,4373262	0,51902586	2,76927666
GATA1 2623	3,1759634	1,0285112	3,087923009
GBP7 388646	1,9383273	0,65805346	2,945546856
GKN2 200504	9,943389	4,5074315	2,205998915
GPA33 10223	8,136766	3,9195318	2,075953562
GPC5 2262	6,043915	2,829931	2,135711083
GPM6A 2823	10,239844	4,1057186	2,494044283
GENE ID	GroupA (normal)	GroupB (early)	FC

	Mean Average	Mean Average	
GPR182 11318	2,9909582	1,3553613	2,20676081
GRIK4 2900	5,1211247	1,8573241	2,757259597
GYPE 2996	6,508371	2,5807354	2,521905578
HEMGN 55363	1,961249	0,49257186	3,981650515
HSPB3 8988	4,002853	1,2937107	3,094086645
IRX1 79192	6,8954196	0,208589	33,05744598
ITLN2 142683	9,015552	0,7992436	11,28010534
KCNA4 3739	5,6196437	1,1358455	4,947542337
KCNIP1 30820	4,059912	1,5854095	2,560797069
KLF17 128209	3,4848986	1,2275481	2,838910019
KLHL33 123103	1,6128268	0,42417312	3,802284313
KRT27 342574	3,3772027	1,0924238	3,091476678
KRT4 3851	9,495609	4,7258854	2,009276188
LGI3 203190	10,864179	4,12859	2,631450205
LHFPL3 375612	6,105437	2,6559932	2,298739696
LIN7A 8825	6,173955	2,86324	2,156282743
LOC257358 257358	2,5526803	0,9995538	2,553819814
LOC283392 283392	3,317297	0,8327328	3,983627161
LOC400804 400804	1,6195064	0,31280133	5,177428114
LOC572558 572558	1,7571205	0,5157032	3,407232106
LOC723809 723809	7,9251947	3,939973	2,01148452
LOC90586 90586	2,8907733	1,1384894	2,539130623
LOXHD1 125336	4,5330167	1,8289431	2,478489735
LRRTM4 80059	3,78329	1,3709701	2,759571489
MAP3K15 389840	4,5975385	1,5455241	2,974743972
MAPK4 5596	8,137145	4,012172	2,028114697
MGC27382 149047	4,5428805	1,0652946	4,264435866
MUSK 4593	3,5123036	1,0806131	3,250287823
NKAPL 222698	4,4268036	2,151486	2,057556312
NRG3 10718	4,8136506	1,4317902	3,361980407
NTNG1 22854	7,0784	3,51211	2,015426624
ODAM 54959	5,7103567	1,2218052	4,673704695
ODF3L1 161753	5,379729	2,6711323	2,014025662
OR2W3 343171	3,0378652	0,52102745	5,830528123
OTC 5009	3,2612658	0,63570565	5,130150723
OVCH1 341350	5,3197083	1,86584115	2,851104608
OVCH2 341277	3,895136	1,7016311	2,289060185
P2RX6 9127	4,0380263	1,8500171	2,182696744
PAK7 57144	2,4745655	0,82410604	3,002727052
GENE ID	GroupA (normal)	GroupB (early)	FC

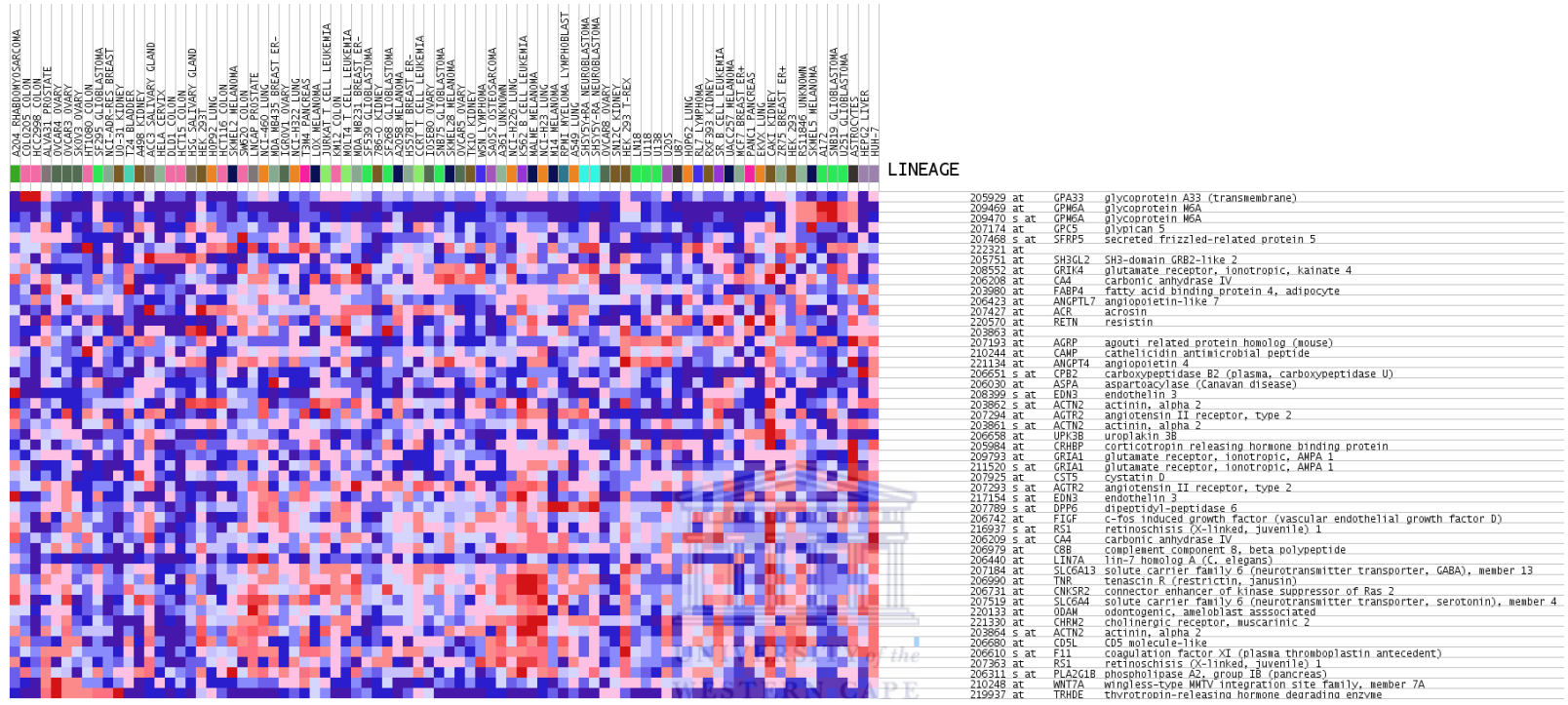
	Mean Average	Mean Average	
PLA2G1B 5319	8,445295	4,0442	2,088248603
PRKG2 5593	7,2551813	3,5093539	2,067383771
PTCRA 171558	4,992306	2,3703027	2,106189222
PTPRQ 374462	6,7849884	1,4416903	4,706273185
PURG 29942	2,1752582	1,0161554	2,140674743
RANBP3L 202151	5,180508	2,4948077	2,076515958
RBP2 5948	5,118596	1,0438054	4,90378379
RETN 56729	7,533858	3,046637	2,472843992
RMST 196475	2,2621553	0,69590604	3,250661972
RPH3A 22895	3,6431863	1,1569144	3,149054329
RPL13AP17 399670	7,67246	2,390377	3,209728005
RPL23AP32 56969	2,078843	0,48285732	4,305294574
RSPO1 284654	6,767636	2,53213	2,672704798
RSPO2 340419	7,202499	3,2611396	2,208583466
RXFP1 59350	6,3622127	2,4175732	2,63165256
RXRG 6258	6,8258986	2,735596	2,495214425
SCUBE1 80274	8,760947	4,111886	2,130639565
SFRP5 6425	6,79853	2,938662	2,313478039
SGCG 6445	6,458385	1,7367709	3,718616543
SH2D4B 387694	4,3255267	1,5447667	2,800116484
SH3GL2 6456	5,9703283	1,7873415	3,340339997
SH3GL3 6457	6,5783734	1,1899475	5,528288769
SIRPD 128646	2,6429155	0,5804424	4,55327781
SLC27A6 28965	5,0158305	1,2780949	3,924458583
SLC5A4 6527	3,5556588	1,1802734	3,012572172
SLC6A13 6540	4,638028	0,2026144	22,89091002
SLC6A4 6532	10,887906	2,64655	4,113999736
SLCO1A2 6579	7,279196	2,0216997	3,60053276
SLITRK2 84631	5,016351	1,5283369	3,282228545
SOSTDC1 25928	9,003033	3,5778239	2,516343244
ST8SIA6 338596	5,702789	1,6818341	3,39081542
SYN2 6854	5,596719	1,6770558	3,337228851
SYNPO2L 79933	4,9549394	1,9781082	2,504887953
TCEAL2 140597	6,9793286	3,1712759	2,200795144
TMEM132C 92293	5,0876803	1,8587906	2,737091687
TNR 7143	3,716323	0,63475776	5,854710622
TRHDE 29953	6,973898	2,420352	2,881356927
TRIM58 25893	6,4378753	2,6731522	2,40834596
UNC45B 146862	4,3746734	1,6354035	2,674981067
GENE ID	GroupA (normal)	GroupB (early)	FC

	Mean Average	Mean Average	
UPK3B 80761	11,248419	5,334326	2,108686083
VWC2 375567	2,0223505	0,47118497	4,292052227
WNT3A 89780	8,14199	2,8524773	2,854357509
WNT7A 7476	7,4504957	3,3099694	2,250925854
ZCCHC5 203430	1,6827285	0,58805156	2,861532244
ZDHHC19 131540	3,0852334	0,8454799	3,649091362
ZNF536 9745	2,849423	0,742615	3,837012449
ZNF705A 440077	0,89204854	0,35725042	2,49698388





Appendix B: Heat map visualisation of oncogenic signatures of candidate genes of interest generated by the Molecular Signatures Database (MSigDB) NCI-60 cell line (National Cancer Institute) with red indicating upregulation and blue depicting downregulation (<http://www.broadinstitute.org/msigdb>).



Appendix C: Heat map visulisation of oncogenic signatures of candidate serum markers generated by the Molecular Signatures Database (MSigDB) NCI-60 cell line (National Cancer Institute) with red indicating upregulation and blue depicting downregulation (<http://www.broadinstitute.org/msigdb>)