EVALUATION OF PROTEOMIC AND TRANSCRIPTOMIC BIOMARKER DISCOVERY TECHNOLOGIES IN OVARIAN CANCER.

CLARE RITA ELIZABETH COVENEY

A thesis submitted in partial fulfilment of the requirements of the Nottingham Trent University for the degree of Doctor of Philosophy

October 2016

Copyright Statement

Acknowledgments

This work was funded by The John Lucille van Geest Foundation and undertaken at the John van Geest Cancer Research Centre, at Nottingham Trent University.

I would like to extend my foremost gratitude to my supervisory team Professor Graham Ball, Dr David Boocock, Professor Robert Rees for their guidance, knowledge and advice throughout the course of this project.

I would also like to show my appreciation of the hard work of Mr Ian Scott, Professor Bob Shaw and Dr Matharoo-Ball, Dr Suman Malhi and later Mr Viren Asher who alongside colleagues at The Nottingham University Medical School and Derby City General Hospital initiated the ovarian serum collection project that lead to this work. I also would like to acknowledge the work of Dr Suha Deen at Queen’s Medical Centre and Professor Andrew Green and Christopher Nolan of the Cancer & Stem Cells Division of the School of Medicine, University of Nottingham for support with the immunohistochemistry.

I am grateful for the colleagueship of the numerous and high-quality scientists, both staff and students I have had the opportunity to work alongside during this time.

I am indebted for the unconditional and unknowing support of my partner, friends and family, especially my parents.

Thank you.

CONTENTS

FIGURES ... i

TABLES ... iii

Abbreviations and Glossary ... iv

HYPOTHESES ... vii

Thesis Abstract ... viii

1. Introduction ... 1

1.1. General Concepts ... 1

1.1.1. Cancer ... 1

1.1.2. Ovarian Cancer ... 4

1.1.3. Biomarkers ... 10

1.1.3.1. Essential and Desirable biomarker properties ... 11

1.1.4. The Need for Effective Screening Strategies ... 12

2. Methodological Overview ... 18

2.1. Proteomic Approaches to Biomarker Discovery and Onco-proteomics ... 18

2.1.1. Proteomic Techniques for Cancer Biomarker Discovery ... 18

2.1.2. Analysis of the Serum Proteome ... 19

2.1.3. Mass Spectrometry and Tandem Mass Spectrometry ... 22

2.1.3.1. An Historic Summary ... 22

2.1.3.2. Soft Ionisation: Matrix Assisted Laser Desorption Ionisation (MALDI) and Electrospray Ionisation (ESI) ... 24

2.1.3.3. Mass Analysers ... 28

2.1.3.4. Tandem Mass Spectrometry ... 30

2.1.4. Mass Spectrometry Approaches/Techniques used for Biomarker Discovery ... 32

2.1.5. Sample Fractionation and Liquid Chromatography ... 33

2.1.5.1. Reducing Sample Complexity ... 33

2.1.5.2. The Mechanics of and Sources of Variability in Liquid Chromatography ... 33

2.1.5.3. Deconvolution of Complex Biological Samples Using Liquid Chromatography ... 36

2.1.6. Advantages or Disadvantages of Tagging in MALDI-TOF-MS.............................38

2.1.7. Label Free Quantitation Techniques 38

2.2. Transcriptomic Approaches to Biomarker Discovery and Onco-genomics...................39

2.2.1. Gene Expression Profiling 40

2.2.1.1. DNA Microarray Experiments..40

2.2.1.2. Next Generation DNA Sequencing...41

2.2.2. Data Mining 42

2.2.2.1. Statistical Analysis for Omics Data 43

2.2.2.2. Categorical and Continuous Variables in Omics Data 44

2.2.2.3. Analysis of Omics Survival Data...44

2.2.2.4. Machine Learning. 48

2.2.2.5. Artificial Neural Networks (ANNs) 49

2.2.3. Curated Data Repositories and Online Tools 54

2.2.3.1. Data Sharing 55

2.2.3.2. Protein interaction databases 55

2.2.3.3. Ontological Databases 59

2.3. Aims of the Project Overall 59

3. Proteomic Evaluation: MALDI-MS Profiling Strategy for Biomarker Discovery in Ovarian Cancer 60

Chapter Abstract 60

3.1. Introduction..61

3.1.1. MALDI Mass spectrometry and ovarian cancer 61

3.1.2 Aims and Hypotheses of the Chapter. 67

3.2. Materials and Methods...68

3.2.1 Materials 68

3.2.1.1. Equipment used..68

3.2.1.2. Reagents used: 69

3.2.1.3. Stock Solutions Made and Used 69

3.2.1.4. Samples..69

3.2.2 Methods 70

3.2.2.1 Sample Preparation and Data Acquisition 70
3.2.2.2 Biomarker Panel Generation...71
3.2.2.3 Identification of *m/z* Values in the Biomarker Panel..71
3.2.2.4 ELISA 72
3.2.2.5 Additional Analyses...72
3.3. Results...73
3.3.1. Generation of Biomarker Panel from MALDI-TOF-MS Data 73
3.3.2. Identification of the Peaks in the Biomarker Panel 74
3.3.3. Validation of the Peaks in the Biomarker Panel...76
3.4. Discussion 78
3.4.1. Samples..80
3.4.2. Identification of the Biomarker Panel Proteins..80
3.5. Conclusion 81
4. Proteomic Evaluation of LC-MALDI-TOF-MS as a Profiling Strategy for Biomarker Discovery in Ovarian Cancer...83
Chapter Abstract 83
4.1 Introduction..84
4.1.1 Need for Identification and Reproducibility of Biomarkers. 84
4.1.2. Liquid Chromatography...85
4.1.3. The need to Address Reproducibility to Progress with the Research in the Area...85
4.1.4. Sample Fractionation Techniques. 85
4.1.4.1. Millipore Zip Tips...86
4.1.4.2. Alkylation and Reduction. 86
4.1.5. Aims and Hypotheses of the Chapter...88
4.2. Materials and Methods...90
4.2.1. Materials 90
4.2.1.1. Equipment used..90
4.2.1.2. Reagents used: 91
4.2.1.3. Stock Solutions Made and Used 91
4.2.2. Methods 92

4.2.2.1. Production of the BSA Standard Curve. 92
4.2.2.2. QC Serum Sample 92
4.2.2.3. Multiple Workflows Tested for Optimisation 93
4.2.2.5 Alkylation and Reduction of Sera....93
4.2.2.6 Pre-digestion C_{18} ZipTip of Sera 94
4.2.2.7. Paired Comparisons 95
4.2.2.8. A Model for use on Clinical Cohort of Samples 97
4.3. Results....99
4.3.1. Semi-quantitative Nature of MALDI-TOF-MS....99
4.3.2. Reproducibility of the Third Dimension; Retention Time of the Analytical Column 101
4.3.3. Reproducibility of Identities Acquired using LC-MALDI-TOF-MS/MS 102
4.3.4. Comparison of the Proteins Identified from each Workflow....103
4.3.5. A Model for use on Valuable Clinical Samples 106
4.3.5.1. Measured Error. 106
4.3.5.2. Investigation of Peaks Calculated to be Significantly Differentially Expressed Between the two Groups. 106
4.4. Discussion....108
4.4.1. Semi-quantitative Nature of MALDI-TOF MS 108
4.4.2. Chromatographic Reproducibility 109
4.4.3. Reproducibility of Identities Acquired Using LC-MALDI-TOF-MS/MS 110
4.4.4. Comparison of the Proteins Identified from each Workflow....111
4.4.5. A Model for use on Valuable Clinical Samples 112
4.4.6. Limitations of the Methods Tested 113
4.5. Conclusion 115
4.6 Review of Findings and Future Direction for Onco-proteomics in MS 116
4.6.1. Future of Protein Mass Spectrometric Biomarker Discovery....117
4.6.2. Future for Biomarkers for Ovarian Cancer....120
5. Transcriptomics: Gene Expression Array Analysis as a Strategy for Biomarker Discovery in Ovarian Cancer 123

Chapter Abstract 123

5.1 Introduction....................123

5.1.1. Known Influences on Survival time from Ovarian Cancer 124

5.1.1.1 Platinum Resistance....................124

5.1.1.2 Epithelial to Mesenchymal Transition (EMT)....................125

5.1.2. Aims and Hypothesis of the Chapter 127

5.2 Materials and Methods....................127

5.2.1. Selection of Data Sets....................127

5.2.1.1. Array-Express Search Parameters for Sample Cohort Selection....................127

5.2.1.2. Two Data sets used for Meta-analysis 128

5.2.2. Pre-analysis Data Evaluation and Processing....................130

5.2.3. Analyses Applied 131

5.2.3.1. ANN of Short versus Long Term Survival....................131

5.2.3.2. Cross Validation with Cox Univariate Survival Analysis 132

5.2.3.3. Cross-comparison of Significant Genes 133

5.2.3.4. T-tests....................133

5.2.3.5 STRING Analysis....................133

5.3. Results....................133

5.3.1. ANN of Short and Long Term Survival....................133

5.3.2. Cox Univariate Survival Analysis....................134

5.3.3. Cross-comparison of Significant Genes 134

5.3.4. T-tests....................137

5.3.5. STRING Analysis 138

5.3.6. Interaction Intact Analysis 139

5.4. Discussion....................139

5.4.1. Comparison of Results with the Data Source Publications 140

5.4.2. Interpretation and Implications of Results....................142

5.4.2.1. Known Mechanisms of Resistance to Platinum Based Chemotherapy 142

5.4.3. Support of the Methods Used 143

5.4.4. Criticisms of the methods used 144

5.4.4.1. The Availability of Additional Information 145

5.4.4.2. Challenges in Studying Ovarian Cancer 145

5.4.4.3. The Array and Data Analysis Methodology 146

5.4.4.4. Downstream Analyses 147

5.4.5. Future work 148

5.5. Conclusion 150

6. Validation Strategies 151

Chapter Abstract 151

6.1. Introduction 151

6.2 Validation of Genes of Interest using Gene microarrays 155

6.2.1. KM Plotter Introduction 155

6.2.2. Kaplan-Meier Methods / Utilisation Strategy 155

6.2.3. Results and Discussion of KM Plotter Reports 157

6.2.3.1. Cautions and Caveats Considered when Interpreting KM Plotter Data 161

6.2.3.2. Indications for Further Research 163

6.3. Translational Validation Strategy using Immunohistochemistry 163

6.3.1. Protein Verification of EDNRA in Ovarian Tissue 176

6.3.2. Immunohistochemistry Method 176

6.3.3. Results and Discussion of Immunohistochemical Staining of EDNRA in Ovarian Tissue. 178

6.4. Discussion of Verification Strategies 183

6.5. Conclusion of Verification Strategies 187

7. Discussions for Future Work and Overall Conclusions 188

7.1. Ongoing Unmet Clinical Need and Future Work Required 188

7.1.1. Targeted Protein Mass Spectrometry Based on Transcriptomic Discovery as a Strategy for Future Work. 190

7.1.2. Clarity and Cohesiveness of Current Resources Challenge Future Work 192

7.1.3. Integration of Cross-Platform Data Challenges Future Work 193

7.2. Overall Summary and Conclusions 194

References..197

FIGURES

Figure 1. Personification of the Hallmarks of Cancer. 2

Figure 2. Histological Anatomy of the Ovary...5

Figure 3. Stages in Ovarian Cancer 7

Figure 4 The Complexity and Challenge of Studying the Human Proteome. 20

Figure 5. The Basic Components of a Mass Spectrometer. 24

Figure 6. Matrix Assisted Laser Desorption Ionisation (MALDI). 25

Figure 7. Electrospray Ionisation (ESI). 27

Figure 8. Linear and Reflectron Mode of the MALDI-TOF-MS. 29

Figure 9. Peptide Backbone Fragmentation...31

Figure 10. Liquid Chromatographic Separation Coupled to Mass Spectrometric Detection. ..37

Figure 11. A Schematic Depicting how Genomic and Proteomic Data Represent a System at Different Perspectives. 39

Figure 12. A Typical Gene Array Experiment Workflow..41

Figure 13. Typical Multilayer Perception Artificial Neural Network Architecture. 50

Figure 14. A Representation of a Hidden Layer Node from a MLP ANN...............................51

Figure 15. Population Chart of the Performance of the Biomarker Panel Discriminating Cancer from Controls. 74

Figure 16. Boxplot of Transferrin Levels as Measured by ELISA. 77

Figure 17. Boxplot of Vitronectin Levels as Measured by ELISA. 77

Figure 18. Vitronectin Levels as Measured ELISA (ug/mL) and Peptide Peak *m/z* 1647.8 Intensity. 78

Figure 19. Multiple Analysis of one Serum Sample via Different Sample Preparation Workflows. 88

Figure 20. Flow Chart of Alkylation and Reduction of QC Sera and Control Workflow.........96

Figure 21. Flow Chart of Pre-digestion Zip-tip QC Sera and Control Workflow.....................97

Figure 22. Increase of Peak Intensity with Increase in Concentration of Sample Loaded from a BSA digest. 100

Figure 23. BSA Standard Curve from Peak Intensity. 101

Figure 24. Retention Time Reproducibility. 102

Figure 25. Histogram of Protein Identity Occurrence. 103

Figure 26. Venn Diagram Comparing the Lists Protein Identities Acquired from the Alkylation and Reduction Sample Preparation Workflow and Control. 104
Figure 27. Venn Diagram Comparing the Lists of Protein Identities Acquired From the Pre-digestion ZipTip Sample Preparation Workflow and Control. 105
Figure 28. Venn Diagram Comparing the Lists of Protein Identities from Peptide IDs Matched from MS-T-test. 108
Figure 29. Number of Investigated Genes in Ovarian Cancer. 122
Figure 30. Histogram of Distribution of Survival Times of Two Cohorts of Patients with Ovarian Cancer 131
Figure 31. Overview of Gene Microarray Meta-analysis Methodology. 135
Figure 32. A Graphical Representation of the Order of Significance of the Genes of Interest. 136
Figure 33. Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) Output Displaying known Associations Between the Genes of Interest. 138
Figure 34. The Biomarker Discovery to Validation Pipeline. 153
Figure 35. Kaplan Meier Plots of the Highest Ranking Genes of Interest. 160
Figure 36. Verified GOIs Location in the Proteoglycans in Cancer pathway. 171
Figure 37. Verified GOIs in the Pathways in Cancer pathway. 174
Figure 38. Range of Staining Intensity Observed in EDNRA Stained Ovarian Tumour Tissue from the Biomax OV6161 TMA. 178
Figure 39. EDNRA Protein Expression in Ovarian Tissues of Different Stage. 179
Figure 40. EDNRA Protein Expression in Ovarian Tissues of Different Grades. 180
Figure 41. EDNRA Protein Expression in Ovarian Tissues of Different Histology. 181

TABLES

Table 1. Past Potential Markers for Ovarian Cancer...14
Table 2. A Summary of Globally Accessible Data Repositories and Resources. 58
Table 3. A Summary of Similar Research...63
Table 4. Equipment Utilised for MALDI-TOF-MS..68
Table 5. Reagents Utilised for MALDI-TOF-MS..69
Table 6. All stock solutions made and used for MALDI-TOF-MS 69
Table 7. Stepwise Analysis Generation of a Biomarker Panel. 73
Table 8. Potential Identities of the *m/z* values of the biomarker panel.. 75
Table 9. Equipment Utilised for LC-MALDI-TOF-MSMS 90
Table 10. Reagents Utilised for LC-MALDI-TOF-MSMS 91
Table 11. All stock solutions made and used for LC-MALDI-TOF-MSMS 91
Table 12. Dilutions made for production of a BSA standard curve...92
Table 13. Summary of the Workflows Applied to Replicates of one Test Sample. 93
Table 14. Numbers of Cases in Short and Long or Short Term Survival Groups. 130
Table 15. Summary of KM Plotter Analysis...158
Table 16. Aliases of the Finalised Seven Genes of Interest. (GeneCards 2015)....................165
Table 17. KEGG Pathways found to be Associated with the Shortlisted Genes of Interest. ..169
Table 18. T-test Results Comparing the Significance of Protein Expression Differences Between Cancer Stages...179
Table 19. T-test Results Comparing the Significance of Protein Expression Differences Between Cancer Grades...180
Table 20. T-test p-values Comparing EDNRA Protein Expression Between Cancer Histologies. 182

Abbreviations and Glossary

ADF	Array Design File
AI	Artificial Intelligence
Albuminome	All isoforms of albumin and bound proteins
ANN	Artificial Neural Network
AOCS	The Australian Ovarian Cancer Study
Ascites	Fluid accumulation in a peritoneal cavity
Biomarker	A naturally occurring molecule or characteristic from which a disease or condition can be identified
Bucket Table	The table of data bins defined within the Bruker software containing, compiled, aligned LC-MALDI-TOF data of mass, retention time and peak intensity.
CA125	Cancer Antigen 125
CAD (HPLC)	Charged Aerosol Detection HPLC
Cancer	A disease resulting from the uncontrolled division and proliferation of cells.
CEA	Carcinoembryonic Antigen
CHCA	Alpha-Cyano-4-hydroxycinnamic acid
CRUK	Cancer Research United Kingdom
CT Scan	Computerised Tomography
DCN	Decorin
DDA	Data Dependent Acquisition
hhH$_2$O	Deionised and distilled water
DESI	Desorption Electrospray Ionisation
DHB	2,5-dihydroxybenzoic acid
DIA	Data Independent Acquisition
EDNRA	Endothelin receptor type A
ELISA	Enzyme Linked Immunosorbent Assay
ELSD (HPLC)	Evaporative Light Screening Detection HPLC
EMBL-EBI	European Bioinformatics Institute part of the European Molecular Biology Laboratory
ESI	Electrospray Ionisation
FAIMS	Field Asymmetric Ion Mobility Spectrometry
FDA	U.S. Food & Drug Administration
FDR	False Discovery Rate
Gene Microarrays	A group of techniques where short strands of nucleic acid are immobilised to an array and through complementary binding are used to ascertain and quantify the expression of target genes.
Genomics	The study of a genome
GEO	Gene Expression Omnibus
GOI	Genes of Interest
HPLC	High-Performance Liquid Chromatography
ICAT	Isotope Coded Affinity Tagging

ICPL	Isotope-coded Protein Labelling
IGF2	Insulin-like growth factor 2
IHC	Immunohistochemistry
IMEX	The International Molecular Exchange Consortium
Immunoassays	Assays including immunoglobulins for antigen recognition
InnateDB	Immune Response pathway database.
IntAct	A molecular interaction database hosted by the EMBL-EBI
iTRAQ	Isobaric tagging for relative and absolute quantitation
I2D	Interlogous Interaction Database
KEGG	Kyoto Encyclopaedia of Genes and Genomes
KNNs	Nearest Neighbour Analysis
LC	Liquid Chromatography
LPA	Lysophosphatidic acid
MALDI	Matrix Assisted Laser Desorption and Ionisation
MatrixDB	Extracellular Matrix Interaction Database
MBInfo	Mechanobiology database
MCCV	Monte Carlo Cross Validation
MINT	Molecular Interaction Database
MLP	Multilayer Perceptron
MS	Mass Spectrometry
MS/MS / MS^2	Tandem mass spectrometry
NAV3	Neuron Navigator 3
NCBI	National Centre for Biotechnology Information
NGS	Next Generation Sequencing
NHS	National Health Trust
NICE	The National Institute for Health and Care Excellence
OCAC	Ovarian Cancer Association Consortium
OCTIPS	Ovarian Cancer Therapy Innovative Models Prolong Survival
Oncogene	A gene coding potentially carcinogenic traits
OTTA	Ovarian Tumour Tissue Analysis Consortium
PCA	Principal Component Analysis
pI	Isoelectric point.
PIR	Protein Information Resource
PPFIB1	PTPRF interacting protein, binding protein 1 (liprin beta 1)
PRIDE	The Proteomics IDEntification database
Proteomics	The study of the proteome
PSA	Prostate Specific Antigen
REIMS	Rapid Evaporation Ionisation Mass Spectrometry
RVM	Relevance Vector Machines
SDRF	Sample and Data Relationship Format
Shotgun proteomics	A proteomic strategy aiming to measure the entire proteome and measure the difference between two groups
SIBm	Swiss Institute of Bioinformatics
SILAC	Stable Isotope Labelling with Amino acids in Cell culture

SPL	Scheduled Precursor List
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVM	Support Vector Machines
Targeted Mass Spectrometry	Mass spectrometric strategy aiming to quantify a predetermined set of proteins
TCGA	The Cancer Genome Atlas
TMA	Tissue Microarray
TMT	Tandem mass tags
TNFAIP6	Tumour Necrosis Factor, Alpha-Induced Protein 6
TOF	Time of Flight
TP53	Tumour Protein 53
Transcriptomics	The study of the transcriptome
UCL-BHF	Cardiovascular Gene Annotation Initiative funded by the British Heart Foundation
UFX	Bruker UltrafleXtreme
UKCTOCS	United Kingdom Collaborative Trial of Ovarian Cancer Screening
UV	Ultraviolet
WHO	World Health Organisation
WTAP	Wilms Tumour 1 associated protein

HYPOTHESES

Section		H_1	H_0	**Page**
3.1.2	i	Unique protein patterns expressed in the sera of ovarian cancer sufferers can be detected using MALDI-TOF-MS with ANNs and can be used to positively identify a blind validation set.	MALDI-TOF-MS with ANN analysis will not be able to detect unique protein patterns expressed in the sera of ovarian cancer sufferers.	67
	ii	The masses of the peptide peaks expressed differentially in the tested serum can be assigned a protein identity by linking the MS-MALDI data with LC-MALDI-MS/MS data.	The masses of the peptide peaks expressed differentially in the tested serum cannot be identified by linking data from LC-MALDI-MS with the MS profiles.	67
	iii	Unique protein patterns expressed in the sera of ovarian cancer sufferers can be detected using MALDI-TOF-MS with ANNs and can be used to positively identify a blind validation set.	No difference in expression will be noted of the proteins demonstrated to be expressed using MALDI-MS.	68
4.1.5	iv	The signal intensity values of a detected protein are relative to the amount of protein loaded.	There is no correlation between protein amount loaded for detection and the signal intensity of protein detected.	88
	v	One sample preparation technique prior to LC-MALDI-MS will yield a greater amount of meaningful protein identities.	All tested sample preparation techniques prior to LC-MALDI-MS produce equal amounts of meaningful protein identities.	88
	vi	Differences will be seen in the LC-MALDI profiles of serum samples prepared under identical conditions.	There will be no significant difference between the LC-MALDI-MS profiles of serum samples prepared under identical conditions.	89
5.1.2.	vii	Genes will be found to be consistently significantly associated with survival time when a complement of statistical strategies are applied in a meta-analysis approach to two separate cohorts of patients measured with two different microarray platforms	None of the gene expression measurements from the in two cohorts will be found to be consistently associated with survival times from ovarian cancer when tested with a complement of statistical strategies.	127
6.2.1.	viii	Genes of interest will be verified to significantly associate with ovarian cancer survival time when investigated on a wider sample cohort.	None of the genes of interest found to significantly associate with ovarian cancer survival time will be verified to do so when investigated on a larger sample cohort.	155
6.3.1	ix	Protein expression of *EDNRA* will be found to be different between different stages, grades and histologies of ovarian cancer samples.	No difference in protein expression will be observed between different stages, grades and histologies of ovarian cancer.	176

Thesis Abstract

Novel, specific and sensitive biomarkers are prerequisite to improve diagnosis and prognosis of patients with ovarian cancer. Firstly, a proteomic bottom-up MALDI-TOF mass spectrometric profiling analysis was conducted on a cohort of sixty serum samples specifically collected for this purpose. An in-house stepwise Artificial Neural Network (ANN) algorithm generated a biomarker panel of *m/z* peaks which differentiated cancer from aged matched controls with an accuracy of 91% and error of 9%, identities were inferred where possible and validation conducted using ELISA on the same cohort. Lack of complete verification, or the ability to verify the full panel lead to an in-depth evaluation of the strategy used with the aim to repeat with an improved methodology. Following this, a feasibility analysis and evaluation was performed on the next generation of equipment for sample fractionation prior to analysis on multiple replicates of stock human serum collected in the same way as the ovarian cohort. The results of which combined with the limited amount of available ovarian cancer sample cohort altered the trajectory of the project to the mining of transcriptomic data acquired from an online data repository. A meta-analysis approach was applied to two carefully selected gene expression microarray data sets ANNs, Cox Univariate Survival analyses and T-tests were used to filter genes whose expression were consistently significantly associated with patient survival times. A list of 56 genes were refined from a potential 37000 gene probes to be taken forward for verification for which more freely available online resources such as SRING, Kaplan Meier Plotter and KEGG were utilised. The list of 56 genes of interest were refined to seven using a larger cohort of transcriptomic data, of the seven one, *EDNRA*, was selected for translational verification using immunohistochemistry of a tissue microarray of ovarian cancer specimens. Significant association is seen with cancer stage, grade and histology. The merits and flaws of the verification are discussed and future work and direction for research is suggested.

1. **Introduction**

1.1. General Concepts

1.1.1. Cancer

Cancer will at some point effect most people in Western society (Scotting 2011), nearly 50% of the population will receive a cancer diagnosis in their lifetime (Ahmad *et al.,* 2015) and the rest will most likely know someone directly affected. In the UK in 2012 there were 338,623 new cases of cancer diagnosed, and 161,823 consequent deaths (Cancer Research UK 2016). Fifty percent of people diagnosed in 2012 were predicted to survive for 10 years or more (Cancer Research UK 2016). It is the second most common fatal disease in the UK, following only heart disease (Scotting 2011).

Cancer is a condition of a cell where it has lost its ability to regulate growth. Cancerous cells are able to migrate to other organs in the body where they may continue to proliferate uncontrollably, eventually interfering with homeostatic cell, system and organ function, until potential complete upheaval then malfunction of tissue function (Cooper 2000).
Mutations in the genetic code can subtly alter genes coding for proteins essential for normal cell growth, regulation and homeostasis resulting in a traits indicative of cancer; oncogenesis. Cells containing oncogenes and translated onco-proteins exhibit cancerous phenotypes involved in the cell regulatory process resulting in uncontrolled growth. Malignant cells are morphologically, genetically and phenotypically distinguishable from normal tissue (Baba and Câtoi 2007).

Cancer is primarily subcategorised by the origin site of a primary tumour. There are over 100 types of cancer by this definition. However, common characteristics are noted between cancers of different origins and sometimes treated with the same therapy (Barretina *et al.,* 2012). Equally, the diversity of pheno- and genotypes of cancers from one origin organ can be wide ranging and most subtypes are continually being defined/clarified, notably breast cancer is now able to be grouped by genotype into specific subtype for a more targeted treatment (Dent *et al.,* 2007, Banerji *et al.,* 2012 and Caldas and Stingl 2007).

A large proportion of people suffering from or affected by cancer are unaware of the complex and conflicting/ complex molecular mechanisms in play. Often human characteristics are used to attain a lay understanding. Figure 1 is a personification of Hanahan and Weinbergs (2000, 2011) widely used/known depiction of cellular attributes hallmarking cancer. Stickmen represent how cells within a cancer cell 'act selfishly' (Scotting 2011) making (cellular) environmental alterations for personal gain. The lack of programmed cell death, senescence or other noble self-limiting traits of non-cancer cells is in disregard to the (cellular) society they are in. Sooner or later the rebellious actions of the cells, like colonising other tissues; metastasis, altering existing resources and supply routes; angiogenesis, is destructive to neighbouring and non-adjacent organs.

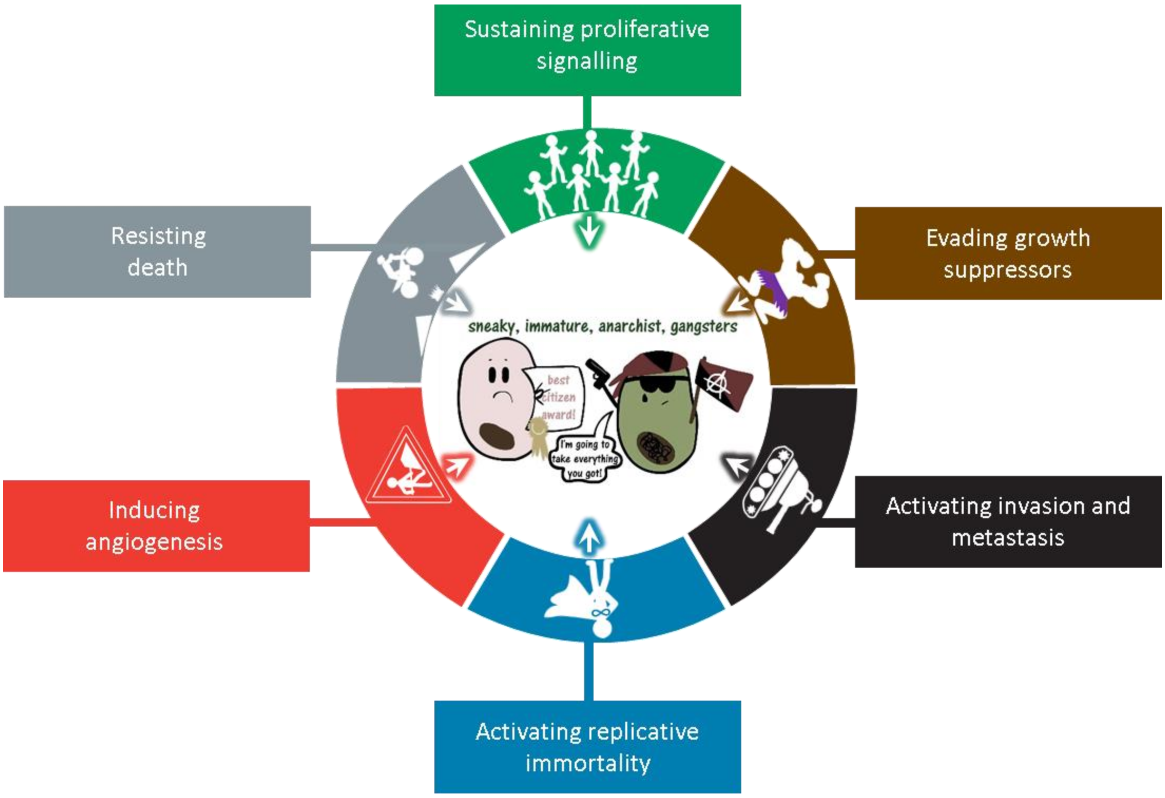

Sustaining proliferative signalling
Evading growth suppressors
Activating invasion and metastasis
Activating replicative immortality
Inducing angiogenesis
Resisting death
sneaky, immature, anarchist, gangsters
best citizen award!
I'm going to take everything you got!

Figure 1. Personification of the Hallmarks of Cancer. An adaptation of Hanahan and Weinburg (2011) with Illustrated Health (2014)
Each of the cell characteristics used to classify cancer are satirised into a bad human characteristic. The analogy being; cancer harms the body as some human characteristics do to a society. In the six sections key characteristics are represented by a pictograms: From the top centre and continuing clockwise: In green multiple stickmen represent limitless proliferation - overcrowding straining recourses, the hulk-like character in brown represents increased growth – greed or an inadvertent overpowering by size, in black the tank driver presents invasion – metastasis, in blue the infinity symbol represents the immortality – lack of a natural death, in red roadworks symbol a sign well associated with disruption of traffic infrastructure to redirect supply routes, finally, in grey a character performing a death defying stunt - resisting death.

In 2000 and again in 2011 Hanahan, and Weinberg compiled cancer literature and defined six hallmarks of cancer, all cancer traits can be categorised as one or more hallmark, phenotypical of cancers; these are summarised below and a brief example outlined for each.

- **Uncontrolled cell proliferation, or the dis-regulation of cell proliferation.** Cancer cells may display up-regulation of cell surface receptors to growth factors, typically tyrosine kinases, the receptors themselves may be onco-proteins, altered, activating independently or change in tertiary structure increasing affinity to ligands, or cells may release the ligands growth factor themselves. Additionally, cancer cells have exhibited altered glycolic metabolism sometimes preferring aerobic glycolysis when oxygen is available. In multiple examples an altered/dysfunctional/onco-protein plays a key role in transmitting or receiving in a negative feedback loop in a cell growth system. Such examples include *PTEN* and *mTOR* kinase, both are normally transducers of a signal which in normal cell growth are triggered to signal for cessation of further growth.

- **Evading growth suppressors and un-controlled cell growth**. A renowned, well characterised example of which is Tumour Protein 53 (TP53) protein, responsible for adjudicating the decision/molecular outcome as to whether a cell proliferates, undergoes senescence or apoptosis. Mutations in, or faulty production of, *TP53* protein products, many of which have been characterised, results in a loss authority/governance within the system.
- **The ability to induce angiogenesis.** Tumours over ~1cm display the ability to induce the growth of neo-vasculature from otherwise quiescent adjacent blood vessels. Like all cells the supply of nutrients and removal of waste products is requisite. Descriptively named -Vascular Endothelial Growth Factor-A (*VEGF-A*) and downstream effectors of it have been noted in numerous tumour types, it's up-regulation is known to be triggered by hypoxia, a natural consequence of excessive tissue growth.
- **The ability to invade and metastasise to other organs/sites in the body**. Metastasis is a multistep process sometimes referred to as the invasion-metastasis cascade. Cancer cells have been shown to release factors, Matrix Metalloproteases, which disrupt the extracellular cellular bonds and status quo. Further to this cancer cells enter the lymphatic system bloodstream where they are transported to distant parts of the body where they settle and continue to mitose/colonise/grow/duplicate.
- **The ability to evade immune detection. Cancer immunology is a wide and growing field**. Cancerous cells are thus harder to detect by immune system than foreign invaders. Additionally, if triggered the immune system can exacerbate a cancerous environment if an inflammatory process is activated releasing cells/biomolecules/creating an environment to promote tumour growth, nurturing conditions for angiogenesis, cell growth and proliferation and invasiveness.
- **Replicative immortality or the lack of programmed cell death**. Telomeres are comprised of repeating hexonuclotides cap each chromosome within a cell nucleus, as well as having a barrier protective role they are shortened every time the cell undergoes

mitosis. Telomerase is able to counter this shortening adding hexonucleotides lengthening the telomeres and increasing the number of mitotic events before irreparable damage of the DNA chromosome ends thus triggering apoptosis. Up-regulation of telomerase has become a common trait in the immortalisation of cancer cell lines.

Eleven years later Hanahan, and Weinberg (2011) narrate the following decades of cancer research to define two more emerging hallmarks and two enabling characteristics of cancer. The additional hallmarks are; deregulating cellular energetics and avoiding immune destruction. The enabling characteristics being genome instability and mutation and tumour promoting inflammation. The reader is referred to Hanahan and Weinberg (2000) and Hanahan and Weinberg (2011) for a detailed benchmarking definition and characterisation of the phenotypes of cancer.

A poignant progression when the literature is summarised, is the change of emphasis from cancer cells alone, to put them in a scene of a cellular microenvironment, and the contribution of and communication with pericytes and paracrine signalling (Hanahan and Weinberg 2011).

Identifying and understanding the specific molecular pathways and mechanisms responsible for the malevolent characteristics of malignant cells will expose ways to detect, treat and even prevent cancer. Under the premise that molecules such as proteins are secreted from, shed by or released in response to the tumour microenvironment into the circulation, cancer research endeavours to detect these molecules for use as a biomarker in serum samples.

1.1.2. Ovarian Cancer

Typical/normal ovarian function: Ovaries are almond shaped structures approximately 2 x 3 x 4 cm located within the female pelvis at the top of the genital tract. Their role is to generate and release germ cells into the reproductive system. They are suspended in the opening to the fallopian tubes by ligament and connective tissues called the tunica albuginea, this is covered by the germinal epithelium which is a simple squamous mesothelium (Peckham *et al.,* 2004). In the endocrine system, ovaries release oestrogen and progesterone and are stimulated by gonadotrophin which is released from the anterior pituitary. The ovary is the female gonad, and is the site of oogenesis within the ovary germ cells mature from Primordial follicles mature to Secondary, to mature then Graaffian Follicle phase to be released as into fallopian tube (Peckham *et al.,* 2004). Other cells found within ovarian tissue include epithelium surrounding the capsule and stroma creating structural foundation to the tissue. Ovum mature and are released from the ovary surface as part of the menstrual cycle, corpus luteum cyst is the term for an ovarian cyst that may burst around the time of menstruation, repair of this action can take up to 3 months (Adam *et al.,* 2012) Follicular and/or Granulosa cells "are somatic cells of the sex cord that are closely associated with the developing female gamete" (Adam *et al.,* 2012) The Anral follicle, also known as a Graafian follicle is the term for the mature ovum cyst prior to rupture and releasing the ovum into the fimbriae and the fallopian tubes, Folicular fluid surrounds the ovum and fills the ovum follicle (Adam *et al.,* 2012)

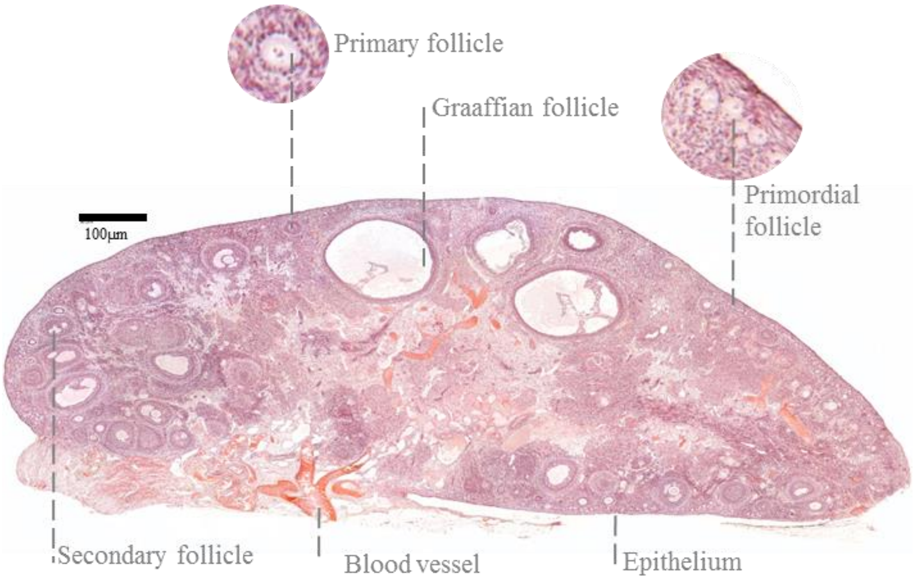


Figure 2. Histological Anatomy of the Ovary
Annotated from (Peckham *et al.*, 2004) A histological cross section of a human ovary. Stages in oogenesis are observed and in different locations within the section: Germ cells mature from Primordial follicles to Secondary, then Mature then Graaffian Follicle phase and are released into the phallopian tube

A subtype of follicle epithelial cells known as border cells are of interest as a cancer model and have been used as a model in studies researching metastasis on account of their unique migratory characteristics and ability to invade adjacent tissue; a number of their characteristic genes have been identified in cancer cell lines (Naora *et al.,* 2005). Primates are often used as an ovarian model as healthy human ovarian samples are in shorter supply (Adam *et al.,* 2012) however cannot fully represent a human genome.

Incidence. With approximately 136 new diagnoses each week in 2011 in the UK alone, ovarian cancer is the 5th most common cancer in the UK (Cancer Research UK 2015), it is the fourth most common cancer in US females aged 40-59 and 5th most in US females aged 60-79 (Siegel *et al.,* 2013).

Survival. The key prognostic for the survival time is the stage and grade at diagnosis (Erickson *et al.,* 2014). A 92%, 5-year survival can be expected from a Stage 1 diagnosis, this drops to 22% at Stage 3. Little changed in 5-year survival rates between 1975 and 2008 (Vaughan *et al*., 2012, Siegel *et al.,* 2013). Unfortunately, due to the asymptomatic nature of the early stages, its insidious growth pattern of the disease and the lack of a sensitive screening tool, over half of ovarian cancer is diagnosed at Stage 3 or above (Cancer Research UK 2012).

When diagnosed, ovarian cancer can be categorised by stage and grade to determine the prognosis and direct treatment. Tumour grade refers to cell morphology with the tumour and the Stage refers to the occurrence and distance of secondary tumours from the primary tumour site; metastasis. Figure 3 below illustrates the typical abdominal distribution metastasis of a Stage 3 ovarian cancer. The high morbidity of ovarian cancer is often attributed to the majority being diagnosed at a later stage. The ability to stratify patients with this heterogeneous disease, based on identification of molecular pathways, would enable precision treatment and improve prognosis. Hundreds of genes have been significantly associated with ovarian cancer yet few have been verified by peer research (Braem *et al.,* 2011).

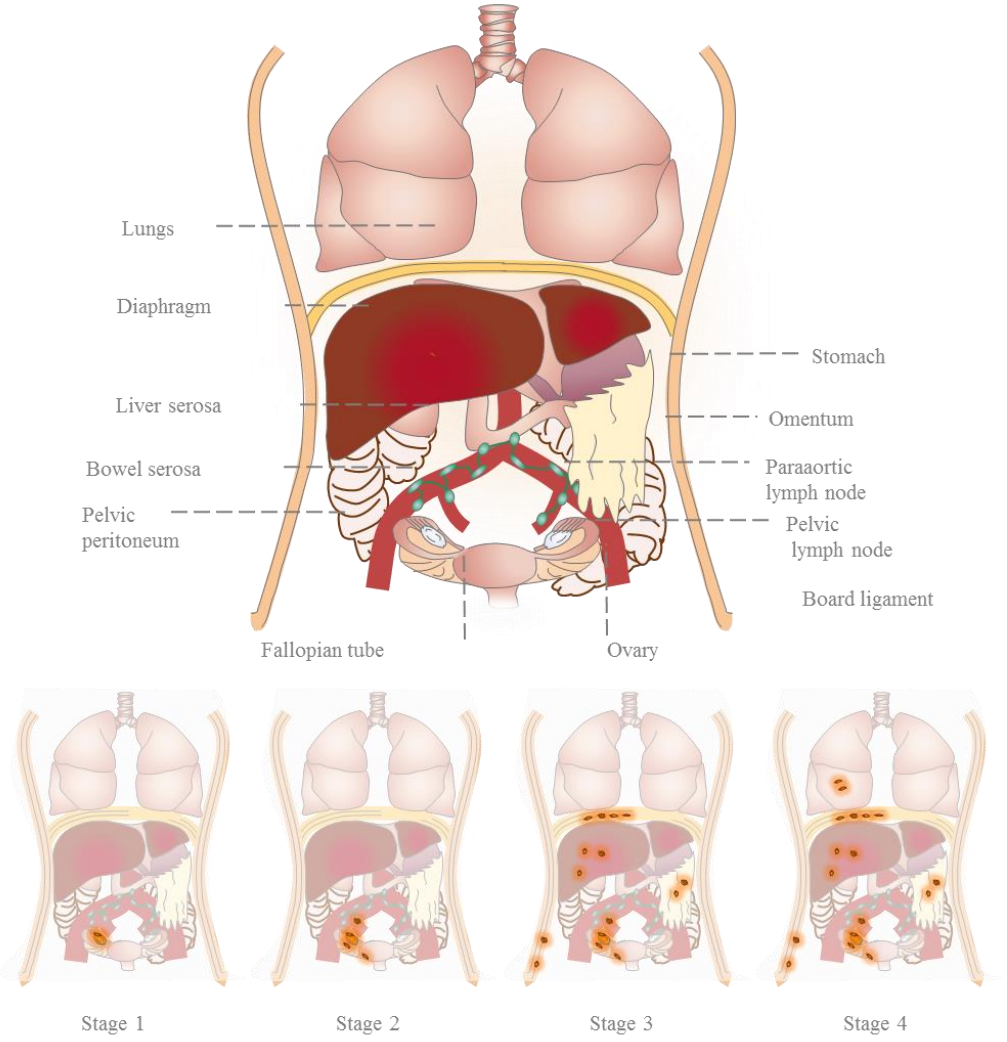


Figure 3. Stages in Ovarian Cancer
Adapted from Naora *et al.*, (2005); cancer cells are confined to one (1a) or both (1b) ovaries and may also be present on the surface of the ovary or ascites (1c). **Stage 2;** local metastasis where the cancer lesions are also found in the fallopian tubes or womb (2a), other local organs such as bladder or bowel (2b) and may also be present in ascites (2c). **Stage 3;** abdominal metastasis, cancer cells (3a) or larger visible lesions (3b) are found on the lining of the abdomen, or in the lymph nodes and upper abdomen and or groin (3c). **Stage 4**; distant metastasis, tumours found outside of the peritoneum or inside other organs for example within the liver or lungs.

The underlying reason for late stage diagnosis is the asymptomatic nature of the early stage disease. Few if any symptoms are expected from Stage 1 and 2 disease and indicators of the later stages often at best vague and easily miss-attributed to general less serious complaints including; back or abdominal pain, bloating or abnormal menstrual patterns.

Currently, factors known to influence a patients' survival time from ovarian cancer include but are not limited to the histology and grade of the tumour (Matuzaki *et al.,* 2015), distance of metastasis or stage and, if the cancer displays resistance to chemotherapy. Some chemo-resistant molecular pathways, mainly involved in DNA repair have been demonstrated in some ovarian cancer cell lines (Marchini *et al.,* 2013) but this has not yet been extrapolated to apply to the general population. Specific pathways are discussed at molecular level in (Chapter 5).

Cytology. As yet, there is no defined pre-malignant stage, as there is in cervical or prostate cancer (cervical/ prostate intraepithelial neoplasia).

Prognosis. Only 22% of patients diagnosed at Stage 3 are expected to live for 5 or more years, this is improved to 92% if diagnosed at Stage 1. Other than the increase in reported incidence in the early part of the 20th century nothing to date has made a dramatic impact on the death rates from ovarian cancer (Siegel *et al.,* 2013).

Currently there is no screening tool with a performance specific or accurate enough to be implemented to the general population.

Current Treatment. Despite the continuing study of ovarian cancer cell lines and patient material with numerous publications implicating novel genes associating with its incidence, little has changed in the treatment and expected outcome of patients presenting with ovarian cancer. Platinum based chemotherapy sometimes administered with an adjuvant. A response to which is seen in approximately 70% of patients, however most will develop a resistance to the therapy and experience a recurrence of tumour some more aggressively than others (Miller *et al.,* 2009). Repeated cycles of platinum therapies are administered for most recurrent disease, however, typically the length of progression free survival shortens due to chemo-resistance until the disease is terminal (Marchini *et al.,* 2013). Additionally, not all patients diagnosed with the disease are eligible for treatment (Erickson *et al.,* 2014)

Chemotherapy. Platinum based chemotherapies act by binding directly to DNA strands and disrupting the cells ability to divide. Historically cisplatin was the original platinum therapy this was replaced with Carboplatin which is less toxic to other organs, more recently Oxaplatin was developed which is still considered an analogue but has been shown to be effective were resistance to Carboplatin or cisplatin has occurred (Martin *et al.,* 2008). This treatment pathway yields 50% 1.5-year progression free survival of patients diagnosed with Stage 3 ovarian cancer 20-30% of these patients will progress after this with 10-year survival rates as low as 10% (Marcus *et al.,* 2014).

More recent therapies target the tumour microenvironment, such as Bevacizumab which inhibits the angeogenic pathway (Kim *et al.,* 2012). Bevacizumab has been administered as an adjuvant in disease recurrence after resistance to platinum chemotherapy has occurred with some improvement in survival, it has also been trialled as an adjuvant to first line therapy alongside cisplatin in platinum-sensitive cases (Vaughan *et al.,* 2012). However, resistance to anti-VEGF agents such as Bevacizumab have been reported (Vaughan *et al.,* 2012).

Preliminary studies have identified some success using immune therapies, were by antigenic stimulation of T-cells the body's natural anti-tumour response and can be stimulated to recognise and eliminate tumour (Vaughan *et al.,* 2012). Immunotherapy strategies are developing quickly for many cancers, however, identifying the immunogenic biomarker is a key prerequisite to this.

Metastatic Pattern, Nomenclature and Peritoneal Cancer. Ovarian cancer metastatic pattern is distinctive from other cancers in that, although spread is seen and defined by its presence in local and distant lymph nodes and blood vessels it also 'seeds' in to adjacent organs via aescetic fluid to form numerous lesions across the abdominal cavity (Naora *et al.,* 2005, Vaughan *et al.,* 2012) as seen in Figure 3. For this reason, the presence of aescitic fluid is associated with a poor prognosis (Rosanò *et al.,* 2011). Surgical removal of innumerable tiny lesions requires radical surgery at least and could be considered near-futile, thus debulking and adjuvant chemotherapy is the best possibility.

It has been agreed among experts that what falls under the label ovarian cancer could originate from a number of tissues of vastly differing in histology. It has been suggested that the term ovarian cancer replaced with "pelvic" or "peritoneal" but it was agreed to be too confusing to change the meanings (Vaughan *et al.,* 2012).
Research has shown that metastatic spread is not a random event and that cancer cells can also be directed by factors such as a chemokine gradient (Scotton *et al.,* 2001).

Immune response in the tumour microenvironment. There is a strong body of evidence uncovering the role of chemokines and the immune system in orchestrating angiogenesis, metastatic patterns as well as directing T-cell directed anti-tumour responses and inhibition of apoptosis in the tumour microenvironment, which is of use for sub-typing and identifying targets for therapies (Obermajer *et al.,* 2011, Balkwill *et al.,* 2004, Vaughan *et al.,* 2012).

Risk Factors.

- First degree female relative with ovarian cancer
- Tobacco smoking

- A postmenopausal status
- age of >50 years
- *BRCA1* and *BRCA2* mutations
- Years of oral contraceptive use
- Other pre-existing conditions such as polycystic ovarian disease
- Parity (number of times a woman has given birth to a foetus with a gestational age of 24 weeks or more)

1.1.3. Biomarkers

A biomarker is defined as "a naturally occurring molecule, gene or characteristic by which a particular pathological or physiological process, disease, etc. can be identified" (Oxford Dictionaries 2015). Or, a measurable factor that is used to represent a clinical end point (Strimbu *et al.,* 2011).

In this context, a biomarker is defined as a measurable biochemical found in bodily tissue believed to be produced by, or in response to, diseased tissue in the body. The objective of biomarker discovery research is to identify non-invasive methods to detect specific, sensitive and accurate markers of disease. A specific, sensitive, reliable biomarker may be applied as a screening tool for the general population to detect early stage disease, or to known sufferers of a disease to stratify the most appropriate treatment or monitor the progression or reoccurrence.

In a standard clinical setting, biomarkers can be grouped as either:

- Diagnostic: The presence or absence of the biomarker can be used a classifier, to diagnose a disease or clinical condition.
- Prognostic. The presence or absence of the biomarker can be used to assign a likely cause of a disease or clinical condition.
- Predictive. Predictive biomarkers can be used to categorise subpopulations of patients and used as a marker of risk or likely hood of an event. For example, a likely response to a given therapy.

Biomarker discovery experiments aim to stratify patients according to clinical parameters or therapeutic response, it can also be the optimal scenario that they also are appropriate target genes / proteins for therapeutic intervention. For example, in breast cancer an overexpression of the Her2/neu receptor correlates with poor prognosis and likelihood of metastasis (Carmen *et al.,* 2008). It is also the target of therapy, trastuzumab (Herceptin). *HAGE* (*DDX43*) has been shown to be overexpressed in sarcoma, testis and breast solid tumours (Abdel-Fatah *et al.,* 2014), and, has also shown immunogenic potential with view to be used as an immunotherapeutic target (Mathieu *et al.,* 2007).

1.1.3.1. Essential and Desirable biomarker properties

A biomarker is only able to progress from scientific discovery to clinical implementation firstly though extensive scientific peer reviewed research, followed by the rigour of all stages of clinical trials (de Gramont *et al.,* 2014, Henry *et al.,* 2012 and Goossens *et al.,* 2015), for this reason there are few new fully approved biomarkers. Anderson (2010) reports the rate at which novel protein analytes are introduced has stabilised and remained the same for 15 years, at an average of 1.5 per year.

A clinically useful biomarker test must be:

- Biochemically stable.
- Specific and sensitive enough to minimise the number of false positives and false negatives respectively. Specificity of >99% and positive predictive value of 10% (Hays *et al.,* 2010). Jacobs *et al*., (2004) state most researchers in the area agree at no more than 1 false positive for every nine true positives and a 99.6% specificity.

A clinically useful biomarker would ideally be:

- Detectable from sample attained from a non-invasive method i.e. urine or blood sample, not tumour biopsy or exploratory surgery.
- Unaffected by natural variations caused by circadian rhythm, seasonal rhythm, diet, lifestyle, sex and race.
- In the case of a combination of biomarkers compiling a clinical test, the biomarker panel must contain no more than four or five biomarkers to make it a marketable tool (NBDA 2016).

The specificity and sensitivity of a biomarker is needed to calculate the risk to potential patients. In a clinical setting, false positive results cause unnecessary harmful exploratory surgery or treatment, false negatives result with disease going undiagnosed or untreated and therefore likely to worsen.

Biomarkers currently used in clinical practice to detect or monitor progression of cancer include Cancer Antigen 15.3 (CA15.3) for breast cancer, Cancer Antigen 19.9 for pancreatic cancer, Prostate Specific Antigen (PSA) for prostate cancer, Cancer Antigen 125 (CA125) for ovarian cancer and Carcinoembryonic Antigen (CEA) for colorectal and other cancers (Engwegen *et al.,* 2006, Hanash *et al.,* 2008).

The predictive performance can sometimes be improved by concurrent measurements, a biomarker panel. However, less than half of FDA approved biomarkers have more than one protein analyte (Anderson 2010). Screening strategies may be based on other factors, such as cytology of a collected specimen for pap smear tests for cervical cancer.

Existing monitoring of ovarian cancer progression or recurrence assays the levels of circulating Cancer Antigen 125 (CA125) and carcino-embryonic antigen (CEA) in blood, however these tests are flawed by the natural variation and fluctuations of these proteins resulting in false positives and unnecessary explorative surgery.

Strimbu *et al.,* (2011) critiques the current conceptual status of biomarkers as clinical diagnostic tools and identifies room for vast improvement. In clinical settings and studies the use of a biomarker is often necessary to make a clinical endpoint measurable, however is a reductionist view and does not allow for consideration of wider influences to the measured system. Strimbu *et al.,* (2011) concludes that we will only be able to use biomarkers to represent clinical endpoints when we fully map out and understand all of the biomolecular interactions within normal physiology which is not currently the case. This notion is also outlined by Hanahan and Weinburg (2011), who in their decennial review of cancer explain how fully mapping heterotypic as well as atypical cellular molecular circuits is central to the understanding of cancer and future personalised or now more realistically “precision” (Goossens *et al.,* 2015) medicine. They predict that over the next decade mapping of cellular mechanisms will “eclipse” current knowledge. These advances should increase the confidence in the measured biomolecules chosen to represent a clinical endpoint.

1.1.4. The Need for Effective Screening Strategies

Nearly 50 years ago the World Health Organisation (WHO) identified the number one priority for ovarian cancer as being a screen for early stage ovarian cancer in the asymptomatic population (Wilson and Junger (1968) in Nossov *et al.,* (2008)), however, one with the required specificity or sensitivity is yet to be verified.

Many currently used diagnostic biomarkers and biomarker panels listed above are not specific or sensitive enough to be implemented as a screening tool; this poor performance also events in misdiagnosis and false positives which further risks the lives of patients. For example, CA-125 has a sensitivity and specificity of (85-90%) and is less sensitive to detection the early, asymptomatic stage ovarian disease where treatment has an enormously greater impact on 5-year survival, or specific enough to distinguish many benign from malignant growths leading to unnecessary and harmful investigative biopsy procedures (Timms *et al.,* 2011, Buys *et al.,* 2011). CA125 is only elevated in 60-80% of ovarian cancer patients, it is more sensitive to the later stage and serous cancers, however does not perform as well to detect Stage 1 (50% sensitivity), or other histological subtypes such as mucinous (Marcus *et al.,* 2014). For every 100 patients with an ovarian cancer screened for CA125 either as follow up for a previous cancer or for suspected new cancer, 15 will not have a serum CA125 level above the normal distribution, thus leaving 15 cancer patients with false negative screen and potentially untreated. Currently, a “high” CA125 blood level (above 35IU/ml) followed by a ultrasonogram indicating ovarian cancer - a positive biomarker screen, would most likely need to be investigated by explorative surgery to attain a biopsy for a conclusive diagnosis (NICE, 2016). Due to the location of the ovaries all surgery, even laparoscopic, has associated risks including general anaesthetic.

Prostate Serum Antigen (PSA) is an example of an unstable biomarker. PSA is a kallikrein protease expressed exclusively in the epithelial cells of normal, benign and malignant prostate, (Oesterling *et al.,* 1991). Its measurable presence in the serum make it a convenient biomarker to detect and monitor prostate cancer and is the main tool utilised for this by the NHS today. Unfortunately, both false negatives and false positives are common, PSA serum levels are increased in benign prostatic hyperplasia, in certain ethnic groups, bacterial prostatitis, and acute urinary retention, all common conditions. Further to this, PSA binds to other circulating serum proteins so is present in multiple forms bound and unbound (Catalona *et al.,* 1996) only the non-bound molecule will be measured. Hence PSA is not accurate enough to rely on alone to monitor cancer progression or recurrence. An invasive biopsy is the only route more conclusive diagnosis, although, often still hold question. It is important to identify accurate biomarkers or biomarker panels to improve diagnosis, monitor progression and predict a patient’s response to a therapy.

Other serum biomarkers investigated as potential ovarian cancer screening tools include CA72-4 or TAG72, CA125, LASA, CA15-3, CA19-9, CA54/61, Serum macrophage colony-stimulating factor (M-CSF), Monoclonal antibody OVX (OVX1), Lysophosphatidic acid (LPA), Prostasin, Osteopontin, HE4 (Homosapiens epididymis specific 4, Inhibin, and various Kallikreins (Jacobs *et al.,* (2004), Nossov *et al.,* (2008)). Table 1 below, summarises some investigated promising biomarkers of interest from a comprehensive review (Jacobs *et al.,* 2004). The reader is referred to Jacobs *et al.,* (2004) for a full review on investigating novel biomarkers, panels combining existing biomarkers and other screening strategies tested in ovarian cancer worldwide.

Table 1. Past Potential Markers for Ovarian Cancer. Summarised from Jacobs *et al.,* (2004) lists some select past potential markers for ovarian cancer.

Abbreviation	Summary
CA72-4 or TAG 72	Cancer antigen 72 (CA72-4) also known as tumour-associated glycoprotein 72 (TAG 72) is a glycoprotein surface antigen found in gastric, colon, and ovarian cancer. Higher expression has been observed in mucinous tumours. It has been investigated as marker panel with CA125 but no conclusive data.
M-CSF	Serum macrophage colony-stimulating factor (M-CSF) is a cytokine released by normal as well as neoplastic ovarian epithelium. Elevated levels have been demonstrated in 68% ovarian cancer compared to 2% of those classified as healthy controls. M-CSF has been shown to be sensitive in ovarian cancer cases where CA125 is not elevated.
OVX1	Monoclonal antibody OVX1 specifically binds an antigenic determinant found in ovarian and breast cells. Combining OVX1 and M-CSF with CA125 yields a higher sensitivity for the detection of earlier ovarian cancer than CA125 alone. However, the methodology used to conclude this is susceptible to sample handling instability.
LPA	Lysophosphatidic acid (LPA) is a bioactive phospholipid has mitogenic potential its functions with similarity to growth factors. LPA has been shown to stimulate the growth of cancer cells. Plasma levels of LPA are under investigation as a biomarker of ovarian as well as other gynaecologic cancer. Increased LPA levels were detected in the plasma 9 of 10 Stage 1 ovarian cancer as well as the later stage disease. This performed with a higher specificity than the cohort tested.
Prostasin	Prostasin is a serine protease found in prostate gland secretions. Identified as a biomarker after discovery from microarray platform. RNA of Prostatsin was found to be overexpressed in ovarian cancer pooled from ovarian cancer and normal human ovarian surface epithelial cell lines. The sensitivity of both CA125 and prostasin is improved when used in conjunction.
Osteopontin	Osteopontin is secreted phosophoprotein. Also, discovered from gene expression profiling. Increased levels of osteopontin were found to be cancers from patients with epithelial ovarian cancer compared with healthy controls, ovarian disease, and other gynecologic cancers
Inhibin	Serum inhibin, a natural ovarian product decreases to levels below detection in post-menopausal women. Some cancers (mucinous, sex cord stromal tumours and granulosa cell) have been shown to secrete Inhibin hence it’s the basis for a diagnostic test for serum.

	Different forms of Inhibin have been found in serum; free, dimer subunit assays that are able to detect both forms have shown promising specificity and sensitivity.
Kallikrein	Kallikreins are serine proteases, there are 15 identified members of the human kallikrein family. One of note is Prostate Specific Antigen (PSA) also known as hK3. Two reports have suggested that hK6 and hK10 have potential a serum biomarkers of ovarian cancer diagnosis.

Several biomarkers for ovarian cancer have been found to be inflammatory markers, such as chemokines and their receptors, though they have been shown to have the sensitivity to detect disease, they lack the specificity to distinguish cancer from benign disease, infection or simple inflammation. For example, overlap has been observed between panels of potential ovarian cancer and other polycystic ovarian syndrome (Galazis *et al.,* 2012). Furthermore, inflammatory markers/inflammation is a characteristic that can exacerbate the cancer environment, tumour cells typically release or stimulate the production of inflammatory cytokines and as a consequence, avoid detection and elimination by the immune system (Vaughan, *et al.,* 2012).

Ultrasonography can be used to visualise the ovaries, malignant growths have distinctive asymmetric, irregular morphology. Doppler imaging can also be used, as reduced blood flow/pressure is commonly observed in malignant growths due to the lack of smooth muscle in the endothelium of blood vessels, formed by cancer-induced angiogenesis. Both have been investigated as a potential screen of the early detection of ovarian cancer. Jacobs *et al.,* (2004), Nossov *et al.,* (2008). However, ultrasonography is subject to inter-observer variability of the radiologists, a study described in (Marcus *et al.,* 2014) found only 25% agreement between radiologists to identify the focal point of the ovarian cancers and 15% variability in measurements. This error may be exacerbated by the diverse morphological and metastatic patterns, tumours surrounded by ascites are viewed more clearly using a CT scan.

Tumour vascularisation has also been used as a prognosticator of survival (Brown *et al.,* 2000). By simply counting the number of blood vessels in tumour sections, it has been shown that an increased number of blood vessels correlates with decreased survival. However, this requires a tumour section and such is thus not suitable for a detection of early disease.

Using combinations of existing screening tools or targeting screening to a high-risk population is believed to reduce the number of false positives and reduce exposure to the associated harms from medical procedures (Buys *et al.,* 2011). A multimodal trial incorporating more than ultrasonography together with CA125 monitored over several time periods termed the Risk of Ovarian Cancer algorithm (ROC) has been shown to increase detection of the diseases (Jacobs *et al.* 2004), but none have as yet proved sensitive or specific enough to change NICE guidelines. For a full historical review of ovarian cancer screening strategies investigated worldwide, including number of cancers per positive screen, the reader is referred to Jacobs *et al.* (2004), the author of which is a key investigator on the United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) trial (discussed below).

There is also a mandate for an effective tool for the monitoring of disease progression or recurrence of treated patients (Marcus *et al.,* 2014), identification of low volume metastasis, and, detection of cancer grade (Vaughan, *et al.,* 2012). However, again, currently CA125 and ultrasonography are the best available tool for this, however, their performance is poor. A small study described in Marcus *et al.,* (2014) of 80 patients undergoing a “second look” surgery, found no correlation between CA125 levels and tumour burden. They also found little or no evidence to show that screening for recurrence using CA125, or physical examination strategies, improves survival over a patient waiting for symptoms (Marcus *et al.,* 2014). Preliminary reports from a United States based screening trial based on CA125 and ultrasonography as a screening strategy, show that screening for detection of early disease does not improve overall survival, due to the high risk involved in following up a positive test resulting from screening tools with such low accuracy, i.e. a positive high CA125 level or irregular ultrasonogram was followed up by a explorative surgery which itself holds significant risk to the said target population (Vaughan, *et al.,* 2012). One such sizeable US study testing screening strategies in the general population in fact found a higher mortally rate in the screening arm (Buys *et al.,* 2011). Buys *et al.,* 2011 attributes an increased mortally to the increased exposure to the associated harms from medical procedures.

It is accepted that ovarian cancer is a relatively rare yet genotypically diverse disease, in fact the term ovarian cancer has been described as “a general term for series of molecularly and etiologically distinct diseases that simply share the same anatomic location” (Vaughan *et al.,* 2012). A tangible risk of screening a general population with a screening tool of dissatisfactory accuracy has been demonstrated at the cost of those screened (Buys *et al.,* 2011). Cooperation and sharing of sample material, data and technology worldwide will speed up the progress of research, Worldwide organisations-supporting such research include: Ovarian Cancer Association Consortium (OCAC), The Cancer Genome Atlas (TCGA), United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS), The Australian Ovarian Cancer

Study (AOCS), OCTIPS (Ovarian Cancer Therapy Innovative Models Prolong Survival) and Ovarian Tumour Tissue Analysis Consortium (OTTA). It is hoped that these efforts will generate novel biomarkers that are prerequisite to an effective screening strategies and have paved the way to access appropriate sample and data cohorts to evaluate emerging biomarkers on a wide scale across the variety of populations.

2. Methodological Overview

2.1. Proteomic Approaches to Biomarker Discovery and Onco-proteomics

2.1.1. Proteomic Techniques for Cancer Biomarker Discovery

Due to the heterogeneity of all types of cancer, and the numerous number of molecular changes that occur in a tumour, it is reasonable to assume that the expression level of one molecule alone would not provide an indication of cancer status with sufficient sensitivity and specificity. It is more logical to assume there will be a change in a combination of protein expressions from, or in response to, a tumour. However, detection and confirmation of multiple smaller changes in protein expression is a far more complex task (Hanash *et al.,* 2008).

Proteomic biomarker discovery workflows can be segregated into two approaches; top-down proteomics or bottom-up proteomics. Bottom-up proteomic biomarker discovery workflows entail recording as much information about a samples proteome as possible (proteome mapping), then comparing and contrasting the recorded proteomes of two sample groups to observe the identifiable differences. A bottom-up approach is most fitting to detect the multiple yet minute changes expected in protein expression, and general biomarker discovery. Top-down approaches are a targeted methodology focusing on changes in the expression of one or more markers of interest, the majority of the information on the samples proteome is disregarded to focus on the changes in presence of this or these key proteins, top-down proteomics is commonly seen in verification and validation stage experiments.

Protein mapping/profiling studies have evolved since the 1930's. The number of proteins identified in serum has exponentially increased as the technology to separate and de-convolute the proteome has become available (Anderson and Anderson 2002). These studies began with separating proteins based on their mass using ultracentrifugation. Electrophoretic separations in liquid, paper, agarose, starch then polyacrylamide followed. The first two dimensional separation of plasma were published in 1977 (Anderson and Anderson 1977); the number of proteins isolated and identified from 2D gel electrophoresis steadily increased since then as it has been coupled with other sample fractionation methods including immune-depletion, size exclusion, lectin binding, ion exchange and hydrophobic interaction (Anderson and Anderson 2002).

Identifying proteins using mass spectrometry to measure and match a proteins fragment masses to those known or calculated in databases of protein amino acid sequences, has become fundamental to proteomics (Nesvizhskii 2007). Mass spectrometers have evolved dramatically since their invention yet since their application to proteomics the reliability and accuracy of mass spectrometry has been unparalleled by any other proteomic identification technique at any given time (Jennings 2012). Firstly, for the quantification and identification of proteins isolated using other techniques such as 2D gels or immuno-precipitation. But also for the discovery/ generation of biomarkers themselves.

Current NICE approved biomarker detection tools used in the NHS are mainly based around the use of labelled antibodies to specifically bind and signal the presence of known biomarkers. A review of the currently used FDA approved diagnostic assays based on protein measurement are predominantly immune-assay (approximately 80%) with the remainder being enzyme assay or in one case a coagulation assay (Anderson 2010). Two key examples include; Enzyme Linked Immunosorbent Assay (ELISA) analysis of blood or Immunohistochemistry (IHC) on sections of biopsy sample. Mass spectrometry is accepted mostly as a research and discovery tool, however, there are a few instances where mass spectrometry and database matching is applied in clinical laboratories, namely the Bruker BioTyper from Bruker Daltonics. This is used to classify strains of bacteria (Buchan *et al.,* 2012). Potential future clinical applications protein biomarkers and mass spectrometry include the iKnife© currently in phase II clinical trials or Rapid Evaporation Ionisation Mass Spectrometry (REIMS) where the vapour from the cuts of an electric surgical knife is used as the ionisation source and directly fed in to a mass spectrometer for near real time detection of cancer biomarkers from tumour reduction surgery's (Balog *et al.,* 2013). Immuno-based techniques such as ELISA and IHC are tried, tested and trusted to measure the presence their known biomarker target protein/s however do not offer the same scope, speed or accuracy or type of measurement of the targeted mass spectrometric techniques mentioned. Incorporation of mass spectrometers to clinical laboratories would require investment in capital equipment and the patient benefit would need to be deemed to offset this cost by appropriate authorities.

2.1.2. Analysis of the Serum Proteome

Blood serum and plasma are a popular source for biomarker discovery investigations as they can be easily sampled non-invasively to a patient. It is logical to expect abnormal or altered expression levels of molecules released from tumours to spread into the circulation, carried around the body and be detectable at lower levels in the blood. Although a higher concentration of a biomarker released from a tumour would be expected nearer the tumour site and diluted levels in the general circulation. Few biomarkers though have currently been first identified in tumours and then shown to be present in serum (Hanash *et al.,* 2008).

The study of the serum proteome is challenged by the huge dynamic range and size of its constituent proteins and the natural inter- and intra- variation in people (Timms *et al.,* 2011).

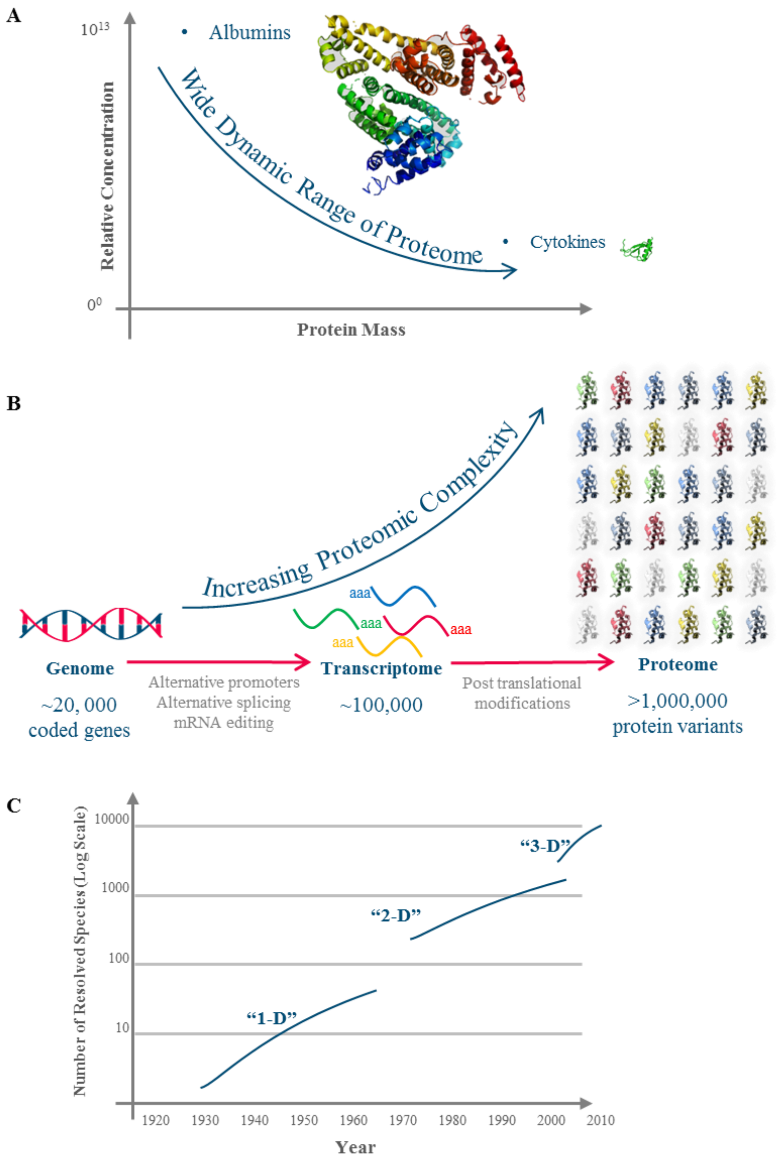


Figure 4 The Complexity and Challenge of Studying the Human Proteome.
A) The Wide Dynamic Range and Mass of Proteins in Serum. Abundance of measured proteins range over 13 orders of magnitude; the graphics of albumin verses cytokines are approximately to scale. **B) Complexity of the Proteome Compared to the Genome.** Splice variants and post translational modifications such as glycosylation or phosphorylation exponentially increase the possible number of protein species to detect. **C) The Increase in Protein Species Resolved and Identified in Plasma over a 70-Year Period.** Adapted from Anderson and Anderson (2002) Illustrates how the number of protein species has increased, as new sample preparation methodologies became available.

In contrast to the calculated number of coding genes in the human genome, which has been steadily decreasing discovery (Harrow *et al.,* 2012) the number of proteins being postulated and discovered in the human proteome is larger and increasing.
As Figure 4a indicates, the abundance of proteins known to be present in the serum, span over thirteen orders of magnitude, further to this, a plethora of potential splice variants and post translational modifications increase the size of the possible protein species in the proteome exponentially. As Figure 4 (inspired from Anderson and Anderson 2002) depicts, the final count/ exact number of proteins in serum is remains unknown and the estimated number is expected to steadily increase as the tools to detect them have become more and more sensitive (Anderson and Anderson 2002).

Any person's serum proteome is a moving target. Measurable clinically relevant chemical analytes in serum, including proteins, have been shown to vary to differing extents both, between different patients within the same demographic group (age and sex) and within an individual when repeat samples are taken at multiple time points (Harris *et al.,* 1970, Williams *et al.,* 1978). Inter and intra individual variability poses a huge challenge to biomarker discovery and validation. This highlights the need for including large numbers of appropriate samples into a cohort to suitably represent the natural variation of any measured component across a population, and, if at all possible repeat measurements could be used to establish intra-variation. However, experimentally the logistics of sampling such appropriate control or comparator groups is sometimes not possible and as close a match as possible is used instead. This can limit the ultimate utility of the biomarkers. A biomarker needs to be robust enough to detect a disease state despite the noise of sample variation to be clinically applicable.

Gil *et al.,* (2015) identified data management to be the current bottle neck in progression of the omic research. It is still not foreseeably possible to fully enumerate or catalogue the human plasma/serum proteome. Each generation of discovery platform offers increased sensitivity and specificity, exponentially advancing computer processing and software provide the capacity to compile and combine measurements with both existing databases and measurements from other platforms. For example, Sciex have created the OneOmics Cloud data processing platform (Sciex 2016, Illumina 2016), where, both gene and protein measurements are compiled from the same samples.
The holistic aim to study 'omics' encompasses the aim to map the full proteome as the ability to do so into a wider body of data emerges (Gil *et al.,* 2015). This approach holds promise to

confront the long-held challenges of proteomics of the wide range in size and abundance of proteins present, and the innumerable permutations of post translational modifications.

Post translation modifications such as ubiquitination or a proteins activation-state i.e. phosphorylation or glycosylation can be investigated by extracting phosphopeptides or glycoproteins using commercially available kits prior to protein identification (Jensen 2004, Thermo Fisher Scientific 2016).

Many initial biomarker studies searching for biomarkers for ovarian cancer, to distinguish cancer from control, applied mass spectrometry to search for differences in the expression of low molecular weight serum proteome by preparing their samples onto specialised surfaces (SELDI) described more fully in Chapter 3 (Petricoin *et al.,* (2002); Kozak *et al.,* (2003); Vlahou *et al.,* (2003); Zhang *et al.,* (2004); Yu *et al.,* (2006); Zhang *et al.,* (2006) Kong *et al.,* (2006)). An *et al.,* 2006 focused on glycans in sera not proteins (An *et al.,* 2006).

2.1.3. Mass Spectrometry and Tandem Mass Spectrometry

A mass spectrometer is an instrument designed to measure the mass of electrically charged molecules; ions, see Figure 5. Mass spectrometry can be used to quantify known molecules, identify unknown molecules and further elucidate their chemical structure and properties.
Molecules need to be charged, either positively or negatively, in order to be measured by a mass spectrometer (Greaves and Roboz 2014).

2.1.3.1. An Historic Summary

It could be said that mass spectrometry was an incidental discovery by Physicist J. J. Thomson whilst researching cathode rays in the late 1800s (1889). The first "mass spectrometers" were invented whilst attempting to prove the existence of electrons, and later to investigate the masses of charged atoms. Since then the technology of the mass spectrometry field has evolved and the breadth of application widened unimaginably (Griffiths 2008).

Fundamentally, all molecules have an electromagnetic characteristic/charge. By applying electromagnetic fields to molecules trapped within a vacuum it is possible to trap or direct /control their trajectory. Hence mass spectrometry is used to separate molecules based on their mass to charge ratio (*m/z*). In lay terms, molecules are weighed using mass spectrometers.

Key developments that advanced the field include:

The development and application to wider fields than physics by Alfred Nier (1911-1994), who, amongst countless accomplishments developed the 60° sector field instrument, the first widely used mass spectrometer, and as a spontaneous collaboration, separated the ^{255}Uranium isotope prompting researchers in the Manhattan Project, which began the nuclear age (Nier (1991), Griffiths 2008).

Increased resolution. Marshall and Camiasow in the 1970's were the first to apply furrier transform (FT) method, a mathematical function, and altered the ion path at detection to enhance deconvolution of wave data to interpret and vastly increase the resolution of recorded MS data (Griffiths 2008).

Identification of unknown compounds. The concept and development of identification of unknown compounds using database matching by McLafferty, Beinman and Djerassi in the 1960's and 70's where systematic experiments data-basing/recording the fragmentation characteristics of each class of organic molecule thus opening the field to the identification of unknown molecules (Griffiths 2008).

Inclusion in proteomics. Most relevantly to this document, the introduction of soft ionisation of analytes via either MALDI or ESI, opening the technique to large organic molecules, proteins, DNA and complex carbohydrates.

It was in the 1980's when the two methods of soft ionisation (section 2.1.3.2) were developed that mass spectrometry was opened to and advanced biology and proteomics (Fenn 2002, Hillenkamp and Karas 2000). Mainly because it enabled the analysis and identification of protein opening the technology to biology - proteomics (Griffiths 2008). Prior to the advent of soft ionisation MS was mostly applied to small organic molecules. Larger organic molecules, proteins, oligonucleotides, lipids and complex carbohydrates were of interest however suffered excessive fragmentation or degradation prior to analysis during the ionisation into a gaseous phase (Griffiths 2008).

ESI came from Fen in the US Fenn (2002). The matrix assisted element of MALDI came from Hillenkamp and Karas of Germany (Hillenkamp and Karas 2000) and the laser desorption part from Tanaka in of Shimadzu Corp (Japan). Although Fenn and Tanaka, not Hillencamp were credited with the 2002 Nobel Prize for soft ionisation techniques (Griffiths 2008).

The fundamental components of MS are as follows: source, analyser detector. Analytes are ionised in the source prior to separation in the mass analyser, the detector counts/senses the frequency of ions as they contact it and converts this information into a digital output.

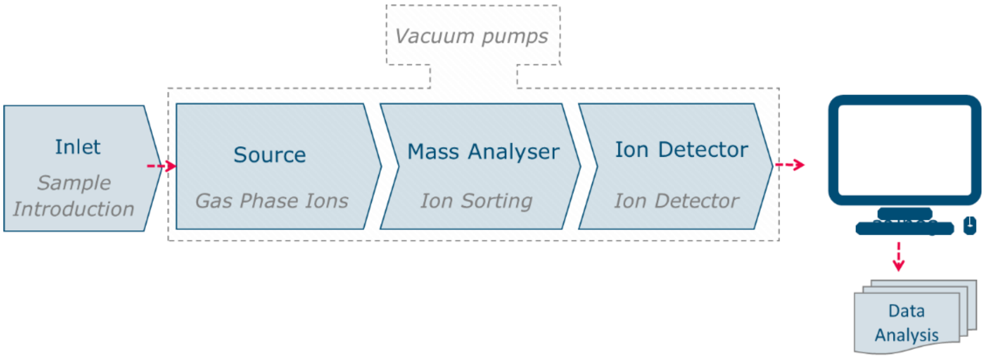


Figure 5. The Basic Components of a Mass Spectrometer.
A schematic diagram of a typical mass spectrometer: samples are introduced in a gaseous phase in at the source, separated based on *m/z* in the mass analyser and counted at the detector the process happens under vacuum and the system is controlled by an online PC.

2.1.3.2. Soft Ionisation: Matrix Assisted Laser Desorption Ionisation (MALDI) and Electrospray Ionisation (ESI)

Matrix Assisted Laser Desorption Ionisation (MALDI):

When using MALDI, the analyte is mixed with a chemical matrix, and dried to a crystal on a metal target plate, a high-power laser is then fired at the crystal, the energy is preferentially absorbed then transferred through chemical matrices causing excitation and ionisation of the matrix-analyte crystal into a gaseous phase plume ready for separation and detection. A progression from the original application of laser desorption method for biological molecules developed by Tanaka *et al.,* 1988, analyte was dissolved in a mixture of glycerol and metal nano-particles to convey the laser energy (Griffiths 2008). The vast majority of MALDI uses matrix the most common of which are 3,5-dimethoxy-4-hydroxycinnamic acid (sinapinic acid), α-cyano-4-hydroxycinnamic acid (CHCA, alpha-cyano or alpha-matrix) and 2,5-dihydroxybenzoic acid (DHB) respectively for small medium and large organic molecules.

In the example of serum proteome analysis using MALDI-TOF-MS, small amounts of sera are combined with the matrix and spotted on to a specialised steel target plate. SA can be used for analysis of intact proteins whereas CHCA can be used for digested peptides.

The matrix-sample amalgam is then pipetted into a small spot and left to dry. Upon close inspection, the spot will have a crystalline formation. The crystal formation is closely related to the amount, concentration, distribution, dissolution and nature of the proteins in the sample. Depending on the make and model of the MS, the target plate can contain any number of spots (up to hundreds of spots), each spot potentially containing a separate sample. The target plate is then inserted into the mass analyser through an air lock and the source area brought to vacuum.

An ultraviolet laser is fired in pulses at each sample location in turn. Each firing of the laser separately desorbs the sample from its crystallised matric into a tiny gaseous plume. It is assumed that for each desorbed ion will acquire one proton from the matrix mixture. Thus, all ions are singly charged in a positive state. As the samples are separated on their mass to charge ratio (*m/z*) if all the molecules of the analytes carry a single positive charge the only property that will cause them to have different flight paths down the flight tube will be their mass.

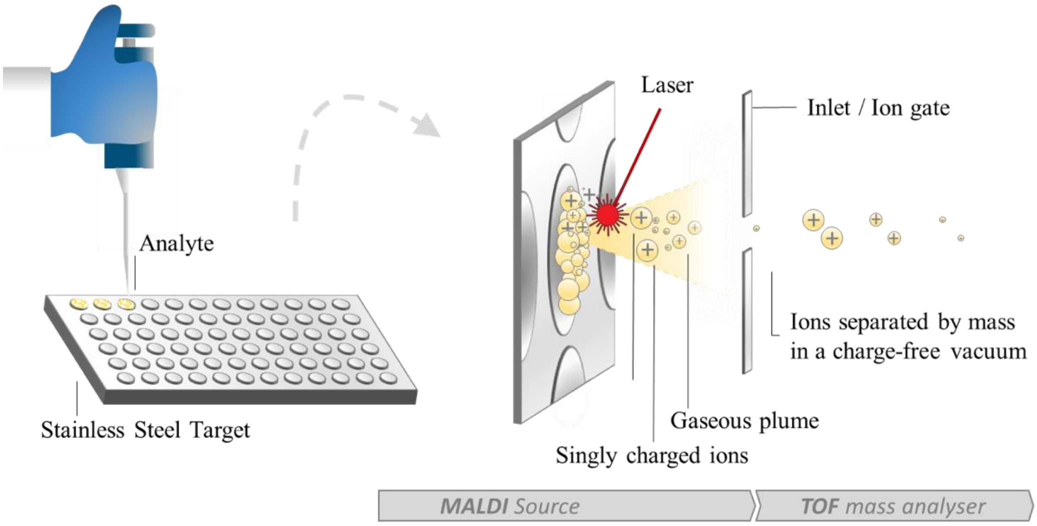


Figure 6. Matrix Assisted Laser Desorption Ionisation (MALDI).
A schematic diagram depicting sample-matrix crystal of one of multiple spots on a steel target. Sample is ionised via a UV laser exciting transferring energy through the chemical matrix. One positive charge is transferred to each ion.

Ions generated from MALDI are all predominantly singly charged (=+1), however occasionally a second or third proton is imparted. This simplifies and almost negates the need for calculation mass using the *m/z* ratio (Perkel 2012).

Surface Enhanced Laser Desorption Ionisation (SELDI), such as Cyphergens's ProteinChip © arrays are a version of MALDI, in which, a sample/analyte/protein is prepared directly onto the specialised surface on a chip platform (Tang *et al.,* 2004). The chip coating is comprised of a factor such as a receptor ligand, antibody, hydrophobic, hydrophilic, cationic or anionic substance to immobilise specific components for analysis. Only components of the sample/analyte/protein complementary to the chip surface will be retained during a wash phase. Examples include Hydrophobic ProteinChips, Weak Cation exchange ProteinChips (WCX2), Strong anion exchange ProteinChips (SAX2), Immobilised metal affinity ProteinChip surfaces

(IMAC) or Immobilised copper ProteinChip surfaces (IMAC3) (Roboz 2005) which will bind different analytes/proteins with greater or less affinity depending on the properties of their constituent amino acids.

Hillencamp, a pioneer of MALDI (Hillenkamp and Karas 2000), views MALDI to be "competing and complementary" to ESI (Griffiths 2008), some experiments could be conducted using either a MALDI or ESI source, however some would be better suited to one. Some molecules would not be identified using one or the other. As analytes are dried using MALDI, it is more tolerant of contaminants such as salts. ESI is better suited to coupling to liquid chromatography increasing the directionality, and data acquired from an experiment (Griffiths 2008). Multiply charged ions are thought to fragment better if singly charged (Perkel 2012), making MALDI a more challenging platform for fragmentation and consequently protein identification. Further to this, the possibility of multiply charging ions provided by ESI, increases the likelihood of detection of those on the cusp of the detection range in their singly charged state. Thus, ESI enhances the analysis of larger molecules (Perkel 2012). Researchers are challenged to choose an instrument best suited to their individual work demands, ESI and MALDI have been the two main soft ionisation options for many years. However, an increasing number of ambient sources are being development for bespoke purposes.

Electrospray Ionisation (ESI):

ESI, was a modification of a current high voltage ionisation where non-volatile solutes analytes such as proteins were first dissolved in solvents (Griffiths 2008). A high voltage is applied to a spray of the solution creating charged micro droplets. The solvent evaporates leaving the gaseous analyte with one or more charges. As depicted below, during the evaporation of the solvent, analyte molecules can be left with varying number of positive charges, multiply charged species of the same molecule may exist +1, +2, +3. The speed an ion will travel through the mass analyser is reduced as the charge it holds increases.

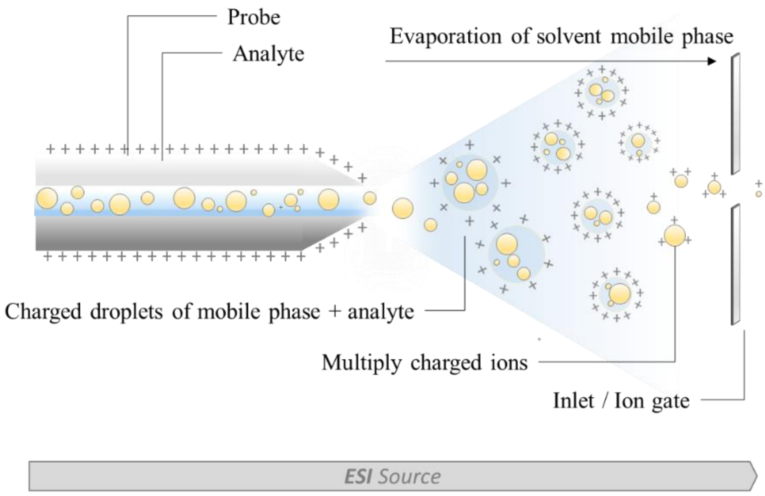

Probe
Analyte
Evaporation of solvent mobile phase
Charged droplets of mobile phase + analyte
Multiply charged ions
Inlet / Ion gate
ESI Source

Figure 7. Electrospray Ionisation (ESI).
A schematic diagram adapted vastly from (Gates 2009) depicting ESI, high voltage is applied to the exterior of a probe through which an analyte dissolved in a solvent is being sprayed. A gaseous plume of droplets is created in front of the inlet to the MS. Solvent evaporates leaving a gas phase analyte which has retained one or more charges.

Ambient Sources

During the popularity and common usage of ESI or MALDI in a biological/proteomic research lab setting, alternative ionisation techniques have continued into emerge, especially in the context of clinical setting and precision medicine. A key goal in the applicability of these technologies is the ideal that a clinical sample could be analysed with minimal or no sample preparation and fed directly into the mass analyser. For this reason, they can be grouped as "ambient" sources, these include:

- Desorption Electrospray Ionisation (DESI), where an electro-statically charged mist is passed over a liquid sample and the pneumatic perturbation of the surface, fires sample particulates into the analyser inlet (Takáts *et al.,* 2014).
- Acoustic mass spectrometry, where sound waves are used to eject a droplet from a wave and disrupt it into a spray with partials fine enough to effectively release analyte molecules into the mass analyser (Ho *et al,* 2011).
- Rapid Evaporative Ionisation Mass Spectrometry (REIMS) where a sample heated and the vapour from a sample is directed into the mass analyser. One such example is the iKnife© described above (Balog *et al.,* 2013).

2.1.3.3. Mass Analysers

Separation of ions occurs in the mass analyser, after ionisation and before detection. Separation of ions is done using at least one of the following; Time of Flight (TOF), Quadrupole, Quadrupole ion Trap (QTrap), Ion Cyclotron Resonance (ICR), Electrostatic Sector Mass Analyser or a Magnetic Sector Mass Analyser (Greaves and Roboz 2014).

TOF, Quadrupole and ion traps are most commonly seen in proteomics/clinical research laboratories and will be summarised below.

Time of flight (TOF):
Using TOF, gas phase ions are accelerated down a flight tube. If all the ions carry the same charge their kinetic energy will be equal and the only thing effecting their velocity is mass. Ions with a lower mass will reach the end of a flight tube faster than ions with a larger mass and equal charge. Thus, they have a lower *m/z* ratio (Greaves and Roboz 2014).

Most TOF analysers can be run in linear or reflectron mode. In linear mode, desorbed ions travel down a straight flight path to a detector located at the opposite end to the source. In reflectron mode, an ion mirror at the end of a linear flight tube deflects their trajectory toward a detector in a different location, this may be at an <90° angle flight tube or back at the base of the flight tube near the source. Reflection is applied to extend the flight path enabling better resolving power. The separation distance between ions of a similar mass will increase the further their flight path, the longer the flight tube the higher the resolution and sensitivity to smaller masses.

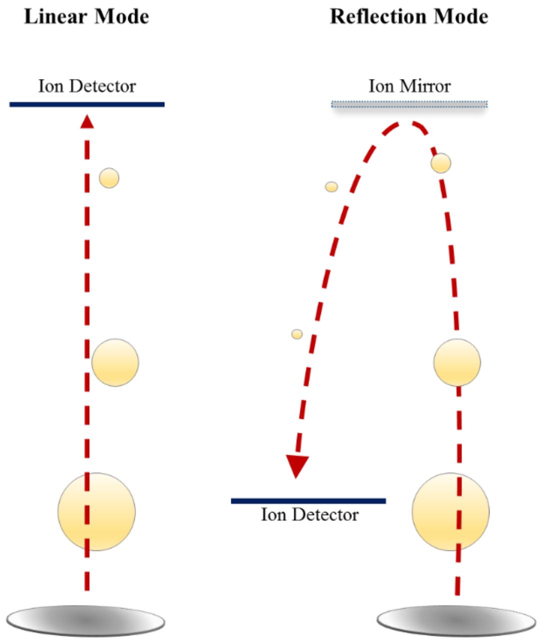


Figure 8. Linear and Reflectron Mode of the MALDI-TOF-MS.
A schematic diagram of the ion mirror increasing the flight path in reflectron mode allowing amplified separation then analyses of the smaller molecules.

The laser of the MALDI-TOF-MS is often fired in a raster (at regular points across the dried spot of sample) to ensure even sampling from all areas of the spot. Alternatively, some models, including the Bruker UltrafleXtreme offer an auto-quality function, in which, the areas of the crystallised sample that yield the highest signal ‘hot spots’ are automatically ascertained and the laser is focused on these locations for data acquisition. Spectra acquired in auto-quality mode will have lower background noise and higher intensity signal in comparison to that acquired in Rasta mode, where data is acquired from evenly spaced locations across the sample spot. A good signal to noise ratio is necessary to distinguish low intensity peptide species.

Magnetic and electrostatic sector field mass analysers are similar to TOF in that ions are directed down a flight tube with a detector at the far end. However, in this case the flight tube is not straight, a magnetic/electric force is applied at the bend of the flight tube. The force will affect each ion differently depending on its *m/z*. Only ions of a specific *m/z* flight path will be altered at the correct trajectory to reach the far end and hit the detector (Greaves and Roboz 2014).

The principal of using specific electric forces to direct and separate ions by influencing their flight path is key to most mass analysers.

Quadrupoles: As its name suggests a quadrupole is structure of four steel poles positioned longitudinally and equidistant along an ion flight path. Electrical currents are applied to the poles and surrounding environment to separate ions based on their mass and charge properties. A DC is applied at the base of the poles to give the ions a forward trajectory down the centre of the four poles. The quad of poles act as two pairs, each pole is twinned with the one on the opposite side of the ion flight path. Specific electrical currents are applied to each pair and alternately switched from positive to negative creating two sinusoidal electrical fields at a 180-degree orientation as they oscillate create a circle of electrical force. Ions will have a spiral motion path as they are attracted to and repelled from each pole as they pass. The amount the poles voltage influences the flight path of the ion is dependent on its *m/z* charge. At any given set of voltages, only ions of a specific *m/z* will be directed to the end of the flight path without hitting the edge or being thrown out. Thus, the quadrupole selectively filters out ions that are not the intended *m/z* (Greaves and Roboz 2014).

Ion traps work using the same principal where instead of directing in ions along forward path, the electrical fields trap them in repeating shape or orbit (Greaves and Roboz 2014).

2.1.3.4. Tandem Mass Spectrometry

Within a mass spectrometer, ions can be deliberately broken down/fragmented to measure the masses of their resulting fragments (McLafferty 1981, Hoffman 1996, Greaves and Roboz 2014). The masses of these fragments can then be used to ascertain the chemical makeup and infer the identity of the parent ion (Hoffman 1996, Nesvizhskii 2007). For small molecule ions or short peptide ions the fragment masses can only be matched (within an appropriate tolerance) to the known mass of one or more element/compound. Thus, making its identity unequivocal. For larger molecules and the majority of protein/peptide experiments, this needs to be done by best matching the mass difference between fragment peaks to known masses of elements and compounds and incrementally compiling these in the order the parent ion fragmented them. This method is based on assumptions such as; the ion has been successfully and thoroughly fragmented to create a fragment at each stage of destruction/deterioration of the parent ion, or, in a peptide experiment that the beginning and end fragment is signified by a peak difference the same mass as the amino acid cleavage site target of the proteolytic enzyme used (Nesvizhskii 2007).

Ions are fragmented within a mass spectrometer by different means depending on the model and type of the mass spectrometer. These include, an increased laser power when using laser desorption ionisation, within the vacuum of the mass spectrometer a high voltage can induce fragmentation to ions isolated along the flight path, or ions can be directed through a "collision cell” where ions traveling along the flight path current collide with molecules of a known gas which inhabits the collision cell, the force physical force of the collision fragments the ions (McLafferty 1981, Hoffman 1996, Greaves and Roboz 2014). Additionally, ions may be inadvertently fragment at the source, ion source decay.

For MALDI-TOF-MS/MS ions are desorbed from the target using a significantly higher laser power than that used for the spectra of parent ions. Once desorbed the fragmented MS/MS ions are detected the same was as they are in MS some designs include devices to boost detection of fragment ions. The Bruker Ultraflex III utilises such technology, where a "LIFT" cell is inserted into the ion flight path in MS/MS mode to add velocity to fragment ions, their improved acceleration increased the number that reach the detector for measurement (Suckau *et al.,* 2003). In tandem mass spectrometry, each peak detected has two identifying features; its mass, and the parent mass from which is was fragmented.

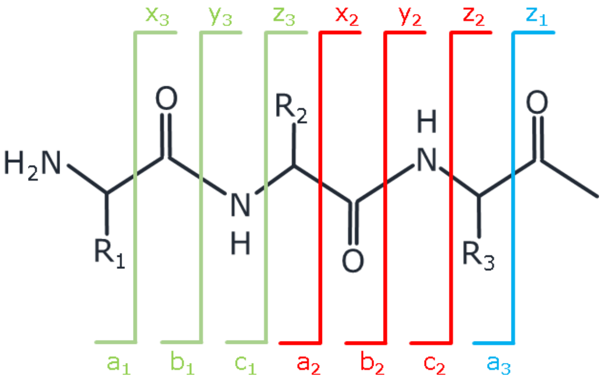


Figure 9. Peptide Backbone Fragmentation.
A schematic diagram depicting ion fragmentation. Coloured lines depict where the molecule would fragment and the common nomenclature for the fragment ions until they are matched to a known mass and assigned an identity.

Figure 9 (above) depicts a peptide ion, the coloured lines depict where the peptide would fragment in the mass spectrometer. Until they are matched either manually or automatically to a known entity the common terminology for the fragmented ions is from the carboxyl terminus from “z, y, x” and from the N terminus “a, b, c” A consensus between the distance of each of the “b” ions and each of the “y” ion is used to measure assign the unique amino acid identity.

2.1.4. Mass Spectrometry Approaches/Techniques used for Biomarker Discovery

Since its development in the 1980's MALDI mass spectrometry has been applied to protein samples to separate them based on their mass to charge ratio. Identifying proteins using mass spectrometry relies on databases of hypothetical protein masses derived from genomic sequence information.

Protein Identification Using Mass Spectrometry. Inside a mass spectrometer charged ions, for example enzymatically digested peptides, can be isolated and fragmented further i.e. to the amino acid level. The identity of the amino acids sequence can be derived from the fragment masses and possible protein identity calculated using online databases such as Mascot (Perkins *et al.,* 1999).

ESI and MALDI, are the two MS techniques applicable to biological molecules, they work on different concepts, variation in the properties of analyte fragment ions mean some are better detected using one platform than the other, thus their results are complimentary.

MALDI: Popularity of biomarker discovery via MALDI or SELDI-MS peaked in early to mid-2000s; numerous groups published mass values of peptides identified from mass spectra that were significantly differentially expressed in ovarian cancer, control or benign (see Table 3 section 3). Failure to reproduce findings or give meaning to the mass values of ions that discriminated the cancer cohorts damaged the image and trust/confidence in its use. This is reflected by a drop in the number of publications from MALDI-TOF-MS data (Albrethsen 2011). To confirm any potential novel biomarker findings results must be reproduced on either or both of; a second technological platform and a separate sample cohort.

MALDI mass spectrometers do not produce numerically quantitative data. Amino acids behave differently in the ionisation and desorption phase depending on their efficacy to transfer energy when excited. Consequently, some proteins will ionise and be detected more easily than others depending on their amino acid composition (Benk and Roesli 2012). However, the data can be treated as relatively quantitative.

2.1.5. Sample Fractionation and Liquid Chromatography

2.1.5.1. Reducing Sample Complexity

Fractionating a serum sample, prior to a bottom-up analysis, allows access to identify lower abundant proteins (Faca *et al.,* 2007), the total protein content of each fraction will be lower than the whole sample thus decreasing the overlap of proteins in any one dimension, which potentially masks lower abundant proteins which share the same properties. In mass spectrometry, this is termed “ion suppression" (Mallick *et al.,* 2010). The strength of the signal from detection of highly abundant proteins, expressed at several orders of magnitude more than lower abundant proteins, masks the detection of the latter. Fractionating samples reduces the complexity of each fraction of a sample to be analysed, thus increasing the potential to see the proteins expressed in smaller amounts. However, fractionation reduces the sample throughput. Firstly, due to the time taken to perform the fractionation steps to the sample and, secondly each fraction then needs to be analysed separately.

Albumin, the most abundant protein in human serum, can be extracted using commercially available kits (Margulies and Shevack 1996). This vastly reduces signal suppression of lower abundant proteins allowing their investigation, evaluation and analyses. However, anything bound to the albumin fraction is lost, some researchers have chosen to focus on the albumin and its bound proteins; the albuminome (Gundry and Cotter 2007, Gundry *et al.,* 2007)

Other fractionation techniques used to deconvolute a sample include; immunodepletion of the high abundance proteins (Tang *et al.,* 2013), separation by pI isoelectric focusing (Stein *et al.,* 2013), this can be done to whole proteins after proteolytic digestion and most commonly reverse phase liquid chromatography using C_4, C_{18} or strong or weak cation exchange (SCX WCX) media packed into a column or pipette tip (Chen *et al.,* 2009).

2.1.5.2. The Mechanics of and Sources of Variability in Liquid Chromatography

Liquid chromatography (LC) or liquid chromatographic gradients are often employed upstream of mass spectrometric analysis as a means to fractionate or deconvolute samples. As defined by the International Union of Pure and Applied Chemistry (IUPAC) Chromatography is “a physical method of separation in which the components to be separated are distributed between two phases, one of which is stationary and the other moves in direction” (Ettre 1993). High

Performance Liquid Chromatography (HPLC), developed in the 1970s (Hager 2008), is a process where analytes are passed through a column packed with microscopic particles with a deliberately engineered surface chemistry; the stationary phase. The column is washed through with an analyte followed by solvents of differing strengths/properties; the mobile phase. The components of the analyte interact with the particle surfaces and retains components with affinity, changes in the composition of the mobile phase as it is washed through alters the chemistry of the environment surrounding the interactions. Components within the analyte are separated based on their affinity to bind in the altering environment. Polar compounds prefer the mobile phase environment whereas non-polar compounds favour the stationary phase, the mobile phase flow may be isocratic or a gradient of continuously increasing solvent. As the solvent concentration of a mobile phase increases it becomes a favourable environment for more of the molecules bound to the surface, thus components of the analyte are separated based on their differential interactions between the mobile and stationary phase. The chromatographic separation does not need to be coupled to a mass spectrometer, detection of the eluted sample may also be via UV, fluorescence, Refractive Index, Conductivity, Evaporative Light Screening Detection (ELSD) and Charged Aerosol Detection (CAD). The retention time (t_R) of a compound is the measured time from injection of the sample into the system to the elution of the peak at its highest point. Retention factor (*k*), is the amount of a molecule that remains bound in the solid phase in proportion to that eluted into the mobile phase. A goal of liquid chromatography is for all identical molecules within an analyte to elute from the column with the same t_R both within one experiment and when measured on multiple occasions; a reproducible t_R. Factors effecting t_R variation include: the differing length of possible route identical molecules take around and through the packed particles in the column known as "the multipath" effect or Eddy Diffusion, the dispersion via random molecular diffusion of identical molecules within the mobile phase after they have eluted, the strength of the interactions from the particles surface chemistry and temperature (Snyder, Kirkland and Dolan., 2010).

Since the 1970's technological developments in column production, namely the engineering of silica particle attributes, have made HPLC a robust and reproducible platform with wide reaching applications. Purity of the silica from metal ions, associated with the production process, has steadily improved since the 1980s. Regulation of particle size and consistency minimise the multipath effect by standardising the path lengths taken by compounds within an analyte as they travel through the column thus increasing the reproducibility of their t_R. The move to solid core/ superficially porous silica particles has the same result, analyte will only travel the prescribed distance throughout the outer layer of the particle. Designing the terrain of the particle surface to make be a smooth sphere with pores of regulated shape can firstly, further standardise the path lengths and number of interactions between the analyte and stationary phase and secondly, as a means to filter analyte based on size exclusion. Compounds too large to enter the pores are unable to bind with the surface within them. Smaller particles increase the overall column surface, as does pore size. The pores throughout silica particles provide over 99% of the surface area. The longer the column length the larger the chromatographic effect and distance between different molecules within an analyte. Column ovens regulate column temperature ensuring the consistency of the environment sample injections. Increasing flow rates minimises the opportunity for identically eluted molecules to spread by random molecular diffusion. Ultra-HPLC (UHPLC) describes the use of HPLC at a lower volumes and higher pressure to increase sensitivity and resolution for working with smaller volumes. The van Deemter Equation (van Deemter *et al.,* 1956) is used to compute all the potential factors effecting retention time spread, and column efficiency. Efficiency is measured in Theoretical plates (Snyder, Kirkland and Dolan, 2010)

The van Deemter Equation:
Theoretical Plate Height = Eddy Diffusion (multi-path effect) + (random molecular diffusion / velocity) + (mass transfer between phases x velocity)

Theoretical plates are a model to measure column efficiency and performance. The various factors effecting a columns efficacy and the result can be calculated and conceptualised as layers within the column, theoretical plates. The more plates the column has the greater its efficacy at separating compounds within a sample.

Different modes of HPLC include: reversed-phase where the sample is bound to and then eluted from a column, ion exchange where the bonds between the analyte require and anion or cation exchange, size exclusion, Hydrophilic Interaction Chromatography (HILIC) where surface chemistry is added to another to encourage water molecules to the surface for a hydrophilic environment, Chiral, Affinity, Hydrophobic Interaction. Ion pairing is when an intermediate chemical with a higher affinity to an analyte of interest is used to bind to the column and act as an intermediary (Majors 2012). However, reverse-phase is the most commonly used by far 93% (Majors 2012). Hydrocarbon chains of 18, 8 or 4 are a commonly used surface chemistry for the solid phase. Majors 2012 states C_{18} to be the most commonly used (38%), followed by C_8 (22%), Phenyl (16%), CN (10%), Fluorinated (6%), C_4 (4%), Graphitised carbon (2%) and longer than C_{18} (2%).

2.1.5.3. Deconvolution of Complex Biological Samples Using Liquid Chromatography

In a typical proteomic reverse-phase C_{18} LC-MS experiment aiming to fractionate a complex biological protein extract, a complex sample is passed through the analytical column containing C_{18} media, molecules with a corresponding surface chemistry i.e. hydrophobic peptides bind with varying degrees to the resin within the column. The proteins bind to, and are eluted from, the column at increasing concentrations of an organic solvent within the mobile phase. More relevantly, using LC, proteins within a serum sample can be bound to and then separated from a C_{18} column based on their affinity to bind to the chains of 18 carbons attached to the surface of the silica in the presence of a solvent such as acetonitrile or methanol.

Depending on the downstream analysis the column eluate can either be monitored in real time or in periodical fractions collected and analysed, now at a greater detail than the original sample.

LC systems are often coupled to an electrospray mass spectrometer for a continuous mass spectral analysis of the sample as it is eluted at an increasing concentration from the column.

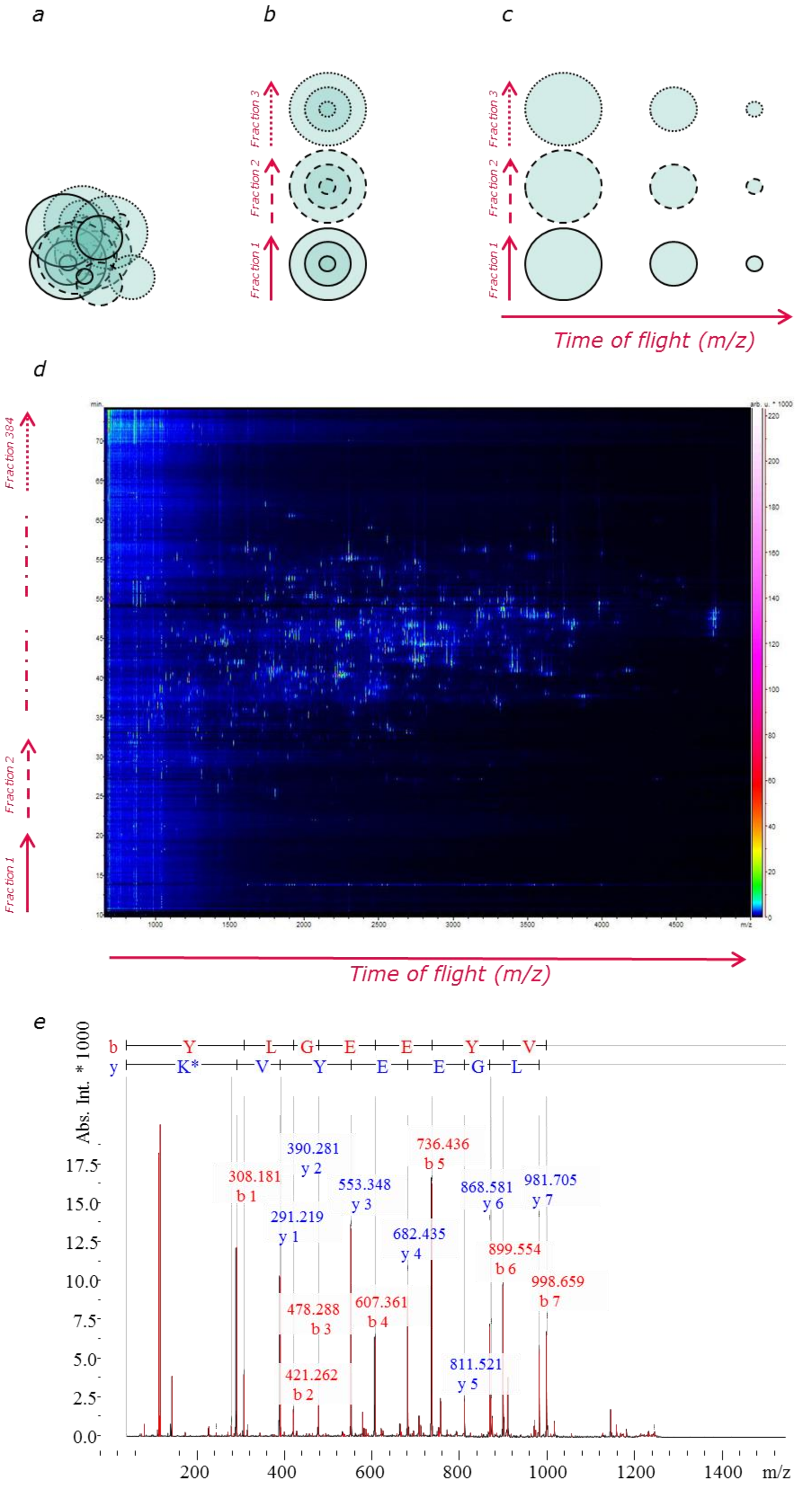


Figure 10. Liquid Chromatographic Separation Coupled to Mass Spectrometric Detection.
a) a representation of a complex protein sample containing proteins of varying sizes and biochemical properties represented by their size and dotted/dashed/solid border; **b)** depicts how the peptides are first separated to fractions based on their biochemical properties by their affinity to a solid phase inside an LC column **c)** in the mass spectrometer fractionated proteins within the sample are separated by mass **d)** Heat map generated by Bruker software of a serum sample analysed by LC-MALDI-TOF. The 384 fractions spotted over an 80-minute gradient of increasing solvent eluting proteins from a C_{18} solid phase is plotted on the Y axis, mass to charge ratio is plotted on the X axis, brighter colours represent a higher intensity. **e)** is a MS/MS spectrum of a peptide fragmentation with annotations assigning amino acid identities derived from the fragment masses, these could be imagined as the Z dimension of the heat map in **d**. See also Figure 9.

When using LC-ESI a sample analyte is in liquid form and is directly fed into the mass spectrometer as it ionises at the source inline/online as it is eluted from the analytic column. This is different to LC-MALDI as the sample needs to be mixed with a chemical matrix and dried to a crystal as a spot on a MALDI target plate first, this is prepared separately to entering the mass spectrometer so can be referred to as an offline separation.

2.1.6. Advantages or Disadvantages of Tagging in MALDI-TOF-MS.

Proteins can be labelled before MALDI analysis to enable a quantitation of results. These methods include: Isotope tagging, isotope coded affinity tagging (ICAT) late 1990s (Benk and Roesli 2012), Stable isotope labelling with amino acids in cell culture (SILAC), Isobaric tagging for relative and absolute quantitation (iTRAQ), Isotope-coded protein labelling (ICPL), Tandem mass tags (TMT).

Other than the obvious advantage that tagging produces exact quantitation of peaks proteins, tagging protocols also have high reproducibility as samples are normally pooled to be analysed Christoforou and Lilley 2011). On the other hand, not only are the tagging systems listed above costly, there e is a limit of how many comparisons can be made in one experiment. Pooling samples is disadvantageous as it dilutes signal from each sample. Labelling protocols increase sample handling which gives opportunity for variation, error and loss of sample material. Isobaric tagging has been shown to decrease accuracy if not incorporated into methods appropriately (Christoforou and Lilley 2011).

2.1.7. Label Free Quantitation Techniques

Quantitative mass spectrometry data from methods not using labelling has been gathered using one of two approaches (Benk and Roesli 2012):

- Area under the curve (AUC) of precursor ion peaks. This approach assumes that the amount of protein present for detection from a sample is directly correlated with the amount detected.

- Methods counting MS^2 spectra of parent ions calculated to belong to one protein. Again, this makes the assumption the amount of protein detected is correlated with the amount of sample loaded, and, the time the spectra is collected for.

A trade-off of a more accurate quantitation in exchange for less sample processing, lower costs and no limit on the amount of possible protein identities and comparisons is made.

2.2. Transcriptomic Approaches to Biomarker Discovery and Onco-genomics

Above, methods of recording protein expression, a gene's translated information is described. Transcriptional information from mRNA is also mined from biomarker discovery.

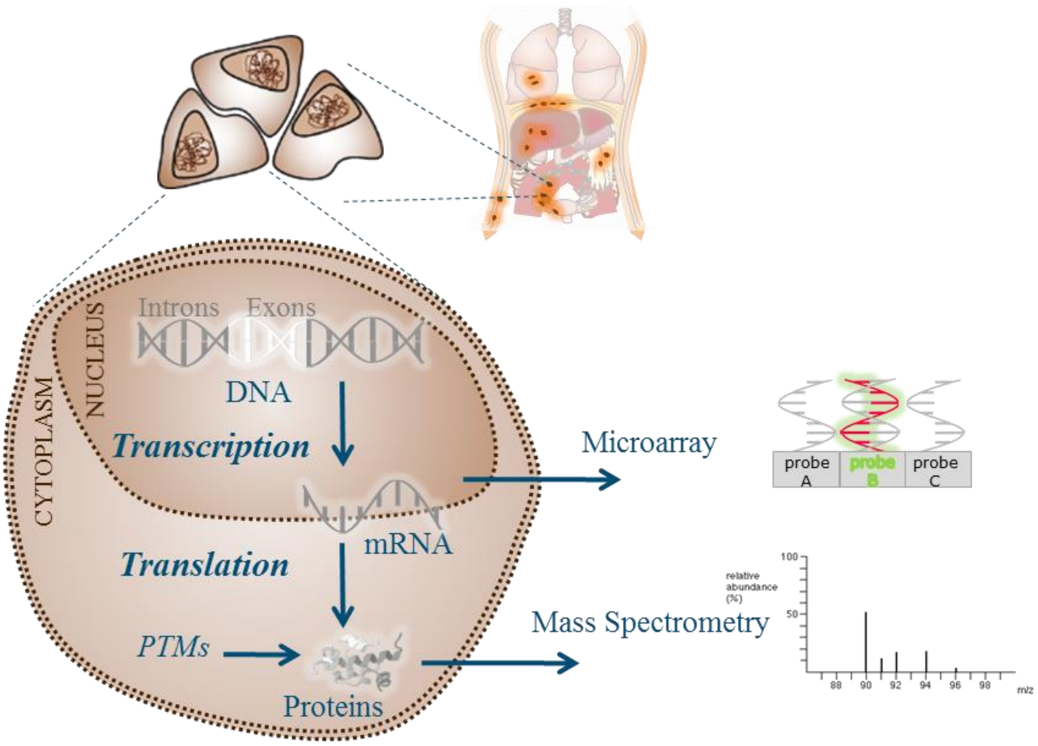


Figure 11. A Schematic Depicting How Genomic and Proteomic Data Represent a System at Different Perspectives.
The difference between transcriptional and translational data is depicted. A transcriptomic measurement is of a biologically upstream factor and is therefore less complicated by the possibility of posttranslational modifications. The current common method to measure each proteins is by database matching of measured masses whereas genes are measured via specific complementary binding.

Figure 11 depicts the difference between transcriptional and translational data. Firstly, a transcriptomic measurement is of a biologically upstream factor and therefore is less complicated by the possibility of downstream modifications. Secondly, the current common methods to measure each; proteins are currently measured by database matching of their measured masses whereas genes are measured via specific complementary binding. Proteomic data measures only what is active in a system at a given time and it is complicated by a plethora of potential posttranslational modifications (PTMs) such as glycosylation or phosphorylation. Genomic data (i.e. of mRNA fragments) can be less ambiguous yet still not conclusive due to splice variation or other factors controlling transcription.

2.2.1. Gene Expression Profiling

2.2.1.1. DNA Microarray Experiments

DNA Microarray experiments, also referred to as RNA transcriptomic, mRNA chip, GeneChip® DNA array microarray and gene expression array, allow determination of the expression of large numbers of genes in nucleic acid extracted from biological samples. Expression patterns of entire genomes, known splice variants, Single Nucleotide polymorphisms (SNPs) or collections of genes are generated and compared (Heller 2002). Early versions of the technique were first reported in the 1980's (Taub *et al.,* 1983)

Fundamentally, microarrays consist of short sections of single stranded nucleic acid (referred to as oligonucleotides or probes) immobilised to a glass or Nylon surface which are used as bait to bind and measure genetic material with a complementary sequence extracted from a biological sample. The oligonucleotide probes are designed and manufactured to be complementary in sequence to the coding regions of genes, and in some cases exons, the expression of which are to me measured (Koboldt *et al*., 2013, Life Technologies 2013).

In a typical microarray experiment, genetic material is extracted from a biological sample, fragmented using restriction enzymes, labelled and hybridised to a gene chip to assign an expression value to each probe representing a transcribed gene. The immobilised oligonucleotide probes are complimentary in sequence to their target coding gene and will therefore bind to a corresponding sequence of mRNA if present in the analysed genetic material. A well-designed gene chip has multiple probes corresponding to each gene (Life Technologies 2013, spread at random locations across the surface to minimise experimental bias. Microarrays can contain probes for a small number of genes or a represented portion of an entire genome. As long as the sequence is known (Koboldt *et al*., 2013, Life Technologies 2013).

Figure 12 below summarises the fundamental steps in the generation of microarray data: RNA is extracted from a biological sample such as a tumour tissue lysate, it is then transcribed into cDNA and labelled with a fluorophore; more than one fluorophore is used for comparative studies. The labelled cDNA is hybridised with the microarray, cDNA binds with the corresponding single stranded DNA probes. The remaining unbound tissue labelled material is removed and a reading is taken. Each probe is assigned a value depending on the fluorescence emitted which will be higher the more corresponding cDNA has bound to the complimentary probes (Life Technologies 2013).

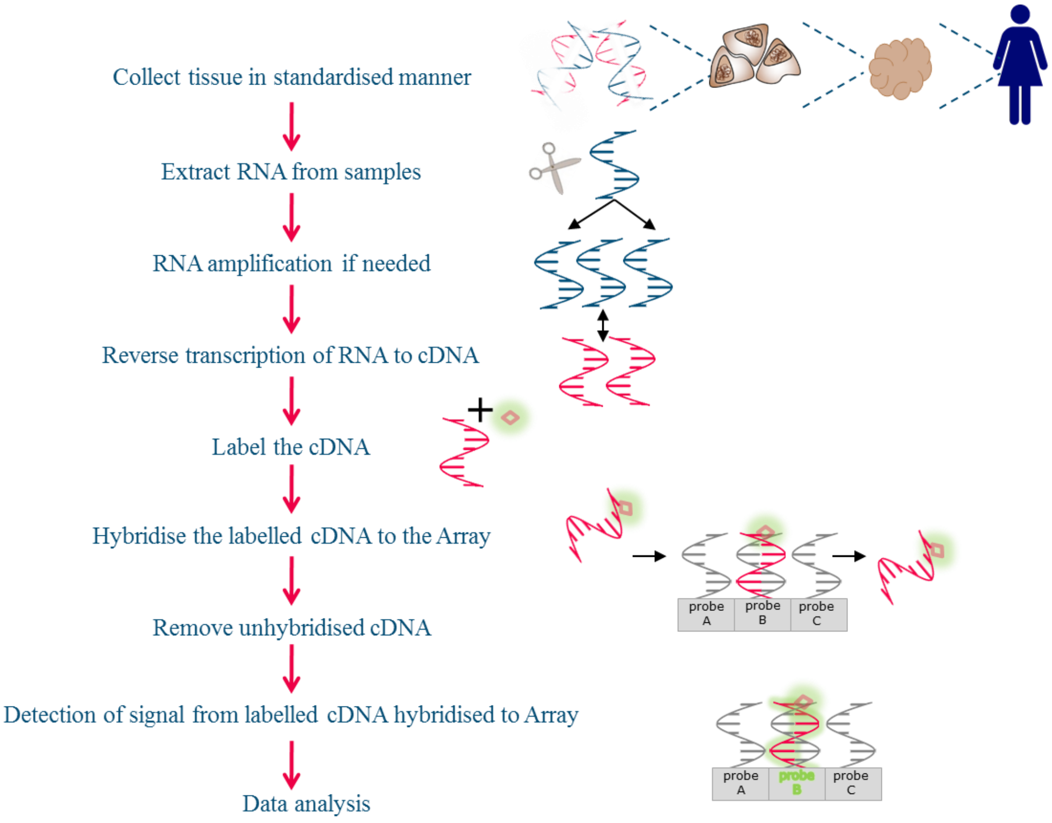

Collect tissue in standardised manner
Extract RNA from samples
RNA amplification if needed
Reverse transcription of RNA to cDNA
Label the cDNA
+
Hybridise the labelled cDNA to the Array
probe A
probe B
probe C
Remove unhybridised cDNA
Detection of signal from labelled cDNA hybridised to Array
probe A
probe B
probe C
Data analysis

Figure 12. A Typical Gene Array Experiment Workflow.
Adapted from Life technologies 2013.Tissue containing genetic material is collected *in vivo* (as above) or grown *in vitro*, from this RNA is extracted cut in specific locations, restriction sites, using restriction enzymes (depicted above with scissors). The resulting short strands of RNA are amplified using polymerases strands with a complementary nucleotide sequence. The cDNA strands are fluorescently labelled at the signature beginning/end sequence used to cut them. The cDNA is then hybridised against a gene chip array consisting of specifically designed short strands of nucleotides called probes or oligo’s which are positioned in specific locations across a surface. After a wash step a fluorescent reading is taken the location of any fluorescent emission is inferred as the presence of a cDNA with a specific sequence match to the probe. Multiple probes are used to constitute a representative proportion sequence of a gene.

2.2.1.2. Next Generation DNA Sequencing

Aptly termed, Next-generation DNA sequencing (NGS), is a more recently developed methodology for sequencing DNA or RNA (EMBL-EBI 2015). Instead of binding the cDNA sample to probes of denoting predetermined known genes sequences, the DNA is cleaved into shorter strands approximately between 100 bp and 1 kb, the strands are then annealed to a position on a platform, the probes are then exposed to a mix of nucleotides, a DNA polymerase and a terminator which ensures only one nucleotide is bound at a time. The first nucleotide is bound and its character recorded, the terminator is unbound, and the process repeated. The exact method of recording what nucleotide varies depends on the platform used; Illumina anneal the probes to a slide, then fluorescently label each nucleotide with a different colour, the colour emitted from each location is recorded each cycle. In 454 Sequencing the DNA fragments are bound to beads and the beads are immobilised into wells on a platform, the slides are flooded with one nucleotide species at a time (A or G or T or A) which emit a light for detection, the strength of the light signal emitted from each location denotes how many of that nucleotides base pair is next in the sequence. For example, if the slide was washed with a Guanine species bead location X a recodes a signal of 1 out of 4 and bead location Y records a signal of 4 out of 4, the next part of the sequence for bead location X is T (thymine) and bead location Y is TTTT (thymine, thymine, thymine, thymine) (Life Technologies 2013).

Ion Torrent or Proton sequencing is similar to Illumina and 454 in that DNA is fragmented and one end is then immobilised to a slide, however it does not record emitted signal with light. When a nucleotide species is bound to its counterpart the reaction releases a hydrogen ion (H^+) thus changing the pH within the well. The pH change is measured for each probe between each wash to determine the strength of signal i.e. how many nucleotides were bound after that wash. Like 454 sequencing the slides is washed in one nucleotide species at a time and the strength of signal emitted recorded (Life Technologies 2013).

NGS sequencing is preferable to using microarrays as it more accurately reveals the genomic sequence within each sample (EMBL-EBI 2015). The use of microarrays limits researchers to detection of predetermined probes. This is sufficient to interrogate a sample of presence/ absence or changes in expression of known genes. Whereas NGS allows a wider scope for investigation of single nucleotide polymorphisms, mutations and more (Koboldt *et al*., 2013) thus are a closer biological representation of any a given disease state.

None the less, gene microarrays are able to report the presence or absence of a gene within biological samples and assign a quantitative value thus are still popular investigative tool (Koboldt *et al*., 2013, Life Technologies 2013).

2.2.2. Data Mining

Gene microarray, protein profiling and many omics protocols produce highly dimensional noisy data requiring in-depth bioinformatics approaches to interpret (Lancashire *et al.,* 2009). Moreover, prerequisite data pre-processing steps such as alignment, normalisation, baseline subtraction or feature extraction add to an expanding maze of possible data processing workflows (Allison *et al.,* 2006). Suppliers of omic technology platforms commonly integrate a data processing workflow within their software, however, interrogation of complex multidimensional data using different bioinformatic processes can yield different answers. Often researchers merit from re-exploration of such data by alternative analyses. Combine these challenges with the often-limited sample numbers and wide heterogenetic inter-sample variation from biological samples. The challenge of data mining in biomarker discovery is to filter through the meaningless variation between samples from natural biodiversity and identify only those relating to the disease (Allison *et al.,* 2006).

2.2.2.1. Statistical Analysis for Omics Data

After the overall research goal, the parameters/variables available with clinical samples primarily dictate the appropriate analyses for any given data set. Common examples of controlled variables from cancer research include

- Paired categorical: - cancer verses control
- Categorical: – Stage 1 verses 2, 3 and 4.
- Continuous: – time to event data time until relapse or death after a particular treatment, commonly referred to as survival data.

In a clinical research setting, it can occur that the ideally desired information to answer the research question is not available and has to be accommodated for post analysis during result data interpretation.

For each data type, there are numerous statistical analyses that would be considered appropriate, each with benefits and disadvantages and each yielding slightly, or not so slightly, different findings from the same data (Allison *et al.,* 2006). An ongoing scrutiny of published data and debate as to what is the most appropriate set of analyses for each data set encourages evaluation of statistical methodologies and encourages development of new or more widely accepted data handling procedures (Allison *et al.,* 2006). One could argue a robust approach would incorporate more than one type of data analysis, focuses on results common between the two and evaluates the reasons behind differing results.

2.2.2.2. Categorical and Continuous Variables in Omics Data

T-test

Used for over a century the T-test is a widely-accepted test to assign a significance to the difference between two populations of data. For example, the differential expression of a gene across two cohorts. Incorporated into most software packages of omics technology platforms a T-test can be applied to thousands of inputs simultaneously. Limitations of using T-test include that they only compute two-dimensional data, and do not account for sample variation, or noisy data which inherent problems of large omics data sets. With respect to biomarker discovery, a T-test is a rudimentary assessment.

False Discovery Rate (FDR) control such as the Benjamini Hochberg or the more stringent Bonferroni are normally applied to multiple hypothesis testing procedures of data sets with large variables compared to sample sizes. The purpose is to minimise the number of falsely rejected null hypotheses; Type 1 errors (Devlin *et al.,* 2003).

Principal Component Analysis (PCA), is another analysis/algorithm often written into the software of omics technology platforms. PCA identifies which variables of a data set to plot the sample cohort against to explain the maximum variance. The top principal components are separated recorded and the process is repeated thus finding the successive combinations of variables which best separate a sample cohort. PCA can be applied to biomarker discovery where in two sample subpopulations are identified when the data is separated on its principal components (Ringnér 2008).

Cluster analyses is an exploratory grouping analysis applied to group samples within a cohort with similar expression patterns of particular subsets of variables. This is done by identifying similar attributes between samples. Most commonly in gene microarrays this would be applied to ontology or pathway categories when these descriptors are available (Eisen *et al.,* 1999).

2.2.2.3. Analysis of Omics Survival Data

Time to event data in the biomedical research setting is commonly termed survival data. Survival data is continuous numeric unpaired data. The demand for analysis of survival data is predominantly from medical data were the measurement of time to event, often death is poignant (Machin, Cheung and Parmar 2006). Referring to the following particular set of collections of statistical analysis as ‘survival analyses’ is believed to have become popular from the mid-20th century with an increase in commercial interest into safety (Singh and Mukhopadhyay 2011).

The statistical methods used to analyse survival data are constantly evolving as the availability of increasingly powerful software is developed to handle data of escalating size and dimensionality (Machin, Cheung and Parmar 2006).

Numerous software packages including SPSS and Statistica enable non-statisticians to perform survival analysis without needing to know the mechanics of the calculations.

A life table is a rudimentary way to look at survival data (Singh and Mukhopadhyay 2011). The survival results are divided displayed separated by their variables and set time intervals. From this, further calculations can be made; the proportion of samples not survived/ failed, the portion surviving, the hazard function and the hazard ratio.

The hazard function is the calculated probability that the event will happen at any particular given time point. The hazard ratio is the proportional difference between the calculated hazard function between two groups.

Regression Analysis for Survival Data

Regression analysis is applied to investigate the relationship between risk factors and various events, logistic regression for categorical variables and linear regression for continuous variables like survival data. A correlation plot of time versus continuous variable Y, the potential risk factor, can be plotted, the distance between each data plot and the line of best fit, the residual value is used to assess the fit of the correlation i.e. the strength of the association, or, the likely hood that the risk factor influences the survival time. Basic regression models cannot compute survival time as a viable or censored data (Singh and Mukhopadhyay 2011).

Censored data is incomplete survival data. Censoring is problematic to survival data analyses, however commonly occurs in the biomedical setting for the following reasons; if the study ends before all the participants experience the measured event, if something happens so a participant is unable to experience the measured event or if a participant leaves the study. Respective examples for a study of patient survival after a test treatment; if a patient outlives the study period, if a patient dies of a reason irrelevant to the study and if a patient chooses to leave the study. Censoring can have a left or right bias termed left censored or right censored data. In biomedical research, right censoring is more common as after a deliberately designed study start event is almost always known the end date may not be not be possible to collect due to reasons listed above (Singh and Mukhopadhyay 2011).

Kaplan Meier plot

A Kaplan Meier plot is survival table data graphically depicted as a Kaplan Meier plot (Kaplan and Meier 1958). Time is seen on the X axis and Y axis is the number of survivors usually as a percent. A line is plotted for each sample group which decreases/increases at every point an event occurs, often taking the appearance of a set of stairs with varying sized steps. The difference between the two groups plots can be observed. Visualisation of the survival on a Kaplan Meier Plot can be more informative than a test statistic from a set time point, or a comparison of overall influence of the risk factor on survival as these will not illustrate or work calculate indicate a lower significance if the survival curves may cross at any point or if there is a far greater difference in survival within one sub time period.

The Cox Proportional Hazard (Cox 1972), followed by the Log-rank Test and Wilcoxen Rank test are the most common survival analyses used in the biomedical setting. The Cox proportional is the most popular as it is accommodating of censored data (Singh and Mukhopadhyay 2011).

The Log-rank Test

The Log-rank test is commonly implemented to compare survival data of two or more groups (Peto and Peto 1972). Using ovarian cancer survival data as the example, the Log-rank calculation tests the null hypothesis that there is no difference between the probabilities of a death occurring at a time point between two groups of patients: The calculation compares the difference of the observed and expected survival values and X^2 is used to assign a significance value:

For each group the following calculation is made at each event time point, where A is the number of patients in each group, B is the number of events that has occurred and C the number still alive.

$$A(B/C)$$

The sum of the observed (O) and expected (E) values for each group from all event time points are compared

$$(O_1 - E_1)^2/E_1 + (O_2 - E_2)^2/E_2$$

This value and the degrees of freedom, which is the number of groups minus one, is applied to a X^2 distribution table to return a p-value describing the significance of the difference between the two groups (Bland and Altman 2004).

The advantage of the Log-rank test is it takes the trend of the entire follow up period into account which can be a more representative analysis compared to looking at survival rates at a particular time point, for example 3 or 5-year survival rates.

Alone, the limitation of the Log-rank test is it only describes the significance of the difference between the two groups. Hazard ratios need to be incorporated to comment on the size and trend of the differential (Bland and Altman 2004). It is not essential to visualise the shape of the survival curve to perform a Log-rank Test, however may be beneficial to understand the trends of the data, the Log-rank test is less likely to find significance if the survival curves overlap. Overlapping survival cures are common in medical research trials with surgical intervention as the surgery itself has a high risk within a small time interval (Bland and Altman 2004). Bias can occur if censored data is not evenly distributed between the two groups. (Bland and Altman 2004)

A limitation to parametric methods is that assumptions are made about the data. Both the Log-rank Test and the Kaplan Meier make the following assumptions about the data to which they are applied

- Censoring has no effect on prognosis
- Cases recruited at the beginning and the end of the study have the same survival probabilities as those at the beginning
- That the time periods recorded are accurate (Bland and Altman 2004)

Cox Proportional Hazard Regression Model

The Cox proportional hazard regression model measures and compares the length of time between two marked events in two or more cohorts of samples. Introduced in 1972 (Cox 1972), it is a regression model applied to survival data, it can be used with data with and without censoring (Singh and Mukhopadhyay 2011). It compares the difference between the group's measurements and assigns a significance, whilst at the same time accounting for influential covariates, i.e. censored data. Most commonly this is the time from the commencement of observation to and event of relevance namely death, disease recurrence or recovery, hence the name ‘survival analysis’ (Singh and Mukhopadhyay 2011).

Cox model assumes two parameters; that the hazard ratios of each data point/ patient are independent of time, and are only for time-independent covariates (Singh and Mukhopadhyay 2011).

2.2.2.4. Machine Learning.

Machine learning, is a division of artificial intelligence (AI) relating to the creation and development of pattern recognition algorithms. The use of AI in place of conventional parametric linear based statistics is believed to improve the probability of identifying novel biomarkers from omics data via iterative processes that are designed to accommodate highly dimensional, noisy data. Data derived from nature is inherently ‘fuzzy’ (Lec and Guĕgan 2000), AI such as ANNs allow computational biologists to create a bespoke interrogation method for each data set, that will accommodate the non-linear relationships between variables.

Lancashire *et al.,* (2009) reports omics data to be considered be challenging due to the high number of input variables and limited number of cases, an inherent characteristic of biological data sets. Yet, in an in-depth comparison of MLP-ANNs with other statistical methods applied to large, multidimensional non-linear omics data sets was found to be the optimal analysis to apply, mainly due to their architecture giving ability to cope with highly dimensional or noisy data (Lancashire *et al.,* 2009). This however may not apply to all data sets/ data types.

The overall goal of machine learning or AI applied to data mining is to identify and interpret meaning from large highly dimensional data sets that it would be impossible to calculate manually. AI is applied to feature extraction and data pre-processing stages of omics data as well as the key applications which include supervised learning, such as relevance vector machines (RVM), decision trees, neural networks and Support Vector Machines (SVM). Decision trees are a predictive model used to make predictions of data based on sequential observations of a model data set. SVMs are an AI model used to find an optimal multidimensional line, a hyper-plane, to define subgroups or clusters of multi-dimensional data as if data were plotted in an artificial highly dimensional space (Dreiseitl *et al.,* 2001). RVMs are akin to SVMs however include probability into the classification (Tipping*,* 2001).

“The Curse of Dimensionality” a term coined by Richard E. Bellman (1966) is a phrase used to acknowledge the complexity of studying large data with numerous variables using algorithms, where the dimensionality of the data has the potential to mask the key features driving the structure of the data.

2.2.2.5. Artificial Neural Networks (ANNs)

ANNS, the most popularly used AI in medicine and molecular biology (Lec and Guĕgan 2000) are a statistical application of AI that can be applied in classification or predictive modelling. “A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation” (Caudil 1987). The term neural network is used in reference to biological neurons in that, the conditioned plasticity of synapses dictate future behaviour.

Multilayer feed-forward neuronal networks, also known as back propagation or multilayer perceptron (MLP) neural networks were first described by Rumelheart *et al.,* (1986). Who, details a then new procedure where connections within networks of neurone like processing units were continually being adjusted based on the measurements performance in comparison to an output vector.

MLP ANNs are a form of supervised learning, meaning that a training step is required from a set of model data with known outputs. The association between the input and output layer needs to be derived from training information or example data. A relevant example would be using measured gene or protein levels as input and the categorisation of cancer or control as output.

The work described below will focus on multilayer perception (MLP) ANNs, a basic form of ANN used for generalisation.

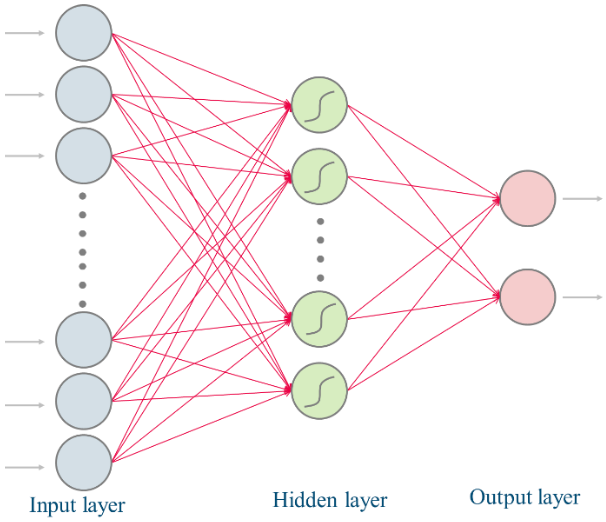


Figure 13. Typical Multilayer Perception Artificial Neural Network Architecture.
A fixed number of input and output neurons and hidden layer intermediating communication. A calculation made in each hidden layer node tailors the output based on learned information from each set of inputs.

Figure 13 displays an example architecture of an ANN. Each circle represents a calculation centre, also termed a node. MLP-ANNs consist of three layers of nodes; an input layer, an output layer and a hidden layer. Or, two layers of variables sandwiching a layer of calculation neurons. The input nodes represent predictor variables, the hidden layer nodes are where the weighting for each variable is calculated and the output layer is the calculated predicted outcome. The MLP-ANNS used below are tailored for categorical data (i.e. cancer vs control), however the output is in a decimal place format ranging from 0 to 1, anything above 0.5 is rounded up and below is rounded down to create a categorical output. The input and output variable nodes will vary depending on the nature of the data under scrutiny, a basic example model would be where the input variables are genes from an expression microarray derived from cancer or control tissue.

The ANN is perceived to ‘learn’ when pattern recognition algorithms in the hidden layer calculate the relative importance of one input/output node in relation to another, then, assigns the connection between them a respective weighting i.e. the strengthening or weakening of the *in silico* synapse.

MLP-ANNs can have an internal training, testing and validation phase, used to prevent overfitting. Sample cases of data sent to the ANN are allocated, often randomly, to train test or validate the network in a process termed cross validation.

Data used to train the ANN is not blinded, hence this is termed supervised learning. Weighting of the network is attributed to the input/output node connections, based on the two variables correlation. In the example of microarray data from cancer and control samples, if one genes expression is consistently high in the cancer group and consistently low in the control group the algorithm in the hidden layer node will increase the weighting of this genes outbound influence, thus giving it more influence for the test and validation stages. During the testing stage the now trained model is applied to the test sample cohort predicting what their output would be based on the training set. The measured error between the test groups predicted and actual output is determined, and the error is fed back into the model to adjust weightings. The adjusted model can then be applied to the validation set, the performance of the model can be assessed on its ability to correctly categorise the data from the blind validation set.

Within each node is a sigmoidal activation function which calculates whether to feedback positively or negatively, either increasing or decreasing the weight of the preceding *in-silico* synapse.

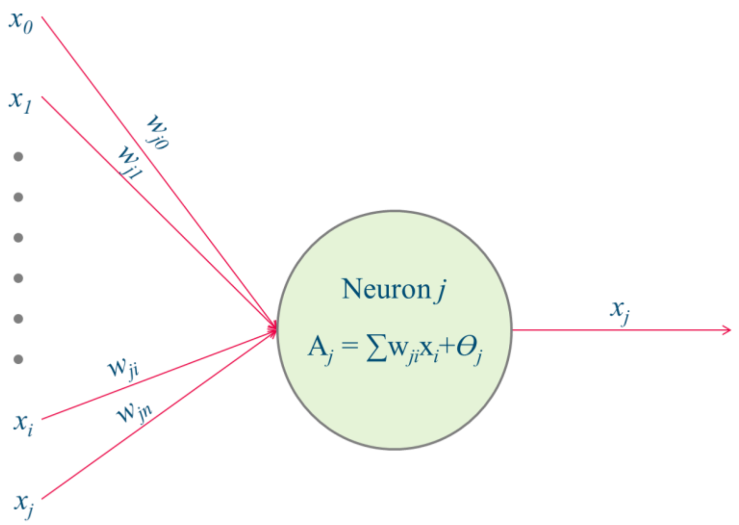


Figure 14. A Representation of a Hidden Layer Node from a MLP ANN
Adapted from Lec and Guĕgan (2000)**.** Each input synapse value (x_i) is associated with a weight (w_{ji}). The output (x_j) is calculated in the node. The node may be connected to more than one output, this diagram has only one output to signify the node is only able to produce one output value per input.

Figure 14, adapted from Lec and Guĕgan (2000), depicts how the MLP ANN node mirrors the physiology of a biological neuron. Signal is received from dendrites converging from multiple, separate locations or inputs (*x*), the information is processed and transduced in the cell body or calculation node (Neuron j) and one single output signal is emitted, which may create a positive or negative feedback loop.

The weight of the effect of the *i*th neuron to the *j*th neuron is represented $w_{ij.}$ The formula is expressed as:

$$A_j = \sum w_{ji} x_i + \Theta_j$$

Where; *i* is the totalled count of nodes in the preceding layer. *Θj* is a bias term, this influences the calculations horizontal offset. If *Θj* is used as the weight from the modified output unit Lec and Guĕgan (2000). After the association between the input and output layer is established, the predictive error, the output value is determined.

Not only can MPL-ANNs be implemented in different ways, the calculations can be extracted at different stages for interpretation as needed. In the example of an MLP-ANN applied to a gene microarray to classify cancer samples from control, the ANN can be used in a stepwise manner to generate a biomarker panel or simply to rank genes in order of their significance. To generate a biomarker panel, the MLP-ANN can be used in a stepwise manner wherein after one cycle of training, testing and validation the most influential gene is selected based on the lowest error, the data relating to that gene is set aside and the training stage is repeated. At the validation stage of the second loop the predictive performance is measured on all the top ranking from this training cycle together with those of all preceding cycles. This process is repeated until the performance of the model stops increasing, thus yielding a panel of the most predictive markers. Alternatively, the genes can be ranked based on their predictive performance in the first cycle, or an average performance over several cycles to assign a predictive value to all genes

Benefits of using MLP-ANNs include:

- They are able to perform generalisation that is that they can be used to make predictions of new data based on training data.
- They are able to handle highly dimensional data and do not rely on a normal data distribution.
- Calculations or results can be extracted at several junctures for an in-depth and flexible analysis. These are 1) analysis of the interconnecting network weights 2) sensitivity analysis and 3) rule extraction
- They are able to process complex non-linear relationships or interactions within data that are too complex to de-convolute using conventional linear methods.

- The detrimental effect of noisy data is dampened by the redistribution of samples across multiple cycles of training i.e. they are fault tolerant.

The limitations and caveats of using MLP-ANNs include:

- They are limited by the quality of their training data. The inherently wide heterogeneity found in biological samples will hinder the ability to isolate potential biomarkers from noisy inputs. Predictive algorithms, modelled from poor quality data, with poorly controlled extraneous variables will perform badly at validation.
- They can take a long time to complete the training stage calculations. The higher the number of hidden layers the longer the model can take to train. The number of hidden layers required is determined by the number of data features needing to be captured i.e. the complexity of the data.
- Over fitting (Hawkins 2004) can occur and will be detrimental to performance at the test and validation stage. Over fitting is a ‘memorisation’ of the training data and is most commonly caused by a small training data set, a common challenge with clinical data sets.
- ANNs, and other AI have been dis-affectionately labelled “black boxes” as the calculations and algorithms within them are often not seen or available for scrutiny. The MLP-ANNs utilised in the following work were created specifically with intent to apply to omics data from biomedical sample cohorts and successfully applied as such (Lancashire *et al*., 2008, Dhondalay *et al.,* 2011 Kafetzopoulou *et al.,* 2013) and a researcher familiar with their programming oversaw their application to the following work.
- Another common adage associated with analyses of complex, fuzzy high noise data ‘rubbish in rubbish out’ refers to the propensity of models to perform poorly if the quality of the training data is poor or high in extraneous variables. This highlights the necessity to control any extraneous variables everywhere possible. A specific example is seen in the reporting of MALDI-TOF based studies into novel serum biomarkers of ovarian cancer. Subsequent acclaimed novel biomarker panels were later accused as being an artificial product of not controlling sample handling extraneous variables prior to a predictive modelling analysis thus producing potentially artefactual interpretations See section 3.1.1. (Petricoin *et al.,* 2002, Baggerley *et al.,* 2005, Vaughan *et al.,* 2012). Hence standardised sample handling and quality control and data pre-processing protocols must become part of data prep (Tong *et al.,* 2012, Allison *et al.,* 2006).

ANNs can also be applied to visual spatial data such as ultrasonography (Jacobs *et al*. 2004) to do this measurements and image intensity is broken down to numerical format.

The reader is referred to Lancashire *et al.,* (2009) for a full evaluation of MLP-ANNs against other machine learning approaches such as support vector machines, logistic regression, nearest neighbour analysis (kNNs) and other computational strategies for analysing large multidimensional data sets.

2.2.3. Curated Data Repositories and Online Tools

Sharing of information is key to progress scientific knowledge. Establishing databases of scientific measurements/information/sequence enables for peer critique/review allows for dispute/conflicting interpretations to be addressed and resolved the ease and speed this can happen will dictated the maturity and reliability of the data within. Ultimately/eventually/ eventuating in libraries of trusted/widely accepted almost-facts.

One potentially problematic observation explored below is when multiple databases of the same information exist. This will naturally happen in the event novel types of information, or curation of existing information, creating a need for novel databases, these will most likely begin at a small scale potentially in parallel timescale with peer researchers at different institutes.

A simple example of this and how this problem is resolved with the time/evolution of a database is that of Uniprot a world-wide trusted library of proteins sequence data. The need for a proteomics database was first identified and published in 1969 (Dayhof 1969). In the prevailing 40 years, the creation of various databases of protein knowledge has evolved with discovery and demand for their use. Uniprot, is now comprised of a consortium of three major institutes the European Bioinformatics Institute (EBI), the Protein Information Resource (PIR) and the Swiss Institute of Bioinformatics, each are responsible for a key aspect of Uniprot. These were previously separately available containing contrary sequence data and annotation priorities.

A more complex example of this is databases of protein interaction, several exist and are explored below. Although some draw their source information from multiple sources, a consortium of all protein interaction database providers does not currently exist, multiple are available. Some are painstakingly manually curated by researchers, others are created using algorithms searching and matching key words from literature. The former contains a qualitative data that has been read and interpreted in the context it was meant, however, may not be as comprehensive as a search performed by a computer. The latter is a more comprehensive search however the data must be considered with the caveat that is the way that is was created – without human thought.

2.2.3.1. Data Sharing

Array Express

The European Bioinformatics Institute, part of the European Molecular Biology Laboratory (EMBL-EBI) website hosts ArrayExpress; an online database of functional genomic experiments. All publications using gene arrays are encouraged to upload data generated for fellow researchers to scrutinise and attempt to validate published findings.

Regulations enforcing the availability details of each published experiment are in place. Details of the samples source, collection protocol, and arrays used data acquisition protocol and data normalisation and must be listed to ensure as meaning full interpretation as possible can be made (ArrayExpress 2013).

Gene Expression Omnibus (GEO) hosted by the National Centre for Biotechnology Information (NCBI) is the US counterpart to Array Express (Gene Expression Omnibus 2015).

There is a crossover between Array Express and GIO, most importantly data submitted to either must meet the Minimum Information About a Microarray Experiment (MIAME) guidelines. Having this template as a basis means that incomplete data cannot be uploaded, and that variables are not recorded in an ambiguous, confusing or misleading format. This is essential as the purpose of sharing the data is for peer scrutiny, it is essential that the way the data was collected and recorded is clear.

The Proteomics IDEntification database (PRIDE), also hosted by EMBL-EBI is a database of proteomic mass spectrometry data experiments where data from such experiments can be shared among peers for reanalysis (Vizcano *et al.,* 2016).

2.2.3.2. Protein interaction databases

STRING 9.05

Search Tool for the Retrieval of Interacting Genes/Proteins 9.05 (STRING) (Snel *et al.,* 2000) is a user friendly online resource containing listings of proteins linked by; localisation, homology, text-mining, databases, experiments, co-expression, co-occurrence and gene fusion. Lists of gene or protein identifiers can be entered and interactions between those listed are generated and displayed diagrammatically. Filters can be applied to control the nature of the interactions listed (STRING 2013).

Reactome

Reactome is a freely available curated database of protein interactions from EMBL-EBI (Croft *et al.,* 2010). It prides itself on being manually curated and peer reviewed. Stringent regulation of what qualifies as a protein interaction allows for a higher confidence to be put in any interaction identified via Reactome compared to that of an automated, non-curated database such as STRING or IMEx. However, any new, missed or misinterpreted interactions by the curation team will not be identified.

IMEX

The International Molecular Exchange Consortium (IMEx) is another project to centralise knowledge of protein interactions IMEx (Orchard *et al.,* 2012), it sources its information from curated databases however itself does not state to be manually curated. The constituent collaborators of IMEx include; Cardiovascular Gene Annotation Initiative funded by the British Heart Foundation (UCL-BHF), The SIB Swiss Institute of Bioinformatics (SIBm), Uniprot, Molecular Connections plc, , Extracellular Matrix Interaction Database (MatrixDB), Interlogous Interaction Database (I2D), Molecular Interaction Database (MINT), Database of Interacting Proteins (DIP) an immune response pathway database (InnateDB), a mechanobiology database (MBInfo) and IntAct the molecular interaction database hosted by EMBL-EBI (Orchard *et al.,* 2012). Although IMEx is of a larger collaboration than STRING, search results are links to the origin database thus necessitating a deeper, analytical review of search results to use. Both STRING and Reactome facilitate searches of multiple terms.

IntAct is a database of molecular interactions, it is hosted by EMBL-EBI and is curated by their web-based curation tool (Kerrien *et al.,* 2012).

MINT is a manually curated protein-protein interaction database with a focus on experimentally proven interactions (Licata *et al.,* 2012).

More recently, considerations have been made to combine IntAct and Mint for a more comprehensive database (Orchard *et al.,* 2014).

I2D (Brown and Jurisica 2007) is a database of experimental, predicted and known protein-protein interactions. It was developed in the Ontario Cancer Institute.

UniProtKB Interactions is a search tool subpage of UniProt that facilitates searching for interacting proteins. Its information is derived from IntAact and is updated monthly. However, it is limited to binary interactions, and, when tested on two known interacting proteins IGF2 and IGF2BP, a value of "no match found" is returned.

Genes and proteins commonly have multiple aliases, formatting of search terms for each database is different and information can be lost by mislabelling a search term especially when databases are searching each other.

KEGG

Kyoto Encyclopaedia of Genes and Genomes (KEGG) is a database curated of genomic and molecular level information, containing schematic diagrams of systems and pathways that were manually created using bespoke software and algorithms (Kanehisa and Goto 2000, KEGG 2016)

More Pathway data bases can be found on

- Qiagen
- Tocris from R&D systems.
- Thomson Reuters pathways, also known as METACORE Life Science Research

Other protein interaction or pathway databases services that are not freely available include

- Cell Signalling Technologies (CST), a curated pathway database from CST
- Extra Cellular Matrix Database Interaction Database (Matrix DB) (faulty website) (Matrix DB, 2015)
- Qiagen Pathways, a curated pathway database from Qiagen
- Pathway guide and iPathwayGuide from Adviata,
- NextBio Research from Illumina
- BioSystems pathways, incorporates NCBI, Entrez and KEGG, no fee but a membership credentials are required for use.

Table 2 (below), summarises the nature of the contents and key features in the curation of some widely used major data repositories and resources. For a full, detailed report on currently available data resources in the context of cancer the reader is referred to Pavlopoulou *et al.,* (2014), who, highlights how, in the area of cancer alone, the production of exponentially increasing amount of data from studies focused on genes, proteins, immunomics protein-protein or gene-gene interactions are being compiled into various repositories. There are repositories available for individual cancers these include lung, breast, osteosarcoma, pancreatic, renal, cervical prostate but most relevantly ovarian (Ganzfried *et al.,* 2013).

Table 2. A Summary of Globally Accessible Data Repositories and Resources.

Type	**Database name**	**Description**	Curation	Protein data	Gene data	miRNA	Interaction data	Pathway visualisation
Repository	PRIDE	Archive - proteomics data repository	Manual	Y				
Repository	Array Express	Database of functional genomic experiments	Manual		Y			
Repository	GEO	The Gene Expression Omnibus			Y			
Resource	STRING	Search Tool for the Retrieval of Interacting Genes/Proteins	Algorithm	Y	Y		Y	Y
Resource	Reactome	Reactome Database of Protein Interactions	Manual	Y			Y	Y
Resource	IMEX	The International Molecular Exchange Consortium	Algorithm	Y	Y		Y	
Resource	IntAct	Molecular interaction database hosted by EMBL-EBI	Algorithm	Y			Y	
Resource	MINT	Molecular interaction database hosted by EMBL-EBI	Manual	Y			Y	
Resource	I2D	Database of experimental, predicted and known protein-protein interactions	Hypothetical					
Resource	KEGG	Kyoto Encyclopaedia of Genes and Genomes	?		Y		Y	Y

2.2.3.3. Ontological Databases

Databases annotating classifications of genes, gene products and sequences have been evolving in parallel with gene and protein characterisation. Two examples are the Protein ANalysis THrough Evolutionary Relationships (PANTHER) Classification System, where proteins are grouped by subfamily/family, molecular function, biological process or a pathway (Thomas *et al.,* 2003, PANTHER 2016). Or, the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang *et al.,* 2008, DAVID 2016) which hosts tools to annotate and visualise refined gene or proteins result lists based on function, functional classification, biological theme, interactors, and more (Huang *et al.,* 2007, DAVID 2016).

Importantly, large scale centralised collaborative efforts exist (Gene Ontology Consortium 2004) however, in the case of ontology although source data needs to be clarified and not duplicated, there is more of a justification to have multiple, individual niche tools which may focus on a specialist classification. The categories listed between the two examples above are similar yet notably different.

2.3. Aims of the Project Overall

- To identify novel biomarkers that can improve the prognosis of ovarian cancer suffers.
- To discover novel protein biomarkers detectable in serum using MALDI-MS that could be used as a screening tool for the early detection and monitoring of progression of disease.
- To find a reproducible sample preparation and MALDI-MS workflow which enables identification of novel protein biomarker/s in serum.
- To identify genes that are associated with the progression and survival of ovarian disease.

3. Proteomic Evaluation: MALDI-MS Profiling Strategy for Biomarker Discovery in Ovarian Cancer

Chapter Abstract

The work described in this chapter aims to identify novel serum protein biomarkers using MALDI-TOF mass spectrometry. A MALDI-TOF MS profiling comparison was conducted and detailed. Briefly; MALDI-TOF-MS profiles were acquired from serum samples collected from 30 patients with ovarian cancer and 30 age matched controls. Tight regulations were applied to the sample collection and processing to address criticisms of previous similar studies and prevent the degradation of proteins within the serum. MALDI-TOF-MS data was generated, processed, exported to Excel and analysed using Artificial Neural Networks (ANN's). A biomarker panel of five peptide masses was found to differentiate cancer from aged matched controls with an accuracy of 91% and error of 9%. To assign a protein identity to each of the five peptide mass values to a protein identity a pool of cancer serum samples and a pool of the age matched controls were created. The two sample pools were separated by liquid chromatography (LC) on a C_{18} column into 384 fractions. Tandem mass spectrometry of the 384 fractions produced a matched peptide sequence to every possible mass value present across the two samples (LC-MALDI-MS/MS). These masses values of the biomarker panel were cross referenced with those of the LC-MALDI-MS/MS sequence data to assign an identity based on the matched mass. This produced a list of several possible identities for each peptide value of the biomarker panel and for some no potential identities were matched. The lists of possible identities of the biomarker panels were cross-referenced with literature, two selections were made to attempt a validation of finding on a separate platform. ELISA was used to assay the expression of these two proteins, Transferrin and Vitronectin, in the serum. The expression patterns seen in the ELISAs did not correlate with the trends observed in the MS data and literature.

On evaluation of the available data it was noted that, although the best available option, an ELISA assay of two proteins is an inadequate to validate the biomarker panel of the five protein identities generated via ANN analysis. Individually, the peptide identities held little statistical significance. Unfortunately, not all peptide mass values were able to be matched to a possible protein identity, thus it was not possible to attempt to fully validate the panel

Further to this, the use of the immunological approach assumes the link between the peptide and the whole protein. The statistical significance of identifying a peak based only on its mass is questionable.
In conclusion, more information on identity of mass values significantly differently expressed between cancer and control groups would be needed for this data to be successfully validated.

3.1. Introduction

The aim of the following work was to identify novel protein signatures of serum proteins that could be used as a diagnostic or prognostic tool in ovarian cancer.

3.1.1. MALDI Mass spectrometry and ovarian cancer

Matrix Assisted Laser Desorption and Ionisation Time of Flight Mass Spectrometry (MALDI-TOF-MS) is (was – when this work was conducted) a potentially powerful novel biomarker discovery tool in bottom-up proteomics and is introduced in section 2.1.3.2. Briefly, an analyte of interest such as blood serum containing complex mixture of proteins is combined with an acid matrix and dried to a crystal on a target, inside a mass spectrometer a laser is fired on the crystal, laser energy is transferred through the matrix and into the protein sample, ionised proteins are desorbed from the target and travel through a vacuum towards a detector, over the length of a flight tube proteins are separated massed on their mass; smaller molecules will have a shorter flight time than larger ones. Signal from the detector is commonly presented as spectra, a graph with time of flight convert to mass to charge ratio on the x-axis plotted against the intensity of signal detection at the time. Using MALDI-TOF-MS hundreds of samples can be spotted to one target, unique profiles of from all generated simultaneously and compared. Using this bottom-up approach minimises experimental bias as data from all samples for a comparison is acquired in a small time frame. Due to the sensitivity of the instrumentation MALDI-MS is susceptible to bias introduced from the variation in sample preparation and handling.

Popularity of biomarker discovery via MALDI/SELDI-TOF-MS peaked in early to mid-2000s; numerous groups published mass values of peptides identified from mass spectra that were significantly differentially expressed in ovarian cancer, control or benign (see Table 3). Failure to reproduce findings or give meaning to the mass values of ions that discriminated the cancer cohorts damaged the image and trust/confidence in its use. This is reflected by a drop in publications from MALDI/SELDI-TOF-MS data. To confirm any potential novel biomarker findings results must be reproduced on either or both of; a second technological platform and a separate sample cohort.

Table 3. A Summary of Similar Research. An overview of the sample selection, preparation, purification, methods and data analysis used by researchers investigating specific proteomic fingerprints in the sera of ovarian cancer patients. Numerous mass to charge values found to be significantly linked to ovarian disease detection. Few mass values are assigned to a protein identity.

Researcher	Sample source	Mass to charge ratio (*m/z*) (or mass in Da where indicated) values where significant differences in peak amplitude occurred in ovarian disease vs control spectra	Sample purification technique utilised	Data analysis technique used	Statistical significance attained.
Petricoin *et al.,* (2002)	The National Ovarian Cancer Early Detection Program Clinic at Northwestern University Hospital (Chicago) (n=100) Simone Protective Cancer Institute (Lawrenceville) (n=17)	534Da, 989Da, 2111Da, 2251Da, 2465Da.	C16 Hydrophobic interaction ProteinChip.	Genetic algorithms combined with cluster analysis. Samples divided into training and validation. (no test set)	100% sensitivity, 95% specificity, positive predictive value of 94%
Sorace and Zhan (2003)	Three sets from The Clinical Proteomics Program Databank: (Data set **2-16-02** used in Petricoin., *et al.,* (2002), **4-3-02** consist of the same samples but run on WCX2, and **8-7-02** a different data set) n=469	2665.397, 3969.46, 3991.844, 4003.645, 4027.3, 4056.967, 4744.889, 6801.495, 7786.054, 8349.266, 14796.14, 15955.47, 17034.05	1st set Ciphergen H4 ProteinChip array (since discontinued). 2nd and 3rd set Ciphergen weak cation exchange 2 ProteinChip array (WCX2)	Nonparametric statistics. No baseline subtraction. Wilcoxon test; then Wilcoxon with stepwise non-discriminant analysis.	100% sensitivity and specificity
Baggerly *et al.,* (2003)	As above	435.46, 465.57, 2760.67, 3497.55, 6631.70, 14051.98, 19643.41	As above	Peak alignment, the 506 most frequent peaks were selected for two-sample T-test analysis, 15 were significant.	94% Correctly classified samples.

Zhu *et al.,* (2003)	National Institutes of Health and Food and Drug Administration Clinical Proteomics Program Databank web site (Ovarian Data Set **4-3-02** and Ovarian Data Set **8-7-02**)	167.8031, 321.4157, 322.4204, 359.6322, 385.5688, 413.1668, 433.9079, 434.6859, 444.4690, 445.2563, 1222.1849, 1528.3431, 3345.7995, 3449.1503, 3473.3084, 3528.5266, 6101.6299 and 6123.5190	Weak cation exchange protein chip (WCX2)	Smoothing via Gaussian filters, pointwise two sample *t/z* test between groups of the training set, random field theory to identify the threshold and validated using the *l*-nearest neighbour method which also attains the sensitivity and specificity.	Sensitivity and specificity of the test are both 100%. The 95% confidence intervals for sensitivity and specificity are (93%, 100%) and (95%, 100%), respectively.
Vlahou *et al.,* (2003)	Division of Gynecologic Oncology, University of Texas, Southwestern Medical Center. n=139	5.54, 6.65, and 11.7 kDa (detected on the IMAC chip) 4.4 and 21.5 kDa (detected on the SAX surface form the main splitters).	Strong anion exchange (SAX) and immobilised-copper (IMAC) chip surfaces.	Samples separated to learning and test set. Five peaks were selected by the BPS algorithm to discriminate cancer from the non-cancer groups.	A specificity of 80% and sensitivity of 84.6% were obtained from the cross-validation set.
Kozak *et al.,* (2003)	Gynaecological Oncology Group and Cooperative Human Tissue Network. n=184	3.1 kDa, 4.5 kDa, 5.1 kDa, 7.8 kDa, 8.2 kDa, 16.9 kDa, and 18.6 kDa (increased expression in the cancer group) 13.9 kDa, 21.0 kDa, 28.0 kDa, 79.0 kDa, 93.0 kDa, and 106.7 kDa contrary	Strong anion-exchange (SAX2) (Ciphergen Biosystems).	Univariate and multivariate statistical analysis applied to protein-profiling data	The individual proteins in the malignant biomarkers group had values for ROC area ranging from 0.617 to 0.851, sensitivities from 48.1% to 81.5%, specificities from 66.1% to 88.1%, and accuracies from 61.3% to 79.3%.
Zhang *et al.,* (2004)	M.D Anderson Cancer Centre, Duke University Medical Centre, Groningen University Hospital, Netherlands, The Royal Hospital for Woman, Australia. (Four different medical institutes) n=503	12828Da 28043Da (decreased expression in cancer group) 3272Da contrary	Bound in triplicate with a randomised chip/spot allocation scheme to IMAC3-Cu, SAX2, H50 and WX2	Divided into test and training set for the derivation and testing of non-linear unified maximum separability analysis. Then further analysed using Mann-Whitney U-test or Krusckal-Wallis test	Sensitivity 83% Specificity 94%

Yu *et al.,* (2005)	Serum banks of Guangxi Medical University. n=61	5881Da, 7564Da, 6044Da (increased expression in the cancer group) 2085 Da, 9422 Da contrary	Hydrophobic surface (H4)	Tenfold cross-validation support vector machine established a diagnostic pattern.	estimated specificity of the test sets was 96.7%, the estimated sensitivity was 96.7%, the estimated positive predictive value was 96.7%
An *et al.,* (2006)	UC Davis Medical Centre Clinical Laboratories n=10	788.545Da and 899.690Da. (peptide peaks)	Dialysed with Pierce Slidealyzers, MWCO 7000-10000 for glycan analysis but some peptises identified.	n/a	n/a
Kong *et al.,* (2006)	Gynaecology and Obstetrics Hospital, Fudan University in Shanghai, China, from October 2002 to November 2003 n=195	7676_21, 11463_8, 11545_9, 11681_2,11706_6, 13790_8, 15908_3).	Immobilised metal affinity capture arrays (IMAC3)	ProPeak software. Calculates and ranks the contribution of each individual peak toward the separation of the two groups. Unified maximum separability analysis (UMSA) combined with CA 125	Combining the three biomarkers and CA125 produced sensitivity and specificity of 97%.
Zhang *et al.,* (2006)	125 samples from Department of Obstetrics and Gynaecology, Qilu Hospital of Shandong University, China Between Apr 2004 and April 2005	6195Da, 6311Da, 6366Da (decreased expression in cancer group) 11498Da contrary	Weak cation exchanger (WCX2)	A decision tree algorithm. The receiver operation characteristic curve found most significant peaks. Candidate biomarkers evaluated by Mann Whitney-U test or Kruskal-Wallis test.	Sensitivity of 87% Specificity of 95%
Current Study	120 serum samples collected from Derby City General Hospital, England Between September 2005-2008	1647.8, 1219.6, 3312.2, 1493.8, and 1820.	Millipore C18 ZipTips before and after a tryptic digest.	ANN used to compare the peptide profile of digested cancer serum verses age matched controls.	Accuracy (test performance) 91% Error (test error) (9%)

A number of the studies detailed in Table 3 use a SELDI approach and could be criticised for doing so, SELDI data only represents the proteome of protein species complementary to the ProteinChip© utilised. Vlahou *et al.,* (2003) demonstrates this by generating multiple biomarker panels from repeating the analysis of one set of samples with two SELDI platforms. A processed or purified sample is not representative the full serum proteome, which, biomarker discovery experiments would ideally aim to represent. For this reason, the data of many of the studies in Table 3 cannot be used to verify, validate or refute each other's generated biomarker panels.

Zhang *et al.,* (2004) attempted this by using a combination of four SELDI surfaces in a randomised layout. Another approach explored below would be to minimise or negate sample preparation with an aim to best represent the entire proteome. In this chapter sample preparation is kept minimal and MALDI is employed to attempt to measure a larger portion of the proteome as possible.

A common criticism of studies in Table 3, namely Petricoin *et al.,* 2002 is that they are biased by artefacts in sample collection and processing (Baggerly *et al.,* 2005, Vaughan *et al.,* 2012). For example, Petricoin *et al.,* 2002 tested samples and control samples were collected from different sources. The importance of control and standardisation of collection and storage of serum and plasma have been reviewed in Engweden *et al.,* 2003. Alterations of known serum cancer biomarkers has since been demonstrated to be altered by diet (Ong *et al.,* 2009). The sera for this study were collected in a tightly controlled manner with this in mind.

Though it should be fairly noted that supporting evidence for the peptidomic signature published in Petricoin *et al.,* (2002) was later reported (Conrads *et al.,* 2004) as summarised in Nossov *et al.,* (2008). The exact overlap in results is not specified in the original paper. Conrads *et al*., (2004) begins by noting that the low resolution TOF-MS used to generate the initial intriguing results is reproducible within runs and over small intervals of time, however week-to-week and machine-to-machine variation was at an unacceptable level for work in a clinical setting. The study expands by applying a higher resolution instrument and several bioinformatics approaches for interpretation of, quality control, and analysis of the data on an expanded double-blinded cohort to yield four models with 100% sensitivity and specificity.

Sample collection was another theme in the critique of this methodology. Protein signatures may be affected by any part of the sample collection procedure. Namely, the about of time samples are stored and at what temperature. Data sets have now become available online, including one of ovarian cancer tumour samples, collected with a 0, 5, 30 and 60 min delay between collection and freezing for storage (National Cancer Institute 2015). However, measuring the effect of degradation is an ambiguous task, as measuring the potential biomarkers themselves, lack of confirmed identity assigned every peak or a quantitative measurement prohibit this clarification. It has also since ben argued that a large enough sample size, though rarely available, should dampen the effect of outliers from unwanted inter-sample variation (Dun *et al.,* 2011). These critiques can be pre-empted by collecting both control and test samples under as similar conditions as is possible.

3.1.2 Aims and Hypotheses of the Chapter.

To identify peptide masses and identities that are significantly differently expressed in the serum of ovarian cancer patients compared to the serum of benign condition and control patients.

H_0-i: MALDI-TOF-MS with ANN analysis will not be able to detect unique protein patterns expressed in the sera of ovarian cancer sufferers.

H_1-i: That unique protein patterns expressed in the sera of ovarian cancer sufferers can be detected using MALDI-TOF-MS with ANNs and can be used to positively identify a blind validation set.

If H_0-i above is rejected: To identify the peaks found to be present in significantly different amounts in the tested serum by matching by matching LC-MALDI-MS/MS of a pool of the samples to MALDI-MS data.

H_0-ii: The masses of the peptide peaks expressed differentially in the tested serum cannot be identified by linking data from LC-MALDI-MS with the MS profiles.

$H_{1\text{-ii}}$: The masses of the peptide peaks expressed differentially in the tested serum can be assigned a protein identity by linking the MS-MALDI data with LC-MALDI-MS/MS data.

To reproduce the difference in expression observed using a different platform; immunoassay.

H_0-iii: No difference in expression will be noted of the proteins demonstrated to be expressed using MALDI-MS.

H_1-iii: That unique protein patterns expressed in the sera of ovarian cancer sufferers can be detected using MALDI-TOF-MS with ANNs and can be used to positively identify a blind validation set.

3.2. Materials and Methods

3.2.1 Materials

3.2.1.1. Equipment used

Table 4. Equipment Utilised for MALDI-TOF-MS

Equipment	Supplier
Bruker UltrafleXtreme Matrix Assisted Laser Desorption Ionisation-Time of Flight Mass Spectrometer (MALDI-TOF)	Bruker Daltonics
Desktop computers with use of SpecAlign, Excel, Statistica and Bruker Software which comprises of: FlexControl 3.3, FlexAnalysis 1.3	Bruker Daltonics
384 spot MALDI-TOF targets (Grounds Steel 384)	Bruker Daltonics
Automated pipetting machine	FluidX
Sonicator	VWR Ultra Sonic Cleaner
-80˚C freezer	New Brunswick
37˚C Incubator	Heraeus
Vortex	Scientific Industries
10 µL Millipore ZipTips C_{18} pipette tips	Millipore
Polypropylene bottles for reagent storage	Nalgene F.E.P.
1.5 mL Eppendorf	Eppendorf
1 mL pipette	Gilson
200 mL pipette	Gilson
2 mL pipette	Gilson
0.2-1 mL tips	Gilson
20-200 µL tips	Gilson
0.5-10 µL tips	Gilson

3.2.1.2. Reagents used:

Table 5. Reagents Utilised for MALDI-TOF-MS

Reagent	Supplier	Grade
Acetone	Sigma	LC MS
Acetonitrile (ACN)	Sigma	LC MS
Alpha-Cyano-4-hydroxycinnamic acid (CHCA)	Bruker Daltonics	MALDI-TOF-MS
Ammonium bicarbonate	Sigma	Laboratory
Distilled water	Barnstead Diamond	nano pure
Methanol	Sigma	LC MS
Peptide Calibrant II Bruker	Bruker Daltonics	MS
Trifluoroacetic acid (TFA)	Fisher Scientific	HPLC
Trypsin – Trypsin Gold, Mass Spectrometry Grade	Promega	MS
Tryptic Digest of Bovine Serum Albumin	Bruker Daltonics	MS

3.2.1.3. Stock Solutions Made and Used

Table 6. All stock solutions made and used for MALDI-TOF-MS

Reagent	Composition
0.1% TFA in H_2O (50 mL)	50 µL TFA 49.95 mL – Distilled water
0.1% TFA in Acetonitrile (50 mL)	49.95 mL – Acetonitrile 50 µL TFA
80% ACN diluted with 0.1%TFA (50 mL)	40 mL – ACN 10 mL - 0.1% TFA solution
Trypsin solution 0.5 mg/mL (200 uL)	100 mg – Trypsin Gold, Mass Spectrometry Grade 200 mL – 100 mM ammonium bicarbonate NH_4HCO_3
Iodoacetamide 200 mM (1 mL)	36 mg Iodoacetamide 1 mL of 50 mM ammonium bicarbonate
Dithiothreitol (DTT) 200 mM (1 mL)	30 mg DTT 1 mL 50 mM ammonium bicarbonate

3.2.1.4. Samples

Two hundred serum samples were selected and categorised by a consultant gynaecologist from a bank of serum samples ethically collected from patients in a Derby City General Hospital

Gynaecology ward between 2004 and 2007 (Southern Derbyshire Local Research Ethics Committee: REC reference number: SDLREC Ref 0205/495).

The four categories were age matched as closely as possible
30 Cancer. (1 Clear cell carcinoma,14 endometrioid adenocarcinoma, 1 mucinous cystadenocarcinoma, 1 mucinous papillary cystadenocarcinoma, 1 poorly differentiated adenocarcinoma, 6 serous adenocarcinomas, 2 serous cystadenocarcinomas, 4 serous papillary cystadenocarcinoma).

- 30 Cancer controls. Treated in the gynaecology ward but not for ovarian malignancy or benign condition aged matched to the cancer group.
- 20 Benign (4 mucinous cystadenoma, 2 serous cysts, 3 serous cystadenofibromas, 18 serous cystademonas and 3 serous papillary cystadenomas.
- 20 Benign controls. Treated in the gynaecology ward but not for ovarian malignancy or benign condition age matched to the benign group cohort.

A full analysis between each group was conducted, however for this report only cancer vs cancer control will be reported.

3.2.2 Methods

3.2.2.1 Sample Preparation and Data Acquisition

Two hundred serum samples were defrosted on ice, diluted 1 in 20 in 0.1% TFA and refined using Millipore C_{18} ZipTips and an automated pipetting robot. Trypsin was manually added to each sample for an overnight 37°C digestion then the automated pipetting robot was used to repeat the C_{18} ZipTip clean up and spot to the ground steel target for MALDI-MS analysis (as described in the section 2.1.3.2).

Samples were placed in randomised order to negate batch effect, and appropriate standards and blanks run alongside. 60% of the samples were processed on one date for biomarker discovery, the remaining 40% were processed on a second date to be used as a validation set, to test any biomarker discoveries.

MALDI-MS spectra profiles 800-3500 *m/z* were acquired for each sample using the Bruker Ultraflex III. All spectra were reviewed visually alongside control samples before progressing to data analysis.

3.2.2.2 Biomarker Panel Generation

Data was exported from the Bruker software to Excel and an in-house designed artificial neural network (ANN) (as described in the section 2.2.2.5) algorithm was used to data mine it and generate set of peptide masses that can discriminate two groups.

A stepwise analysis of 10 steps (repeats) was used.

3.2.2.3 Identification of *m/z* Values in the Biomarker Panel

At the time of defrosting for MALDI-MS profiling 2µL of each of the samples was taken to amass a pooled sample from each category. i.e. one “cancer pooled” and one “control pooled” sample that is made up of all of the other samples

The two pooled serum samples were sent to the collaborative/sales contacts at Bruker (the supplier of the mass spectrometer) in Bremen, Germany to be analysed by the next generation of mass spectrometric technology. At Bruker, each pooled serum sample was deconvoluted via C_{18} liquid chromatography fractionation prior to tandem mass spectrometry (see section 2.1.3 and 2.1.5). This was done in the same model of instrument (Bruker UltrafleXtreme) however using an automated program and different mode within Flex Control 3.3. Using this function an automated run acquires all the *m/z* values detected in each spot and compiles a list, following this the instrument switches to a more sensitive mode/reflection mode (see section 2.1.3) then returns to each spot and selectively isolates, fragments and measures each of the best intensity *m/z* values for each spot. In reflectron mode, the flight path of the ions is increased allowing for separation of the small, fragmented peptides (see 2.1.3.4). The fragment parent and fragment *m/z* values were searched at Bruker and the list of protein identities assigned to each was returned.

The peptide m/z values from the biomarker panel generated above were cross-referenced to the parent *m/z* values of the file returned from LC-MALDI-TOF profiling. A very wide boundary

(1 Da) was applied to allow for machine-to-machine variation and the disconnected nature of this method. Additionally, with the knowledge that thousands of peptides share the same *m/z* value, these potential identities were used to not conclude a match but as a research clue/piece of evidence to guide future experiments. The potential error in identity matching is reduced slightly by this being the same model of instrument analysing the same samples. For this reason, more than one protein identity may be assigned to each peak.

3.2.2.4 ELISA

Two ELISA kits for were purchased and performed as per instructions in the kits.
Genway Vitronectin ELISA KIT Catalog number 40-831-160002.
Immunology Consultants Laboratory Inc Transferrin L11 0-3S1

3.2.2.5 Additional Analyses

The following additional analyses were conducted and not relevant or reported in this line of investigation.

- Comparison of Transferrin and Vitronectin ELISA results against tumour stages and grades.
- Box plots of peaks of interest against stages and grades.
- CA125 and CEA levels supplied from clinical information were
  - Correlated with MS values of interest
  - Correlated against Transferrin and Vitronectin levels as assayed by ELISA
  - Included in a stepwise analysis
- CK10Ab was purchased as there was not an ELISA KIT available. It was used for IHC and western blots and also SILAC work started with CK10.
  - IHC of Biomax ovarian cancer TMA
  - IHC of frozen tissue from same source as
- The ANN analysis was repeated only containing the peptide masses that had matched identities.

3.3. Results

3.3.1. Generation of Biomarker Panel from MALDI-TOF-MS Data

Spectra of the test samples were acquired (as described in the methods section 3.2.2.1) and viewed in FlexAnalysis, any spectra deemed not to be of sufficient quality were removed from analysis at this stage. Intensity values of monoisotopic peak values for each sample were exported to MS Excel format using FlexAnalysis software. Statistica and in house developed software was used to compute the mass values that are differentially expressed between cancer and cancer-control groups.

Table 7 (below) details the results of the stepwise analysis. The performance (Average Test Performance), and error (Average Test Error) of the model with the addition of each peak value (input ID) is seen. Optimal performance of the model is at loop 5, the *m/z* for ions 1647.8, 1219.6, 3312.2, 1493.8 and 1820.0 produce a test performance of 91% and a test error of 9%.

Table 7.Stepwise Analysis Generation of a Biomarker Panel. Summarises each step of the stepwise model. The performance and error of the model is detailed with the addition of each input ID/ peak label.

LOOP 1	Input ID (*m/z*)	Average Train Perf	Average Test Perf	Average Valid. Perf	Average Train Error	Average Test Error	Average Valid. Error	Input Index
1	**1647.8**	0.59	0.64	0.50	0.22	0.22	0.27	4240.00
2	**1219.6**	0.76	0.73	0.75	0.18	0.18	0.18	2099.00
3	**3312.2**	0.82	0.86	0.75	0.14	0.13	0.18	12562.00
4	**1493.8**	0.85	0.91	0.83	0.12	0.12	0.15	3470.00
5	**1820.0**	0.91	0.91	0.83	0.10	0.09	0.13	5101.00
6	3293.8	0.91	0.91	0.83	0.08	0.07	0.15	12470.00
7	2269.2	0.91	0.91	0.83	0.09	0.08	0.14	7347.00
8	2541.4	0.88	0.91	0.83	0.09	0.08	0.13	8708.00
9	1981	0.90	0.91	0.83	0.09	0.08	0.14	5906.00
10	2202.6	0.88	0.91	0.83	0.09	0.08	0.14	7014.00

The ANN model was remade in Statistica software using identical parameters with 60% of the samples, tested using 20% then validated on the final 20%. Then all of the cases were blindly classified using the Statistica ANN fifty times.

Figure 15 (below) is a population chart of the performance of the model in Statistica. Each sample is seen in the x-axis; the y-axis represents the amount of times that case was correctly or incorrectly classified. Three control cases were incorrectly classified as cancer and 7 cancer samples were classified as controls.

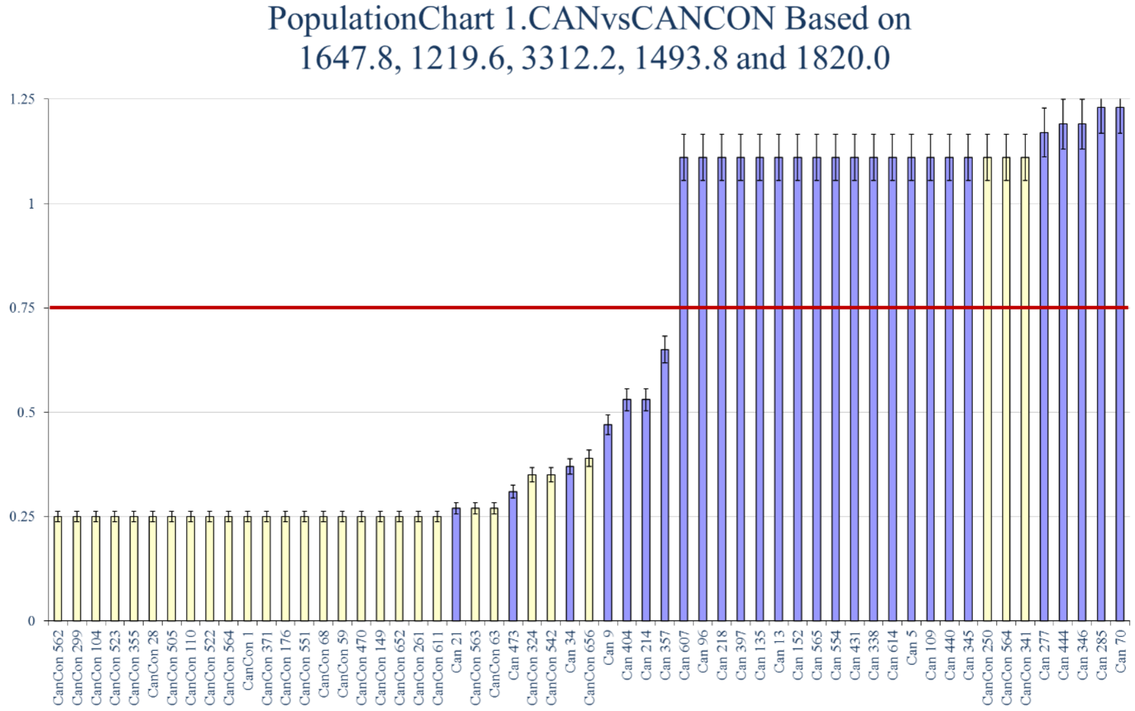


Figure 15. Population Chart of the Performance of the Biomarker Panel Discriminating Cancer from Controls.
All cases with a value above 75/the red line were classified as cancer, all cases below 75/the red line were classified as controls. Three false positives and 7 false negatives are seen.

3.3.2. Identification of the Peaks in the Biomarker Panel

LC-MALDI-TOF-MS/MS was performed on a pool of the cancer and cancer control serum to produce a list of possible identities for each peptide mass identities in the biomarker panel. The possible identities were compared with current literature and Vitronectin and Transferrin were selected to be interesting candidates for validation using immunoassay.

The peptide *m/z* values from the biomarker panel were cross-referenced to the parent *m/z* values of the file returned from LC-MALDI-TOF profiling. Table 8 below details the protein identities nearest to the *m/z* values of interest. This gave more than one possible identity for some of the *m/z* and none for one. A very wide boundary (1 Da) was applied to allow for machine-to-machine variation and the disconnected nature of this method. Additionally, with the knowledge that thousands of peptides share the same *m/z* value, these potential identities were used to not conclude a match but as a research clue/piece of evidence to guide future experiments. The potential error in identity matching was reduced slightly by this being the same model of instrument analysing the same samples.

Table 8 Potential Identities of the *m/z* values of the biomarker panel. The *m/z* values generated from the cancer vs control panel (left column) are cross-referenced against the *m/z* values of the parent ions generated from tandem mass spectrometry of the same samples by Bruker in Bremen, Germany. Empty cells indicate no match.

***m/z* from OvCa vs Con panel**	**Pooled Cancer Run 1of1**		**Pooled Cancer Run 2of1**		**Pooled Control Run 1of1**	
	Identity	***m/z***	**Identity**	***m/z***	**Identity**	***m/z***
1647.8	Homo sapiens **Vitronectin** (sterile-Cell Culture Tested Attachment Factor)	1646.8	APOE_HUMAN Apolipoprotein E Homo sapiens	1647.8	APOE_HUMAN Apolipoprotein E OS=Homo sapiens APOE PE	1647.8
	-	-	VTNC_HUMAN **Vitronectin Homo sapiens**	1646.8	**VTNC_HUMAN Vitronectin Homo sapiens VTN**	1646.82
1219.6	-	-	LV102_HUMAN Ig lambda chain V-I region Homo sapiens	1219.66	-	-
3312.2	Homo sapiens keratin 1 (KRT1)	3312.3	-	-	-	-
1493.8	KRHU0 keratin 10, type I, cytoskeletal - human	1493.7	K1C10_HUMAN Keratin, type I cytoskeletal 10 Homo sapiens	1493.73	**TRFE_HUMAN Serotransferrin Homo sapiens TF**	1494.73
	-	-	**TRFE_HUMAN Serotransferrin Homo sapiens TF**	1494.73	-	-
1820	-	-	-	-	-	-

Table 8 shows, when the samples were re-analysed with the same instrument with peptide matching functionality. The protein identities with the closest *m/z* values to 1647.8 were Apolipoprotein E and Vitronectin, 1219.6 to lambda chain V-I, 3312.2 to Keratin 1, 1493.8 to Serotransferrin and no parent *m/z* values were within 1 Da of 1820.

Keratin 1 is a common contaminant in mass-spectrometric experiments and is generally discounted from results, leached from researcher’s skin and hair despite personal protective equipment can be hard to eliminate entirely, thus is a poor biomarker candidate for further research. Apo lipoprotein is a highly abundant serum protein so without a quantitative application it would be hard to make a publishable novel contribution to current knowledge.

The lists of potential identities were then compared with current literature and two identities were chosen for further investigation based on their relevance, reported link to ovarian cancer, and therefore their potential use as a prognostic biomarker. Transferrin has been reported to be expressed in serum at decreasing levels relating to increasing stage (Nosov *et al.,* 2009). The role of Vitronectin in ovarian cancer progression has been inferred in a cell line study of the effects of its agonist on cell adhesion and motility/ metastatic potential (Beck et al., 2005)

Keratin 10 and lambda chain V-I were chosen as secondary candidates to be investigated at a later stage. Namely CK10, from these results CK10 was chosen to immunohistochemically stain a tissue microarray of ovarian cancers by other researchers however the results are not available to this study.

3.3.3. Validation of the Peaks in the Biomarker Panel

The Figures below show the amount of inferred proteins as measured by immunoassay for each cancer and cancer-control case. The cancer control group show a wider range of Transferrin serum levels which is significantly (p-value = 0.0243) higher in the cancer control group. Figure 17 (below) shows the amount of Vitronectin as measured by immunoassay for the cancer and cancer control group. No significant difference in expression between the two groups is noted (p-value = 0.258)

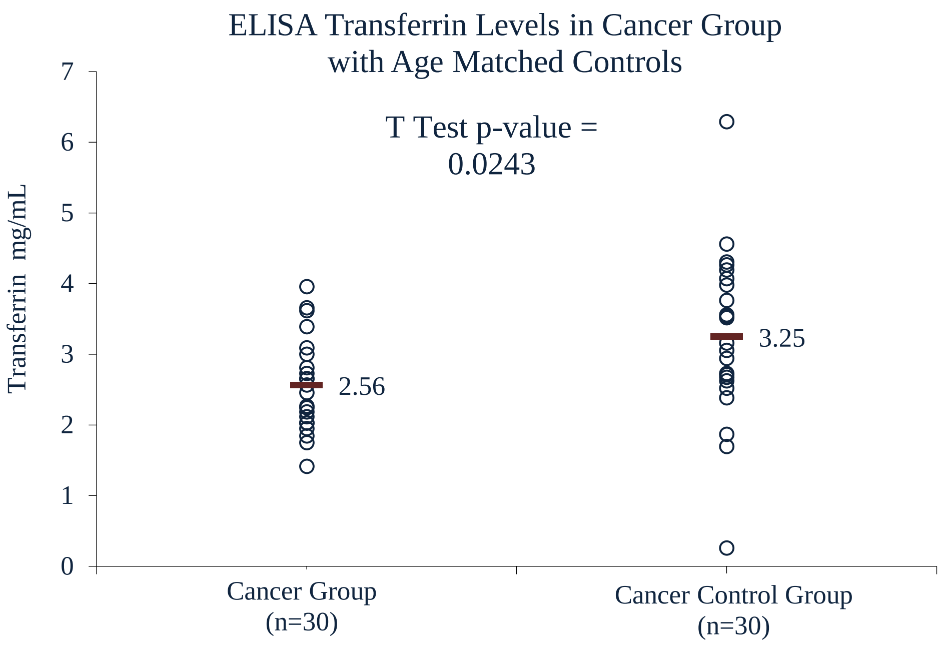


Figure 16. Boxplot of Transferrin Levels as Measured by ELISA.
Transferrin levels of 30 cancer cases and 30 cancer control cases as measured by immunoassay. The average serum concentration of the cancer group is (2.56 mg/mL) is lower than that of the cancer-control group (3.25 mg/mL) with a p-value of 0.0243.

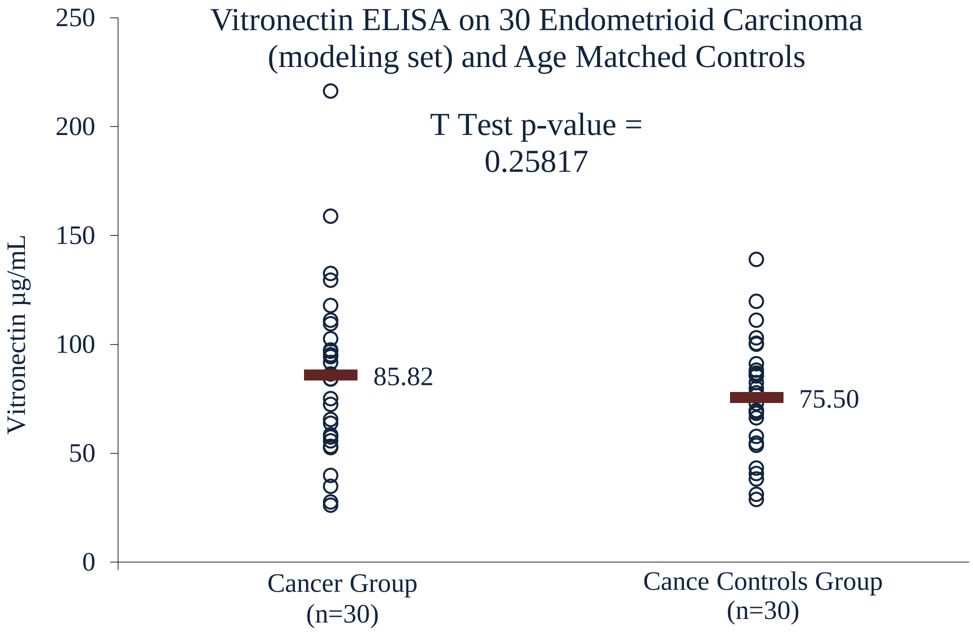


Figure 17. Boxplot of Vitronectin Levels as Measured by ELISA.
Vitronectin levels of 30 cancer cases and cancer control cases as measured by immunoassay. The average serum concentration of the cancer group is 85.8 µg/mL and the control group is 75.5 µg/mL.

The peak intensities of a peak suggested to be Vitronectin were correlated against the immune assay values. Figure 18 (below) shows that no correlation was seen between the level of

Vitronectin as measured by immune assay and peak intensity of the peak suggested to have been Vitronectin. To better demonstrate the poor correlation all 200 serum samples analysed (cancer, cancer control, benign and benign control) are shown, zero values are seen in cases where either no Vitronectin was measured by ELISA (x axis), or, no *m/z* of 1647.8 was detected in the MALDI-TOF spectra (y axis).

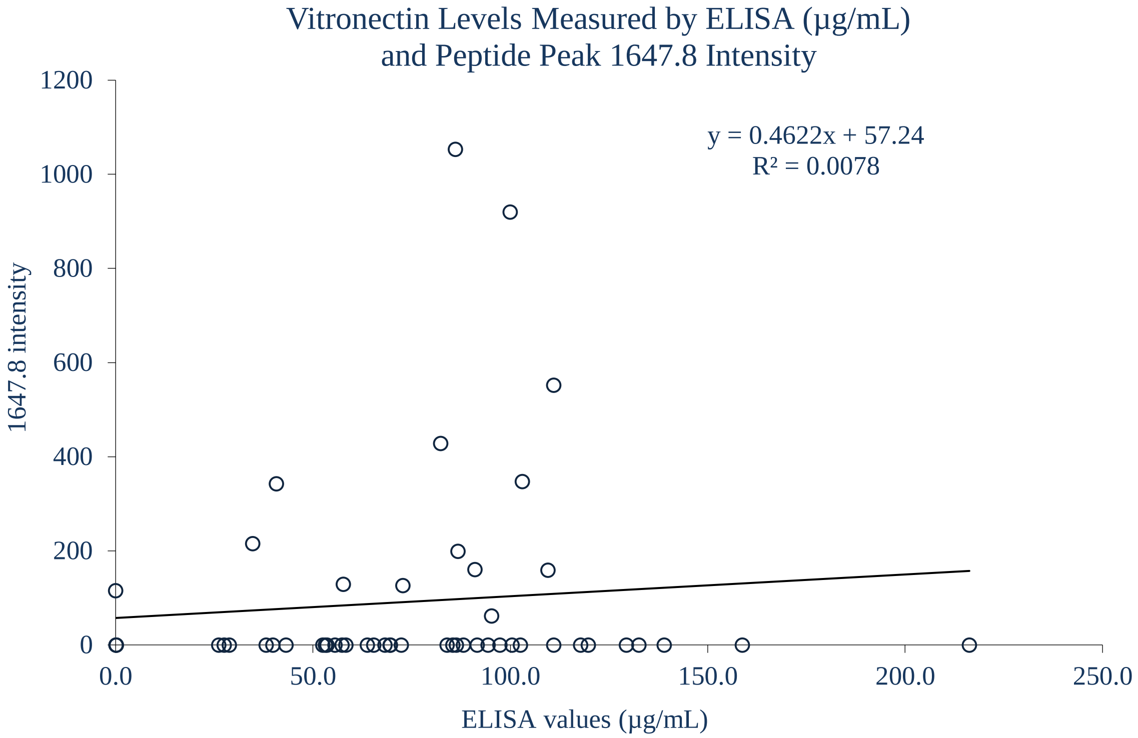


Figure 18. Vitronectin Levels as Measured by ELISA (µg/mL) and Peptide Peak *m/z* 1647.8 Intensity. Correlation of Vitronectin as measured by ELISA against the intensity value of the peptide peak with the suggested identity as Vitronectin.

3.4. Discussion

This chapter aimed to address the need for robust replacements for out-dated low-accuracy diagnostic clinical tools currently used. Despite its popularity, MALDI-MS profiling studies of ovarian cancer patient material have so far not produced any robust biomarker identities to further this field (Cadron *et al.,* 2009, Hays *et al.,* 2010, Timms *et al.,* 2011)

There are instances where the biomarker discovered by MALDI-MS data has been utilised (Yang *et al.,* 2013) and (Timms *et al.,* 2011). Most recently Timms *et al.,* 2011, published a MALDI data profiling study with methodologies similar to those in this report. Combined with CA125 two chemokines improved the prognostic ability of serum analysis. Analysis of peptide yielded peptides which when combined with CA125 improved sensitivity and specificity. Due to the nature of the data identity of the peptides were disputable. They used this identity as a suggestion, and combined with literature and made an inferred identify. Based on the calculated guess a separate platform was used to validate the differential expression of this which worked.

In this study MALDI-MS profiles of serum from cancer patients were acquired and analysed using bioinformatics methodologies. It was demonstrated that a biomarker panel can be generated from MALD-MS data using ANN algorithms with a promising predictive power comparable to that of CA125.

A MLP-ANN used in a stepwise manner was used to generate the biomarker panel. MLP-ANNs have been criticised for their closed, black-box nature, the ANN used to generate the panel was created in-house and conducted under the supervision of one of its creators to ensure its correct use (Lancashire *et al*., 2008)

A biomarker panel of five peptide mass values discriminated cancer from age matched controls with a 91% sensitivity (Table 7). Possible protein identities were assigned to these peptide masses reanalysing a pool of the serum using LC-MALDI-MS/MS and online databases to calculate probable protein sequence information thus protein identities. Immunoassay was performed for the quantitation of Transferrin and Vitronectin, two potential identities of peptide peaks of interest (Figure 16 and Figure 17).

In an ideal scenario, a biomarker panel would be repeated and validated on a different technological platform, and on a different cohort of serum samples from patients with ovarian cancer and age matched controls and the same changes in protein expression observed.

The sensitivity and reproducibility achieved with the biomarker panel generated from the MS data was not reproduced for a series of reasons including

- The heterogeneity of the sample set used
- The lack of a second cohort serum of samples
- The lack of identities of the peptide masses in the biomarker panel
- The allocation of possible identities for a peptide based only on mass the same sample but on different machine or different analyses in the same machine.

3.4.1. Samples

No other cohorts of serum from ovarian cancer patients with aged matched controls, collected in a standardised protocol were available to this group for validation of the biomarker panel produced. Blinding the data and presenting it to the trained ANN is the closest that could be attempted to 'blind validation' of the model. Presenting the same data to the ANN may arguably produce a higher performance. Though it should be noted the results of the validation seen in Table 7 are truly blinded 20%.

Although all samples analysed were selected by a consultant gynaecological oncologist from a biobank collected over three years it was unavoidable to include a variety of stages, grades and histologies in the cancer group. See section 3.2.2.1 sample details. Treating this 'mixed bag' of cancer samples as one group will hinder the detection of biomarkers that that would categorise between the subcategories.

3.4.2. Identification of the Biomarker Panel Proteins

Not all peaks in the biomarker panel were matched to any potential identities. Thus, the full performance of the biomarker panel cannot even be attempted to be repeated using another technological platform. As in other biomarker detection studies using MALDI-MS data (Timms *et al.,* 2011 and Yang *et al.,* 2013), the data served mainly as a guidance, and where a story seemed to match the other literature this was investigated further.

One of the potential identities of the peaks in the biomarker panel was Transferrin. Nosov *et al.,* (2009) published their interest in this protein alongside Transthyretin, Apolipoprotein A-1 and CA125 based on their previous evidence and the hypothesis that all three of these play a role in oxidative stress which links to carcinogenesis (Nossov *et al.,* 2008 and Nossov *et al.,* 2009).

The increased expression of Transferrin in the serum of cancer patients compared to controls produced in this study (Figure 16) did not reflect the trend noted in the literature; where serum Transferrin levels decreased with higher grade of tumour (Nosov *et al.,* 2009). This discrepancy may be explained by the particular subset of samples, it was suggested by Nossov *et al.,* (2008) that these markers were more sensitive to a mucinous histopathological subtype.

The approach of matching peptide mass values from MS MALDI data to LC-MALDI-MS/MS data for an identity was used as it was the best available option, however, is fundamentally flawed.

The exact peptide mass values from the biomarker panel have 4 decimal places and is a result of what the Bruker FlexAnalysis software has recognised to be a monoisotopic peak value and used to export as the raw data. The mass values for the LC-MALDI-MS/MS ‘identity’ run were also generated using the Bruker software suite however on a different instrument in a different laboratory. Some inter and intra instrument variation is to be expected so a very wide window of 1 Da either side of the biomarker was used to generate the possible identities.

Researchers at the time (early 2000s) see Table 3, were less concerned with the identity of biomarkers generated by SELDI and MALDI MS if they were clinically applicable this was all that mattered: “Although knowledge of the identity of a marker is not prerequisite to its utility” Jacobs *et al*. (2004)

Further to this, it is possible that the peptides that were not matched to an identity are novel proteins coded by a cancerous mutation or novel single nucleotide polymorphism (SNP) therefore would not be found by conventional database searching. If a SNP or mutation was suspected the MALDI-TOF-MS/MS data could be searched against an *in-silico* database of hypothetical mutations however, this is not currently available. Moreover, is more likely that the peptides were either not detectable by or not detected on in the MALDI-TOF-MS/MS sample used for the matched identification.

In summary, the inability to identify markers, and the ambiguity of the identities assigned to peptides of the biomarker panel, was a major limitation to this strategy for biomarker discovery.

3.5. Conclusion

In conclusion, the first null hypothesis ($H_{\text{O-i}}$) can be rejected: Unique patterns in protein expression in the sera can be detected by MALD-TOF-MS and ANNs and used to distinguish cancer from control on a blinded validation set.

However insufficient data or evidence leads to the acceptance of the second null hypothesis (H_O-ii): The masses of the peptide peaks expressed differentially in the tested serum cannot be identified by linking data from LC-MALDI-MS with the MS profiles.

Consequently, the third null hypothesis cannot be addressed as the full set of peptides in the biomarker panel were not identified.

Approaching the task with the next generation of technology, where all protein and sequence data is available from the same analysis the biomarker panel is generated from will result in a biomarker panel of peptides of a known identity.

4. Proteomic Evaluation of LC-MALDI-TOF-MS as a Profiling Strategy for Biomarker Discovery in Ovarian Cancer

Chapter Abstract

Spectral features alone, such as the biomarker panel of *m/z* values generated in chapter 3 have limited real world applicability, and are not able to be validated fully without further insight as to what they are. A confident assignment of a protein identity to each data/spectral feature was key to being competitive with current literature.

The work described in this chapter aimed to test and evaluate the capabilities and reproducibility of the (now available) next generation of protein profiling instrumentation LC-MALDI-MS/MS, and upstream sample preparation and fractionation techniques, such as the OFFGEL fractionator and Millipore C_{18} Zip Tips. Several sample preparation protocols were planned to be compared with a view to apply the best performing to the cohort of clinical samples used in chapter 3, of which there were limited stocks.

Multidimensional data was produced from ten replicates of four sample preparation techniques; due to time restrictions, this was not as many as initially outlined. The ten replicate experiments of each of the four methods were compared to assess the overall reproducibility of sample preparation and data acquisition. Of the tested sample preparation methods analysed to completion one was found to consistently yield a higher number of protein identities. Unfortunately, a deeper analysis of the data collected indicated that there was an overall relatively low/poor reproducibility in the data acquisition. This was demonstrated by a retention time shift of the chromatograph however the accuracy is also limited by the experimental design of LC MALDI spotting being an ‘offline’ analysis. It was concluded that, as the capability of the instrumentation had been measured and its limitations now known, a clinical cohort of samples could be analysed in this manner and the resulting data used with the caveat that there is a margin of error of a known size which must be taken into consideration on final analysis. However, this caveat may make it unappealing to apply to a rare sample set of which there is a limited stock.

4.1 Introduction

4.1.1 Need for Identification and Reproducibility of Biomarkers.

The results from chapter 3 as with many studies listed in Table 3 produced results that had potential, if validated, to improve the prognosis of ovarian cancer sufferers. However, these proved to be massively flawed by the lack of conclusive identification of peptide mass values that distinguished cancer from controls.

Previously the identification of the peptide mass values of interest generated from MALDI-TOF-MS data were matched to protein identities from a separate LC-MALDI-TOF data acquisition based on its mass alone (Section 3). LC-MALDI-MS/MS analysis of each test sample individually would provide a three-dimensional peptide map and matched protein identities of peptides present in the sample (as opposed to a two-dimensional MS spectrum with no sequence information), analysis of this data would provide the identity of any proteins found to be differentially expressed between groups. Reproducibility of data and validation of findings needed to be addressed to compete with criticisms in the current literature. Had the technology been available, the samples for biomarker discovery could have all been analysed via LC-MALDI-TOF-MS/MS. The mass values that distinguished cancer from control with protein identities assigned to them could have been generated at this stage, thus more confidence in their identity, and likelihood of validation. When LC-MALDI-TOF technology became available to the project, the ovarian cancer cohort could be reanalysed. Due to limited sample, optimisation of the sample preparation protocols was needed to ensure the samples were analysed under the conditions that would produce the maximum protein identities.

To maximise the information recorded from each test sample, experiments were conducted to test the capabilities of the available Bruker LC-MALDI-MS platform.

Different sample preparation and data extraction methods were tested to uncover which produced the highest number of proteins identified with a high confidence.

4.1.2. Liquid Chromatography

Liquid chromatography (LC) is used to fractionate molecules within a sample based on their affinity to bind to an analytical column (as described in section 2.1.5.). A complex sample is bound to an analytical column and eluted from it at increasing concentrations of a solvent; the mobile phase. For example, using LC, proteins within a serum sample can be bound to then separate from a C_{18} column based on their affinity to bind to C_{18} silica in the presence of the solvent acetonitrile. Periodical fractions of the eluent from the column can be collected and analysed at a greater detail than the original sample. LC systems are often coupled to an electrospray mass spectrometer for a continuous mass spectral analysis of the sample as it is eluted at an increasing concentration from the column.

4.1.3. The need to Address Reproducibility to Progress with the Research in the Area.

Publications based on MALDI data are heavily criticised on the reproducibility of the data and the lack of the identification of the peptide peaks of interest.
It is also seen (Cadron *et al.,* 2009) that the majority of the identities that have been implied are high abundant proteins, or already known acute phase reactants (Diamandis 2004). In theory, a serum protein released by or in response to a tumour would be found at concentrations orders of magnitude lower than these.

4.1.4. Sample Fractionation Techniques.

Removal of the highly abundant proteins, or separation of the proteome on a separate dimension, can be conducted to allow access to the lower abundant proteome (Margulies and Shevack 1996). Additionally, fractionating the proteome on a third dimension and analyses of each separate fraction decreases the complexity of the sample and increases the chance of access to proteins of lower abundance. Two different examples of fractionation techniques include.

- Immuno-depletion proteome (Margulies and Shevack 1996). Relevantly to serum proteomics is the Sigma ProteoPrep20 immuno-depletion column (Sigma-Aldrich 2012); Sera to be analysed is run through a column containing antibodies to the 20 most highly abundant proteins in serum. The abundance proteins are captured and removed or analysed separately. This process removes 90-95% of the total protein from the sample (Sigma-Aldrich 2012).

- Isoelectric focusing. The OFFGEL fractionator 3100 (Agilent Technologies 2010) applies an electric current to the sample loaded. Proteins or peptides within the sample are fractioned into 12 or 24 portions based on their isoelectric point (pI).

Analysed data acquired from each fraction can be compiled post-acquisition to piece together a proteome at an increased resolution than a non-fractionated sample.

4.1.4.1. Millipore Zip Tips.

A ZipTip, or solid phase extraction in a pipette tip (see section 2.1.5 for solid phase extraction), is one such method of sample purification; they are relatively cheap means for liquid chromatography used with an isocratic mobile phase. Millipore Zip Tips are a 10 µL pipette tip with a bed of chromatography media fixed at its end. Drawing in and aspirating a sample through a ZipTip will remove salts and detergents which, due to their charged nature, can hinder spectral quality by increasing signal to noise ratio: biomolecules in the sample bind to the immobilised absorbent resin inside each tip, damaging salts and detergents are washed away resulting in de-salted, purified and concentrated sample. C_{18} and C_4 are two types of ZipTip that can be used depending on the nature of the analyte (Millipore Corporation 2005).

4.1.4.2. Alkylation and Reduction.

Proteins are commonly reduced and alkylated prior to analysis.

Breaking disulphide bonds relaxes and linearises the 3D structure of a protein (Sechi and Chait 1998, Hale *et al.,* 2004 and Wedemeyer *et al.,* 2000). Disulphide bonds between cysteine residues are key to the stability of tertiary structure of proteins (Wedemeyer *et al.,* 2000). Reductive unfolding, is the loss of protein tertiary structure due to the chemical reduction of these bonds (Wedemeyer *et al.,* 2000). The resulting exposed sulphydryl groups are highly reactive so an alkylating agent is often added to oxidise and stop unwanted reactions or refolding within the protein (Hale *et al.,* 2004). In gel-based techniques such as in 1 or 2-Dimensional Polyacrylamide Gel Electrophoresis (1 or 2D-PAGE) this standardises the movement of proteins through medium increasing the resolution of bands or spots (Sechi and Chait 1998). Additionally, the linearisation or the protein increases access of proteolytic enzymes to digestion cleavage sites prerequisite for protein identification from mass spectrometric peptide mass fingerprinting. Improving digestion efficiency should result in more peptides cleaved per protein, and a cleaner mass spectrum with better resolved peaks for optimal protein identification through database matching (Hale *et al.,* 2004).

To be able to publish in the area, and more importantly for confidence in the reproducibility in the data, repeated measurements of the capabilities of the instrumentation, sample preparation, and methods used was a necessary step.

To do this, a large stock of human serum used for regular quality control was analysed multiple times through different sample preparation workflows and data acquisition and extraction to discover the optimal rout for the future analysis of the serum from ovarian cancer patients with controls.

Figure 19 (below) depicts 40 replicates of one sample being analysed via four protocols/ workflows (10 in each). Some requisite procedures are consistent across all 40 replicates i.e. thawing and dilution or tryptic digestion. Other potentially optional steps i.e. alkylation and reduction or pre-digestion C_{18} ZT are conducted on 10 replicates and 10 replicates of a control condition which was identical barring the optional procedure was shown next to it.

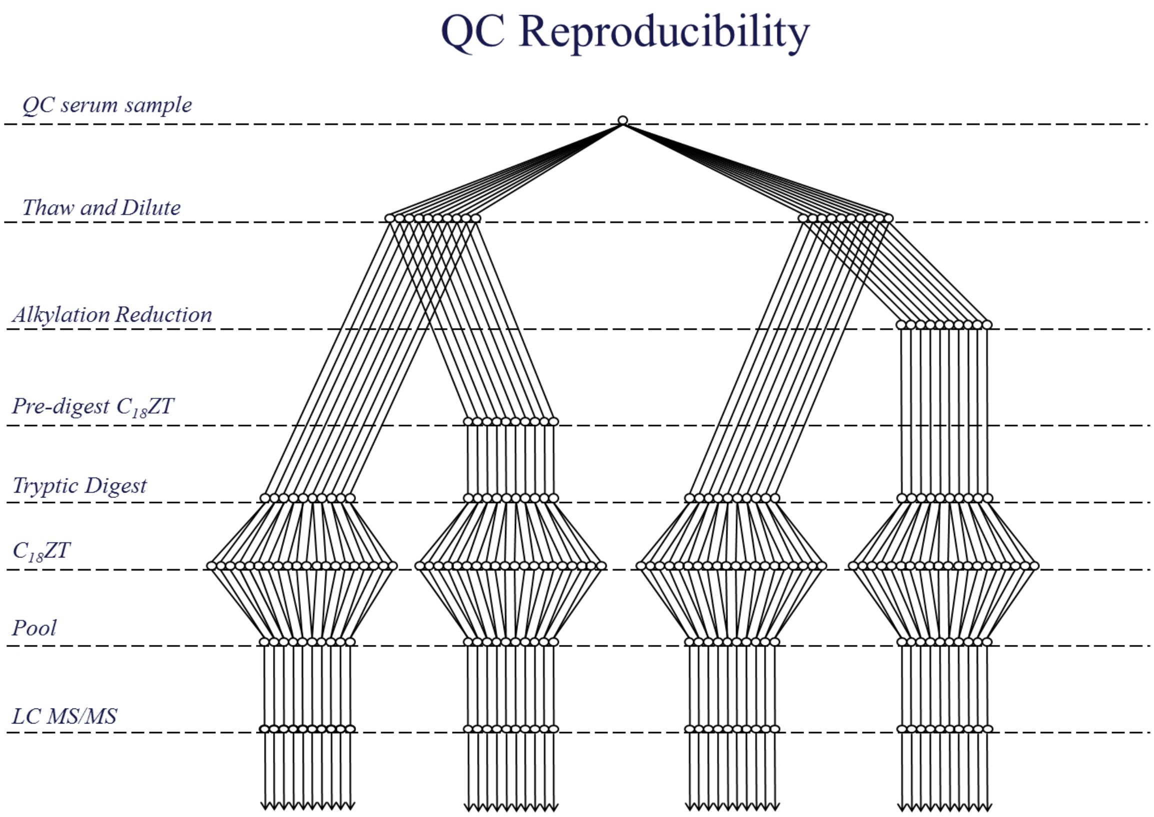


Figure 19. Multiple Analysis of one Serum Sample via Different Sample Preparation Workflows. One serum sample used for quality control seen on the top level was analysed 10 times four each of four workflows. Each line and circle represents one replicate.

4.1.5. Aims and Hypotheses of the Chapter

To assess the quantitative power of MALDI-TOF-MS data.

*H*0-iv: There is no correlation between protein amount loaded for detection and the signal intensity of protein detected.

H_1–iv: The signal intensity values of a detected protein is relative to the amount of protein loaded.

Optimise Sample preparation; identify a satisfactorily reproducible combination of sample fractionation and preparation techniques to maximise the number of meaningful protein identities from one serum sample using MALDI-TOF MS.

*H*0-v: All tested sample preparation techniques prior to LC-MALDI-MS produce equal amounts of meaningful protein identities.

H_1–v: One sample preparation technique prior to LC-MALDI-MS will yield a greater amount of meaningful protein identities.

Method validation. Assess reproducibility and robustness of above protein mapping methods.

To assess the validity of using LC-MALDI-MS data to find differences in protein expression in serum.

H_0–vi: There will be no significant difference between the LC-MALDI-MS profiles of serum samples prepared under identical conditions.

H_1–vi: Differences will be seen in the LC-MALDI profiles of serum samples prepared under identical conditions.

4.2. Materials and Methods

4.2.1. Materials

4.2.1.1. Equipment used

Table 9. Equipment Utilised for LC-MALDI-TOF-MSMS

Equipment	Supplier
Bruker UltrafleXtreme Matrix Assisted Laser Desorption Ionisation-Time of Flight Mass Spectrometer (MALDI-TOF)	Bruker Daltonics
Nano-HPLC-Protineer fc-II Target Spotter	Bruker Proteineer
Desktop computers with use of SpecAlign, Excel, Statistica and Bruker Software which comprises of: FlexControl 3.3, FlexAnalysis 1.3, Profile Analysis 1.1, Biotools 3.2, ClinProTools 2.2 software controlled by WarpLC 1.2 as part of the Compass Series 1.3.	Bruker Daltonics
384 spot MALDI-TOF targets; Grounds Steel 384, Anchor Chip 384 and PAC 384	Bruker Daltonics
Automated pipetting machine	FluidX
Sonicator	VWR Ultra Sonic Cleaner
-80˚C freezer	New Brunswick
37˚C Incubator	Heraeus
Vortex mixer	Scientific Industries
10 µL Millipore ZipTips C_{18} pipette tips	Millipore
Polypropylene bottles for reagent storage	Nalgene F.E.P.
1.5 mL Eppendorfs	Eppendorf
1 mL pipette	Gilson
200 mL pipette	Gilson
2 mL pipette	Gilson
0.2-1 mL tips	Gilson
20-200 µL tips	Gilson
0.5-10 µL tips	Gilson

4.2.1.2. Reagents used:

Table 10. Reagents Utilised for LC-MALDI-TOF-MSMS

Reagent	Supplier	Grade
Acetone	Sigma	LC MS
Acetonitrile (ACN)	Sigma	LC MS
Alpha-Cyano-4-hydroxycinnamic acid (CHCA)	Bruker Daltonics	MALDI-TOF-MS
Ammonium bicarbonate	Sigma	Laboratory
Distilled water	Barnstead Diamond	nano pure
Methanol	Sigma	LC MS
Peptide Calibrant II Bruker	Bruker Daltonics	MS
Trifluoroacetic acid (TFA)	Fisher Scientific	HPLC
Trypsin – Trypsin Gold, Mass Spectrometry Grade	Promega	MS
Tryptic Digest of Bovine Serum Albumin	Bruker Daltonics	MS

4.2.1.3. Stock Solutions Made and Used

Table 11. All stock solutions made and used for LC-MALDI-TOF-MSMS

Reagent	Composition
0.1% TFA in H_2O (50 mL)	50 µL TFA 49.95 mL – Distilled water
0.1% TFA in Acetonitrile (50 mL)	49.95 mL – Acetonitrile 50 µL TFA
80% ACN diluted with 0.1%TFA (50 mL)	40 mL – ACN 10 mL - 0.1% TFA solution
Trypsin solution 0.5 mg/mL (200 uL)	100 mg – Trypsin Gold, Mass Spectrometry Grade 200 mL – 100 mM ammonium bicarbonate NH_4HCO_3
Iodoacetamide 200 mM (1 mL)	36 mg Iodoacetamide 1 mL of 50 mM ammonium bicarbonate
Dithiothreitol (DTT) 200 mM (1 mL)	30 mg DTT 1 mL 50 mM ammonium bicarbonate

4.2.2. Methods

4.2.2.1. Production of the BSA Standard Curve.

The following dilutions (Table 12) were made from an unused tube of Bruker Tryptic Digest of Bovine Serum Albumin (125 µL was added to the new vial containing 500 pMol to create a 4 pMol/µL stock.)

Table 12. Dilutions made for production of a BSA standard curve.

Volume of 0.1%TFA (µL)	Volume of BSA	BSA (fMol/µL)
10	0	0
195	5uL of (4 pMol/µL)	100
5	5 µL (100 fMol/µL)	50
6	4 µL (100 fMol/µL)	40
7	3 µL (100 fMol/µL)	30
8	2 µL (100 fMol/µL)	20
9	1 µL (100 fMol/µL)	10
9.5	0.5 µL (100 fMol/µL)	5
10	0	0

0.5 µL of each dilution were manually spotted to a Bruker PAC target. One of each replicate were used for optimisation of the mass spectrometer parameters i.e. laser power and detector sensitivity then the remaining seven were fired on under identical parameters using an automated data acquisition function within the Bruker FlexControl software. Bruker FlexAnalysis software was used to export the intensities of the calculated monoisotopic peaks present in the sample.

4.2.2.2. QC Serum Sample

QC serum sample was collected in November 2008. Briefly; 50 mL of vein blood was taken from 4 volunteers; 2 female and 2 male fully informed and consenting volunteers. The blood was left to clot for between 30 min and one hour at room temperature until clotted then processed under contamination level 2 conditions. Firstly, the bloods were centrifuged at 22°C for 15 min and at 2000 x rcf, then the serum supernatants were collected with Pasteur pipettes and pooled into one sterile container and gently agitated at 4°C while aliquots were made. Three thousand 30 µL aliquots were made into sterile 0.5 mL micro tubes. Aliquots were stored at -80°C for future use.

Following standard practice for serum samples in the John van Geest Cancer Research Centre proteomics lab, serum samples are used for no more than three freeze thaw cycles before they are discarded.

Later protein assay revealed the QC serum concentration to be 70 mg/mL; (1.061 µL contains 1 mg of protein).

4.2.2.3. Multiple Workflows Tested for Optimisation

Although more workflows were originally planned and would have been insightful, data was collected for 10 replicates of each of four workflows. As depicted in Figure 19 above and Table 13 below, multiple replicates of one test serum sample were tested under each condition.

Table 13. Summary of the Workflows Applied to Replicates of one Test Sample.

	Sample Preparation Condition 1: Pre-digestion ZipTip	Controls run in Parallel to Condition 1	Sample Preparation Condition 2: Alkylation and Reduction	Controls run in Parallel to Condition 2
Sample Number	n=10	n=10	n=10	n=10
Samples were digested prior to digestion	**Yes**	No	No	No
Samples Reduced and Alkylated	No	No	**Yes**	No

See sections 4.2.2.5 and 4.2.2.6 below for the protocols used in each condition.

4.2.2.5 Alkylation and Reduction of Sera

The following protocol was adapted from an online source (http://www.ocbn.ca/insolution.htm 2012). Urea was omitted for compatibility with MALDI-TOF-MS.

5 µL of DTT was added to 1.06 µL of QC serum (containing 1 mg total protein) diluted in 100 µL 50 mM ammonium bicarbonate and the sample thoroughly vortexed. The sample was wrapped in foil to protect from light and incubated at 37°C. After one hour precisely, 20 µL of

200 mM iodoacetamide was added and the sample was again vortexed. 20 µL of 200 mM DTT added and the sample vortexed and left at room temperature for one hour. 10 µL Trypsin (0.5 mg/mL in 100 mM ammonium bicarbonate) was added then the sample incubated at 37°C overnight (18 h). Following the digestion, digests were refined using Millipore C_{18} ZipTips: (see3.1.4). Elutes were diluted in 20 µL 0.1%TFA and chromatographically separated and spotted to Bruker MALDI-TOF targets using the Bruker nLC. Data was acquired using the Bruker UltrafleXtreme

See Table 13 and Figure 20. Flow chart of alkylation and reduction of QC sera and controls.

4.2.2.5.1 Controls for Alkylation and Reduction Protocol

1.06 µL QC serum (containing 1 mg total protein) was diluted in 21.2 µL 0.1% TFA timed to match dilution time of the paired test sample (section 4.2.2.5). The sample was vortexed and kept at 4°C for the duration of the additional steps for alyklation and reduction in section 4.2.2.5. 10ul Trypsin (0.5 mg/mL in 100 mM ammonium bicarbonate) added at the same time as the paired test sample (section 4.2.2.5) and incubated at 37°C overnight (18 h). Digests were refined using Millipore C_{18} ZipTips: see 3.1.4. Elutes were diluted in 20 µL 0.1%TFA and chromatographically separated and spotted to Bruker MALDI-TOF targets using the Bruker nLC. Data was acquired using the Bruker UltrafleXtreme in as close as possible time to the paired test sample (section 4.2.2.5)

See Table 13 and Figure 20. Flow chart of alkylation and reduction of QC sera and controls.

4.2.2.6 Pre-digestion C_{18} ZipTip of Sera

2 µL QC serum diluted in 38 µL 0.1%TFA and vortexed. 10 µL was removed and used for control (see 3.1.6.2), the remaining 30 µL was refined using Millipore C_{18} ZipTip: See 3.1.4 producing 4 µL of refined proteins in 4 µL of 80% ACN in 0.1% TFA.
The following reagents were added to the 4 µL eluate: 7.6 µL HPLC grade water, 16.6 µL 100 mM ammonium bicarbonate and 1 µL of 0.5 mg/mL Promega trypsin gold in 100 mM ammonium bicarbonate and the sample vortexed. Samples were incubated overnight at 37°C. Following this the digests were refined using Millipore C_{18} ZipTips: (see .4.1.4).

The resulting elutes were diluted in 20 µL 0.1%TFA and chromatographically separated and spotted to Bruker MALDI-TOF PAC targets using the Bruker nLC and data was acquired using the Bruker UltrafleXtreme.

See Table 13 and Figure 21. Flow chart of pre-digestion ZipTip of sera versus controls

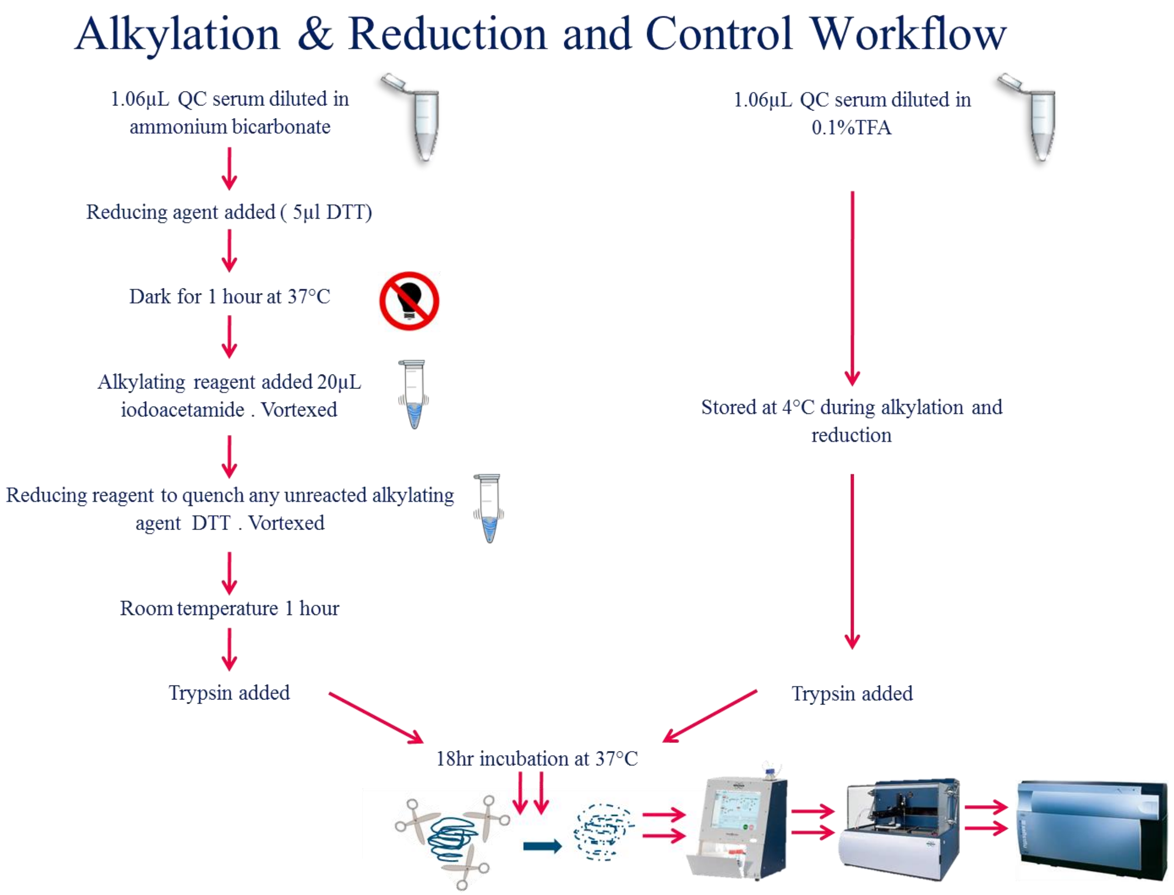
4.2.2.6.1 Controls Pre-digestion C_{18} ZipTip

Firstly 10 µL of a one in 20 dilution of QC serum was prepared; see 3.1.6. After the paired test sample (3.1.6) had been ZipTipped, 7.6 µL HPLC grade water, 16.6 µL 100 mM ammonium bicarbonate and 1 µL of 0.5 mg/mL Promega trypsin gold in 100 mM ammonium bicarbonate were added and the sample vortexed. Samples were incubated overnight at 37°C. The digests were refined using Millipore C_{18} ZipTips (see 4.1.4). The resulting elutes were diluted in 20 µL 0.1%TFA and chromatographically separated and spotted to Bruker MALDI-TOF PAC targets using the Bruker nLC and data was acquired using the Bruker UltrafleXtreme in as close as possible time to the paired test sample.

See Table 13 Figure 21. Flow chart of pre-digestion of sera versus controls

4.2.2.7. Paired Comparisons

To produce LC-MALDI-MS/MS data up to four samples could be analysed each day, this averaged at two per day. To avoid day-to-day variations such as lab temperature affecting the results, each of the 10 replicate test samples were processed in tandem with a control. See Figure 20 and Figure 21 below.



Alkylation & Reduction and Control Workflow

1.06µL QC serum diluted in ammonium bicarbonate

Reducing agent added (5µl DTT)

Dark for 1 hour at 37°C

Alkylating reagent added 20µL iodoacetamide . Vortexed

Reducing reagent to quench any unreacted alkylating agent DTT . Vortexed

Room temperature 1 hour

Trypsin added

1.06µL QC serum diluted in 0.1%TFA

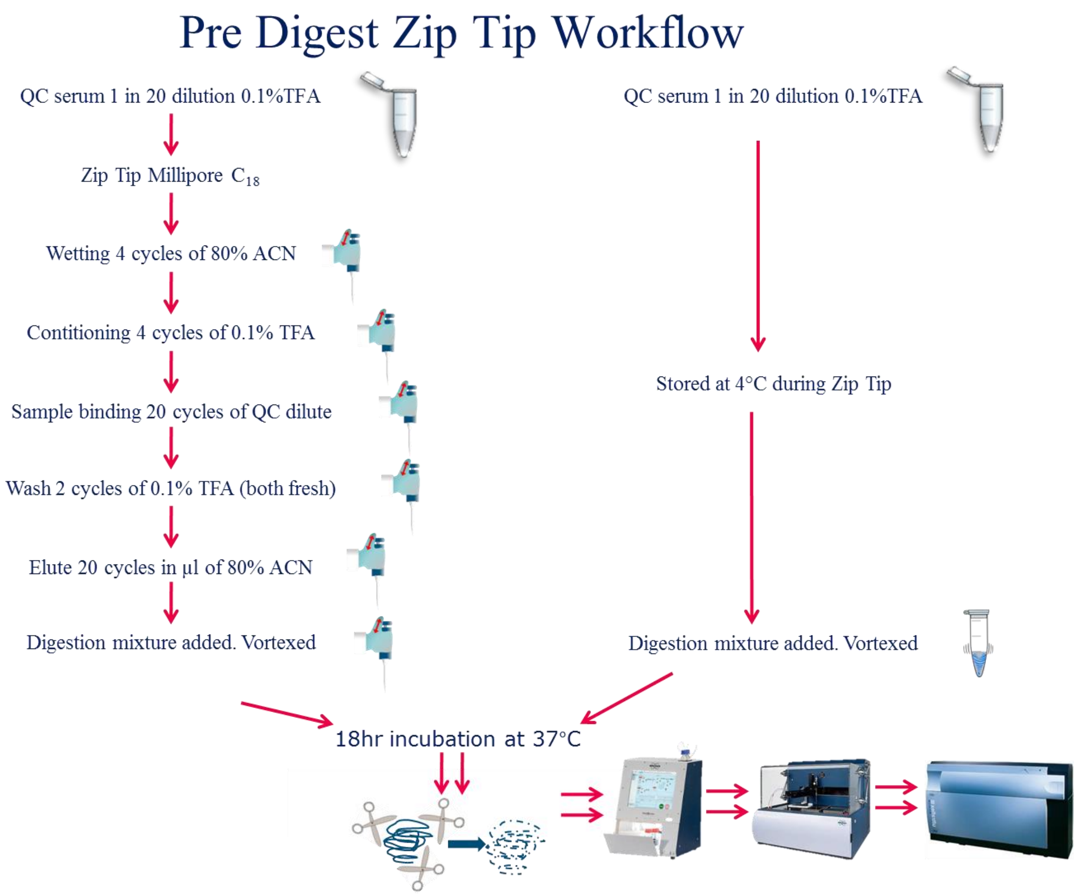
Stored at 4°C during alkylation and reduction

Trypsin added

18hr incubation at 37°C

Figure 20. Flow Chart of Alkylation and Reduction of QC Sera and Control Workflow.
Control and test condition 2 serum samples defrosted, processed and analysed together to minimise external variation.

Each pair were run alternately to negate batch effect i.e. one test sample, one control one test, one control.



Pre Digest Zip Tip Workflow
QC serum 1 in 20 dilution 0.1%TFA
Zip Tip Millipore C_{18}
Wetting 4 cycles of 80% ACN
Contitioning 4 cycles of 0.1% TFA
Sample binding 20 cycles of QC dilute
Wash 2 cycles of 0.1% TFA (both fresh)
Elute 20 cycles in µl of 80% ACN
Digestion mixture added. Vortexed
QC serum 1 in 20 dilution 0.1%TFA
Stored at 4°C during Zip Tip
Digestion mixture added. Vortexed
18hr incubation at 37°C

Figure 21. Flow Chart of Pre-digestion Zip-tip QC Sera and Control Workflow.
Control and test condition 1 serum samples defrosted, processed and analysed together to minimise external variation.

4.2.2.8. A Model for use on Clinical Cohort of Samples

To assess the applicability of the experimental and data-processing set-up ready for application to a clinical cohort of samples, the data from the 10 alkylated and reduced samples with the 10 controls run in parallel were exported from the software and compared. Differences between the groups were investigated in two ways

- Comparison of the protein lists acquired – a qualitative assessment.
- Comparison using Bruker software (see 4.2.2.8.2) – a semi quantitative assessment

4.2.2.8.1 Comparison of the Protein Lists Acquired

The full lists for each of the 20 samples (10 alkylated and reduced and 10 controls) calculated peptide sequences and protein identities were exported from the Bruker WARP-LC and Protein

Viewer software to Excel. Basic Excel functions such as Sort, IF and PiVot-Tables were used to compare the presence or absence of proteins across and between replicates. This is a qualitative analysis as no values are assigned to the protein identities only their presence or absence counted.

4.2.2.8.2. Comparison using Bruker Software

Bruker Profile Analysis software was used to semi-quantitatively compare the data of the ten alkylated and reduced replicates with the 10 controls run in parallel as a model for use on sample with clinical disease and controls.

Profile Analysis software aligns, collates and compares the retention times and mass values to produce bins of data which, to the best of its ability, aims to separate out each peptide peak based on the two dimensions; mass and retention time. Full description of the method of this calculation can be found in (Bruker Profile Analysis User Manual 2.0.).

The Scheduled Precursor List (SPL) is an account of peptide peaks detected and their alignment to a retention time. The bins, referred to as ‘buckets’ are of varied size depending on what it has calculated to be one peak/peptide. It is calculated within the Bruker software and contains the values from the bucket list and the values of error tolerances and the size of the windows and boundaries of the calculated ‘buckets’. The intensity values of the buckets are compared using an MS-T-test within the software to list the buckets in order of the difference in intensity between the two groups.

With great difficulty, the SPL list, the bucket table, and the MS T-test were exported from the software into Excel using Adobe Acrobat Reader and manually checked in word and Excel for further analysis.

The peak *m/z* values from buckets whose intensity values were found through T-test to be significantly differentially expressed between the two groups (p-value of >0.05 and >= 2-fold change) were searched against the peptide sequence/ protein identity tables to match protein identity information.

The SPL list was searched to find the time shift tolerance calculated by the Bruker software.

PiVot Tables in Excel were used to further investigate the occurrence of proteins identified by T-test to be differentially expressed between the two groups. These and the identities of the significant buckets were compared with that of the qualitative analysis as described in section 4.2.2.8.1.

4.3. Results

4.3.1. Semi-quantitative Nature of MALDI-TOF-MS

As a proof-of-principal, to asses if MALDI-TOF data can be treated as semi-quantitative a standard sample of BSA digest was diluted at several concentrations, spotted to a MALDI-TOF target plate and data acquired. Flex Analysis software was used to view spectra (Figure 22) and export the numerical monoisotopic peak values to Excel to plot a standard curve (Figure 23).

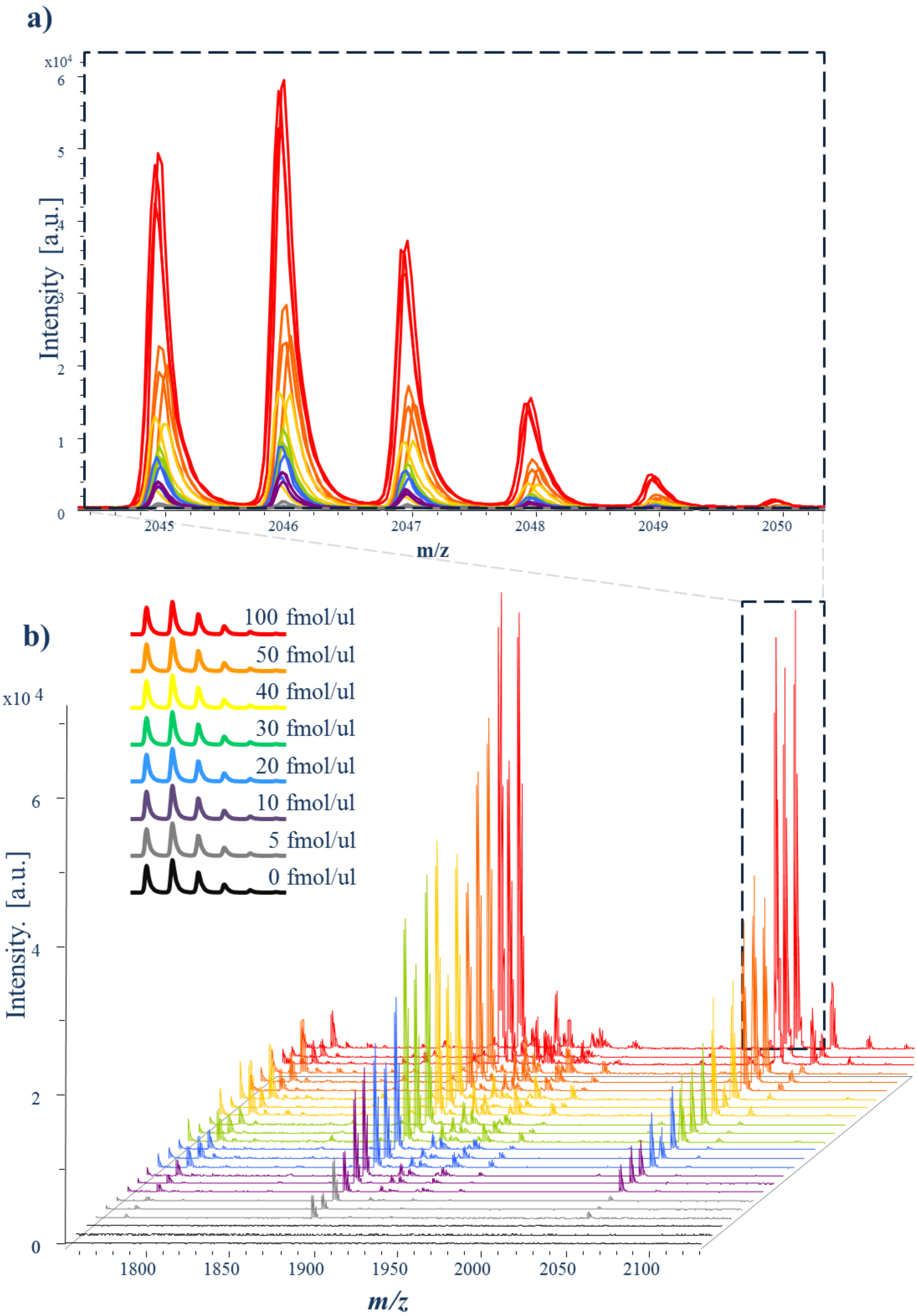


Figure 22. Increase of Peak Intensity with Increase in Concentration of Sample Loaded from a BSA digest. 20a) A monoisotopic distribution of a peptide mass 2044.9 is exported from FlexAnalysis software; mass (*m/z* ratio on the x axis) is plotted against the intensity (arbitrary units on the y axis). The trace in red depicts a sample loaded at 100 fMol/µL; orange 50 fMol/µL, yellow 40 fMol/µL, green 30 fMol/µL, light blue 20 fMol/µL, indigo 10 fMol/µL, grey 5 fMol/µL and black 0 fMol/µL. A stacked view of the same spectra are seen in 8b in visual range of neighbouring peaks.

A standard curve of BSA concentration was plotted against the intensity of one BSA peptide peak measured in the spectra. Linear line of best fit has a R^2 of 0.9567.

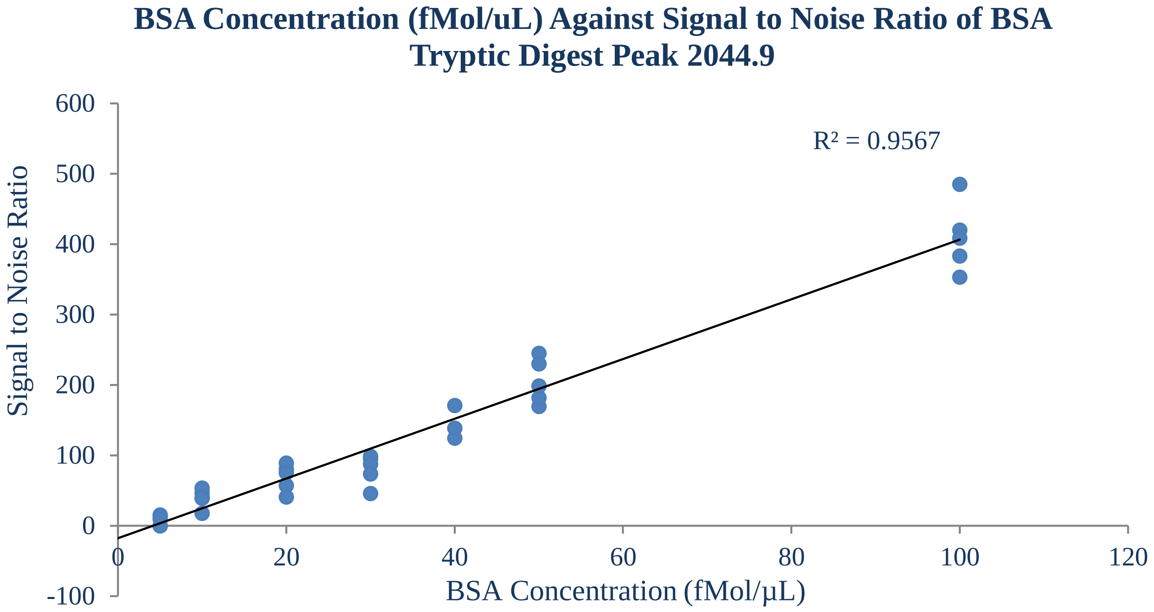


Figure 23. BSA Standard Curve from Peak Intensity.
BSA Concentration (fMol/µL) of the sample loaded on the x axis is plotted against the signal intensity (arbitrary units) detected from the *m/z* value 2044.9 a known BSA digest peak as measured by MALDI-TOF-MS on the y axis. The linear regression curve fitted has a R^2 value of 0.9567.

4.3.2. Reproducibility of the Third Dimension; Retention Time of the Analytical Column

To measure the shift variability of the retention time, the total number of peptides that had eluted from the analytical column were plotted against retention time for all 10 replicates or each workflow. The time at which 50% of the total compounds had eluted was used to assess the variability of retention time. As an example Figure 24 (below) includes the 10 replicates of the controls run in tandem with the alkylated and reduced samples.

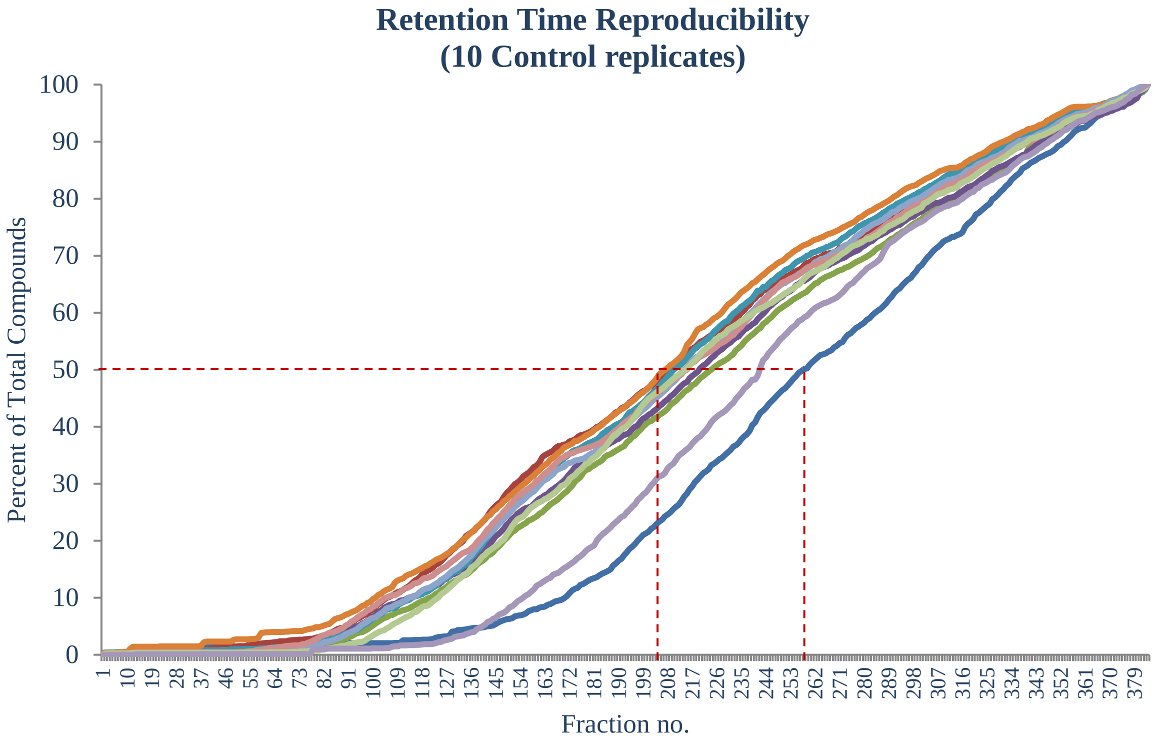


Figure 24. Retention Time Reproducibility.
Total number of eluted peptides from the column is plotted against retention time/fraction (spot over a 384-spotted target plate) for 10 replicates of controls run in parallel with the alkylated and reduced workflow (condition 2 controls, Table 13). Percent of the total compounds eluted is on the y axis plotted against retention time on the x axis Retention time is represented in fractions 10 seconds apart. The retention time/fraction at which 50% of the peptides had been eluted from the column were compared.

For the 10 replicates from the control samples run in parallel with the alkylated and reduced samples (seen in Figure 24). The time/fraction at which 50% of the sample loaded had been eluted had a range of 500 s from 2080 s to 2580 s with a mean average of 2221 s, median and mode of 2160 s and a standard deviation of 159.5 s.

See section 4.3.5.1. for Bruker Software calculations of retention time shift. In summary, retention time windows calculated to correspond to each peptide peak ranged from 0 to 1333 s (22.13 min) with an average of 421 s (7.01 min) preceding the peak and 534 s (8.54 min) following the peak.

4.3.3. Reproducibility of Identities Acquired using LC-MALDI-TOF-MS/MS

The full list of peptides identified from each run for all replicates were exported to Excel using Bruker WARP-LC and ProteinViewer software. For an explanation of the power model the results from the 10 alkylated and reduced replicates are presented in a histogram below; Figure 25; wherein the occurrences of the accession codes across each replicate were examined. A larger proportion of the identities only occurred in one replicate. Based on the distribution noted a power model was used. If a peptide identity occurred in 80% or more replicates it was considered true (Figure 25 solid colour bars). Identities occurring in 70% or less replicates were not considered reproducible and excluded from further analysis (Figure 25 faded bars).

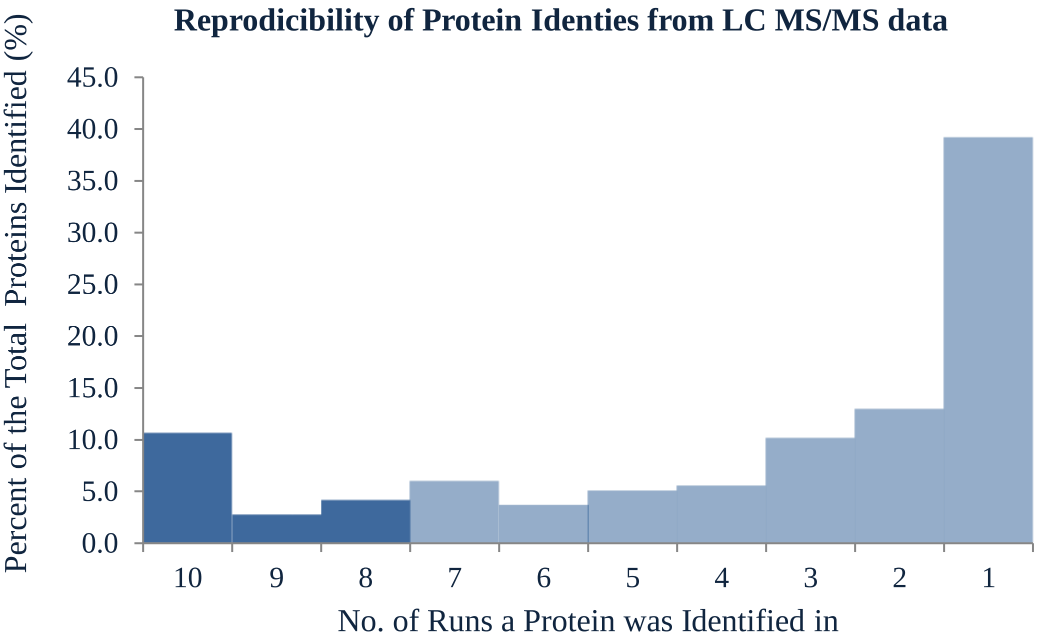


Figure 25. Histogram of Protein Identity Occurrence.
The number of times a protein identity occurred out of the 10 replicates of the alkylated and reduced workflow (condition 2, Table 13) is represented. The largest proportions of identities only occur in one of the replicate. Identities that occurred in 7 or less replicates were regarded as non-reproducible and not included in further analysis (faded bars). The identities that occurred in 8 or more replicates (bold bars) were regarded as reproducible and included in further analysis.

The power model exemplified here on the alkylated and reduced samples was applied to all four workflows, cross-reference and displayed in Venn diagrams below. The application of the power model can be observed within the concentric circles of each workflow. Protein identities not meeting the 80% power model criteria are represented by the seven faded bars above are in the outer seven grey circles. Protein identified considered reproducible represented in the three non-faded bars above are in the central three coloured circles below.

4.3.4. Comparison of the Proteins Identified from each Workflow

The power model as described above was applied to the test samples.

The 10 replicates from the alkylation and reduction workflow detailed in methods section 4.2.2.5 resulted in 38 reproducible identities; the control replicates run in parallel with the pre-digestion ZipTip samples described in section 4.2.2.6. produced 74 reproducible identities (Figure 26). Four identities were only reproducibly identified in the samples that were alkylated and reduced, 40 were reproducibly identified in only the control group and 34 identities were reproducibly identified from both workflows.

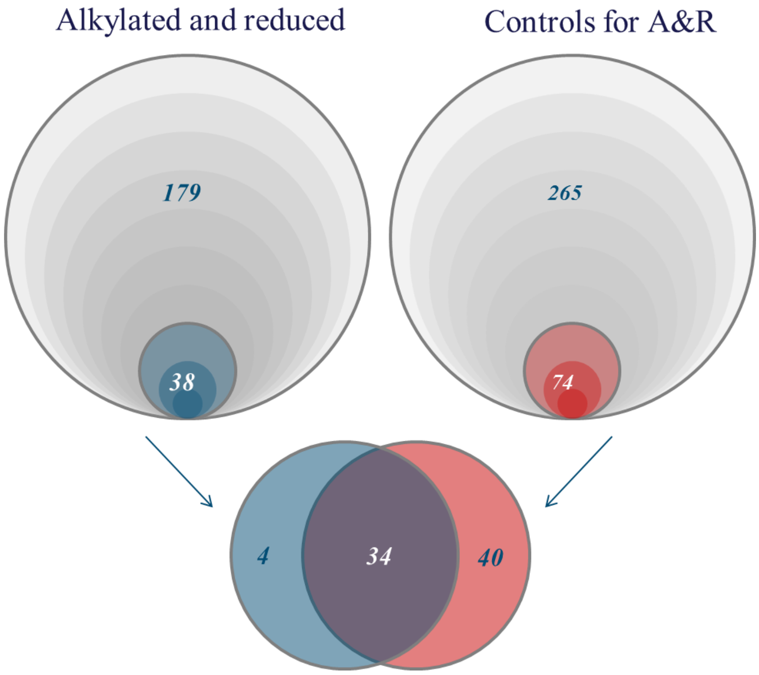


Figure 26. Venn Diagram Comparing the Lists Protein Identities Acquired from the Alkylation and Reduction Sample Preparation Workflow and Control.
The 10 replicates that were alkylated and reduced before digestion produced 217 protein identities; 38 of which occurred in 80% or more replicates and considered reproducible (blue), the remaining 179 occurred in 70% or less replicates and were disregarded from further analysis (grey). For the control samples run in parallel with the pre-digested alkylated and reduced workflow 339 protein identities were acquired; 74 occurred in 80% or more of the replicates (red), the remaining 265 only occurred in 70% or less of the replicates and were removed from further analysis (grey). When the reproducible identities acquired from alkylated and reduced group were cross compared with the controls run in parallel. 4 identities are unique to the alkylated and reduced group, 40 are unique to the controls and 34 are reproducibly found in both.

The 10 replicates from the pre-digestion ZipTip work flow detailed in methods section 4.2.2.6 resulted in 36 reproducible identities; the control replicates run in parallel with the pre-digestion ZipTip samples described in section 4.2.2.6.1. produced 58 reproducible identities (Figure 27).

Four were unique to the pre-digestion ZipTip group, 26 were uniquely found in the control group and 32 were reproducibly identified from both sample preparation workflows.

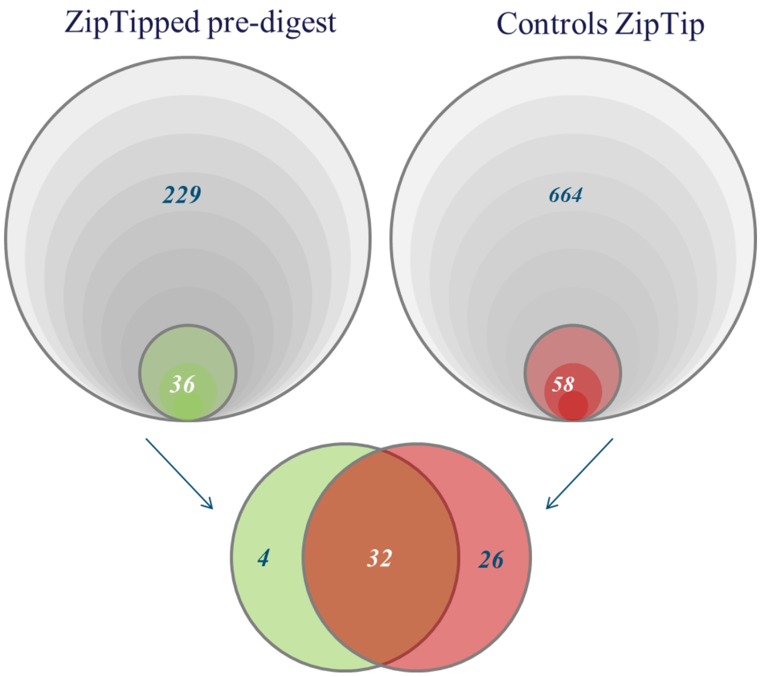


Figure 27. Venn Diagram Comparing the Lists of Protein Identities Acquired from the Pre-digestion ZipTip Sample Preparation Workflow and Control.
The 10 replicates that were ZipTipped before digestion produced 265 protein identities; 36 of which occurred in 80% or more replicates and considered reproducible (green), the remaining 229 occurred in 70% or less replicates and were disregarded from further analysis (grey). For the control samples run in parallel with the pre-digested ZipTip workflow 722 protein identities were acquired; 58 occurred in 80% or more of the replicates (red), the remaining 664 only occurred in 70% or less of the replicates and were removed from further analysis (grey). When the reproducible identities acquired from pre-digest ZipTip group were cross compared with the controls run in parallel. 4 identities are unique to their-digestion ZipTip group, 26 are unique to the controls and 32 are reproducibly found in both.

The identities of the 74 proteins considered to be reproducibly identified from the best performing workflow above; the controls run alongside the alkylated and reduced samples are:

A1AG1, A1AG2, A1AT, A1BG, **A2MG**, AACT, **ALBU**,
ANT3, **APOA1**, APOA4, APOB, APOC1, APOC3, **APOE**,
C1QB, C4BPA, CERU, CFAH, CLUS, FETUA, **FINC**,
GELS, HBA, HBB, HEMO, HEP2, HPT, HPTR,
HRG, HV304, HV305, IC1, IGHA1, IGHG1, IGHG2,
IGHG3, IGHG4, IGHM, IGKC, ITIH1, ITIH2, ITIH4,
KNG1, KV119, KV402, LAC2, LV403, PLMN, PON1,
THRB, **TRFE**, TTHY, VTDB, VTNC, CBG, CFAB,
CO3, KV105, LV302, SC6A2, ZA2G, A2AP, AB12B,
AFAM, ANGT, C1R, CO4B, ITSN1, KV114, KV121,
LUM, PZP, TANC1, and ZCH18

Highly abundant serum proteins are highlighted in bold. Albumin (ALBU), Apolipoprotein A1 (APOA1), Transferrin (TRFE) Alpha 1 Acid Glycoprotein (A1AG1), Complement C1q (C1QB), Fibronectin (FINC) and a2-Macroglobulin (A2M2).

4.3.5. A Model for use on Valuable Clinical Samples

To assess the applicability, data exportability, ease of use and computer processing time of this experimental and data processing set up with an aim to apply it to a to a more valuable clinical cohort of samples, the data from the 10 alkylated and reduced replicates and their 10 controls run in parallel were analysed using Bruker Profile Analysis Software. To re-iterate in this context the comparison of the alkylated and reduced verses control is less relevant, the point of this was to assess the reliability of the available software to process, extract features (i.e. recognise spectral peaks accurately) from the multiple replicates of the multidimensional data and make a comparison.

The software collated and aligned the retention times and mass values to produce a list of peptide peaks that are differentially expressed between the two groups (see section 4.2.2.8.2).

The bucket table consisted of 4458 buckets. The MS T-test within the Profile Analysis software found 553 buckets to have a p-value of <0.05 and a fold changer greater than 2. That is to say the intensity values of 553 areas of spectra aligned by retention time and mass significantly differed between the two groups.

4.3.5.1. Measured Error.

On exporting the SPL list the margin of error of the retention time shift and mass shift calculated by the software are as follows: Retention time windows ranged from 0 to 22.13 min with an average of 7.01 min preceding the peak value and 8.54 min following the peak value. The shift in mass value of peaks calculated to be the same compound ranged from 0 to 0.23 Da with an average of 0.08 Da preceding the peak value and 0.06 Da proceeding the peak value.

4.3.5.2. Investigation of Peaks Calculated to be Significantly Differentially Expressed Between the two Groups.

The mass and retention time values of the ‘buckets’ found to be significantly differentially expressed between the two groups were cross compared with the peptide sequence lists exported from each of the sample runs. The window sizes and boundaries of error for the buckets from the SPL list were used to do this.

Within the 10 alkylated and reduced replicates and the 10 controls run in parallel, there were 30 incidences where one ‘bucket’ value corresponded to more than one peptide identity. That is to say two or more peptide of different sequence identity were shown to have the same mass and retention time within the boundaries/tolerances of the SPL list.

It should also be noted that multiple ‘bucket’ values correspond to one protein identity. From a list of 533 values corresponding to location in spectra aligned by retention time and mass 136 protein identities in total were linked to these values.

Using the SPL list, the peptide sequence and protein identity table for all 20 samples were filtered to only include peptides within a ‘bucket’ values. A consistency confidence of 80% was applied to these peptide lists, in that an identity had to appear in 8 out of the 10 replicates to be included. Sixty-seven identities were matched in the 10 replicates of the alkylated and reduced samples; 6 peptides occur in 80% or more replicates, 61 occurred in 70% or less replicates and are not considered reproducible. One hundred and twenty-two peptide s from the control group remained after filtering with the bucket boundaries 30 peptides occurred in 80% or more of the replicates, 92 occurred in 70% or less. When the reproducible peptides are cross compared 24 were unique to the control group, 6 were found in both and none were unique to the alkylated and reduced samples.

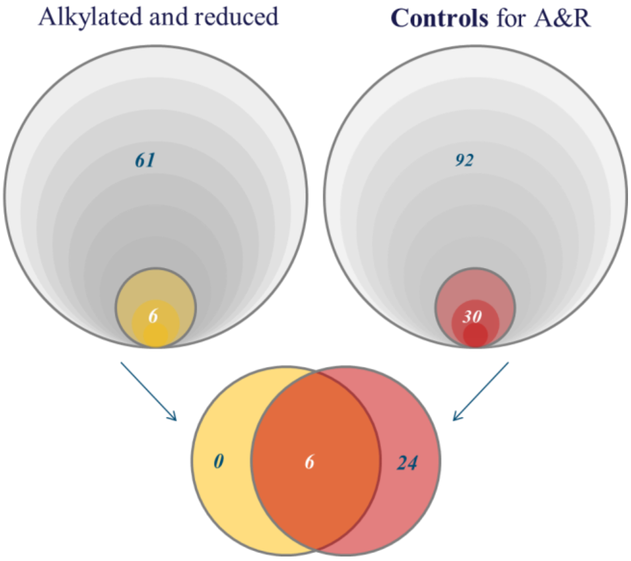


Figure 28. Venn Diagram Comparing the Lists of Protein Identities from Peptide IDs Matched from MS-T-test.
The data from the alkylated and reduced replicates and 10 controls run in parallel were filtered to only include values in the SPL list. Sixty-seven identities were matched in the 10 replicates of the alkylated and reduced samples, 6 of which occur in 80% or more replicates (yellow), 61 occurred in 70% or less replicates (grey). One hundred and twenty-two peptides from the control group remained after filtering with the bucket boundaries 30 peptides occurred in 80% or more of the replicates (red), 92 occurred in 70% or less (grey). When the reproducible peptides are cross-compared 24 were unique to the control group, 6 were found in both and none were unique to the alkylated and reduced samples (yellow).

This provides evidence of 24 proteins detectable in the controls that are not in the alkylated and reduced samples, and 6 proteins that are present at significantly different levels between the two groups.

4.4. Discussion

This chapter aimed to assess the applicability of LC-MALDI-TOF to serum biomarker discovery in ovarian cancer. The Bruker software has no inbuilt feature to corroborate, collate or compare technical replicates, so this was performed where possible using the Bruker Flex software package and for a wider comparison the data was exported and transferred into Excel for analysis.

4.4.1. Semi-quantitative Nature of MALDI-TOF MS

As a proof-of-principal, to asses if MALDI-TOF-MS data can be treated as semi-quantitative a standard sample of BSA digest was diluted at several concentrations, spotted to a MALDI-TOF target, data acquired, exported then reviewed (Figure 22 and Figure 23).

On visual analysis (Figure 22), a clear trend of increased signal intensity with increase in concentration of sample loaded is seen. When the spectra are exported numerically and plotted on a graph (Figure 23) the trend has a R^2 value of 0.9567

Data provided is sufficient to reject the null hypothesis (H_0-iv) and accept the alternate hypothesis: Intensity values of detected proteins is relative to the amount of protein loaded.

4.4.2. Chromatographic Reproducibility

The reproducibility of the chromatography of the C_{18} column in the LC system was assessed. This was done firstly by comparing the amount of the total protein eluted at each fraction time point from 10 replicates (section 4.3.2. and Figure 24). The time at which 50% of the total protein eluted from the column in 10 replicates had a 500s range. As the samples are spotted into 384 10s fractions (3840s), the 500s range represents a 13.02% of the run time in total.

A more accurate representation of the retention time shifts was found in the Bruker SPL list, where each individual peptide peak shift time is being calculated (see section 4.3.5.1.). These had a large range; 1333s, 34% of the total run time. But when averaged they show a slightly smaller range in retention time shift; 421s which represents 10% of the total run time.

In both calculations, the retention time variability compares poorly, is 10 times larger, in comparison to recent literature (Benk and Roesli 2012, Neubert *et al.,* 2008). However, this data was produced on different instrumentation with smaller time intervals between fractions and a lower number of replicates, in Neubert *et al.,* 2008 retention time shift calculations were assessed on 6 peak values. Retention time and area under peaks can be attributed to column packing, column age, contamination, temperature and gradient instability (Hsieh *et al.,* 2014).

When the peptide identity lists were filtered to only include the values in the SPL list there were numerous occurrences of multiple identities being linked to one bucket value. It is possible that numerous peptides have the same *m/z* values and chromatographic properties on a C_{18} column. It is also possible that there is room for error in the bucket window size and boundary calculation and two or more peaks are held within the bucket. The mathematics/algorithms behind the SPL list generation are encrypted in the Bruker program; not known.

Precise assessment of the chromatography of this workflow cannot be conducted with the instrumentation available as the chromatographic unit has no online detection system. Column elute can only be viewed in 10 second fragments. If time allowed the addition of a retention time standard, spiked into a sample before processing could be evaluated. This would address retention time shift problems but is an additional processing step and may suppress signal sample signal of proteins in very low concentration s with similar affinity to the C_{18} column.

Unfortunately, due to the fractionated and uncoupled nature of the LC separation to the MALDI-TOF-MS data acquisition, a thorough investigation of retention time was not possible. Moreover, online monitoring and ad hoc adjustments were not able to be made during data acquisition. Although some of the assessments were made based on the identification of a peptide or bucket table calculation, which holds a potential yet limited possibility of error, it was satisfactory to indicate the size of the retention time variation.

As described in Escher *et al.,* (2012), in LC-MS setups that are directly coupled to a detector, standards of a stable and reproducible retention time can be included or run alongside samples and the variation in their elution time is used to calibrate retention times of sample data as it is being collected.

Bruker software used does not easily accommodate retention time alignment so would have to be done post data acquisition in the data analysis. Although retention time standards were available for purchase, to apply a retention time calibration using them would need to be developed thus would not be simple.

4.4.3. Reproducibility of Identities Acquired Using LC-MALDI-TOF-MS/MS

For a qualitative assessment of protein identities, the full lists of peptides identified from each run for all of the workflows replicates were exported to Excel using Bruker WARP-LC and Protein Viewer software. The occurrences of the accession codes across each set of replicates were observed. Figure 25 is a histogram that displays the typical distribution of protein identity occurrence across 10 replicates.

A large number of protein identities occur only in one replicate. This pattern reflects that shown in similar research where 3 LC-MALDI-MS workflows were compared (Hattan *et al.,* 2005). These are likely to be an error produced from the combination of the size of tolerance in mass shift, algorithms with in Mascot and the probabilities involved with matching when peptides with similar sequences occur. It should be noted that other search databases other than Mascot are available and could be made compatible with the Bruker software output. Mascot is an industry standard and accepted utility in peer research, it has merits and pitfalls, Mascot is not able to incorporate the mass accuracy of the fragment ions when searching, future iterations of the software may do so however on this occasion a measurement made with potential to increase validity of a protein identity match that was not able to be incorporated. Searching the MS data against another database would most likely produce similar, but not identical lists of protein identities for each sample. Although it is possible this may have marginally increased/improved any result adding this extra parameter to explore was not considered relevant expansion of the analysis. The variation of the measured *m/z* values is constant despite the choice of database to deduce protein identities.

When a power model of 80% was applied, in all four workflows the number of protein identities considered 'reproducible' drops dramatically: Of the 10 alkylated and reduced replicates, just 17.5% of total identities were found reproducible; 21.8% in that of the controls run in parallel (Figure 26). Of the 10 replicates that were ZipTipped before digestion, 13.6% of the total proteins identified were found to be reproducible and just 8% for the controls run in parallel (Figure 27).

The purpose of this work was to evaluate sample preparation and data acquisition workflows with a view to apply the best to a cohort of clinical samples. This data demonstrates the importance of running test samples in replicate. However, it would be impracticable/ a drain on sample volume and instrument time to run each of a cohort of test samples in duplicate (10 times). The difference in number of protein identities that occur in three or more replicates warrant running a test sample in as least triplicate.

4.4.4. Comparison of the Proteins Identified from each Workflow

When using the power model of 80% in both sample work flows tested the control group produced more reproducible identities.

Thirty-eight proteins were reproducibly identified in the 10 alkylated and reduced replicates compared to 74 for that of the controls run in parallel (Figure 26). Thirty-six proteins were reproducibly identified from the 10 replicates that were ZipTipped before tryptic digestion, compared to 58 for that of the controls (Figure 27).

Contrary to published work stressing the importance of alkylating and reducing samples prior to aid digestion and therefore identification (Sechi and Chait 1998, Hale *et al.,* 2004 and Wedemeyer *et al.,* 2000), the evidence from this work (Figure 26) shows a marked reduction of proteins identified from serum samples reduced and alkylated prior to digestion.

A C_{18} ZipTip purification prior to digestion was performed to reduce the amount of noise and signal suppression from salt and large, overabundant proteins such as albumin. However, data from this work showed that performing that C_{18} ZipTip step prior to digestion decreases the number of achievable reproducible protein identities (Figure 27)

The protocols selected were tried and tested (Vafadar-Isfahani *et al.,* 2010, Ontario Cancer Biomarker Network 2012). However, if time allowed it would be worth investigating the ZipTip, alkylation and reduction procedure further to assure they were successful before disregarding literature and concluding they do not increase the number of proteins identified. It is possible that the alkylation and reduction procedure used did in fact reduce and alkylate the samples, however confirmation the protocols were successful would support the findings from this work.

Evidence is provided to reject the null hypothesis (H_0-v) and accept the hypothesis that: One sample preparation technique will produce greater amount of meaningful protein identities. This was the method for the controls for the alkylated and reduced samples (section 4.2.2.5.1)

4.4.5. A Model for use on Valuable Clinical Samples

To assess the ease of use, applicability and computer processing time of the experimental/data processing setup ready for use on a more valuable cohort of clinical samples, the data from the 10 alkylated and reduced replicates and their 10 controls run in parallel were compared using Bruker Profile Analysis Software. A function within the Bruker software package, which appropriately handles and performs semi-quantitative comparison of the multidimensional data generated.

This was conducted to confirm the reliability and ability of the software to recognise and extract spectral features (i.e. consistently recognise spectral peaks accurately) from the multi-replicate multidimensional data, and, to investigate the exportability of any data generated. To confirm, in this context the comparison of alkylated and reduced versus control is less relevant, this was conducted on these data, as this was the only data set with multiple replicates generated so far.

Data provided in section 4.3.4. and 4.3.5. provides evidence to reject the null hypothesis (H_0--vi) and accept the hypothesis that: Differences will be seen in the LC-MALDI profiles of serum samples processed through different sample preparation conditions. However, the reproducibility of this difference is questionable. Two different sets of proteins were found to be significantly different between the two groups when the data was exported and analysed in two different ways; qualitatively (section 4.4.4. above), and semi quantitatively (section 4.2.2.8.2.)

When the power model of 80% confidence (see section 4.4.3) was applied to the potential peptide identities from the semi quantitative comparison (Figure 28); 24 proteins were shown to be significantly expressed in the controls and not the alkylated and reduced samples, 6 were found to be expressed at different levels, and none were found to be expressed in the test samples and not the controls. This is contrary to the qualitative comparison (section 4.3.4); where 4 proteins were demonstrated to be expressed more in the alkylated and reduced replicates compared to the controls.

This investigation has shown that the data from samples run in LC-MALDI-TOF–MS/MS has low reproducibility and can easily be interpreted in multiple ways to draw contradicting conclusions.

Furthermore, the final number of protein identities accepted with 80% confidence is low. In comparison to hypothesised size of the serum proteome (see section 2.1.2.) it is miniscule evidencing this to be an insufficient model to represent the system and thus a questionable platform for biomarker discovery.

4.4.6. Limitations of the Methods Tested

The comparison was made to firstly corroborate or refute any differences already seen using each method. Secondly, to rehearse a method and unearth any difficulties that may arise when applied to valuable clinical samples.

To match the areas of spectra aligned by mass value and retention time found to be significantly different between the two groups to a peptide identity with the software resources available the data needed to be exported and matched to the peptide identity lists manually; outside of the Bruker software. This in its self is a source to introduce error. The difficulty of exporting the processed data in the form of a SPL list was noted (section 4.2.2.8.2.) The Bruker software prohibits direct export of the numerical values. The values of the table were exported as an image and converted into text using Adobe reader. The values of the exported table were checked for errors against the original manually. The size of the table leaves large opportunity for human error. This increases the time involved in analysis, decreases the validity of any findings and therefore reduces the utility of this method to biomarker discovery on a cohort of valuable clinical samples.

There were 30 incidences of one ‘bucket value’ corresponding to more than one identity (section 4.3.5.2.). Separating the protein identities from the peptide mass on a third dimension (retention time) was introduced to reduce the ambiguity that has previously been problematic in Chapter 3. As ambiguity in identity still exists the sample preparation workflows and advance in technology tested adds little to the potential of previous work (Chapter 3).

Limitations found with the workflows tested are linked mostly to the flexibility of software provided with the instrumentation, the necessity to export data for use in another, or, the unknown parameters embedded in the software’s coding such as the Bruker ‘bucket’ boundary generation or those within Mascot matching masses to sequences. This view is shared in a recent review, Benk and Roesli 2012 state the capabilities of LC-MALDI-MS have not been realised due to the lack of suitable computer programs. The difficulty in using LC-MALDI-MS data is aligning the data with confidence on both dimensions and normalisation before it is analysed, the way this is conducted has a massive influence on any results (Van den Berg *et al.,* 2006, Podwojski *et al.,* 2009). The purpose of adding the LC separation is to add another value to a peptide of one *m/z* so it can be differentiated. The Bruker software does not take this value into consideration when calculating sequence identity based on probability. This leaves the users of the technology responsible to add in this extra value to the data. Researchers must choose between using the software provided which contains unknown parameters of feature selection which are a possible source of error, or, developing bioinformatic analyses tailored to the nature of the data, samples and study design, which though achievable (Tong *et al.,* 2012, Timms *et al.,* 2011, Shin *et al.,* 2008) in itself is overly demanding on time and a questionable devotion of time and resources compared to overarching goals of research; cancer biomarker detection.

Additionally, due to the high number of possible amino acid combinations making any particular pre-cursor ion (*m/z* value), this methodology is fundamentally challenged. The lack of resolution by chromatographic separation means that multiple peptides with the same mass to charge ratio will likely overlay in a MS spectrum so multiple identities can be inferred from, leaving the question: Which is the correct protein identity? As highlighted in Figure 4, the limited number of separable data points in each spectrum verses the number of proteins or peptides expected to be present in each sample, thus making it impossible to represent this data in the space of a single mass spectrum.

Peak area is a widely-accepted measurement in mass spectrometry and arguably better represents the ion measurements quantity compared to overall intensity as used above. Unfortunately, within the Bruker Flex software package ‘peak area’ data was available but not clearly defined or readily accessible, so, deducts from this methods ability to be intemperate as semi-quantitative measurement.

It is also noted in section 4.3.4 that a number of the proteins significantly consistently identified from all workflows were common, high abundant serum proteins. The utility of these proteins expression in serum is already doubted to have value in disease detection and have been shown to supress the detection of differential expression of lower abundant proteins believed to be of more importance as biomarkers (section 2.1.5).

4.5. Conclusion

To conclude, all null hypotheses listed in section 4.1.5. $H_{0\text{-iv,v,vi}}$ can be rejected in favour of their alternate hypotheses.

- Intensity values of MALDI-TOF-MS data can be used to indicate the relative protein quantity within a sample, however it was noted the accuracy of this is low.

- One sample preparation technique produced more reproducible peptide identities than others.
- Difference in the LC-MALDI profiles of serum samples produced under different conditions was shown to be different using a qualitative and semi-quantitative method.

Evidence is also provided to suggest

- Both sample processing workflows tested reduced the number of reproducible identities attained from samples.
- Clinical samples should be run in at least triplicate to reduce the number of false identities attained
- Of the two methods of data export and analysis conducted, different conclusions can be drawn from the raw data collected.

4.6 Review of Findings and Future Direction for Onco-proteomics in MS

During the time of this study, advances of instrumentation improved considerably allowing a more reliable output of larger numbers of protein identities and more accurate quantitation. At the time of this study it was not possible to generate accurate quantitative data which subsequent generations of mass spectrometer were capable of doing. The precious ovarian cohort of samples was therefore not analysed using the workflows investigated in chapter 4. Considering the rarity and diminished volume of the clinical cohort of ovarian cancer patient serum, the lack of confidence in the potential yield of the LC-MALDI-MS approach and the likelihood that a more accurate way of conducting serum protein biomarker discovery by MS existed (Marx 2013), and would soon be available, these samples would be saved for when they can be employed in a more meaningful way. Meanwhile, the now-evaluated LC-MALDI workflow could be applied to more appropriate cohorts with more abundant sample volumes, the measured error in retention time and protein identity used as a caveat to include upon analysis of results. Other sources can be mined for ovarian cancer biomarkers, namely gene array databases available online.

In the below chapter, an alternative source of data is explored. Gene microarray data sets available through online repositories, are freely available, offer larger sample numbers, and higher accuracy in the measurement of gene targets that lead to protein production.

A repository of gene expression data sets was searched as an alternative source to discover relevant biomarkers. The line of questioning aimed to interrogate data acquired from the cohort of serum samples collected for use in chapter 3 could not necessarily be continued due to the availability of clinical information. Available variables for the cohort of serum samples in chapter 3 categorise the patients by cancer or control, supporting only categorical comparisons designed to detect serum biomarkers differentially expressed between the two groups. The best data sets available in the online repository were derived from genetic material of grade 3 tumours with gene expression and survival time available, both continuous variables. Therefore, a study design to compare gene expression with patient survival time was put in place.

4.6.1. Future of Protein Mass Spectrometric Biomarker Discovery

Shotgun proteomics; is a metaphorical description of a close-range wide-target approach to analyse the entire proteome. As much proteomic information as possible is catalogued from samples then conclusions or further hypotheses are drawn from these. This strategy was used in chapter 3 and 4 of this document and a large portion of mass spectrometry protein biomarker discovery research since the early 2000's (Table 3).

The results suggest that this approach is flawed for the following reasons:

- Identification of the peptides/proteins present are generated from matching masses of hypothetical sequence calculated from genomic information on online databases (i.e. Mascot) to the MS/MS measurements collected from the mass spectrometer.
    - The MS/MS measurements are dependent on the type of mass spectrometer itself, some peptides ionise better under different conditions; i.e. ESI or MALDI (Benk and Roseli 2012), thus not all will be catalogued from one mass spectrometer and generate identity.
    - The tolerances and margins of error in the matching algorithms of the mass spectrometer software and the database itself.
- Data dependent acquisition. The MS/MS data acquisition procedure within the Bruker UltrafleXtreme is similar to other mass spectrometers of its generation and now can be termed Data Dependent Acquisition (DDA) (Law *et al.,* 2013). When data is acquired in this manner the list of peptides which are selected for fragmentation for identification acquisition is dynamic as it is dependent on the detection of the peptide. There are numerous reasons a peptide would not be detected consistently including sample

preparation, its intensity sitting on the threshold of detection or in the case of MALDI inconsistent distribution of the peptide across the matrix crystals of the dried spot. In the Bruker UFX detailed above, the instrument would take an MS scan from each of the 384 fractions, pause while compiling a list of peptides present and assign each peptide to the fraction in which it is expressed the highest, then move on to MS/MS each of the peptides in its assigned fraction. Which *m/z* are selected for MS/MS data acquisition from each run are selected based on the consistency of their ability to ionise within the mass spectrometer, and, algorithms within the mass spectrometer software recognising their consistency (Picotti *et al.,* 2013). In the Bruker software detailed in chapter 4 the peptide precursor list is dependent on what is recognised to be one peptide value based on its retention time to a C_{18} column and its *m/z* values. The lists were shown not to separate out peptides individually, and were shown to have irreproducible variations between samples.

The generations of mass spectrometers produced after the Bruker UltrafleXtreme address this inconsistency by changing the order the sample is fragmented and detected within the mass spectrometer, termed Data Independent Acquisition (DIA) (Law *et al.,* 2013, Chapman *et al.,* 2014). Using DIA quantitative measurements are independent of the detection of the precursor (see Figure 9 and Figure 10 for parent and fragment ion information). DIA approaches fragment the entire sample prior to detection. The detected fragment quantitative measurements are summated and matched to their parent ion using databases or a separate run of MS data. DIA approaches are preferable as they offer increased sensitivity as less sample within a run is lost to the MS scan, improved reproducibility as the detection of fragments are more consistent over replicate samples and have the potential to detect theoretical proteins with use of theoretical ion databases. DIA approach for analysing complex protein mixtures include; Waters$^{©}$ instrumentation which separate the fragments in another dimension, drift time, using the Synapt (Distiller *et al.,* 2014), Thermo who combined existing quadrupole and Orbitrap constituents in the Q-star Exactive (Hao *et al.,* 2012), and ABSciex SWATH (Gillet *et al.,* 2012) which again uses existing technology but detects the sample in a different order (Griffiths *et al.,* 2014, Ziqi *et al*., 2014). Currently SWATH is an emerging popular and increasingly referenced DIA approach (Biognosis 2014).

- With hindsight, MALDI-MS offers a fast, instantaneous measurement of the protein content of a sample, which is an attractive concept for biomarker research, however speed of analysis is of limited effect for discovery and may only be relevant at the clinical implementation stage. As a discovery platform although speed of acquiring data increases the amount of samples that can be analysed in the same time frame as each other however the difference between instant or minutes to an hour when coupled to and LC does not impact results of a typical discovery cohort of 50-100 samples. In fact, for this reason desorption ionisation is popular in developing mass spectrometric techniques emerging for precision medicine/clinical proteomics such as FAIMS and DESI (ELRIG 2016, Takáts *et al.,* 2014, Balog *et al.,* 2013).

- It is still not currently possible to catalogue the entire serum proteome although proteomic discovery technology has experienced extraordinary technological advances in recent years. The exponentially increasing sensitivity and specificity of novel technologies and data processing algorithms, together with the ever-increasing capabilities and solutions in computing, provide a promising future for not only characterising the proteome but combining data and technology platforms creating a holistic aim to study 'omics' (Gil *et al.,* 2015). This conceptually also holds promise to confront additional challenges posed in proteomics which add to the dynamicity of the proteome, including the measurement/quantification of global phosphorylations, glycosylations, or any post translational modifications. However, this is not currently the case.
- This sentiment is echoed by Anderson (2010). Who, in commentary on general protein biomarker discovery (not only cancer biomarkers) identifies a similar futility in protein biomarker discovery using the currently (in 2010) available technological platforms. They also site the difficulty in obtaining access to high quality sample sets, the absence of an organised development pipeline and a lack of a "useful theory of biomarkers".

To refer to the shotgun analogy, so far technology available allowed researchers to undertake studies using in the correct range to hit a portion of a large target, with little aim to reproduce the result. The new fashionable term is Targeted Proteomics; crowned method of the year 2012 by Nature Methods (Nature Editorial 2013), focuses on how technology available is best suited to quantifying a smaller subset of proteins/peptides of interest based on an hypothesis, rather than profiling the whole proteome in all of its complexity repeatedly and in more depth every time new technology is available (Marx 2013).

A quantitative mass spectrometer capable of targeting multiple ions is all that is needed to apply a 'targeted' tactic. So far this has typically been a triple quadrupole (Marx 2013). Multiple Reaction Monitoring (MRM), also called Selected Reaction Monitoring (SRM) (Hoffman 1996) is the isolation and quantification of fragments of a peptide based on their characteristics of mass, flight and behaviour in a collision cell. MRM data can provide 'absolute quantitation' of protein content, it is reproducible, selective and robust. MRM has been referred to as the mass spectrometrists ELISA (Picotti *et al.,* 2013) and has been professed to supersede immuno-based protein detection solutions. Quantitative mass spectrometry data of a number of proteins fragments could potentially one day be used in place of a multiplex of ELISA of other immune-technique (Picotti *et al.,* 2013).

However, in the case of ovarian cancer, which protein fragments need to be quantified, remain to be found. The targeted approach has since been applied to early detection of ovarian cancer Tang *et al.,* (2013), however is not a discovery platform which is still needed in this field. Some suppliers have incorporated bioinformatics/ software-solutions to process MRM-type measurements with the potential to be used for quantitative biomarker verification, for example ABSciex SWATH analysis (Gillet *et al.,* 2012, Marx 2013).

4.6.2. Future for Biomarkers for Ovarian Cancer

Despite a wealth of data and information being produced from ovarian cancer patient material little has changed in the diagnostic, prognostic or treatment care for patients with ovarian cancers (Hays *et al.,* 2010, Siegel *et al.,* 2013 Vaughan *et al.,* 2012). This represents an unmet need in patient management.

Detection of ovarian cancer disease in its early stages is accepted to be the ideal route to improved survival. A biomarker from a non-invasive screening test is the awaited discovery. However, firstly; the ideal sample is not available, secondly; if it were there is no confirmed technological analysis platform available with a proven reproducible sensitivity to detect the subtle differences (if any) expected. This view is supported by Jacobs *et al.,* 2004, who acknowledges the majority of cohorts are flawed as they are from late stage disease. A biomarker of late stage disease may be of use to detect recurrences and response to therapy yet may be completely different to early stage. It is possible that metabolic and molecular events are completely different in early or premalignant disease and that an accurate marker of this could have decreased specificity or sensitivity at detecting later stage disease (Jacobs *et al.,* 2004).There are very limited samples from preclinical/pre-diagnosed patients in existence (Jacobs *et al.,* 2004), the samples from UKCTOCs screening trial is one of a handful worldwide.

The majority of samples collected from patients volunteering for research are of those already admitted to hospital and already on a treatment pathway. The only biological samples currently available for study are from the late stage disease. The tissue samples donated by patients of later stage disease can be used with their clinical information to stratify subgroups within them then used to predict future patients' likely response outcome and response to treatment: the concept of precision or personalised medicine.

Ovarian cancer is most commonly diagnosed in Stage 3. For which the prevailing treatment is cyto-reductive surgery proceeded by platinum based chemotherapy. Although 70% of patients respond at first, a majority will develop a resistance to platinum based therapy (Miller *et al.,* 2009). The ability so segregate the patients who are likely to develop resistance may aid treatment.

For the current body of work represented in this document evidence has been produced to suggest that pursuit of the goal of an early stage biomarker is currently not an effective use of funds and samples. Using a "targeted approach" and the next generation of mass spectrometry technology such as those listed in Marx (2013), which offer significant confidence of protein identities shown quantifiably to be differentially expressed between two samples. If proven to yield reproducible results when extensively tested on more freely available human samples would then provide further information.

The human genome is far better characterised than the proteome, thus making its analysis more likely to produce results with lower ambiguity and higher reproducibility. In the following section an alternative approach to biomarker discovery is taken, using genomic array data from in-silica sources online, and literature already available.

This change in tactic offers the project the opportunity to move from evaluation of technological capabilities and methodologies to an evaluation of biological measurements acquired from a pear reviewed source relevant to the hypothesis allowing the clinical question to be addressed.

Figure 29 taken from (Braem *et al.,* 2011) was generated from an extensive investigation and review of potential biomarkers associated with ovarian cancer. The figure highlights the absence of validation or refuting of potential biomarkers published to date rather than generating new. Also mentioned by (Vaughan *et al.,* 2012).

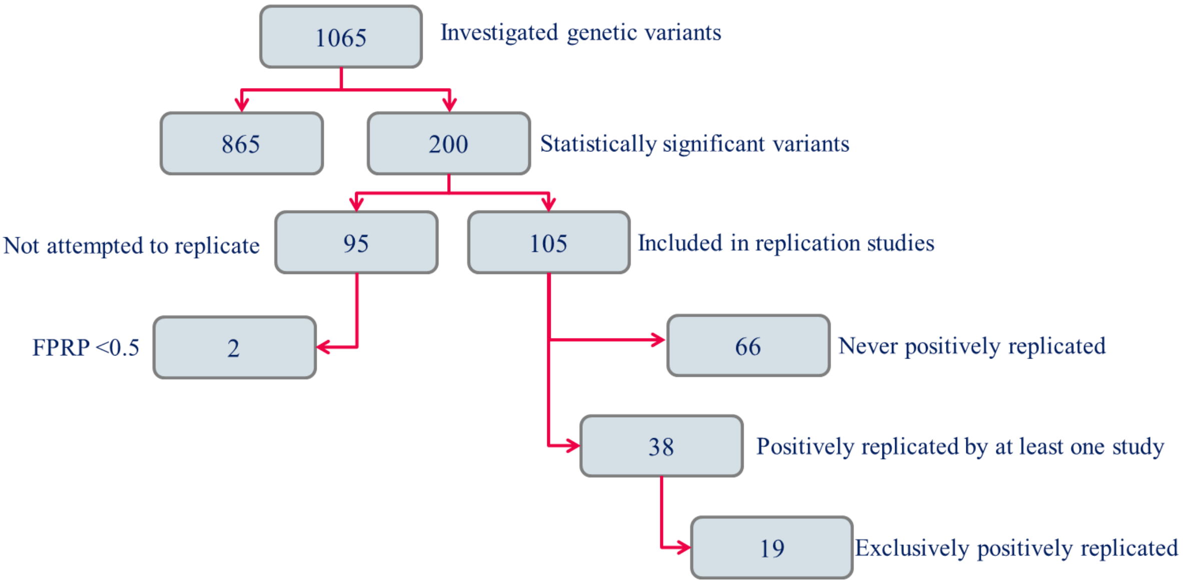

1065
Investigated genetic variants
865
200
Statistically significant variants
Not attempted to replicate
95
105
Included in replication studies
FPRP <0.5
2
66
Never positively replicated
38
Positively replicated by at least one study
19
Exclusively positively replicated

Figure 29. Number of Investigated Genes in Ovarian Cancer.
Remade from Braem *et al*., (2011) lays out numbers accounting for the high number of investigated ovarian cancer genetic markers (>1000), no attempt has been made to replicate a large number of them (865) the rest have been replicated to differing extents.

In a recent collaborative report (Vaughan *et al.,* 2012), strategising effective research on ovarian cancer, it was accepted that the sharing of data, results, methods and samples is crucial to narrowing down common active cellular mechanisms in what is a relatively rare yet genotypically diverse disease. Thus, reinvestigating published genes and data is a worthy endeavour.

5. Transcriptomics: Gene Expression Array Analysis as a Strategy for Biomarker Discovery in Ovarian Cancer

Chapter Abstract

Stratification of patients with the demonstrably heterogeneous disease ovarian cancer, based on evident active molecular pathways, would aid a targeted treatment and improve prognosis. Hundreds of genes have been significantly associated with ovarian cancer, although few have yet been fully verified by peer reviewed research, or clinical trials.

A meta-analysis approach was applied to two carefully selected gene expression microarray data sets (E-GEOD-13876 and E-GEOD-26712) downloaded from ArrayExpress, a freely available repository of microarray experimental data. In both cases the data was collected from full genome arrays applied to Stage 3 serous ovarian carcinoma and tumour samples collected and processed under regulated conditions. Artificial Neural Networks, Cox Univariate Survival analyses and T-tests were used to filter genes whose expression were consistently significantly associated with patient survival times.

A list of 56 genes were distilled from a potential 37000 gene probes to be taken forward for validation. The rigour of combining a meta-analysis approach with multiple testing using a variety of statistical procedures, increases power and confidence in the relevance of genes found to be of interest. Encouragingly, a number of the 54 are already reported to have an association with ovarian cancer survival. Further investigation and validation of the genes that are not yet reported to associate with survival may be clinical of interest and have potential to predict a patient’s response to treatment or be used as a novel target for therapy.

5.1 Introduction

RNA Microarray experiments, allow determination of the expression of entire genomes from nucleic acid extracted from biological samples (see Figure 12). To obtain the data in the current study RNA acquired from ovarian tumours was hybridised against microarray gene chips designed to detect expression levels of the entire human genome, multiple probes corresponding to different sequences within each gene are measured. These large, multidimensional data could be interpreted using endless analytical strategies to draw different conclusions. The debate and discussion of which is the appropriate statistical analysis for different types of data sets is open

(Allison *et al.,* 2006) and the huge numbers of genes reported to have an association to ovarian cancer that have not yet been replicated, warrant reanalysis of data where available (Braem *et al.,* 2011, Vaughan *et al*., 2012). Array Express is an online repository of microarray data which facilitates researches to share raw data for scrutiny and validation.

In this chapter, two cohorts of data, publicly available on ArrayExress were selected, downloaded and analysed using a different strategy to that in their accompanying original publications; Crijns *et al.,* 2009 and Bonome *et al.,* 2008.

- Crijns *et al.,* 2009 used a continuous prediction algorithm to publish a panel of 86 genes that were shown to be strong predictors of survival in women with late stage ovarian cancer. Within the paper some of these, but not all, were validated on other data sets.
- Bonome *et al.,* 2008 differed from Crijns *et al.,* 2009 by first categorising their patients based on their assigned debulking status, then used a Cox regression analysis published a prognostic gene expression signature of 57 probes which they validated on a separate blinded data set.

The content focus is different in the two papers, in that Bonome *et al.,* 2008 includes patient dependent variables and risk factors in analysis where Crijns *et al.,* 2009 centre around survival time and gene expression. However for these purposes they both generated data from tumours from late stage ovarian cancer patients who then followed a similar treatment pathway of debulking surgery and platinum based chemotherapy (where appropriate) so were considered comparable.

The statistical strategies used to meta-analyse the two cohorts of data consist of Cox univariate survival analysis, MLP-ANNs and T-tests (see section 2.2.2.).

Parts of the work reported in the following chapters were published in (Coveney *et al.,* 2015) see Appendix A.

5.1.1. Known Influences on Survival time from Ovarian Cancer

5.1.1.1 Platinum Resistance

Platinum based chemotherapies are used to treat a wide range of cancers with varying effect (Martin *et al*., 2008, Eckstein 2011). Their mechanism of action is to bind covalently to both strands of the DNA helix thus preventing the separation of the two strands prerequisite for translation and cell division. Platinum resistance and evasion of this damage depends on the tumour cells ability to recognise this as DNA damage and repair it or adapt in another way. Different cancer cell lines have been shown to both have this ability inherently, and to acquire it (Marchini *et al.,* 2013). It is still to be proven whether resistant the cells are present in smaller subpopulations within the cancer prior to platinum therapy or whether they are a consequence of it (Marchini *et al.,* 2013).

Clear cell ovarian carcinomas, which are identifiable by histology, are already known to be a more aggressive phenotype of ovarian cancer that are less likely to respond to platinum therapy (Matsuzaki *et al.,* 2015).

Two of the five known DNA repair mechanisms have been reportedly linked to platinum resistance. These are:

- Nucleotide Excision Repair (NER), where abnormalities in the helical structure of the DNA are recognised and enzymatically excised (Chang *et al.,* 1999).
- Mismatch Repair (MMR) in which unmatched, mismatched, inserted deleted base pairs are recognised then enzymatically excised (Kelland 2000).

For a full review of DNA repair and platinum resistance in ovarian cancer and more the reader is referred to Martin *et al.* (2008)

5.1.1.2 Epithelial to Mesenchymal Transition (EMT)

Cell line studies have also implicated the phenomenon of epithelial to mesenchymal transition (EMT) in platinum based drug resistance in epithelial ovarian cancer (Rosanò *et al.,* 2011). However, the exact mechanisms by which this happens are unconfirmed, in fact conflicting results have been reported from both *in vivo* and *in vitro* studies (Miow *et al.,* 2014). The presence of markers of EMT such as *SNAIL* and E-cadherin have been linked with ovarian cancer invasiveness (Rosanò *et al.,* 2011) and the activation of anti-apoptotic pathways such as NF-kB have been observed in cisplatin resistant cell lines (Miow *et al.,* 2104).

Contrary to prior evidence, Miow *et al.,* (2014) found that cisplatin had a higher efficacy on ovarian cell lines with mesenchymal status than those with an epithelial one.

Interestingly, EMT may be inherent or acquired, different chemoresistant phenotypes have been described between cells that have naturally undergone EMT-like changes and those that have undergone EMT-like changes after exposure to a platinum based drug therapy (Miow *et al.,* 2014). The clarification between inherent and acquired EMT is relevant to unearthing molecular pathways involved in chemoresistance, however it is not always discussed when reporting results linking to EMT.

Though there is a lot of research into the mechanisms of platinum resistance (Martin *et al.,* 2008), nothing has yet aided treatment in the clinic. Pathways directing clonal diversity, tumour adaptation and acquisition of resistance need to be verified.

There are a number of geno- and phenotypes documented to correlate survival times from ovarian cancer including:

- Mutations in the *PI3K* subunit, *ARID1A* and *PIK3CA* are linked to clear cell and endometreoide cancers (Jones *et al.,* 2011, Kuo et al., 2009).
- CyclinE1 (*CCNE*) associated with poor outcome (Farley *et al.,* 2003).
- There is an increased statistical likelihood of survival via clonal selection due to the large number of smaller peritoneal metastatic ‘seed’s’ recognised to be the metastatic pattern of ovarian cancer (Vaughan *et al*., 2012).

There are numerous other factors affecting a patient’s survival time that are not specific to ovarian cancer, but would be a consequence of a specific onco-phonotype or characteristic of a type of cancer micro-environment the reader is referred to (introduction) and Hanahan and Weinberg (2011).

In this chapter biomarker investigations were undertaken using mining of gene expression microarray data and is in contrast to the analysis of the previous chapter’s protein serum biomarker investigation in that the analyte measured is biologically “upstream”. The reader is referred to Figure 4 and Figure 11 in the introduction. Hypothetically the RNA measured here may well code for proteins that could be measured by mass spectrometry and the two results could be used to provide complementary evidence of one system. However, by necessity the cohort of patients observed in the two chapters is different. Interestingly in theory gene expression microarrays should be provide a smaller number of variables to compare as the genome is smaller than the estimated proteome (Anderson and Anderson 2002 and Harrow *et al.,* 2012), however, due to the technical challenges of analysing serum protein described above, far more variables are measured in gene expression microarrays.

It is reiterated here the purpose of this change is to tailor the analysis to best suit the available samples, so the data is interrogated with the most relevant scientific question.

5.1.2. Aims and Hypothesis of the Chapter

This chapter aims to characterise genomic differences between tumours from patients with Stage 3 ovarian cancer that responded well to therapy and those which did not, based on the patient’s survival times.

H_0 vii: None of the gene expression measurements from the two cohorts will be found to be consistently associated with survival times from ovarian cancer when tested with a complement of statistical strategies.

H_1 vii. Genes will be found to be consistently significantly associated with survival time when a complement of statistical strategies are applied in a meta-analysis approach to two separate cohorts of patients measured with two different microarray platforms

5.2 Materials and Methods

5.2.1. Selection of Data Sets

5.2.1.1. Array-Express Search Parameters for Sample Cohort Selection

Factors that were considered when selecting data cohorts included:

- The number of patient samples within the data set, this needed to be as large as possible to best represent the population of cancer cases studied. The larger a cohort is the higher confidence can be assigned to any results or conclusions drawn, in this context a sample size of 44 versus 44 is the minimum for statistical significance (Abdel-fatah *et al.,* 2016).

- The sample source. Only studies using patient tumour samples were considered, studies using cell lines were rejected.
- The completeness of the data. In particular the Sample and Data Relationship Format (SDRF) file, those which did not contain clear data for all files available were not considered.
- The focus of the study. A 'fair' meta-analysis needs utilise as similar sample cohorts as is possible. For example, a cohort of data generated from a trial of a novel drug/therapy cannot be fairly compared alongside cohort with patients on a standard/different treatment pathway.
- The depth and detail of the data available. To take a meta-analysis approach, the same variable needs to be available for all data sets included. For example, time to relapse was measured in one data set is not comparable to survival time in anther data set.
- The Array Design File (ADF) files needed to be cross-referenceable, one element of the adf table i.e. gene code needs to be in all data sets for a meta-analysis. The available data sets are generated from different gene chip platforms i.e. Affymetrix, Illumina, Agilent, each with their own probe design and number to represent a genome. Though possible, it is not practical to search and annotate this manually.
- Full genomic representation. Only data generated from gene microarrays representing the full genome were considered. Those including only a subset or set of mutations were discounted.

Survival time was the only dependent variable available in both the study cohorts selected for the analysis. Patients in both studies selected were subject to the same general treatment strategy of a possible debulking surgery, followed by platinum based chemotherapy where necessary.

5.2.1.2. Two Data sets used for Meta-analysis

In this context the term meta-analysis is used to describe a comparison looking for concordance across more than one data set, using more than one statistical analysis.

Gene array data was downloaded from Array Express, the data set was derived from analysed tissue from patients with ovarian cancer who have been treated with the same care pathway. Full data and information is available at http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-13876/ and http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-26712/ (ArrayExpress accessed 2011).

From the variables and data available this data could be mined for more genes that are expressed with significance in relation to survival time from Stage 3 serous ovarian cancer, and, to validate or refute any genes recently reported to be linked to ovarian cancer but not fully validated.

Cohort 1:

Full data and information is available at http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-13876/ (ArrayExpress 2011)
Array: A-GEOD-7759 - Operon human v3 ~35K 70-mer two-color oligonucleotide microarrays.
Sample information: 157 consecutive patients with advanced stage (3, 4) disease donated tumour from cyto-reductive surgery prior to platinum based chemotherapy treated at University Medical Center Groningen (UMCG, Groningen, The Netherlands) in the period 1990–2003.
Accompanying Publication: Crijns *et al.,* (2009).

Cohort 2:

Full data and information is available at http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-26712/ (ArrayExpress 2011)
Array: A-AFFY-33 - Affymetrix GeneChip Human Genome HG-U133A [HG-U133A]
Sample information: 185 late-stage (3, 4) high-grade (2, 3) ovarian cancer tumours donated from previously untreated patients at Memorial Sloan-Kettering Cancer Center between 1990 and 2003.
Accompanying Publication: Bonome *et al.,* (2008).

5.2.2. Pre-analysis Data Evaluation and Processing

Firstly, the survival distributions of the population of the two data sets were observed, the survival times ranged from 1 to 234 months and 0.7 to 130.4 months with a mean range of 25 and 39 months in GEOD13876 and GEOD26712 respectively. The survival distribution was observed to be left skewed and similar between the data sets (See Figure 30).

The MLP-ANN algorithm utilised requires a categorical variable, a cut off defining a short and long term survival group needed to be defined on survival (which, is a continuous variable). To minimise bias introduced from fitting a cut off to a continuous variable, the process was repeated at three possible time points. These were; above and below 16, 23 and 30 months (See Table 14 and Figure 30).

The cut off points were fitted as closely to median, upper and lower quartiles survival time as possible without unbalancing the sample populations more than a ratio of 1:3.

Table 14. Numbers of Cases in Short and Long or Short Term Survival Groups. Group sizes when of short and long term cut off are applied

Survival cut-off		No. of patients E-GEOD- 13876	No. of patients E-GEOD-26712
Cut off 1 (16 months)	Lower	58	32
	Upper	55	97
Cut off 2 (23 months)	Lower	75	48
	Upper	38	81
Cut off 3 (30 months)	Lower	84	62
	Upper	29	67
Total samples		**113**	**129**
Minimum cohort size		**29**	**32**

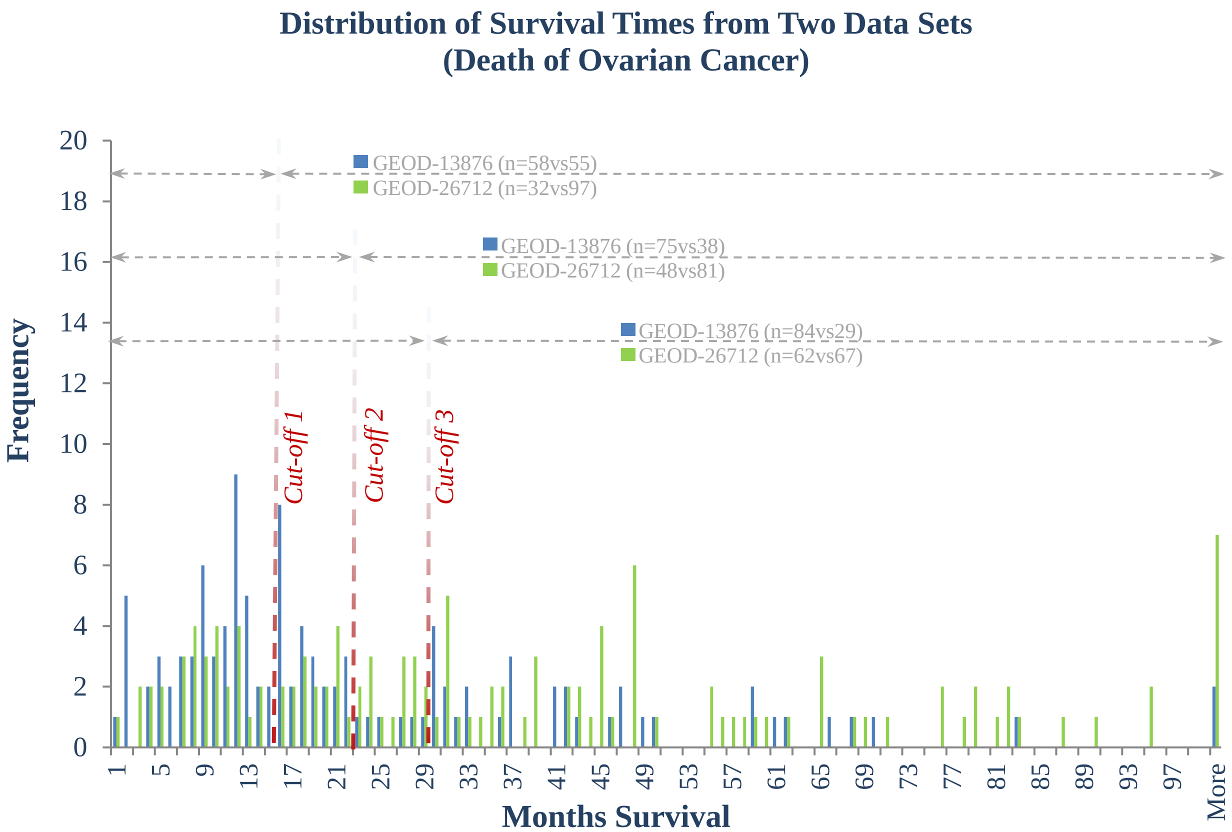


Figure 30. Histogram of Distribution of Survival Times of Two Cohorts of Patients with Ovarian Cancer
To enable a categorical analysis artificial cut-off points defining long and short term survivors were made. This was done at three different time points; Cut-off 1, 2 and 3 these were made at 16, 23 and 30 months.

5.2.3. Analyses Applied

5.2.3.1. ANN of Short versus Long Term Survival

The ANN analysis does not accommodate censored variables i.e. those categorised as "death not from ovarian cancer" or "alive with disease", so were excluded from this analysis.

An in-house designed, multilayer, back propagation ANN algorithm (Lancashire *et al.,* 2009, Lancashire *et al.,* 2010 Kafetzopoulou *et al.,* 2013), with an architecture of 1-2-1 was utilised. Within this a Monte Carlo Cross Validation (MCCV) was applied; the population is randomly divided into training, test and validation cohort with a ratio of 3-1-1 (60, 20 and 20%).

For each gene probe, the gene expression values for a randomly selected 60% of the patient population are used to train the model, 20% to test, then 20% are applied as a blind validation This cycle is repeated 50 times and a report of the averaged predictive performance created;

this includes training, test and validation performance and error. This loop is repeated ten times for each gene probe. At the end, the ten reports for each gene probe were compiled in Excel and a mean average of the ten reports calculated. The average performance over the ten was calculated. All the gene probes were then sorted by the Test Error.

Using the three-time point cut-offs ANNs was conducted on the two data sets. Each ANN was used to rank the gene probes in order of the predictive performance to distinguish short and long term survival on the blind validation subset. Two ANNs were conducted for each time point for each data set, a total of twelve analyses. Within each of the twelve analyses the gene probes were ranked by their predictive performance on an internal blind validation step and gene probes ranking below 0.05% were disregarded. The gene codes of these shortlisted gene probes were cross-referenced across the six ANNs from each time point in each data set. Multiple cross comparison systems were explored. Gene codes were weighted based on the frequency of their presence in the twelve ANNS.

The list of weighted gene codes with a consistent predictive performance between long and short term survival were taken forward to the meta-analysis.

5.2.3.2. Cross Validation with Cox Univariate Survival Analysis

Cox proportional hazard model has the capacity to compute both censored and non-censored cases (Singh and Mukhopadhyay 2011) so “death not from ovarian cancer” or “alive with disease” were included increasing the sample numbers.

A Cox univariate survival analysis was conducted on every gene probe in each data set individually to determine if the is expression significantly correlated with survival. To do this a macro (see Digital Appendix A) was created within Statistica 8 software that cycled round each of the thousands of gene probes within each data set and produced a report for each one. The reports were exported to Word, transferred to Excel and a macro function used to compile the results. Gene probes were ranked by their p-value and any below 0.05 were disregarded.

The gene codes of the gene probes with a p-value of ≤0.05 were taken forward for the meta-analysis.

5.2.3.3. Cross-comparison of Significant Genes

The PiVot table function within Excel was used to cross-compare the gene codes that performed well as predictors in the MLP-ANNs and had a significant p-value in the Cox univariate survival analysis. Gene probes that did not occur in all four categories were disregarded.

5.2.3.4. T-tests

Two tailed type two Student's T-tests were conducted in Excel applying the same time point cut-offs described above (Table 14 and Figure 30) to find a categorical analysis to a continuous variable. Genes that did not have a significant T-test p-value for one or more probe in both data sets were disregarded. Finally the averages of each were compared. Genes whose expression trends when correlated with survival differed between the data sets were disregarded.

5.2.3.5 STRING Analysis

The final list of 56 genes were searched in STRING 9.0 (2013) (see section 2.2.3.) to uncover any already published knowledge of association or interactions between them.

5.3. Results

A meta-analysis approach was applied to two carefully selected gene expression microarray data sets (E-GEOD-13876 and E-GEOD-26712) downloaded from ArrayExpress. In both cases the data was collected from full genome arrays applied to Stage 3 serous ovarian carcinoma and tumour samples collected and processed under regulated conditions.

5.3.1. ANN of Short and Long Term Survival

As described above the only available variable for analysis was survival time, a continuous variable. The ANN algorithm used requires a categorical variable. For this reason a cut off had to be made separating short from long survival. As described in section 5.2.2 multiple ANNs were conducted to best accommodate a categorical analysis around a continuous variable.

Using the three time point cut-offs ANNs were conducted on the two data sets to generate six sets of gene codes of interest. Within each of the 6 ANNs analysis the gene probes were ranked by their predictive performance on an internal blind validation step and gene probes ranking below 0.05% were disregarded. See Digital Appendix B for the ranking of each probe from both data sets ranked by their performance (Average Test Error) at all three cut-off points tested.

5.3.2. Cox Univariate Survival Analysis

Cox univariate survival analyses was conducted on every gene probe individually to determine if its expression significantly correlated with survival. See Digital Appendix B for the full listings of p-values of each Cox Univariate Analysis.

5.3.3. Cross-comparison of Significant Genes

When the gene codes of the gene probes found to be statistically significant from the ANN analysis and the Cox univariate survival analysis from the two data sets were cross compared there was an overlap of 126 gene codes, see Figure 31. These were:

AASS, *ACHE*, *ACOXL*, *ANGPTL2*, *ANKMY1*, *ARHGAP26*, *ATG4B*,
ATP2A3, *BACH1*, *BACH2*, *BLMH*, *BMP4*, *BNC2*, *C19orf42*,
CACNA1E, *CACNB2*, *CDC25B*, *CEP152*, *CLIP3*, *COL13A1*, *COLEC12*,
CSDC2, *CTBP2*, *DCN*, *DCTD*, *DECR2*, *DHPS*, *DNAJC4*,
DOM3Z, *EDNRA*, *EFNB3*, *EIF1AY*, *EPS8L1*, *EXOSC7*, *FAM32A*,
FAM60A, *FGFR1*, *FHOD3*, *FKBP14*, *FYN*, *FZD7*, *GJB1*,
GLP1R, *GLT8D2*, ***GULP1***, *H2AFV*, *HBD*, *HIST1H3C*, *HNRPDL*,
HSD17B14, *IDE*, *IGF2*, *IGFBP3*, *IGFBP6*, *IL17B*, *INTS5*,
KCNC2, *KCNJ15*, *KIAA0528*, *KLHL23*, *LDB2*, *LIMA1*, ***LRRC17***,
MAP4K4, *MATK*, *MFAP4*, *M ME*, *MORC2*, *MPG*, *MTERF*,
MYCN, *MYH6*, *MYO7A*, *NAV3*, *NCOR1*, *NDN*, *NEBL*,
NFX1, *NOL11*, *NSUN6*, *NTRK3*, *OLFM1*, *OLFML3*, *PCDH17*,
PDZRN3, *PHIP*, *PJA2*, *PKD2*, *POGZ*, *POLL*, *PPFIBP1*,
PPP3CA, *PTK2*, *PTPRE*, *RABGAP1*, *RARRES2*, *RBM17*, *RBM6*,
RPL10, *SCAMP1*, *SCN2B*, *SEMA3C*, *SERPINE1*, *SFRP4*, *SLC11A2*,
SMARCA4, *SMARCD3*, *SMC3*, *SMG5*, *SPAG9*, *SPCS3*, ***TMEM45A***,
TNFAIP6, *TNFRSF14*, *TPM2*, *TPPP*, ***TRO***, *TRPM4*, *TUSC2*,
WDR59, *WTAP*, *WWC1*, *ZFHX4*, *ZMYM5*, *ZNF133*, *ZNF45*.

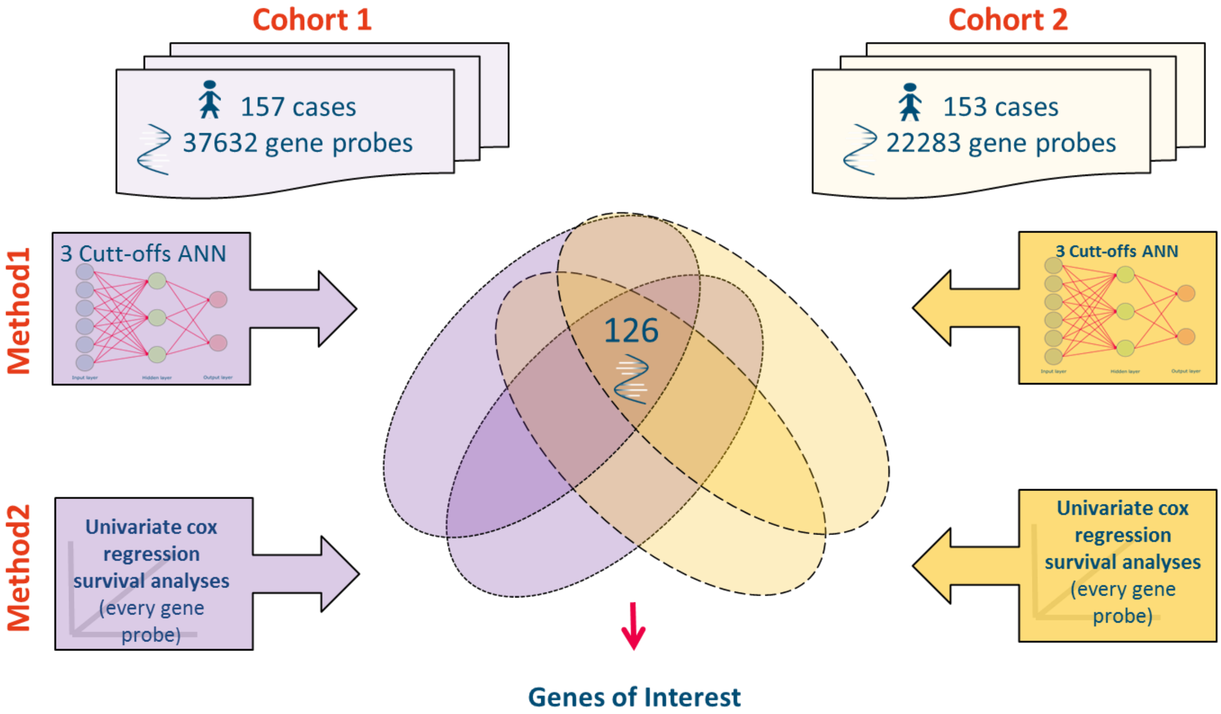


Figure 31. Overview of Gene Microarray Meta-analysis Methodology.
Two data sets (Cohort 1 containing 157 cases and 37632 gene probes, Cohort 2 containing 153 cases and 22283 gene probes) were mined for gene expression values significantly associating with ovarian cancer survival using two statistical approaches. Method 1: a set of three ANNs using differing time point cut offs to define short and long term survival, Method 2; a Cox univariate survival analysis performed on every gene. Upon cross comparison of statistically interesting genes 126 gene probes were selected from a potential 37632 for further analysis.

The list of GOIs was cross reference with the lists reported to be of interest by the initial investigators who generated the data GEOD 13876 (Crijins *et al.,* 2009) and GEOD 26712 (Bonome *et al.,* 2008). Four genes were found to overlap; these are *GULP* from Bonome *et al.,* (2008) and *LRRC17*, *TMEM45A* and *TRO* from Crijins *et al.,* (2009).

Gene codes were weighted based on the consistency of their performance to predict survival times of a blind validation set in the twelve ANNS performed at the three time point cut offs on two data sets see section 5.2.3.1. This is visualised in Figure 32, genes that occurred multiple times carry a higher weighting thus positioned higher in a pyramid of interest.

When compiling the data for each quadrant of the overall meta-analysis depicted above, the lists from each set of three ANN analyses (depicted as method 1 above) could be combined/sub-cross-compared at increasing levels of stringencies. For example if a gene occurs in the highest-ranking portion if any of the three survival time cut offs are applied, or if it had to occur in two or more (more restive), or if it had to occur in all three (most restrictive). The overarching stringency in this meta-analysis is attained from testing using multiple, different methods. For this reason, the least stringent combination was applied to maximise the genes taken forward for analysis by a different method. However the as the analysis had been conducted using all levels this information was compiled to rank the genes in order of occurrence across the increasingly stringent repeats. This may serve as an approximate rank in confidence.

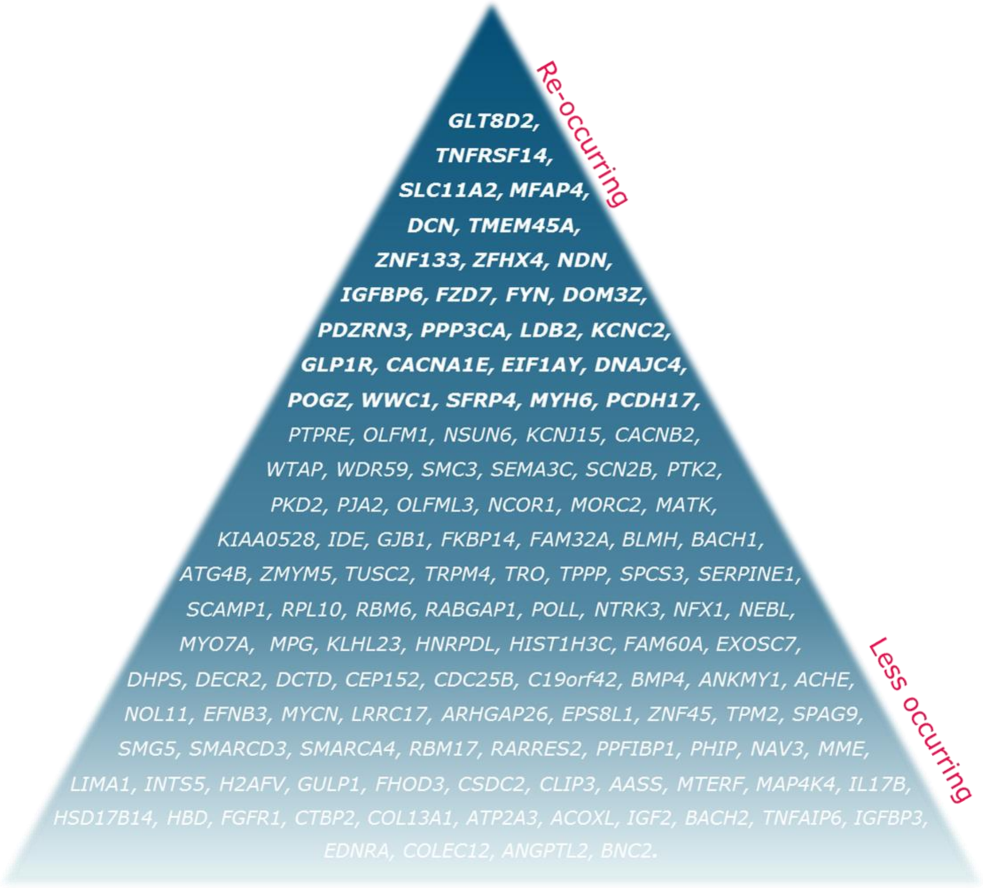

Re-occuring
GLT8D2,
TNFRSF14,
SLC11A2, MFAP4,
DCN, TMEM45A,
ZNF133, ZFHX4, NDN,
IGFBP6, FZD7, FYN, DOM3Z,
PDZRN3, PPP3CA, LDB2, KCNC2,
GLP1R, CACNA1E, EIF1AY, DNAJC4,
POGZ, WWC1, SFRP4, MYH6, PCDH17,
PTPRE, OLFM1, NSUN6, KCNJ15, CACNB2,
WTAP, WDR59, SMC3, SEMA3C, SCN2B, PTK2,
PKD2, PJA2, OLFML3, NCOR1, MORC2, MATK,
KIAA0528, IDE, GJB1, FKBP14, FAM32A, BLMH, BACH1,
ATG4B, ZMYM5, TUSC2, TRPM4, TRO, TPPP, SPCS3, SERPINE1,
SCAMP1, RPL10, RBM6, RABGAP1, POLL, NTRK3, NFX1, NEBL,
MYO7A, MPG, KLHL23, HNRPDL, HIST1H3C, FAM60A, EXOSC7,
DHPS, DECR2, DCTD, CEP152, CDC25B, C19orf42, BMP4, ANKMY1, ACHE,
NOL11, EFNB3, MYCN, LRRC17, ARHGAP26, EPS8L1, ZNF45, TPM2, SPAG9,
SMG5, SMARCD3, SMARCA4, RBM17, RARRES2, PPFIBP1, PHIP, NAV3, MME,
LIMA1, INTS5, H2AFV, GULP1, FHOD3, CSDC2, CLIP3, AASS, MTERF, MAP4K4, IL17B,
HSD17B14, HBD, FGFR1, CTBP2, COL13A1, ATP2A3, ACOXL, IGF2, BACH2, TNFAIP6, IGFBP3,
EDNRA, COLEC12, ANGPTL2, BNC2.
Less occurring

Figure 32. A Graphical Representation of the Order of Significance of the Genes of Interest.
The 126 genes of interest were weighted based on the frequence of occurrence in the twelve ANNs. Genes at the top of the pyramid were seen more frequently than those at the bottom (see See Digital Appendix B for full gene rankings).

All of the genes in the triangle were found to be of significant interest via meta-analysis from two data sets by both univariate cox regression survival analysis and ANN. The gene codes at the top of the triangle re-occur in multiple, and higher stringency options for combining lists prior to meta-analysis. The gene probe for *GLT8D2* was the most consistent and high-ranking probe thus is positioned at the top of the pyramid; gene codes toward the bottom of the pyramid may have only appeared in the least stringent compilation of the pre-meta-analysis ANN lists however do meet all the criteria for the final analysis.

5.3.4. T-tests

Using excel the data for all 126 genes listed above underwent T-tests using the same cut offs to define short and long term survival as described in section 5.2.2. Genes that did not have a significant p-value for one or more probe in both data sets were removed from the list.

After the T-test elimination a data trend comparison (described in section 5.2.3.4.) was conducted. The purpose of this was to remove genes whose significant differential expression disagreed between the two data sets. Genes that were removed at this stage include *ACHE, ATP2A3, COL13A1, EIF1AY, EPS8L1, FGFR1, KCNC2, KIAA0528, KLHL23, MATK, MYO7A, NEBL, NFX1, PTK2, RABGAP1, RPL10, SCN2B, SMARCA4, TPPP, TRO,* and *WWC1* which were all discounted because the significant differential expression between long and short term survival was observed to be opposed between the two data sets. For example, ACHE was in this comparison observed to be expressed at a significantly higher level in the tissue of short term survivors in data set GEOD13876, however a significantly lower expression observed in the tissue of short term survivors in data set GEOD26712.
Genes whose expression trends were not consistent between the two data sets were removed, reducing the list of 126 genes of interest were refined to 56. These were:

BACH1, BACH2, BMP4, CDC25B, CLIP3, COLEC12, CTBP2,
DCN, DCTD, EDNRA, EFNB3, FHOD3, FKBP14, FYN,
FZD7, GJB1, GLT8D2, GULP1, H2AFV, HBD, HIST1H3C,
HNRPDL, ***IGF2****,* ***IGFBP3****,* ***IGFBP6****, INTS5, LDB2, LRRC17,*
MAP4K4, MFAP4, NAV3, NCOR1, NDN, OLFML3, PCDH17,
PDZRN3, PJA2, PKD2, PPFIBP1, PPP3CA, PTPRE, RARRES2,
SCAMP1, SEMA3C, SFRP4, SLC11A2, SMC3, SPCS3, TMEM45A,
TNFAIP6, TNFRSF14, *TPM2,* ***WTAP****, ZFHX4, ZNF45.*

A superficial observation, even without deep research, it was apparent at this stage that several of the remaining genes were already associated with cancer by name; **Tumour** Necrosis Factor, Alpha-Induced Protein 6 (*TNFAIP6*), **Tumour** Necrosis Factor Receptor Superfamily, Member 14 (*TNFRSF14*), Wilms **Tumour** Associated Protein (*WTAP*). Additionally apparent pathway associates in Insulin Growth Factor Binding Protein 1 (*IGFBP3*) and Insulin Growth Factor Binding Protein 6 (*IGFBP6*) and Insulin Growth Factor 2 (*IGF2*).

More relevantly, a brief exploration of literature found *NAV3, SPAG9, SMC, IGFBP6* and more have been described/ implicated in cancer studies (Carlsson *et al.,* 2012, Garg *et al.,* 2008, Ghiselli 2006, Fu *et al.,* 2007). The most pertinent are *IGF2* and *BMP4*, which have been reported with relevance to ovarian cancer survival time (Sayer *et al.,* 2005, Shepherd *et al.,* 2008 and Thériault *et al.,* 2007).

5.3.5. STRING Analysis

For observation purposes only, the final list of 56 genes were searched in STRING 9.05 (see section 2.2.3.) to uncover any obvious or already published association or interactions between them. This version of string was current between March 3^{rd,} 2013 and December 27^{th,} 2013, it lists 5,214,234 proteins from 1133 organisms (although only human was searched) and holds information of 336,561,678 interactions and is still available through archived databases within the website. Fourteen of the genes are reported to be linked by co-mention in literature and five by co-expression see Figure 33.

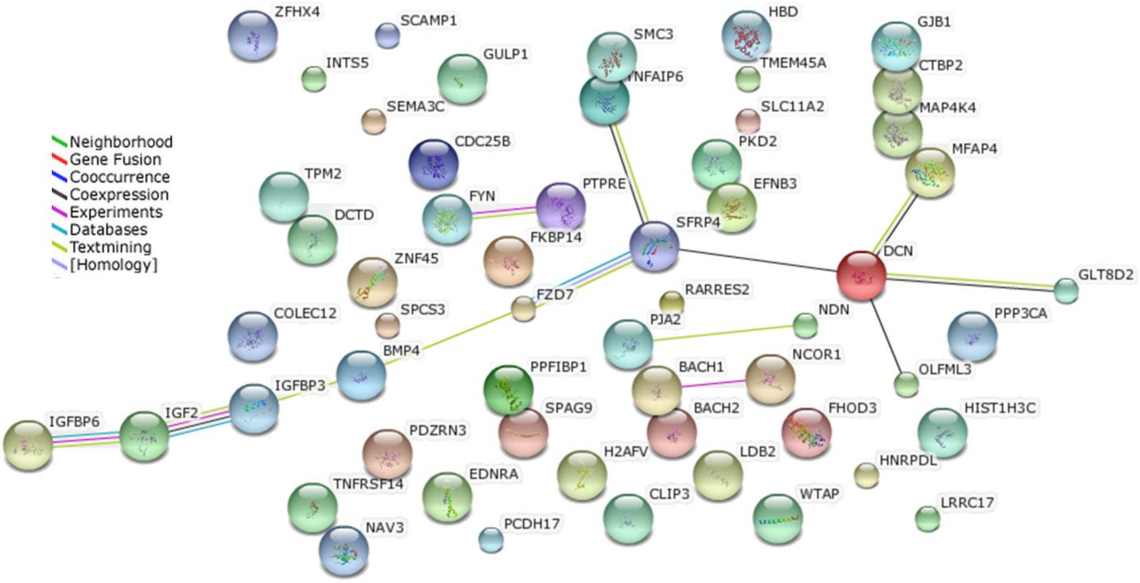


Figure 33. Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) Output Displaying known Associations between the Genes of Interest.
56 gene codes found to be significantly relevant to ovarian cancer survival were entered into STRING. Any relations are represented as a colour coded connection between genes represented by a ball. Fourteen of the genes are reported to be linked by co-mention in literature (yellow lines) and five by co-expression (black lines).

These links can act as leads to the publications linking creating the link.

5.3.6. Interaction Intact Analysis

To follow up on the links observed using STRING, Reactome (Reactome 2013) and IntAct (IntAct 2013) were explored to find interactions or pathways for any demonstrated physical links between the 56 genes. These databases are both manually curated hence this is a more rigours search, only the gene codes for which there is published evidence relating to proteins with confirmed physical interaction should be listed. Using IntAct, evidence was found of translational interaction between IGF2 and Decorin. In that, the translated protein Decorin may act as a stimulatory competitive ligand to IGF2 (Morcavallo *et al.,* 2014).

5.4. Discussion

This chapter aimed to characterise genomic differences between tumours from patients with Stage 3 ovarian cancer that responded well to therapy and those which did not.

Using survival time as a measure of response to therapy based on the patient's survival times a list of 56 genes were distilled from a potential 37000 gene probes to be consistently significantly expressed in relation to survival times from ovarian cancer measured from two separate populations of patients, on two microarray platforms measured in two laboratories.

The rigour of combining a meta-analysis approach with multiple testing using a variety of statistical approaches, increases power and confidence in the relevance of genes found to be of interest. Encouragingly, a number of the 56 were immediately recognised as having known association with ovarian cancer survival (*IGF2* and *BMP4* and more). Further investigation and validation of the genes that are not yet reported to associate with survival may have clinical relevance and have potential to predict a patient's response to treatment or be used as a novel target for therapy. These results warrant genomic and or proteomic validation, for example, using immunohistochemistry on tissue micro array. Moreover using the genes in combination with each other as a biomarker panel and clarifying the nature of these commonalities using more, freely available online resources such as STRING (Figure 33), KEGG, Reactome, BioGrid, Panther and HeTop. This could begin to unearth molecular pathways with potential to characterise and categorise the nature of an individual tumour and enable more tailored treatment.

The most ‘robust’ biomarkers remained.

5.4.1. Comparison of Results with the Data Source Publications

Based on a continuous prediction analysis the owners of the data have published a list of 86 genes they calculated expression to correlate with time of survival (Crijns *et al.,* 2009), some of which had not previously been linked to ovarian cancer. Crijns *et al.,* 2009 used a continuous prediction analysis to list 86 genes whose expression was strongly correlated with survival time. They were able to find and use data sets containing 57 of these genes for validation of their findings.

To validate the finding of any biomarker study the results must be reproduced with a different sample set or on a different technological platform. Crijns *et al.,* (2009) were unable to attempt to reproduce 31 of 88 genes found to be significantly associated with ovarian cancer survival times as the validation data sets though having twin experimental design produced data using a different microarray technology which did not contain probes corresponding to these genes. Equivalently the 56 gene of interest listed above are biased to those genes whose probes on both Operon and Affymetrix platform, the nature of the meta-analysis will have filtered out genes only present on one.

When compared to the accompanying publications of the data sets there were four gene overlaps; three from GEOD 13876 (Crijns *et al.,* 2009) , and one with GEOD 26712 (Bonome *et al.,* 2008) These are *LRRC17*, *TMEM45A* and *TRO* from Crijns *et al.,* (2009) and *GULP1* from Bonome *et al.,* (2008). Their appearance in this meta-analysis acts as a second or third step of validation for each marker from the point of view of each paper. However the lack of crossover is more poignant. The lack of association of the other 83 genes listed to be of interest by Crijns *et al.,* (2009) with survival, and 54 by Bonome *et al.,* (2008) exemplifies the point made by Braem *et al.,* (2011) and Devlin *et al.,* (2003) that the need for reanalysis and meta-analysis of existing data and how different data processing and analysis applied to the same data can yield different results. Completely different gene sets and numbers of genes can be shown to be significantly differentially expressed between two data sets depending on the data mining methods applied to the same data (Devlin *et al.,* 2003).

Interestingly commonalities with other larger meta-analyses were discovered (Ganzfried *et al.,* 2013 or Yang *et al.,* 2014), the results of which were published after the initiation and majority of the work described above (Chapter 5).

Ganzfried *et al.,* (2013) used the R statistical package to compile data from 2973 cases from 23 manually selected gene array data set. The emphasis of their study was to curate the larger resource, as it was not focused on biomarker discovery only reported the identification of one single marker (CXCL12) as an independent predictor of survival time.

The analysis above (Chapter 5) based on two data sets contributes as well as any comparable study, (Ganzfried *et al.,* 2013, Yang *et al.,* 2014, Crijns *et al.,* 2009 and Bonome *et al.,* 2008) and contributes to knowledge with novelty by its unique combination of samples and analytical methods. Meta-analyses utilising a larger number of sub-cohorts to increase sample number, such as Ganzfried *et al.,* (2013) n=2973, will experience detrimental effects consequent from extraneous variables introduced when combining cohorts from different sources such as, sample collection procedure, microarray platform or majority ethnicity. The base studies generating the data (such as Crijns *et al.,* 2009 and Bonome *et al.,* 2008) contain smaller, yet notably sized sample numbers inherently avoid such problematic extraneous variable influencing findings. The analysis in Chapter 5 falls between these two extremes. A larger sample number n=310 was achieved by the use of two data sets (157 + 153) which were carefully, manually selected. Genes of Interest (GOI) were ranked based on their performance in each cohort in parallel /discretely/simultaneously and the highest ranking taken forward for the comparison. Any GOIs that concord with similar studies findings will add a level of validity to an existing body of evidence implicating that genes role. Any discrepancy, are candidates for further investigation. If it were possible, a deeper investigation of each sample, data processing and analysis method used to draw each conclusion may in itself lead to identifying reveal new knowledge (e.g. if one cohort had a higher number of one ethnicity – ethnicity could be investigated as a potential effector variable) however lack of wider sharing and availability of raw data and software used to generate results prohibits this.

As discussed above the same data can be re-analysed to draw different conclusions depending on the analysis applied (Allison *et al.,* 2006). As Ganzfried *et al.,* (2013) combined numerous data sets generated from different microarray platforms a global normalisation would have been applied. Normalisation of data can itself influence downstream results depending on the method applied (Zyprych-Walczak *et al.,* 2015). It may be insight full to analyse the performance of each gene in each constituent cohort, to its performance across the combined, normalised data set as a whole.

5.4.2. Interpretation and Implications of Results

This chapter's analysis infers that the variation of survival times is a consequence of different genes activations acting to either make the tumour more aggressive or able to evade platinum based chemotherapy. Though there are numerous non-recorded uncontrollable extraneous variables that could also determine patient survival times, this assumption must be made in order to hypothesise and derive possible meaning from the results.

A key observation was that the *IGF2* gene, already accepted to be implicated in ovarian cancer was identified to be present in higher amounts in the short term survivors, together with a stimulatory ligand (Morcavallo *et al.,* 2014). This strongly implicates activation of the growth pathway downstream of IFG2 in the cancers from the short term survivors.

5.4.2.1. Known Mechanisms of Resistance to Platinum Based Chemotherapy

Proteins that are reported to play a role in platinum chemoresistance include Excision Repair Cross-Complementation group 1 (ERCC1), xeroderma pigmentosum complementation group F of the NER pathway; increased in acquired chemo resistance, and Mut and Mut associated proteins of the MMR pathway as reviewed in (Martin *et al.,* 2008). See section 5.1.1.

At a superficial level *ERCC1*, though representative probes were on both of the arrays was not found to be amongst the top ranking gene probes associated with survival times in the meta-analysis of two patient cohorts using these the above described analysis. This suggests that different/other cell signalling or/chemoresistance pathways are responsible for this particular difference in survival times observed in the above meta-analysis.

On closer examination, *ERCC1* was found to be in the significant portion (p=0.03744) of highest ranking genes from the Cox univariate analyses of the GEOD 13876 data set if censored cases are removed. This could be taken to suggest that in this patient subset the *ERCC1* mechanism of DNA repair was responsible for their shorter survival, however as recurrence data is not available no conclusion could be drawn. Or, it is possible that there is a subgroup within or outlier within the GEOD 13876 cohort influencing this. This was just a preliminary comparison, as discussed below a deeper investigation into all the genes associated with the ERCC1, Mut and any other known pathways of mechanisms of chemoresistance would be insightful, however is a separate top-down/reductionist/targeted type of data analysis entirely. To be done properly is beyond the scope and resources of the current study. Fishing for one gene of interest is not the appropriate use for this analysis data.

However, any conclusions or indications derived from comparison with current knowledge can be considered to be restricted. The main limitation being the majority of current ovarian cancer chemo resistance knowledge is based from studies of cell lines with acquired resistance (Marchini *et al,* 2013). Though insightful to delve into specific mechanisms, cell line models do not incorporate the heterogeneous character of tumours and the tumour microenvironment as variables, which are accepted to be a significant component of ovarian tumours. Many cell line studies fail to investigate the histopathological origins of the cell they are drawing conclusions from (Vaughan *et al.,* 2012).

The genes found to significantly associate with survival times in the above chapter were are linked to mechanisms of chemo resistance, for example EMT pathways; *EDNRA* (Rosanò *et al.,* 2011)

5.4.3. Support of the Methods Used

Incorporating false discovery testing to an analysis increases the confidence of any deductions, thus, yields results with a higher validity (Devlin *et al.,* 2003). False discovery refers to the phenomena that over any number of observed measurements a proportion of them will have been discovered by random chance. Strategies such as multiple testing, meta-analysis, or adding in decoy data mitigate false discovery, or by its measurement allows researchers to crop result to only include those with minimal probability of having occurred by chance. Gene array studies finding differential expression of a handful of interested genes have been criticised for insufficient hypothesis testing and rejecting the null hypothesis too readily (Devlin *et al.,* 2003). Encouragingly, the approach used in this study has been adopted by other researchers. Marchini *et al.,* (2013) used a similar filtering approach; a series of statistical analyses of different types to gradually refine a list of differentially expressed genes to investigate further.

Using more than one analytical method increased the rigor of a test, in this chapter we have used three different approaches to filter genes based on their association with survival time from ovarian cancer. The differing characteristics of each compile a stringent filter and enhance the meticulousness of the analysis/ made a really stringent shortlist.

Cox proportional hazard analysis was applied to determine if the continuous independent variable of each genes expression levels associated with survival time (see section .2.) The Cox univariate analysis added the capacity to include censored cases/incomplete data, which neither ANN nor T-tests have. ANNs are a form of machine learning applied to non-linear data to assess the predictive power of each variable. Thus adding a predictive element to the finalised list. (See section 2.2.2 for details of each). The T-test is a widely accepted test to assign a significance to the difference between two populations this process coupled with a trend analysis added an extra, fundamental/widely accepted level of confidence in the finalised genes.

Two of the approaches, ANN and T-tests are best suited to data with categorical variables where here they are applied to a continuous independent variable: survival time. Performing these analyses thrice at each of three definitions of what is long or short term survival (Figure 30) tailored these analysis for this purpose.

The observation that a number of the condensed list of 56 GOI contains genes and proteins already associated with cancer is encouraging: *TNFAIP6, TNFRSF14* and *WTAP* are implicated with tumours by name, *NAV3, SPAG9, SMC, IGFBP6* and more have all been reported or specifically implicated in cancer studies (Carlsson *et al.,* 2012, Garg *et al.,* 2008, Ghiselli 2006, Fu *et al.,* 2007), and some associate with ovarian cancer survival, namely, IGF2 (Sayer *et al.,* 2005) and *BMP4* (Shepherd *et al.,* 2008, Thériault *et al.,* 2007)

5.4.4. Criticisms of the methods used

Multicentre studies increase the opportunity for operator bias, even from sample collection. The meta-analysis in this this thesis limited sample numbers to only include samples that could be defended to be described as comparable, however, are they? From the documented evidence available, the samples considered for this analysis were comparable, however, if the full sample collection and patient information were scrutinised in depth there is likely a parameter that would separate them. Due to the nature of experimental design at the different centres there will always be some fundamental differences.

5.4.4.1. The Availability of Additional Information

Non-controlled, non-recorded extraneous or confounding variables that may have also influenced the patient's survival time from ovarian cancer. Namely overall health, smoking status or family cancer history. Despite having more than the MIAME requirements, limited information about the samples used, processing of samples and data acquisition was available. None of the known risk factors associated with ovarian cancer are available such as: *BRCA1* and 2 status, oral contraceptive use, parity and menopausal status (though this could be deduced by age).

Other factors that affect survival time that may differ between patients in the above data sets and patients in data sets used for validation, include the experience and expertise of the care givers of the patients in the two centres (Erickson *et al.,* 2014).

This data is only based on patients who were entered into a treatment pathway. Though based on a different population, a recent US report found that nearly half of patients diagnosed with ovarian cancer did not receive the 'standard' NCNN endorsed treatment pathway (Erickson *et al.,* 2014). Reasons for this included the overall severity of the condition combined with the average age of patient at diagnosis and co-morbidities. Thus, nearly half of the patients that this data appears to be drawn from are not genomically represented in the above analysis.

5.4.4.2. Challenges in Studying Ovarian Cancer

Survival time is a common measurement applied to assess treatment efficacy or subcategorise patient groups in clinical study. It is used as it is an accessible measure, however, can not include any number of extraneous factors effecting each patient's vitality, neither does it incorporate the quality of the life measured. There is still a need to address the way success in treatment is evaluated. As with this body of work, researchers are limited to the samples and variables recorded. Time of survival is a common measure, however, an increase in life span is still normally less than 5 years and not a solution or cure to the disease, the quality of life of a patient undergoing drastic treatment is not considered. A recent collaborative focus group (Vaughan *et* *al.,* 2012) called for the inclusion of quality of life and symptom benefit analysis to be included as a measure of success or parameter/ variable in such research.

According to Machin, Cheung and Parmar (2006), survival analysis has a predetermined ideal sample size which should ideally be determined prior to data collection. As this was a retrospective analysis of data this was not possible however the minimum ideal sample number, based on a log rank power model to achieve a power of 0.8 and a p-value of 0.05 is 65. Using the selection criteria of Stage 3 serous the data utilised in this analysis is nearly double the minimum. However, due to the relative rarity incidence of samples cohorts of such a size are rare. A cohort of comparable size consisting of only Stage 1 tumour would be of great scientific value but near impossible to acquire.

Ovarian tumours are known for their wide ranging reported cellular histology compared to other cancers. This may be due to the tumour microenvironment accounting for a larger proportion of the tumour burden, although this is not quantified (Vaughan *et al.,* 2012).

The data analysed above is from gene arrays are based on lysates of ovarian tissue, Crijns *et al.,* 2009 details the strict requirement for tumour tissue included in the study, however, even within the tumour microenvironment multiple cell types with different activated gene pathways are present. Even with these inclusion criteria in place, the genetic information available represents the entire tumour microenvironment, not just tumour cells, this includes the host reaction to tumour cells. Not all data sets available on ArrayExpress detail their inclusion criteria for tissue, comparing samples with an unknown mixture of both within them is not a fair comparison, and will increase the probability of creating irreproducible results. Laser capture micro-dissection prior to microarray analysis to extract tumour cells from the surrounding stoma cells increases the accuracy of the measurement to represent the genome of the cancer would to some extent mitigate this , as performed in Sayer *et al.,* 2005. However, is not always practicable, the yield from micro-dissected sections are comparatively small and the addition of an extra sample handling protocol may add variation.

5.4.4.3. The Array and Data Analysis Methodology

The data used was selected on criteria that it represented genes from the whole known human genome. However they do not contain probes to all known genes, additionally as the two were from different gene chip platforms a small proportion of genes were not represented on the other chip.

One challenge in meta-analysis of gene microarray data, also disapprovingly reported by Ganzfried *et al.,* (2013), is the disambiguation of in syntax, semantics abbreviations used to annotate variables in both the sample information files and the microarray platform annotation. For example the gene *TNFAIP6* has eleven possible aliases in the format of gene codes alone. On a smaller scale using a search and match function in Excel "*IGF2*" will not be matched to "*IGF-2*". Curated databases may have basic formats for information to be in before publishing however are not completely standardised.

It is commonly accepted that, database and clinical annotations formats need to be standardised to maximise their utility to meta-analyses (Array Express 2010, Wu *et al.,* 2014 and Carey *et al.,* 2008 in Ganzfried *at al.,* 2013). However they are each tailored to their own purpose and still may make using multiple resources a challenge.

In this chapter, a concordance strategy was applied to across multiple analyses to mitigate false discovery. This strategy was reasoned to be more appropriate than other adjustments for false discovery such as the Benjamini-Hochberg or Bonferroni corrections where the p-value threshold of significance is adjusted based on the number of variables and or sample size. Yang *et al.,* (2013) in a study of common cancers used p-values of ≤0.05 and did not correct for false discovery.

5.4.4.4. Downstream Analyses

STRING and all protein text mining interaction databases are inherently bias to well-studied entities/genes. As acknowledged by Yang *et al.,* (2013). The encouraging web of association observed in Figure 33 needs to be considered in context of the colour coding between each point and how it was generated within STRING. Yellow lines represent associations drawn from algorithm text mining, black and pink represent co-expression and experimental evidence. It is possible that all the genes represented as linked with yellow lines are not co-expressed/ have no relation to one and other. Additionally, it is possible that all the genes represented as separate are connected or part of a pathway/family but there is as yet no evidence. STRING analysis is not a result as such, more a tool to direct further investigation.

In summary, 56 genes were found to have a consistent significance throughout the meta-analyses. The chances of having focused on these genes by chance was is minimal due to the number of, and differing statistical filtration steps applied. The inclusion of some genes already implicated in ovarian cancer adds to the confidence in the strategy applied however the study is limited due to the data sets and information available. There are still multiple ways the existing data could be re-examined to add even more rigor to the existing analysis, or, derive different information.

5.4.5. Future work

Re-analysis of the data available

- All the gene probes could be collapsed into one per gene. This would reduce the influence of each probe.
- An interaction analysis could be performed so observe the influence of each genes of interest on all others in the data set. Findings would be interesting to compare to (Yang *et al.,* 2014), who in a similar study, found that the genes with the most prognostic power tended not to hub i.e. be influential of the expression of large numbers of other genes.
- **Hypothesis lead data mining.** Prior analyses could give evidence that the genes found to be of interest share a common characteristic. All gene probes could then be separated and grouped by an ontological annotation i.e. cellular location, function, or pathway involvement, then then each category could be given an overall significance score based what proportion of the genes of interest are has a significant T-test, Cox univariate survival or high predictive performance. In particular; if pathway annotations for both sets of gene probes were available for example, it would be possible to deduce a really informative conclusion. Which ontological category is overrepresented in the set of genes that significantly associate with survival times? For example: DNA repair / angiogenesis pathway contained x% probes that significantly associated with survival, whereas housekeeping pathway had none. However, though possible there is no short way to annotate the gene probes from the two platforms (Affymetrix and Operon) with common pathway search terms. Also see Table 2 for discussion on interaction and pathway databases. Marchini *et al.,* (2013) exemplifies hypothesis lead gene data mining whilst investigating chemo-resistance in ovarian cancer; analysis of quantitative

PCR and microarray analysis of a test set of tumour tissue generated a list of significantly relevant genes. A hypothesis was derived from this list that EMT /MET signalling pathways were significantly associated with chemo resistance in ovarian cancer. Following this, signalling pathway category annotations were assigned to their gene interest list using Reactome and KEGG to assess which pathways are overrepresented in each listed pathway. This approach can be criticised as being bias, commonly reported and investigated genes will be better documented an more likely to come up in a pathway analyses than their newly discovered counterparts. Constant re-use and basing conclusions from databases of existing pathway records such as KEGG may drown out newly discovered interaction and pathways.

- Evaluation of existing evidence. Namely Transferrin, Vitronectin and ApolipoproteinA1 as published by Nosov *et al.,* (2009), though these markers are for early detection of the disease there role in survival. Their lower expression was shown to negatively correlate with early disease detection. Are their increased expression associated with improved survival? (though not replicated by this study).
- Re-divide the full genome to pathways and see which have the highest proportion as significant to survival time however this is currently not feasible due to no ready/ fast way of assigning pathway to each gene probe and ensuring it is officially correct.
- Molecular pathways involved in angiogenesis are increasingly well characterised and implicated in survival from and incidence of ovarian cancer (Zhang *et al.,* 2003), a deeper analysis of the genes in these pathways expressions relation to survival time could make an interesting contribution to this body of research.
- When using online databases of information, great care must be taken to ensure they are being used both to their full potential and appropriately so not to misunderstand the information displayed in them. In the example of STRING there is little room for error conducting a search however the interpretation of the results must be done with the known caveats that, well reported genes will display more connectivity, the database is based on algorithms literature searching the internet i.e. not curated by specialists. Connections between proteins categorised “databases” or “text mining” is vague and may have little scientific meaning. Additionally understanding who has sponsored or curated a search engine or database may influence the interpretation of results. This is often the explanation as to the reasons behind differing findings from different databases.

- A current barrier to the progress of the work is the lack of factual, accepted circuitry diagrams of heterotypic inter and intra cellular interactions, identified by Hanahan and Weinburg (2011) to be an obstacle to current research, but they also predict these circuitry diagrams to develop exponentially over the next decade. Some progress is already noted and discussed (chapter 5 and 6.1) and however not yet truly centralised. Although until these exist in a 100% validated format development of these database are a reductionist approaches that do not truly represent biology. Current existing versions such as KEGG are based from multiple sources of evidence, it could be argued that it is over simplistic impossible to accurately to knit together information separate research sources into something as intricate and individual and variable as a cellular molecular pathway. Obviously, these resources need to be developed somehow but it is up to the scrutiny of a researcher to understand the strength/foundation of these resources whilst under construction.

In the above chapter a number of calculations, interpretations and analyses of the of two microarray data sets are made. These could still be reanalysed using any combination of known or novel data mining strategies. Evidence from other sources could confirm or refute such observations.

5.5. Conclusion

Within the listed boundaries of what was available for analyses, 56 genes were found to significantly and, consistently associate with survival from ovarian cancer. The meta-analysis tactic means these findings are less likely to be biased by sample cohort, collection centre, gene array platform or statistical analysis computation and the risk of false discovery is reduced. The RNA expression micro array platform by nature limited this discovery to a known gene to which a predesigned oligonucleotide probe existed on both arrays used, so a novel onco-sequence was never a potential discovery despite being likely occurrence. Additionally, although expression of mRNA implies translation into protein expression would occur this is not guaranteed, and if it does then any number of post translational modifications could take effect between this measurement and an onco-phenotype. None the less, these findings need to be confirmed on other patient samples and technological platforms.

6. Validation Strategies

Chapter Abstract

A meta-analysis in chapter 5 ranked probes from genome wide microarrays by their relevance to survival time from ovarian cancer. A potential list of 37632 genes were whittled down using increasing levels of stringency into a list of 56 with an encouraging body of evidence indicating they are expressed at different levels between samples from patients with short and long survival times.

In this chapter, after reintroducing and defining validation and verification, the list is further refined based on additional evidence of their differential expression. Firstly each of the genes of interest generated in chapter 5 were individually verified on a larger sample set of gene microarray data using Kaplan Meier Plotter, a freely available online resource which accesses a range of additional microarray data sets. This verification step reduced the list of 56 to 7 genes with observations corroborating the findings of the meta-analysis in chapter 5. Verification of the translated proteins of these 7 genes could give insight into their role at a cellular level within the tumour environment. One of the candidates *EDNRA* was selected to be the first for verification at a protein level by immunohistochemistry on a tissue microarray of ovarian tissue. Significant trends association the expression of EDNRA with cancer stage, grade and histology are observed. The merits, limitations and direction for further research are discussed.

6.1. Introduction

As eluded to in the introductory chapters (section 1.1.3.1.) for any potential biomarker (defined as “a naturally occurring molecule, gene or characteristic by which a particular pathological or physiological process, disease, etc. can be identified” Oxford Dictionaries, 2015) to progress from a discovery stage to a clinical setting it must first undergo challenging and rigorous stages of peer review, verification, validation before clinical trials to satisfy firstly the scientific community regarding discovery methodology, then the clinical community for the safety of is application in testing the general public (de Gramont *et al.,* 2014, Henry *et al.,* 2012 and Goossens *et al.,* 2015), for this reason there are few new fully approved biomarkers.

The word “validation” has been used with different meaning between clinical or research settings (Suresh *et al.,* 2011, Halling *at el.,* 2012). It is therefore poignant to clarify where the results from chapter 4 and 5 fall in this continuum. The verb “validate” means to confirm or substantiate: in a research laboratory setting, a discovery generated from one methodological platform are reported as validated when reproduced on another, however from a wider perspective reproducing results on a second platform is actually “verification” of a finding, in that is not conclusive but supportive evidence warranting further research. Full “clinical validation” requires the rigor of multistage clinical trials including thousands of samples and multiple methodological platforms.

Figure 34 (below) adapted from the National Cancer Institute (NCI 2016) portrays biomarker discovery as a continuum roughly divided into three stages. This encompasses both the clinical and academic/research perspective of either a stage 1 approach for discovery or a stage 2 approach for clinical validation. By definition to verify is “to make sure of” or “demonstrate an accuracy or truth” in practicality verification step biomarker discovery is an analytical validation.

Like Braem *et al.,* (2011) The Office of Cancer Clinical Proteomics also recognise the imbalance in the past decade of thousands of reported biomarker discoveries compared to the handful of those that are clinically validated for any cancer, not just ovarian. As discussed in section 3.6 they also imply the reason for this to be due to the current status/model of research at the moment. Although larger multicentre cohorts or meta-analyses exist, the majority of the 1000s of biomarker discoveries derive from independent research groups operating independently across the globe with seemingly large but in-fact insufficient sample numbers which are the maximum resources available to them.

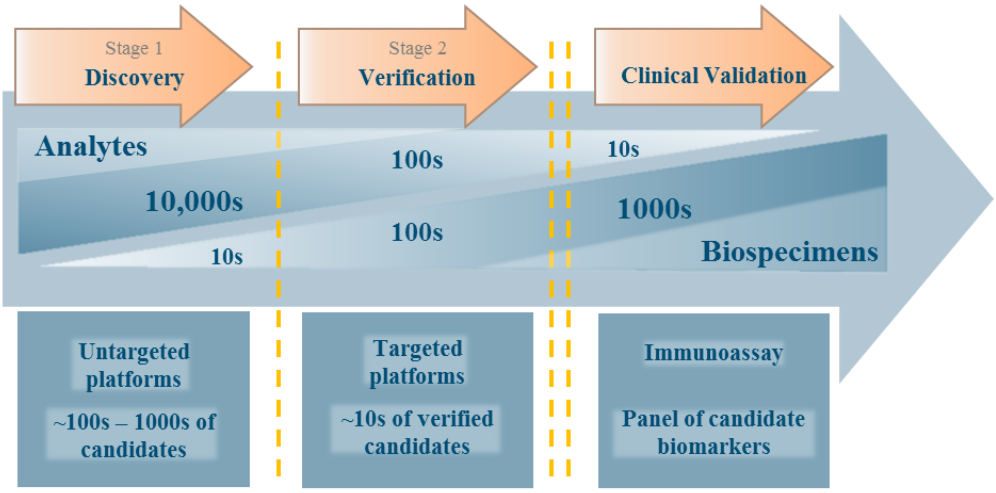


Figure 34. The Biomarker Discovery to Validation Pipeline.
Adapted from NCI, 2016). Stage 1 discovery stage data is commonly generated in a research lab from a non-targeted bottom-up approach of thousands of potential variables on low sample numbers. Stage 2 Verification experiments entail hundreds of samples on a more targeted platform able to measure 10 – 100s of variables. Finally if evidence to substantiate a clinical trial results, the potential individual or sometimes panels of biomarkers are clinically validated via the standard four phase clinical trials, on thousands of prospectively collected patient samples.

Figure 34 above summarises the biomarker discovery to validation pipeline as outlined by the NCI (NCI 2016). Stage 1 discovery stage research is commonly generated in a research lab from a non-targeted bottom-up approach of thousands of potential variables on low sample numbers such as; gene chip microarrays, mass spectrometry, 2D-PAGE perhaps on cell lines, or animal models. Variables indicating signatures with potential diagnostic value are taken forward to scale-up experiments to verify findings. Stage 2 verification experiments entail hundreds of samples often sourced from clinical patients conducted in a research laboratory on a more targeted platform able to measure 10-100s of analytes more accurately for example ELISA, Western blot, targeted mass spectrometry, PCR. Finally, if evidence to substantiate a clinical trial results the potential biomarkers may undergo full four phase clinical validation on prospectively collected clinical samples as part of a four phase clinical trial.

It is at this point and in this context worth highlighting the rarity and value of the large cohort of samples collected prospectively for the UKCTOCs ovarian cancer screening study (Jacobs *et al.,* 2004, Menon *et al.,* 2014, Jacobs *et al.,* 2015), this is a rare example of wide scale standardised sample collection. So far, these samples have been used for their primary purpose; the clinical validation of the Risk of Ovarian Cancer Algorithm (Jacobs *et al.,* 2015).Additionally, banked surplus samples have been utilised for the discovery and verification of more ovarian biomarkers (Russell *et al.,* 2016). Successful completion of such a collaboration of this scale in this context is rare (NCI 2016), the value of these specimens should be highlighted as specimens are suitable for all 3 stages of the biomarker discovery pipeline above, the samples and any data acquired from them should be utilised to its maximum potential.

The analyses described herein (chapter 6), progress the Stage 1 potential biomarkers discovered in chapter 5, through Stage 2 analytical validation alternatively termed biomarker verification and their clinical implication investigated. Firstly the biomarkers discovered in chapter 5 are filtered for accuracy by validation on the same technological platform they were discovered: gene microarray expanded to a considerably larger sample cohort, the consequent/ filtered verified GIOs are investigated at a protein level using immunohistochemistry.

Encouragingly, the 56 genes of interest generated from Chapter 5 include both known and novel candidates associating with ovarian cancer survival. Notably, overexpression of IGF2 in ovarian, and many cancers in general, is well documented. IGF2 and BMP4 are both independent predictors of survival in ovarian cancer (Sayer *et al.,* 2005, Laatio *et al.,* 2011). Increased IFG2 ligand binding/activation is seen in ovarian cystic fluid (Kanety *et al.,* 1996), which eventuates inactivation of molecular pathways key to cell invasion (Lee *et al.,* 2005). IGFBP6 and IGFBP3 are part of these pathways and the latter is downstream of a p53 cascade, TP53 being described as a “near-invariant” feature in ovarian cancer (Bowtell *et al.,* 2015). BMP4 is a recognised mediator of ovarian metastasis and cell invasion (Thériault *et al.,* 2007), overexpression indicative of poor prognosis, and, has been implicated in cisplatin Chemoresistance (Laatio *et al.,* 2011). Others, namely NAV3, WTAP and MAPK, have been discovered in or implicated in cancers of other organs, but less so for ovarian (Carlsson *et al.,* 2012, Little *et al.,* 2000 and Wagner *et al.,* 2009). In a large scale system-level meta-analysis of genomic data from The Cancer Genome Atlas (TGCA) TNFAIP6 and NAV3 amongst a panel of 15 other genes were implicated in ovarian cancer survival and possibly other cancers namely glioblastoma breast and kidney cancer in the context of their signalling pathways (Yang *et al.*, 2013).

6.2 Validation of Genes of Interest using Gene microarrays

6.2.1. KM Plotter Introduction

The Kaplan Meier Plotter (KM plotter) is a free meta-analyses web tool which allows researchers to assess the performance of their biomarkers *in silica* on a manually curated data bank of multiple cohorts of micro-array data (Gyorffy *et al.,* 2103 and KM Plotter 2014).
The curators source their databank from published micro array data and created a simple web interface, researchers enter the gene of interest and selected relevant parameters to tailor the sample subset to their needs. The service is available for breast (n=4142), lung (n=1464) and ovarian cancer (n=1715), although the final sample number is dependent on the parameters selected.

Hypothesis

H_1-viii. Genes of interest will be verified to significantly associate with ovarian cancer survival time when investigated on a wider sample cohort.

H_0-viii. None of the genes of interest found to significantly associate with ovarian cancer survival time will be verified to do so when investigated on a larger sample cohort.

6.2.2. Kaplan-Meier Methods / Utilisation Strategy

A broad perspective was taken with regards to sample selection. The largest number of available samples were included in this analysis. This is primarily; due to the lack of ability to manipulate and select cases within the separate samples sets in the KM plotter.

The gene code of each of the genes of interest was entered into the web-tool. For the first layer of analysis (round of elimination) the default search parameters were used. These were; to produce a report for all available gene probes, to auto-select the best cut off to separate short and long term survivors and to measure the output in progression free survival. No restrictions were made based on stage, histology, grade, CA125 levels, debulking status or chemotherapy/treatment pathway. The latest version of the web tool (version 2013) was used. One thousand one hundred and seventy one samples were available using these parameters.

A report was generated for every available gene probe, some probe gene codes only had one available probe set. The output was available and saved as a .pdf of the Kaplan Meier plot and a .txt file of what is reportedly the raw plot data.

KM plotter gives a score for the reliability of the gene probes in the array that represent that gene, this is based on their criteria; probes are ranked as excellent, medium or poor. For clarity and to minimise any ambiguity in this analysis any probes ranked by KM plotter as “poor” were recorded by default however discounted for the purpose of evaluating each genes validator performance.

Genes of interest were ranked by how many probes separated short and long term survival that matched the pattern observed in the analyses in chapter 5, i.e. higher expression associated with shorter term survival. The number of probes available for each gene of interest were considered. Those which were represented by more than one probe were given extra weight when choosing which were the best candidates for protein validation.

See Appendix (Digital Appendix B) for evaluation of each gene and Table 15 for the summary of this evaluation.

As a second round of analysis the exact process was repeated however with the patient group split cut-off set at median. This is a more stringent evaluation of the performance of the gene. A report generated from a gene represented by multiple gene probes may have all been significant predictors of survival using different time point cut offs. However if they are also all significant using the same cut-off this is a better representation of the whole gene.

See Table 15 for a summary report of the performance of each of the genes of interest from chapter 5 ability to predict progression free survival in the KM plotter assembled cohort described above.

KM plotter does not optimise the cut point based on p-value, at best it can find the best of five (lower quartile, lower tertile, median, upper tertile, upper quartile). For this reason the p-value presented may not represent the optimal p-value for each probe. Given these uncertainties, and other caveats discussed above and below a p-value of 0.05 was used as the criteria to label a result “of interest”. A Bonferroni correction was considered (resulting in a p-value of 0.00089) of which most probes still pass; however some may be falsely excluded. More weight was placed on the qualitative aspects of the comparisons; number of probes, trends matching hypothesis and probe quality. The results presented here are as is without correction or further exclusion.

As a third round of analysis the above process was repeated with filters applied. These were selected to best fit the two data sets in chapter 5. Overall survival with a follow up threshold of 5 years, Stage 3 and 4, serous with possible platin therapy.

The sample cohorts utilised in KM plotter are; GSE14764 (n=80), GSE15622 (n=35), GSE18520 (n=63), GSE19829 (n=28), GSE23554 (n=28), GSE26712 (n=195), GSE30161 (n=58), GSE3149 (n=116), GSE9891 (n=285), TCGA (n=565).

6.2.3. Results and Discussion of KM Plotter Reports

The first, low stringency, round of analysis removed genes whose probes expression did not match the trend observed in the data from chapter 5 and did not distinctly separate short and long term survivors using any group cut off point. A lenient p-value of <0.05 was used to maximise the number of genes that could remain in the analysis, given the caveats of using KM plotter, which are discussed below, the hierarchy of qualitative considerations were deemed to be of more value to evaluate a genes performance. Using the auto-select best cut off function, KM plotter does not reveal which group cut off was used for each individual probe leaving the possibility that they may be different from each other. Hence, information from this round may not appropriately represent, and is of limited use to assess, the performance of genes represented by multiple probes. For genes who had multiple probes strongly separating short and long term survival groups using an undisclosed group cut off the process was repeated using the median as a set group cut off for all probes. It is possible that the genes that were discounted at this level probes best performing cut off points were all the same but not the median, but this was not investigated further at this stage. The genes whose median expression used as a cut-off point to group patients support the findings of chapter 5 are *DCN, EDNRA, IGF3, NAV3*, *TNFAIP6*, *WTAP* and *PPFIBP1*. Genes that strongly supported trends in chapter 5 yet were only represented by one probe are *BMP4, COLEC12, GJB1, GLT8D2, LDB2, MFAP4*, *OLFML3*, *PDZRN3*, *PJA2, RARRES2, TMEM45A, TNFRSF14, ZFHX4*.

Using the high stringency approach *MAP4K4* was the only gene represented by multiple probes, and *GJB1, GLT8D2, LDB2, MFAP4, OLFML3, PDZRN3, PJA2, RARRES2, TMEM45A, TNFRSF14, ZFHX4* by single probes.

Table 15. **Summary of KM Plotter Analysi**s. The performance of each gene of interest (rows) in KM plotter is noted against increasingly stringent criteria (columns), a tick marks the criteria has been met.

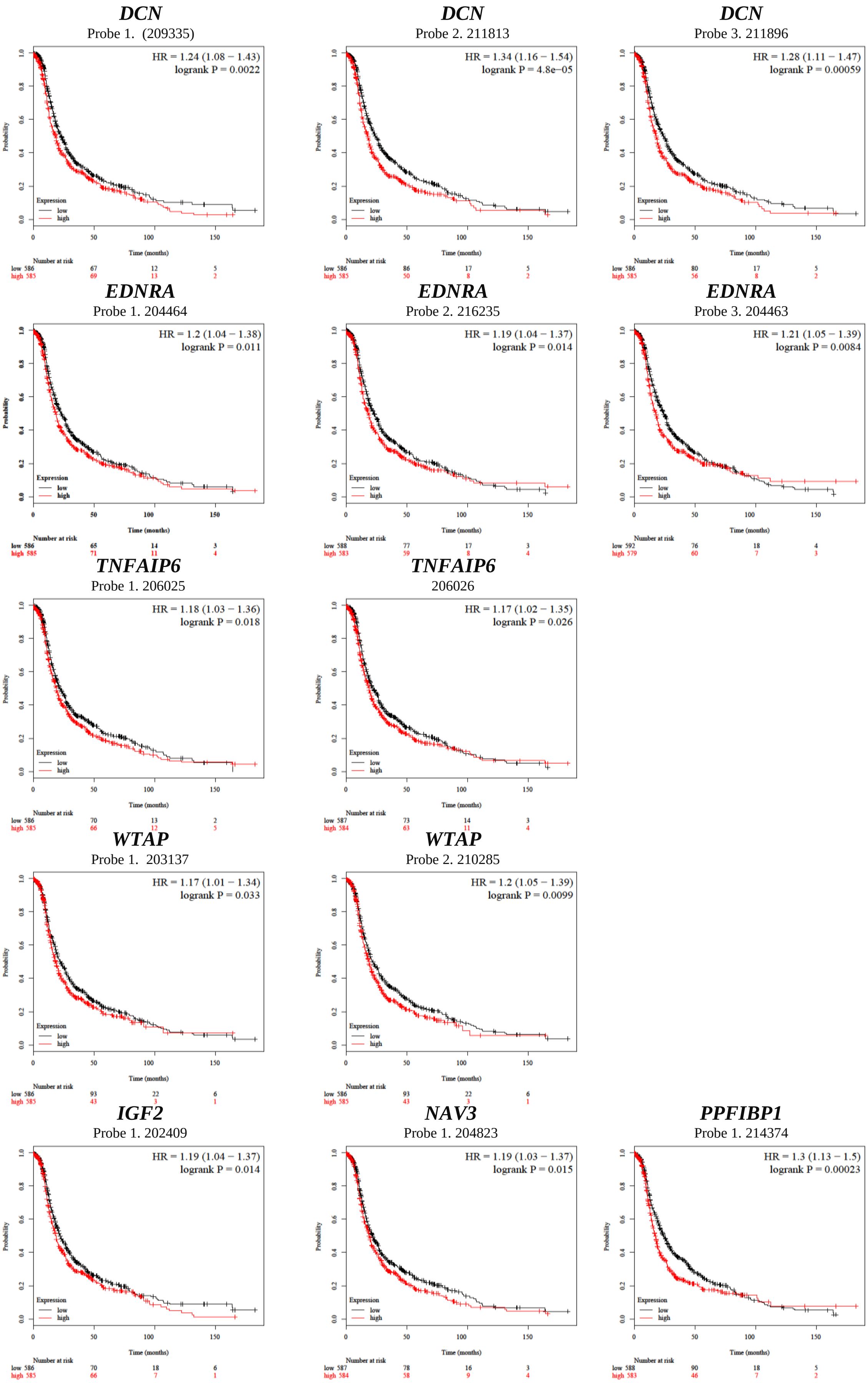
Comment [Gene Symbol]	Is there more than one probe set available?	Is the same trend in correlation with short or long term survival observed in all "reliable" probe sets for the gene? AND Does the trend match that observed in the meta-analysis?	Is there a strong p-value for all available "reliable" probes? Using any cut-off	Is there a strong p-value for all available "reliable" probes? Using median cut-off	Score
DCN	✓	✓	✓	✓	4
EDNRA	✓	✓	✓	✓	4
IGF2	✓	✓	✓	✓	4
NAV3	✓	✓	✓	✓	4
TNFAIP6	✓	✓	✓	✓	4
WTAP	✓	✓	✓	✓	4
PPFIBP1	✓	✓	✓	✓	4
GJB1		✓	✓	✓	3
LDB2		✓	✓	✓	3
OLFML3		✓	✓	✓	3
PDZRN3		✓	✓	✓	3
PJA2		✓	✓	✓	3
TMEM45A		✓	✓	✓	3
GULP	✓	✓	✓		3
IGFBP3	✓	✓	✓		3
INTS5	✓	✓	✓		3
SMC3	✓	✓	✓		3
FYN	✓	✓			2
HNRPDL	✓	✓			2
MAP4K4	✓	✓			2
PPP3CA	✓	✓			2
SEMA3C	✓	✓			2
SPAG9	✓	✓			2
TPM2	✓	✓			2
ZNF45	✓	✓			2
BMP4		✓	✓		2
COLEC12		✓	✓		2
GLT8D2		✓	✓		2
MFAP4		✓	✓		2
RARRES2		✓	✓		2
TNFRSF14		✓	✓		2
ZFHX4		✓	✓		2
BACH1	✓				1
CTBP2	✓				1
DCTD	✓				1
EFNB3	✓				1
FZD7	✓				1
H2AFV	✓				1
HBD	✓				1
NCOR1	✓				1
PKD2	✓				1
SCAMP1	✓				1
SFRP4	✓				1
SLC11A2	✓				1
SPCS3	✓				1
BACH2					0
CDC25B					0
CLIP3					0
FHOD3					0
FKBP14					0
HIST1H3C					0
IGFBP6					0
LRRC17					0
NDN					0
PCDH17					0
PTPRE					0

Table 15 above lays out how GOIS were ranked based on the consistency of the significance in their performance under increasingly stringent criteria.

Genes were eliminated or progressed based on; the number of probes available to represent their identity, the reliability of these probes (as evaluated by KM plotter), if the trend of expression difference observation noted matches all the probes for each gene from the KM plotter cohort, and that of the data generated in Chapter 5.

This evaluation was performed in two rounds, firstly using the "find best cut-off" option, meaning different probes may have been found to be of interest based on different survival time cut-offs. This was considered and although not specified by the software can be deduced based on the number of cases assigned to each group, which is shown. The second round was repeated using the median cut-off only. This narrowed this list to a manageable number to take forward to the next round of validation/verification.

Figure 35 displays the Kaplan Meier plots exported from KM plotter for each of the finalised genes. These were produced at the final elimination round where the median cut-off was applied to all.



DCN
Probe 1. (209335)

DCN
Probe 2. 211813

DCN
Probe 3. 211896

EDNRA
Probe 1. 204464

EDNRA
Probe 2. 216235

EDNRA
Probe 3. 204463

TNFAIP6
Probe 1. 206025

TNFAIP6
206026

WTAP
Probe 1. 203137

WTAP
Probe 2. 210285

IGF2
Probe 1. 202409

NAV3
Probe 1. 204823

PPFIBP1
Probe 1. 214374

Figure 35. Kaplan Meier Plots of the Highest Ranking Genes of Interest.

From the KM plotter validation all using a median cut off. Only high quality probes were considered. Strong differential expression between long and short term survivors was seen probes for the following genes *DCN* p= 0.0022, p=0.0000045, p=0.00059, *EDNRA* p=0.0011, p=0.014, p=0.0084, Two for *TNFAIP6* p=0.018 and 0.026, *WTAP* p= 0.033 and 0.0099, *IGF* p=0.014, *NAV3* p=0.015 and *PPFIBP1* p=0.00023

Figure 35 above visualises the significance of the difference in survival time of patients categorised by the expression of the GOIs in their tumours.

6.2.3.1. Cautions and Caveats Considered when Interpreting KM Plotter Data

KM plotter is a powerful tool however has limited functionality and its calculations are closed, that is, researchers input a question based on limited selection variable parameter selection and a statistical output is given. The rigidity of the analysis meant that several cautions and caveats were considered alongside the data generated from KM plotter.

An important point to consider was if the microarray data of the multiple cohorts accessed by KM plotter was truly representative of the target population. Here, a broad perspective was taken with regards to sample selection. This means the samples used are not as tightly controlled as in chapter 5, where a thorough search of available data sets on Array express was performed and the two data set with the most comparable data patients selected. All available samples were included in this analysis, this is not ideal as different cancer stages, grades, histologies and treatment pathways are inadvertently/involuntarily included. The effect of this was limited by the decision to measure the output as progression free survival and not overall survival as this would have exacerbated the effect of effective/successful treatment pathways. This is primarily due to the lack of ability to manipulate and select within the separate samples sets. Certainly, the next step to fully validate these findings would be to download and explore all the data sets that are included in KM plotter however the data sets are sourced from various repositories, not all are freely available, and the scale of this is beyond the remit of this project. A specific example of the scale of the task is the KM plotter has the option to search each sample cohort separately, however, each gene code (of which there are often multiple aliases) must be manually entered and searched separately. This is very time consuming and may be a more worthwhile exercise in an investigation with a different, more specific, hypothesis for example an investigation of a cell-signalling pathway exploring how a gene is orchestrating a response. A limited number of genes could be manually tested in each selected data cohort ant the concordance of their performance observed.

This analysis is bias towards the genes represented by multiple gene probes within the arrays. For the current purpose of selecting a gene for protein validation, the simplest standard of multiple probes yielding the same trend was used to best indicate up or down-regulation of the gene. Future experiments could bring insight into genes only represented by one gene probe, or for genes with probes yielding highly significant yet conflicting trends, it is possible that these represent mutations or single nucleotide polymorphisms, however, verification of the suggestion of mutation is outside the remit of this project. Genes represented by one gene probe may be highly biologically meaningful however did not have the potential to perform well in the high stringency analysis. This shows how one probe is not best to represent the gene and would need to be addressed by a different array design, or, next generation sequencing (section 2.2.1.2).

This verification could be considered subtly bias as GEO26712 was not able to be removed from the sample set, however, although exactly how many overlapping cases is unknown and not able to be found out, the potential problematic subset can only account for a small portion of the total number n=1171. Evidence also strongly indicates that when the filters to stage were applied the GEO26712 samples were removed. Curiously, this raises concern as to why the curators of the database interpreted the data differently, as this information is clearly stated in the publication this data set was generated from (Bonome *et al.,* 2008). This may suggest the database curators had access to more sample information than what was available to this study, or interpreted the given information in a different way. An unexpected benefit of discovering this was that it makes a cohort of n=808 that does not include GEO26712, a selection process the web tool was not able to do deliberately. From this observation and testing, *MAP4K4* was the only gene with multiple probes that strongly separated short and long term survivors when the median cut off was applied. This could suggest that *MAP4K4* is an indicator of poor overall survival in this particular sample subset: stage 3 and 4 serous ovarian cancers treated with platins. If more time was available repeating this on all the iterations of sample subsets may add to building a profile of its relevance to each category available i.e. gene stage, grade or treatment pathway.

Five categories of group cut off are available. The web tool does not indicate which samples are allocated to each category when each cut of is applied. An “auto select best cut off” option is available and was utilised here, however it is possible that the best performing place, or most appropriate place to divide the cohort is in between one of these 5. A desirable future feature for the web tool may include more iterations of cut off values, or even a sliding scale.

It is also worth noting that at least one of the data sets available in KM plotter cohorts met the stringent criteria and were considered for the meta-analysis in chapter 5, however, rejected on the basis of lack of availability of sample information. The major shortcoming in using KM plotter to verify the findings of chapter 5 is that no considerations and adjustments to sample data are traceable/visible.

Nonetheless, KM plotter provided a platform to perform a powerful analysis to begin the verification of observations generated in the meta-analysis of Chapter 5. Data has been compiled to evaluate and refine the list of GOIs to a manageable number for the next available method of verification.

6.2.3.2. Indications for Further Research

Encouragingly, the finalised genes include those whose expression are already reported to associate with ovarian cancer survival, such as *IGF2* (Sayer *et al.,* 2005, Kanety *et al.,* 1996, Lee *et al.,* 2005) *NAV3* and *TNFAIP6*. Strong association of the expression of *NAV3* and *TNFAIP6* and survival time supports findings of Yang *et al.,* (2013) who in a systems level analysis found these genes amongst 13 others to be prognostically significant in the context pathways involved in multicellular organisational development.

From this data mining *EDNRA* and *DCN* gathered the strongest evidence to take forward for further validation; although their probes did not have the highest significance values, they were both represented by multiple probes graded as high quality, the measured expression of all of these show a consistent and strong difference for all to separate long and short term survivors. The next best performing GOIs; *TNFAIP6*; *WTAP*; *IGF2*; *NAV3*; *PPFIBP1* were represented by fewer (two or one) probes.

6.3. Translational Validation Strategy using Immunohistochemistry

Rationalisation for selection of one GOI for Immunohistochemical validation.

The interest list was reduced from 56 using a larger cohort of gene array data (Chapter 6.2.). The members of the short list could vary depending on the choices of stringencies applied used however put succinctly, *IGF2, WTAP, TNFAIP6, NAV3, DCN, PPFIBP1* and *EDNRA* were of interest as their expression was verified to correlate with survival time from ovarian cancer in an larger, independent sample. Progressing from these genomic findings, evidence of translated protein in ovarian tissue were considered.

Two tissue microarrays of ovarian cancer were available to this project. One sourced from a clinical collaborator, an additional TMA was also sourced from Biomax. The former had survival data that would make comparable analysis of protein expression to the analysis of gene microarray data in chapter 5, the latter was the largest commercially available TMA containing ovarian cancer and accompanying information about cancer stage and grade, however lacked patient survival data. The clinical TMA would provide a better validation strategy however delays were encountered using this TMA inhibiting its use in this project.

Although the visualisation of protein expression matching all finalised genes of interest would provide an informative, comprehensive/ holistic validation of the finding so far and insight for further work. It was unfortunately only feasible to investigate one marker further at first.

A report was compiled for each of the seven finalised GOI. Various data sources were mined for more information namely GeneCards (GeneCards 2013) and The Human Protein Atlas (Pontén *et al.,* 2008, Uhlen *et al.,* 2010, Human Protein Atlas., 2014). GeneCards (2013) was particularly useful for this purpose, its comprehensive compilation of all aliases decreased the chance of overlooking important information related to each gene due to incomplete nomenclature information. Additionally the summarised representation of numerous other databases allowed insight into associated genes and proteins, genetics, domains, function, cellular location, ontology, known pathways, pharmacology of associated drugs, othologs, paralogs, variants, associated disorders and publications. A thorough review of all of these was outside the remit of this project and again many may have been overlooked entirely if it were not for the continual growth of this resource. The holistic view of information allowed hypothesising and insight of the genes of interest in a wider setting than the ovary or cancer tissue.

Table 16 below, lists each of the finalised seven GOI pseudonyms and briefly outlines how different nomenclature can hinder a thorough literature search and how information could be overlooked if it were in not for attempts to catalogue and centralise information. There are at least 11 aliases for each of the seven genes of interest, the number of search terms increases from 7 to over 77.

Table 16. Aliases of the Finalised Seven Genes of Interest. (GeneCards 2015)

Gene Code	Full Name	Aliases
DCN	Decorin	Decorin Proteoglycan, Bone Proteoglycan II , SLRR1B, PG-S2 , CSCD, PG40, Dermatan Sulphate Proteoglycans II, Small Leucine-Rich Protein 1B, Proteoglycan Core Protein, DSPG2, PGII, PGS2.
WTAP	Wilms tumour 1 associated protein	Wilms Tumour 1 Associated Protein, Wilms Tumour 1-Associating Protein, Female-Lethal(2)D Homolog, WT1-Associated Protein, HFL(2)D, Putative Pre-MRNA Splicing Regulator Female-Lethal(2D, Wilms Tumour 1-Associating Protein, Pre-MRNA-Splicing Regulator WTAP, PNAS-132, KIAA0105, MUM2.
IGF2	Insulin-like growth factor 2	Insulin-Like Growth Factor, T3M-11-Derived Growth Factor, IGF-II, Insulin-Like Growth Factor 2 (Somatomedin A), Chromosome 11 Open Reading Frame 43, Insulin-Like Growth Factor Type 2, Insulin-Like Growth Factor II, Somatomedin A, Somatomedin-A, C11orf43, PP9974.
TNFAIP6	Tumour Necrosis Factor, Alpha-Induced Protein 6	Tumour Necrosis Factor, Alpha-Induced Protein 6, TNF-Stimulated Gene 6 Protein , Hyaluronate-Binding Protein, TNF Alpha-Induced Protein 6, TSG-6, TSG6, Tumor Necrosis Factor Alpha-Inducible Protein 6, Tumour Necrosis Factor-Stimulated Gene-6 Protein, Tumour Necrosis Factor-Inducible Gene 6 Protein, Tumour Necrosis Factor Alpha-Induces Protein 6.
NAV3	Neuron Navigator 3	Neuron Navigator, POMFIL1, Pore Membrane And/Or Filament Interacting Like Protein, Unc-53 Homolog, STEERIN3 , KIAA0938 , Unc53H3, Pore Membrane And/Or Filament-Interacting-Like Protein 1, Steerin 3, Steerin-3.
PPFIB1	PTPRF interacting protein, binding protein 1 (liprin beta 1)	PTPRF Interacting Protein, Binding Protein 1 (Liprin Beta 1), Protein Tyrosine Phosphatase Receptor Type F Polypeptide-Interacting Protein-Binding Protein 1, PTPRF-Interacting Protein-Binding Protein 1, HSGT2 , Protein-Tyrosine Phosphatase Receptor-Type F Polypeptide-Interacting Protein-Binding Protein 1, Liprin Related Protein-1, Liprin-Beta 1, Liprin-Beta-1, KIAA1230, HSgt2p, SGT2, L2,
EDNRA	Endothelin receptor type A	Endothelin Receptor Type A , HET-AR, ETA-R, ET-A, ETRA, ETA , Endothelin-1 Specific Receptor, Endothelin Receptor Subtype A, G Protein-Coupled Receptor, Endothelin-1 Receptor, Endothelin A Receptor, ETAR,

Insulin growth factor 2 (IGF2) is a growth factor from the insulin family, has already been reported in colon lung and breast cancer (Sayer *et al.,* 2005, Kanety., *et al.,* 1996). Higher gene expression of IGF2 has already been shown to be a poor predictor of survival in ovarian cancer (Sayer *et al.,* 2005).

Sayer *et al.,* (2005) lists four possible mechanisms which lead to higher levels being seen in short term survivors of ovarian cancer. These are; alteration of downstream binding proteins, loss of transitional suppression or increased transcriptional activation and loss of imprinting. IGF2's synergistic interactions with endothelin's and the endothelin axis which themselves have been characterised in the growth and neovascularisation of a number of cancers namely ovarian (Nelson *et al.,* 2003).

IGF2 overexpression has been shown in relation to the cancer progression (Sayer *et al.,* 2005), thus it is less likely to be a good biomarker of the early stage disease and current investigations are underway investigating its potential to guide treatment by predicting poor responders who develop a resistance to platinum therapy. Evidence from cell line studies have shown IGF2 as a possible therapeutic target (Sayer *et al.,* 2005).

Endothelin receptor type A (EDNRA) is the primary receptor for endothelin-1. Activation of EDNRA initiates G protein coupled receptor (GPCR) mediated activation of phosophatidylinositol-calcium second messenger system. In animal studies it has been shown to be involved in, but not dependent on ovulation via its activation of Endothelin-2 (EDN2) (Bridges *et al*., 2010). Endothelin 1 is reported to be involved in long lasting vasoconstriction, polymorphisms in EDNRA are believed to be linked to migraines and scarlet fever (GeneCards 2014). More relevantly, expression of EDNRA is reported to be associated with poor survival from ovarian cancer, cell line studies have directly implicated activation of EDNRA with EMT in ovarian cancer cell lines. Rosanò *et al.,* (2011) detail endothilin-1 activation of EDNRA as having anti-apoptotic results via downstream activation of the PI3K Akt pathway. In an earlier study the group report the receptor agonist endothelin-1 to be present in high concentrations in ascites (Rosanò, *et al.,* 2003). Both *in vivo* and *in vitro* inhibition of EDNRA increased sensitivity to platinum chemotherapies (Rosanò *et al.,* 2011) making it a potential drug target. Antagonists of EDNRA have been shown in vitro and in vivo to increase tumour cell apoptosis (Rosanò, *et al.,* 2003 and 2006) and inhibits mutagenic effect (Rosanò, *et al.,* 2003), however, has not emerged from clinical trial.

Decorin. There are multiple known splice variant transcripts for the *DCN* gene. The protein coded for is a small peptidoglycan found intracellularly and as part of the extracellular matrix. Decorin is involved in normal cellular matrix assembly binding to collagen fibrils. Decorin has been shown to be expressed in healthy human ovarian stroma and granulosa cells around ovulation and in increased amounts in the corpus luteum at the time of mensuration and the levels detected were dependent on hormone signalling (Adam *et al.,* 2012) Despite being listed as having growth suppressor activity on tumour cell lines (GenCards 2014, Nash *et al.,* 1999) a recent review has compiled evidence from several sources/studies implying that Decorin is actually a tumour promoting factor itself, or by interaction with members of pathways involved in cell survival, cell growth, metastasis, and neovascularisation/angiogenesis (Bi and Yang 2013). This juxtaposition/conflict of evidence may be explained by Bi and Yang (2013) being a collaboration of a wider body of information, whereas Nash *et al.,* (1999) evidence is based solely on two cell lines. In another cell line study, Sherman-Baust *et al.,* (2003) isolated cisplatin resistant ovarian cancer cell lines and explored collagen IVs role in chemoresistance, decorin alongside other genes were found to be contributing factors. More recently in a functional genomic study investigating myofibroblast from gastric cancer, increased amounts of DCN, alongside other proteins, were found to be consequent after stimulation by growth factors, one of which was IFG2, and that loss of ability to regulate secretion of these were associated with advanced cancer (Waugh *et al*., 2015). Similarly, the direct interaction of Decorin as a ligand to IGF2 has been demonstrated (Morcavallo *et al.,* 2014). Most relevantly, in a quantitative proteomic analysis of ovarian cancer specimens sourced from primary debulking surgery, listed decorin, alongside 44 other proteins, to have significantly higher expression in the cases that would go on to develop chemo resistance (Pan *et al*., 2009) this evidence is most complementary to the findings above where increased expression of gene DCN significantly associate with a poorer prognosis from ovarian cancer. This evidence could also be used to support the suggestion that the shorter survival observed in chapter 5 and 6.2 is indeed due to chemo resistance. Further to this some of the additional co-expressing proteins of interest listed by Pan *et al*., (2009) were found to be significant in the meta-analysis of chapter 5 however eliminated in a round of increasing statistical stringency when refining to the genes with the strongest association with survival, namely IGFBP2.

NAV3. The protein coded for by the Neuron Navigator 3 (NAV3) gene is expressed pronominally in cells of the nervous system. The NAV proteins as a group are involved in cytoskeletal dynamics (Carlsson *et al.,* 2012) the *C.elegans* equivalent of human NAV3 is responsible for axon guidance during neuronal growth. Several point mutations in the NAV3 gene have been identified which are missense and lead to increased or decreased activity. Maliniemi *et al.,* (2011) demonstrate its silencing in keratinocytes events in upregulation of up to 20 genes involved in inflammation cell signalling and associates mutations in NAV3 to Basal and Squamous cell carcinomas. Deletion of the *NAV3* gene correlated with increased tumour metastasis (Carlsson *et al.,* 2012). Bleeker *et al,* (2009) narrate *NAV3* amongst a wider list of genes found to have mutated variants in melanoma and pancreatic carcinoma but not glioblastoma. Carlsson *et al.,* (2012) demonstrate two mechanisms decreased NAV3 expression could promote colorectal tumour growth; firstly, colorectal cancer cells became less susceptible to growth control mechanisms and secondly became more sensitive to cell growth signals from inflammatory signals. Though this is at odds with the trend observed in the current data (derived from ovarian cancers) which shows increased expression of normal *NAV3* associating with poor survival, is likely attributed to, thus not worth commenting on, the different tumour origins colorectal and ovarian.

PTPRF Interacting Protein, Binding Protein 1 (Liprin Beta 1) (*PPFIBP1*) gene codes for a protein which is a member of the LAR protein-tyrosine phosphatase-interacting protein (liprin) family. This family of proteins are known to interact with transmembrane protein tyrosine phosphatases who are involved in mammary gland development and axon guidance. PPFIBP1 has been shown to bind to and inhibit a calcium binding protein implicated in tumour invasiveness and metastasis (GeneCards 2014, NCBI 2015) and binds 14-3-3 a known onco-protein (Benzinger *et al.,* 2005). Most relevantly *PPFIBP1* was listed amongst over 100 genes to be significantly differentially expressed between 19 serous papillary ovarian tumours and 15 controls as measured by microarray (Bignottii *et al.,* 2006). However this particular link was not followed up in any subsequent publication.

Tumour Necrosis Factor, Alpha-Induced Protein 6 (*TNFAIP6*) gene codes for a protein of the same name, a member of the hyalouron-binding protein family. This domain is involved in cell migration and extracellular matrix stability. TNFAIP6 expression is induced by tumour necrosis factor alpha and interlukin-1, pro-inflammatory cytokines, and makes its inflammatory effects via its interaction with inter-alpha-inhibitor (GeneCards 2014, NCBI 2015). Despite its name it is involved in the normal function of the ovary in the expansion and fertility of the oocyte (Irving-Rodgers and Rogers 2006). *TNFAIP6*, alongside *NAV 3* and 13 other genes, was included in Yang *et al.* (2013) in a list of genes with high centrality gene expression network analysis in a wide reaching study of glioblastoma, breast, kidney and ovarian cancer. Zhang *et al.,* (2011) finds a strong correlation between the expression of TNFAIP6 and YAP which was demonstrated to confer the characteristics of chemo-resistance in ovarian cell lines although its association was not statistically significant on its own after correcting for multiple testing this observation was still noteworthy.

Wilms tumour suppressor gene (*WTAP*) codes for the protein Wilms tumour suppressor associated protein expressed across the nucleoplasm, implication in transcription and post-transcriptional regulation of other cellular genes via mRNA splicing regulator (GeneCards 2014). WTAP specifically interacts with Wilms Tumour suppressor gene (WT1), Wilms tumours are a rare paediatric condition in which nephroblastomas occur as a consequence gene mutation causing abnormal development of the kidney or genitourinary system (Little *et al.,* 2000). Jin *et al.,* (2012) identify *WTAP* to be overexpressed in glioblastomas as well as demonstrating via knockdown experimentation its regulatory role in cell migration and invasion. *WTAP* has been included in a panel of genes representing stem cell characteristics in a study aiming to isolate subpopulations of breast cancers (Christgen *et al.,* 2007). Helleman *et al.,* (2006) listed *WTAP* as of interest in a study investigating genes responsible for chemo-resistance in ovarian cancer, however was not shortlisted it for further investigation.

For an ortho- and ontologic summary, each of the short listed GOIs were searched using the KEGG web tool (see section 2.2.3) the results of which are summarised in Table 17. Of the 7 GOIs three (*DCN, EDNRA* and *IGF2*) were included in at least one functional pathway. Pathways relating to cancer were adapted and included below:

Table 17. KEGG Pathways found to be Associated with the Shortlisted Genes of Interest. Bold text indicates any which include cancer and are adapted below. Genes are in alphabetical order.

Gene Code	KEGG Orology Code	KEGG Pathway Code	Participant Pathways
DCN	K04660	ko04350 ko05205	TGF-beta signalling pathway **Proteoglycans in cancer**
EDNRA	K04197	ko04020 ko04022 ko04024 ko04080 ko04270 ko04924 **ko05200**	Calcium signalling pathway cGMP-PKG signalling pathway cAMP signalling pathway Neuroactive ligand-receptor interaction Vascular smooth muscle contraction Renin secretion **Pathways in cancer**
IGF2	K13769	**ko05205**	**Proteoglycans in cancer**
NAV3	89795	None found	No KEGG pathways listed
PPFIB1	None found	n/a	
TNFAIP6	K19018	None found	No KEGG pathways listed
WTAP	None found	n/a	No KEGG othologs listed

Both DCN and IGF2 pathways were categorised as Proteoglycans in Cancer: DCNs pathway code Ko05205 is subcategorised as Chondroitin sulphate/ Dernatab sulphate proteoglycan (GSPG/DSPG) (Figure 36), IGF2 orthologue code ko05205 is subcategorised as of Heparan sulphate proteoglycans (HSPGs) (Figure 36). EDNRA pathway code is ko05200 under the category Pathways in Cancer.

The following figures have been adapted from the KEGG web tool (KEGG 2015), the limitations of which have been mentioned above - that the reductionist view to construct such diagrams from currently available data cannot yet accurately represent biology. Never the less, putting the genes of interest into context of current knowledge known pathways holds some insight: the colour coding of the pathway end points is that of the hallmarks of cancer diagram from Hanahan and Weinberg (2011). Green referring to sustaining proliferative signalling, brown; evasion of growth suppressors, black; activation invasion and metastasis, blue; immortality, red; angiogenesis, grey; resisting cell death, purple; deregulating cellular energetics, pink; avoiding immune destruction orange; tumour-promoting inflammation and blue genome instability.

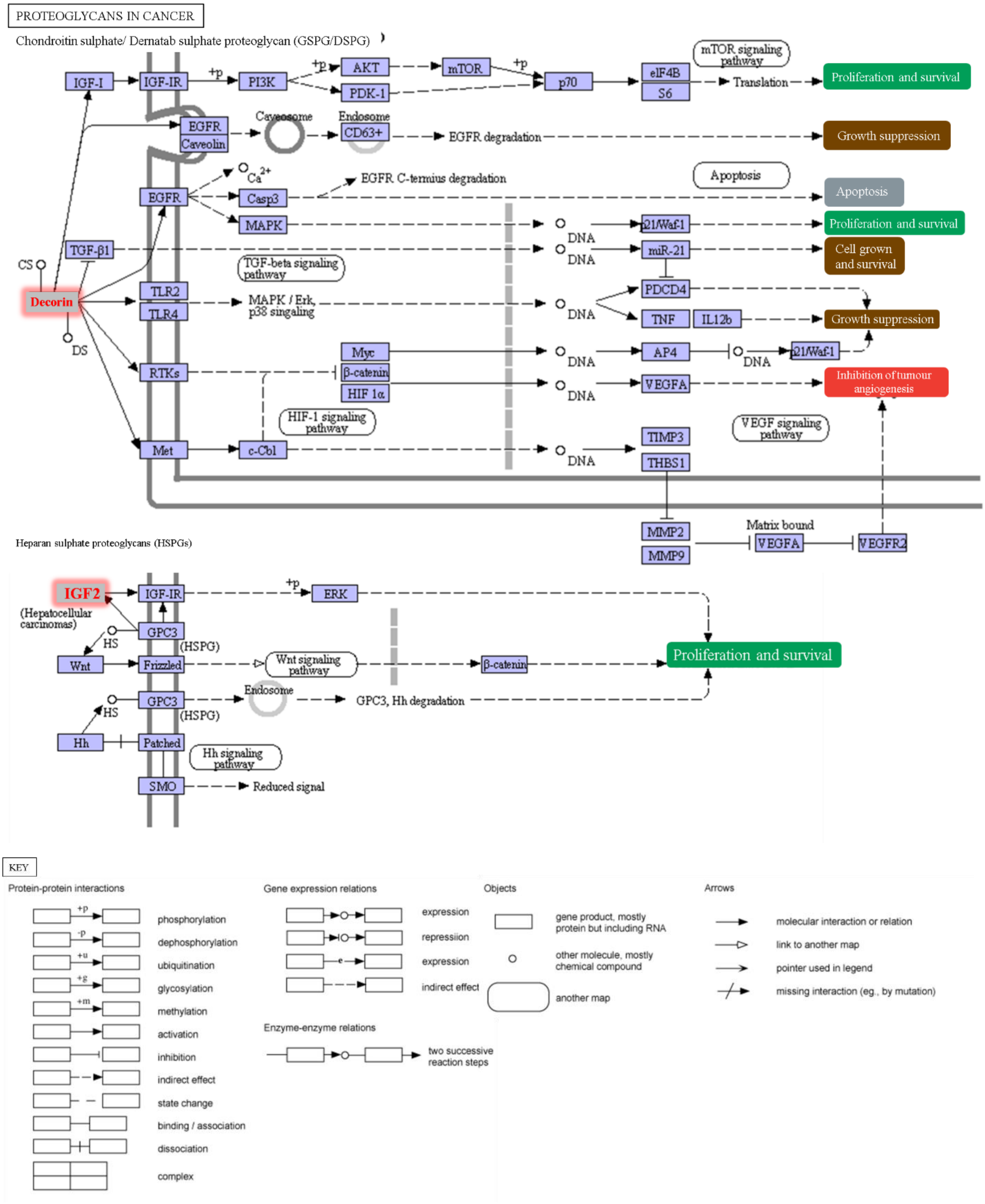


Figure 36. Verified GOIs Location in the Proteoglycans in Cancer pathway.
Adapted from KEGG 2015, Genes of interest have been highlighted in red text. Hallmarks endpoints that are relevant to cancer have been highlighted with white text and a background linking the characteristic to a hallmark of cancer. In-between is known cellular processes as annotated by KEGG (2016). Refer to the key above for detail:

The KEGG pathway figures depict the direct interactors with the GOIs as well as the consequent downstream effects in the context of cancer pathways. The reader is reminded that any gene or protein or factor depicted in the diagram is not limited to the interactions seen. For focus, these are KEGG pathways of cancer only.

Direct interaction of the GOIs: In the top left of Figure 36, Upon ligation, DCN (red text) is seen to activate IGF-1, TLR2 and 4, RTK's and Met, or to have an inhibitory effect on TGF-B1. IGF2 (red text lower left of Figure 36) binds and activates IGF-1R or GPC3. As Figure 37 (below) is an adaptation of the KEGG Pathways in Cancer map; the tag EDNRA (red text) has been added onto the map in the place of a g-generic protein coupled receptor (GPCR) label that could be any or all of EDNRA, EDNRB, BDKRB2 or BDKRB1. These are shown to be ligated by Thrombin, Endothelin 1 or Angiotensin II but all effect by activation of GNAQ or GNA11 a common transducer of transmembrane signals

In the context of cancerous phenotypes: IGF2 signalling events to increased cell proliferation and cell survival cell signal cascade involving direct interaction and activation with IGF-1R which downstream signalling events in phosphorylation of ERK, a mitogen activated protein kinase, which has multiple actions causing a cellular activation cascade as mentioned above and documented in Hanahan and Weinburg (2011).

DCN is also shown to activate cell survival and proliferation via either indirect activation of IGF-1R, PI3K then the mTOR signalling pathway, or by activation of EGFR, MAPKinases. Confusingly activation of the same pathway also initiates apoptosis, evading apoptosis is a hallmark of cancer. Finally DCN is seen to inhibit the VEGF angiogenesis signalling pathway, both by activation RTK to inhibit the activation of B-cadherin which would stimulation production of VEGFA, or by activation of THBS1 inhibiting MMP2 oro 9 then consequently VEGFA. This mixture of tumorigenic and anti-tumour characteristics echoes the conflicting findings of reported by Nash *et al.,* (1999) and Bi and Yang (2013) when trying to understand the role of DCN in ovarian cancer.

Figure 37 The GPCR denoted by EDNRA is seen to transduce an extracellular signal to intracellular signal cascade ending with sustained angiogenesis and cell proliferation. The cascade of activated proteins/ molecules triggered by EDNRA ligation diverges at several points however converge along the Ras – Raf - MEK - ERK cell activation pathway where phosphorylated ERK phosphorylates c-Jun, c-Fos, c-Myc or Ets1 which in turn glycosylates either VEGGF, MMPs or IL8 to stimulate angiogenesis, or, Cyclin D1 CDK4 to add to cell

proliferation signals. RAS also directly activates RASSF1 mutations in which have been linked to downstream mechanisms evading apoptosis

All three KEGG figures highlight the complexity of systems of gene, protein and small molecule interactions they portray, and how being forced into a reductionist model for focus on one or a few genes or molecules to represent a large system for biomarker discovery is challenged to represent the full picture.

Additionally, cropped from Figure below, KEGG has the function to limit the view of the pathway to one based on evidence for some select, specific cancers; colorectal, pancreatic, glioma, thyroid acute myeloid leukaemia, chronic myeloid leukaemia basal cell sarcoma, melanoma, renal cell carcinoma, bladder cancer, prostate cancer, endometrial cancer small cell lung cancer and non-small cell lung cancer. Ovarian cancer has not yet been mapped in this way. All KEGG figures here are generic to cancer.

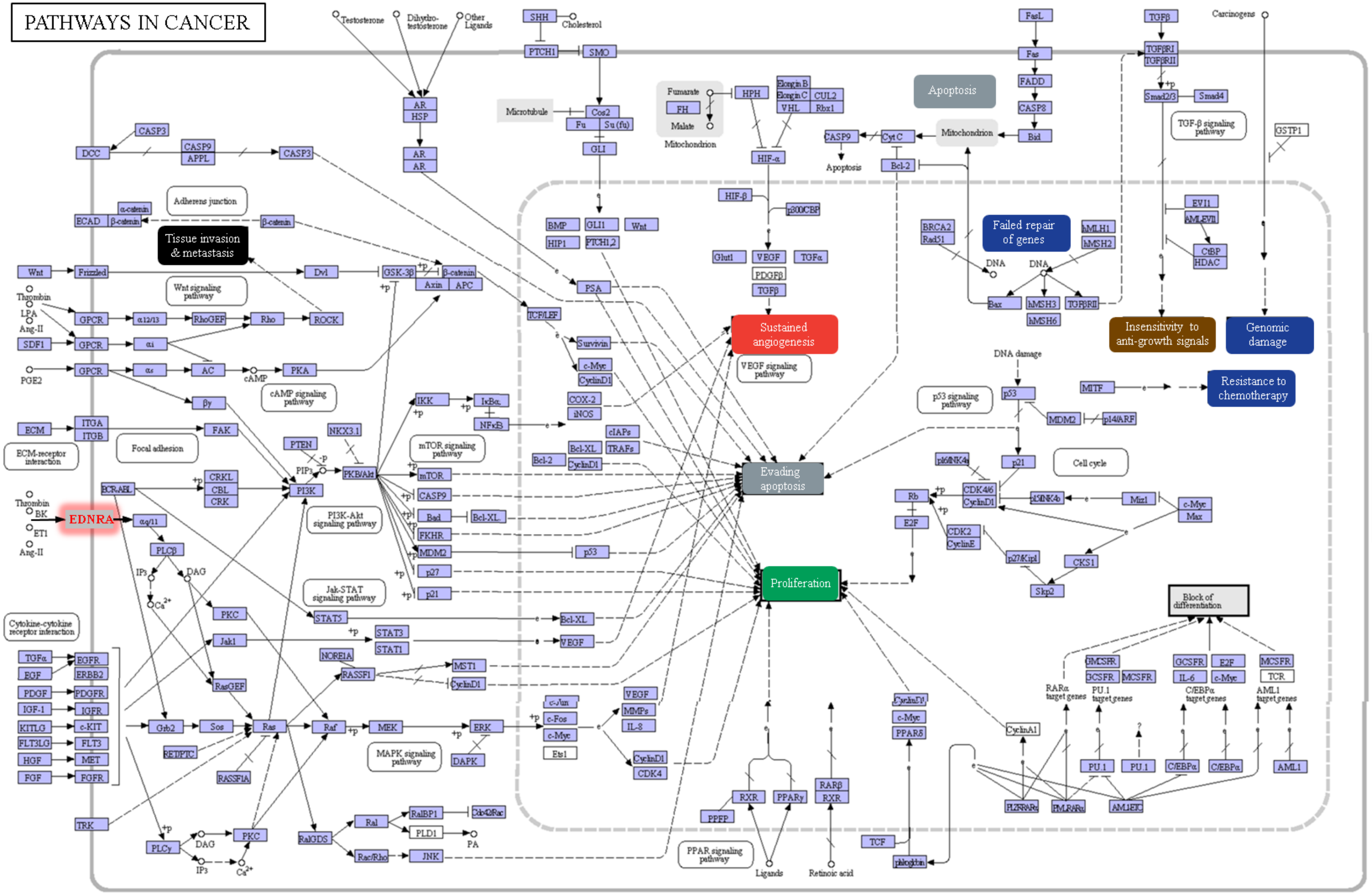


Figure 37. Verified GOIs in the Pathways in Cancer pathway. Adapted from KEGG GOI and Hallmarks have been highlighted as in Figure 36 above.

In a wider, bioinformatic study of microarray data akin to that in Chapter 5, Zhang *et al.,* (2013) conducted network analysis of gene expression of ovarian cancer, two of the genes found to be of interest *DCN* and *EDNRA* overlapped with the findings however focused on a different gene to verify their findings, *FBN1*. One data set used by Zhang *et al.,* (2013) overlapped with those used above. This could be interpreted as promising, further validation of the differential expression of DCN and EDNRA together based on evidence from other methodologies and analysis techniques.

From this brief review, Decorin and EDNRA were the favoured candidates for preliminary future work. The Human Protein Atlas (Pontén *et al.,* 2008, Uhlen *et al.,* 2010, Human Protein Atlas., 2014) was searched for existing experimental evidence of protein expression of the respective genes, specifically in normal ovarian, ovarian cancer or any other cancer.

EDNRA was finalised as the available antibody contained the most validation literature with it. Proof of the specificity of an antibody to its ligand increases the confidence any further findings infer.

Additionally, it was noted that EDNRA has several approved and one experimental antagonist drugs (GeneCards 2016, Drug Bank) where DCN only has one experimental listing. Non-of the PubMed links for any of the EDNRA drugs/compounds were considered with ovarian cancer although in a slightly expanded search, acetylsalicylic acid (Aspirin) (the mechanism of action on EDNRA is not listed) has been extensively investigated epidemiologically protective effects against several cancers (Tavani *et al.,* 2000). These provide a direction for future work in the event that protein validation were to yield evidence supporting a link to ovarian cancer survival. If EDNRA were over activated in some ovarian cancers hypothetically this could be exploited: competitive antagonism of could block the binding of the native stimulatory ligand Endothelin-1, stifling the angiogenic and cell proliferative downstream effects adding to Rosanò *et al.,* (2011) cell line study reporting inhibition of EDNRA renders cell more susceptible to chemotherapy.

EDNRA was rationalised to be the first candidate for verification of its protein expression in ovarian tissue: IGF2 is already well documented in ovarian cancer (Kanety *et al.,* 1996, Sayer *et al.,* 2005), as is BMP4 (Thériault *et al.,* 2007, Laatio *et al.,* 2011) thus further experimentation was unlikely to yield novel information in relation to ovarian cancer. Thus EDNRA would provide interesting addition to current knowledge however.

A common theme observed when collating information of the 7 genes of interest was EMT. The phenomena EMT, in which well differentiated epithelial cells revert back to a less differentiated form has previously been implicated cisplatin based drug resistance in ovarian cancer in cell line studies (Rosanò *et al.,* 2011). The exact role of EMT in ovarian cancer and cancer in general is not fully characterised. Conflicting evidence has come to light in ovarian cancer, the presence of known EMT markers E-cadherin and SNAIL have been linked with invasive phenotypes of ovarian cancer (Park *et al.,* 2008), whereas Miow *et al.,* (2014) found that ovarian cancer cell lines with a more mesenchymal status were more susceptible to cisplatin than their epithelial counterparts.

6.3.1. Protein Verification of EDNRA in Ovarian Tissue

Hypotheses:

H_1-ix. Protein expression of EDNRA will be found to be different between different stages, grades and histologies of ovarian cancer samples.

H_0-ix. No difference in protein expression will be observed between different stages, grades and histology's of ovarian cancer.

6.3.2. Immunohistochemistry Method

EDNRA was selected through a review of literature and web based databases (see section 6.3 and Table 16 above). Anti-EDNRA antibody HPA014087 (Atlas Antibodies Stockholm, Sweden 2012) was sourced via the Human Protein Atlas Website (Human Protein Atlas, 2014), where, antibody quality and specificity is requisite. Product information document included data from IHC and western blot using a human cell lysate to demonstrate the antibodies quality and specificity (Atlas Antibodies 2012).

The largest commercially available tissue MicroArray of ovarian cancer samples was purchased from Biomax (OV6161 from, UD Biomax inc). Further information is available at http://www.biomax.us/tissue-arrays/Ovary/OV6161 (Biomax 2014). This high density microarray contains 616 specimens of paraffin embedded ovarian tissue mounded onto one glass slide. It contains; 280 cases of adenocarcinoma of varying stage and grade; 28 normal or normal adjacent tissue, 13 clear cell carcinomas and 1 transitional cell carcinoma.

Alongside appropriate controls the TMA was deparaffinised then dehydrated by heating to 60°C for 10 min using a hot plate, then within 10 min the slides were transferred into a rack to fit the Leica Autostainer XL and pre-set function automated the following washes with gentle agitation and rinsing: two 5 min Xylene, followed by three two min washes in Industrial Methylated Spirit finished with 5 min in ddH$_2$0. For antigen retrieval, slides were boiled in in a citrate buffer (pH 6). Following this, slides were gradually cooled and gently re-introduced into ddH$_2$0 where the slides were carefully and manually loaded into the Sequenza staining system. As per the manufacturer's recommendations care was taken to ensure no part of the slide ever dried, no micro-air bubbles appeared between the slide and the cover and a check was conducted that the speed the liquid drained past each slide was in a similar rate for all of the slides. The Novolink Polymer detection system (RE7200-CE, Leica Biosystems, Bucks, UK) was used for staining: This consisted of the following incubations each interspersed with two 5 min washes of tris-buffered saline (TBS). Non-specific binding was minimised with room temperature 5 min peroxidase block, an 80 min room temperature incubation with the primary antibody HPA014087 (Atlas Antibodies Stockholm, Sweden) at a 1 in 40 dilution exposed the antigen to the antibody, a 30 min room temperature incubation in the post primary reagent containing secondary antibodies with conjugate peroxidase enzyme amplified the signal from any antigen-bound primary antibody. The antigen-bound primary-secondary-conjugate construct catalysed the hydrogen peroxide into a localised brown stain during a 5 min room temperature exposure to freshly prepared 1 in 20 dilution of diaminobenzidine (DAB) working solution. Finally a 6 min incubation with the haematoxylin reagent as a counterstain enabled visualisation of nucleic material and cell architecture. After staining the slides were fixed using sequential alcohol washes in reverse order as those described above for de-waxing and de-parafinisation. A glass coverslip was sealed on with DPX Mountant for permanent storage and histology.

The concentration of the Anti-EDNRA primary antibody was determined by a prior optimisation experiment conducted on incomplete offcuts of a breast tissue TMA and one additional test slide of ovarian tissue purchased from Biomax. The negative control processed alongside omitted primary antibody using purely antibody diluent in its place. This control ensured any staining observed in the test slide was a consequence of the primary antibody.

To meet the criteria for the stained TMA to be scored a range of staining intensities was seen in tumour cells in cores across the slide. Cores had to contain 100 tumour cells to be considered viable for scoring. A trained technician blindly scored the TMA on a categorical basis assigning a number to the overall intensity of stain seen in the tumour (0 negative, 1 weak, 2 moderate and 3 intense). The scores were blindly validated and accepted by a pathologist familiar with ovarian malignancies who scored a proportion and found a significant consensus.

6.3.3. Results and Discussion of Immunohistochemical Staining of EDNRA in Ovarian Tissue.

To recap, a tissue microarray containing 616 cores of ovarian tissue was immunohistochemically stained for the coded protein of EDNRA; a gene whose expression was found to significantly associate with ovarian cancer patient survival time. A full range of staining intensity was observed across the TMA (Figure 38). The inclusion of appropriate controls ensured that the intensity of stain observed could be interpreted as the relative expression of the EDNRA protein.

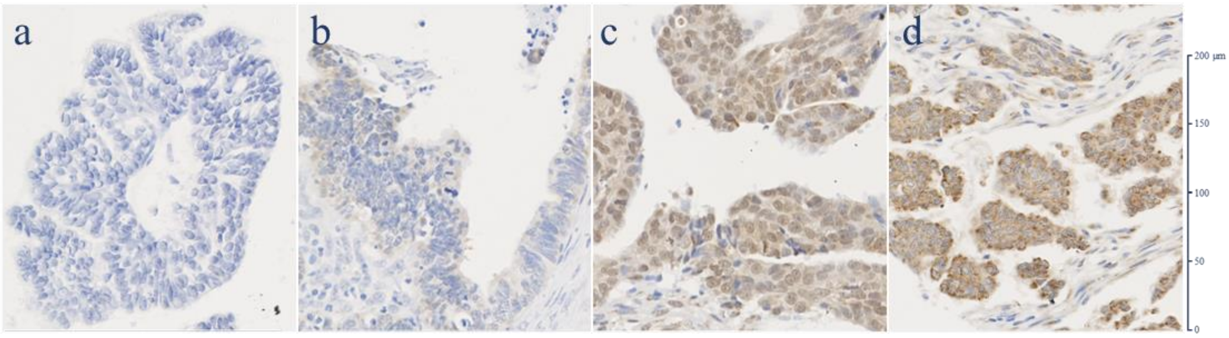


Figure 38. Range of Staining Intensity Observed in EDNRA Stained Ovarian Tumour Tissue from the Biomax OV6161 TMA.
a) b1 adenocarcinoma scored as no stain seen b) b3 serous adenocarcinoma scored as weak c)b28 serous papillary adenocarcinoma scored as moderate d)b17 serous papillary adenocarcinoma scored as strong

In brief, EDNRA protein expression was clearly increased in the later stage, higher grade disease (Figure 39,Figure 40 and Table 18 and Table 19). When separated by histology it was also an apparent difference in expression, however inconsistent sample numbers limited the conclusions that could be drawn from this. Additionally it was noted that distinct staining patterns were seen.

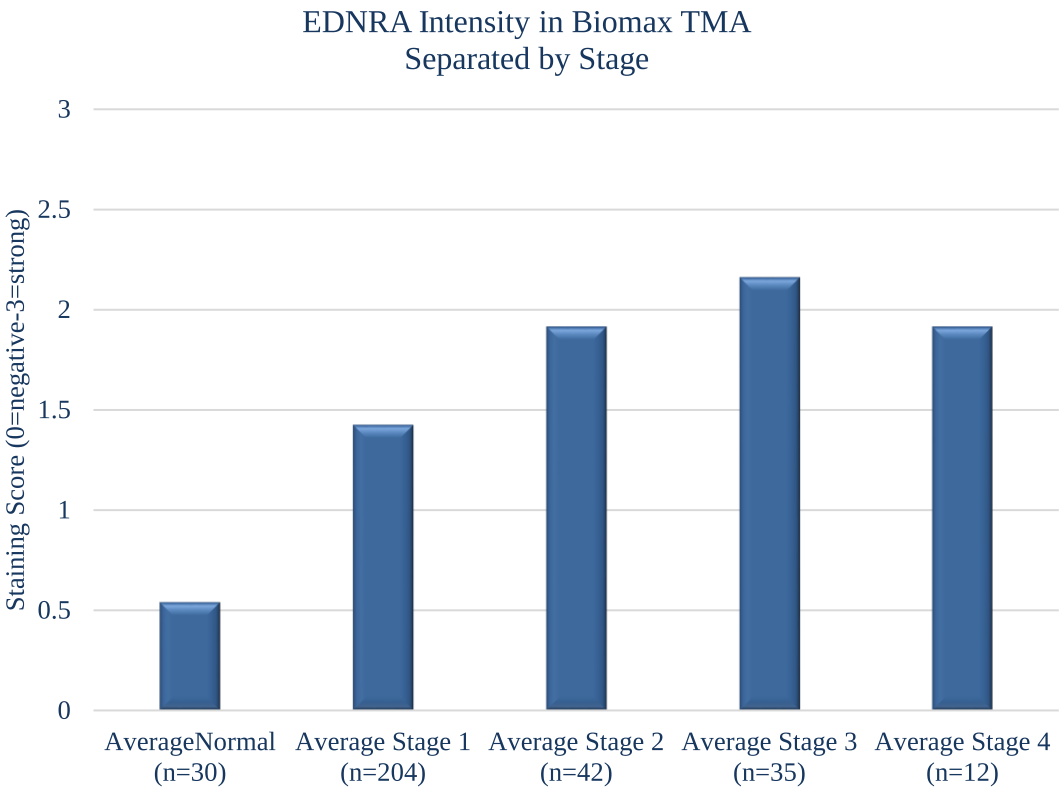


Figure 39. EDNRA Protein Expression in Ovarian Tissues of Different Stage.
Bar graph of protein expression score and cancer stages**.**

Table 18. T-test Results Comparing the Significance of Protein Expression Differences Between Cancer Stages.

	Normal	Stage 1	Stage 2	Stage 3	Stage 4
Normal	-	**2.19741E-05**	**1.01E-08**	**2.20733E-11**	**1E-06**
Stage 1	-	-	**0.000137**	**8.50814E-08**	**0.000137**
Stage 2	-	-	-	0.150605215	0.998291
Stage 3	-	-	-	-	0.316994
Stage 4	-	-	-	-	-

Figure 39 and Figure 40 demonstrates that significantly more EDNRA was present in all cancerous ovarian tissue when compared with normal ovarian tissue. Stage 1 has significantly less than Stage 2, 3 and 4. In this data a clear increase in expression is observed from normal to Stage 1, 2 and 3, this trend plateaus, or marginally decreases at Stage 4. The change in trend observed in Stage 4 could be genuine/represent the larger population, however, the considerably smaller sample number in this group should be noted devaluing the confidence in representing a larger population.

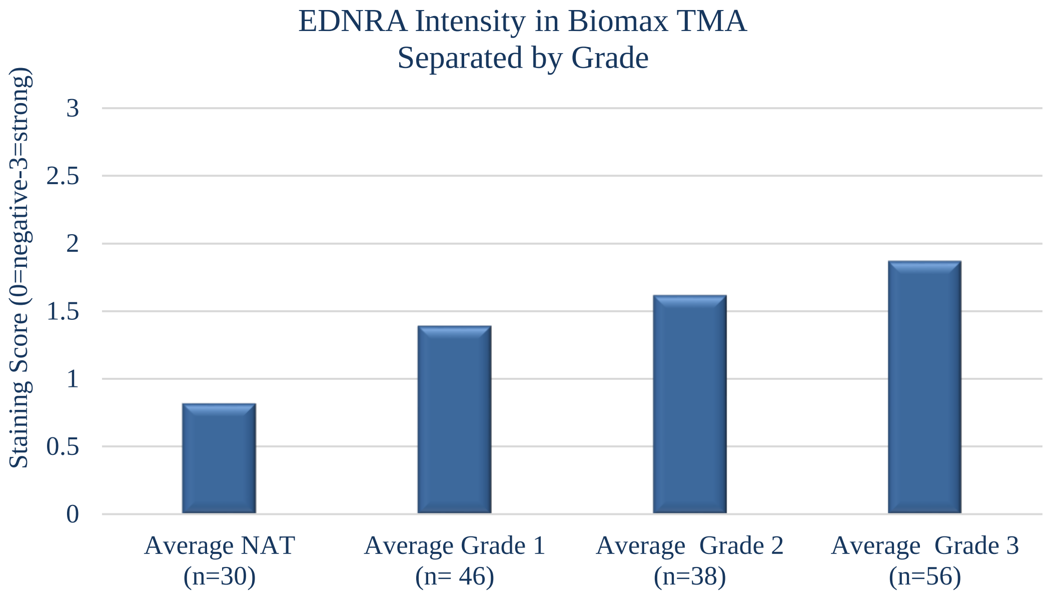


Figure 40. EDNRA Protein Expression in Ovarian Tissues of Different Grades.
A bar graph of protein expression score grades NAT=normal ovarian tissue**.**

Table 19. T-test Results Comparing the Significance of Protein Expression Differences Between Cancer Grades.

p-value	All NAT	All Grade 1	All Grade 2	All Grade 3
All NAT	-	**0.0053026**	**4.648E-06**	**1.36E-10**
All Grade 1	-	-	0.244156689	**0.0075964**
All Grade 2	-	-	-	0.079898109
All Grade 3	-	-	-	-

When separated by grade, only the difference between Grade 1 and Grade 3 is statistically significant, however a clear upward trend is observed progressing through the grades (see Figure 40).

A comprehensive set of T-tests comparing each of the cancer subtypes revealed a wide variation in expression (see Figure 41 and Table 20).

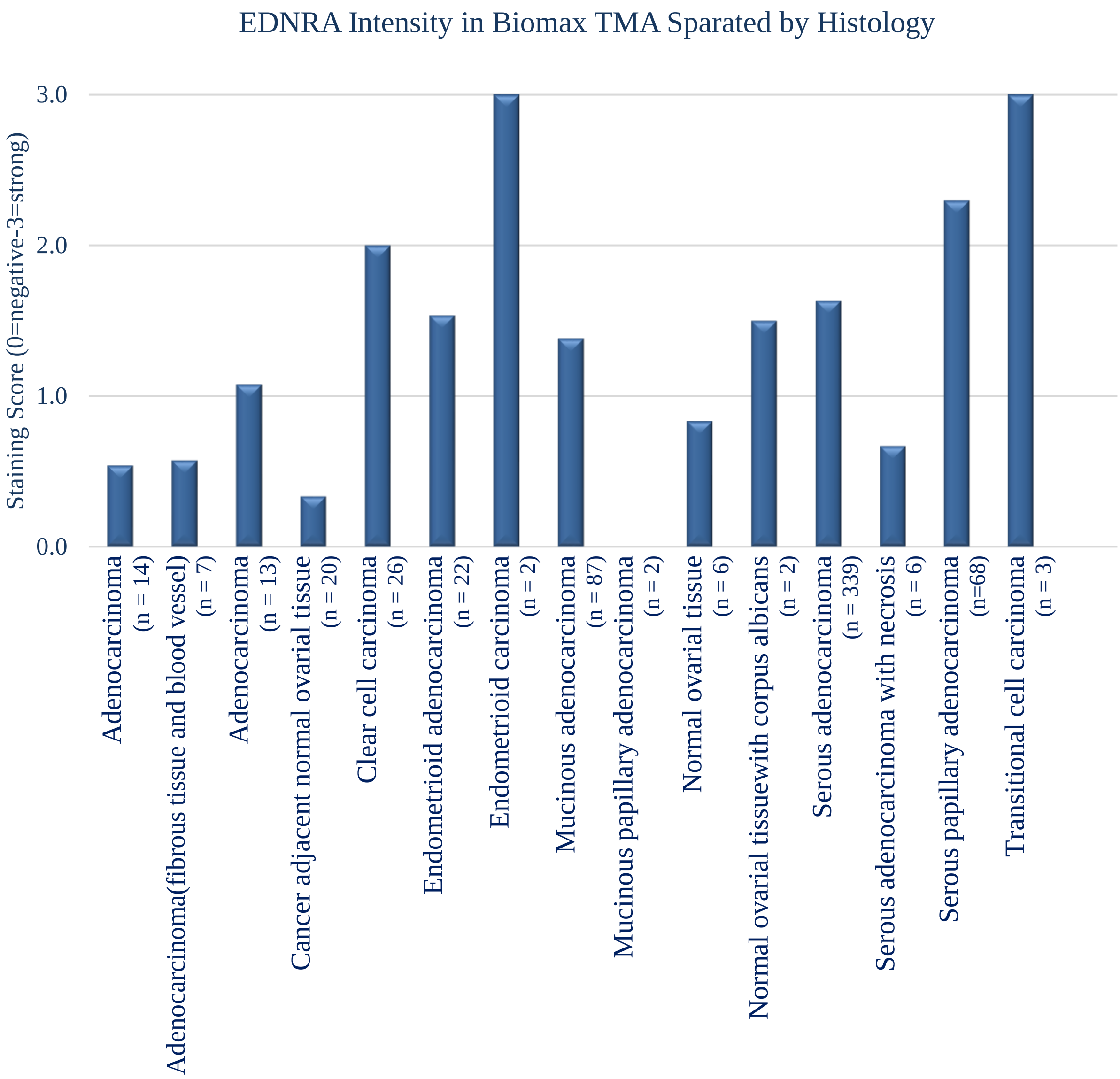


Figure 41. EDNRA Protein Expression in Ovarian Tissues of Different Histology.
A bar graph of protein expression score and cancer histology (x axis).

Table 20. T-test p-values Comparing EDNRA Protein Expression Between Cancer Histologies. p-values of ≤ 0.05 are indicated in green italics, underline is added if still significant after Bonferroni correction is considered for the 98 tests p-values ≤ 0.000512

	Adenocarcinoma	Adenocarcinoma (fibrous tissue and blood vessel)	Adenocarcinoma sparce (n=13)	Cancer adjacent normal ovarian tissue	Clear cell carcinoma (n=26)	Endometrioid adenocarcinoma (n=22)	Endometrioid carcinoma (n=2)	Mucinous adenocarcinoma (n=87)	Mucinous papillary adenocarcinoma	Normal ovarian tissue (n=6)	Normal ovarian tissue with corpus albicans	Serous adenocarcinoma (n=339)	Serous adenocarcinoma ith necrosis	Serous papillary adenocarcinoma (n=68)	Transitional cell carcinoma (n=3)
Adenocarcinoma fibrous (n=14)	-	0.9111	0.1777	0.3656	*0.0068*	0.0551	*0.0005*	0.0708	0.1930	0.3490	0.0785	*0.0058*	0.8052	*0.0000*	*0.0005*
Adenocarcinoma (fibrous tissue and blood vessel) (n=7)	-	-	0.0865	0.3726	*0.0004*	*0.0115*	*0.0002*	*0.0124*	0.2840	0.3312	0.0789	*0.0001*	0.7189	*0.0000*	*0.0002*
Adenocarcinoma sparce (n=13)	-	-	-	*0.0080*	*0.0236*	0.3437	*0.0092*	0.3667	0.1108	0.5234	0.5244	0.0510	0.3413	*0.0000*	*0.0092*
Cancer adjacent normal ovarian tissue (n=20)	-	-	-	-	*0.0000*	*0.0004*	*0.0000*	*0.0003*	0.4486	0.0702	*0.0179*	*0.0000*	0.2894	*0.0000*	*0.0000*
Clear cell carcinoma (n=26)	-	-	-	-	-	0.0917	0.2783	*0.0272*	*0.0362*	*0.0340*	0.5868	0.0911	*0.0201*	0.2286	0.2783
Endometrioid adenocarcinoma (n=22)	-	-	-	-	-	-	*0.0482*	0.9203	0.0773	0.2060	0.5868	0.3165	0.1234	*0.0003*	0.0482
Endometrioid carcinoma (n=2)	-	-	-	-	-	-	-	0.0525	-	*0.0004*	0.0955	0.0561	*0.0086*	0.2568	-
Mucinous adenocarcinoma (n=87)	-	-	-	-	-	-	-	-	0.0967	0.2528	0.8860	0.0586	0.1418	*0.0000*	0.0525
Mucinous papillary adenocarcinoma (n=2)	-	-	-	-	-	-	-	-	-	0.0338	0.0955	*0.0228*	0.3153	*0.0004*	-
Normal ovarian tissue (n=6)	-	-	-	-	-	-	-	-	-	-	0.1340	0.0535	0.6643	*0.0001*	*0.0004*
Normal ovarian tissue with corpus albicans (n=2)	-	-	-	-	-	-	-	-	-	-	-	0.8530	0.2488	0.2030	0.0955
Serous adenocarcinoma (n=339)	-	-	-	-	-	-	-	-	-	-	-	-	*0.0202*	*0.0000*	0.0561
Serous adenocarcinoma with necrosis (n=6)	-	-	-	-	-	-	-	-	-	-	-	-	-	*0.0000*	*0.0086*
Serous papillary adenocarcinoma (n=68)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.2568
Transitional cell carcinoma (n=3)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

The small sample number in some subgroups detracts from the confidence of them to represent a larger population. Notable significant *(p≤0.5)* expression difference between subgroups of more sizable/comparable populations numbers include; the significant higher expression in Clear cell carcinoma (n=26) than Mucinous adenocarcinoma (n=87) and Adenocarcinoma (n=14). As patients with clear cell cancers have a poorer prognosis than those with a serous carcinoma (Matsuzaki *et al.,* 2015), this result suggests a potential causative link between higher EDNRA expression and poor prognosis consequent of increased cell proliferation, resistance to platinum based chemotherapy and traits of EMT also observed in clear cell (Matsuzaki *et al.,* 2015). It should be reiterated here that the data in (Chapter 5) selection criteria did not include clear cell carcinoma so this observational link can only be speculative.

Significantly *(p≤0.5)* there was higher expression in Serous papillary adenocarcinoma (n=68) than Mucinous adenocarcinoma (n=87), Non-serous both fibrous and sparse adenocarcinomas (n=14 and n=13 respectively) and Non-papillary serous adenocarcinomas (n=339). Thus, a link between EDNRA expression and papillary formation could be deduced from this, though is also likely explained by the higher ratio of endothelial cells present in in papillae.

IHC has so far been used for proteomic validation of one GOI, this is a very reductionist evaluation of what was a panel of genes, all found to be of interest and ranked accordingly. In time they could all be validated via IHC but at financial and significant resource cost. Additionally as mentioned earlier, Picotti *et al.,* (2013) warns not to put immune-techniques on a pedestal and criticises that immune based or antibody reliant results are not reproducible when tested extensively as cross reactivity is common.

A platform for proteomic validation that could negate large cost, lengthy time and low confidence/variability of antibodies would be more ideal, such as a mass spectrometry MRM based assay of all the GOIs on a large and appropriate data set available via collaboration.

6.4. Discussion of Verification Strategies

The aim of verification was to confirm or refute the significant association of 56 genes of interest to survival time from ovarian cancer identified in chapter 5.

Fifty six genes identified to be of interest from a biomarker discovery experiment (Chapter 5) were verified in a two stage approach. The majority were filtered out over a validation at genomic level using KM plotter to reproduce the findings in a larger sample cohort (Chapter 6.2). The remaining 7 were then evaluated to prioritise an order of importance for verification by IHC. EDNRA was selected and IHC carried out one and work has begun on a second ovarian TMA.

If any of the genes considered to be "of interest" in chapter 5 had been detected in the LC-MALDI-MS analysis in chapter 3, a retrospective assessment of their presence / absence in cancer, control and benign groups could have made an interesting, though extremely limited proteomic validation. Considering the restricted number of potentially identifiable proteins (120) the chances of this list containing the coding proteins of the GOIs were slim. Further to this, the questionable confidence in those identities (Chapter 3 and 4) would have unfortunately further limited their value.

Validation via KM-Plotter was a valuable approach, to access, collate and annotate this data without this web tool would not have been possible. However, the limitations of this are those that are inherent with microarray technology and combining cohorts of different origins. Put concisely, microarray technologies are limited to detect genes which match predesigned probes on the microarray chip. A novel oncogene will not be detected this way, only an association of a known gene to an available variable i.e. survival. Any gene is only represented by a number of probes on the array, and these vary in number and quality from array to array. The "full genome" microarray chips only represent the known genome, and do not accommodate detection of anything other than that on the chip such as single nucleotide polymorphisms (SNP) which may be key in identifying a patient subgroup within a cohort. Next generation sequencing of the same sample cohorts would be extremely insightful, especially if commonalities between microarray and next gen were present, however such data sets a far rarer and more complex. The sample cohorts available in KM-Plotter are extensive, selection options facilitate the filtering of the sample group and statistical parameter i.e. cut-off, however, full patient data is not accessible, infinite extraneous variable may unknowingly be effecting conclusions.

For example, in KM plotter, when selecting an option to compare one treatment in a meta-analysis of two cohorts there is no clarification of the proportion of which cohort has been treated, thus the measurement may be of something between the cohorts – an extraneous variable e.g. ethnicity, and not the desired test parameter e.g. treatment used.

IHC of EDNRA yielded significant findings associating its expression with ovarian cancer stage, grade and papillary histology. The TMA stained in chapter 6.3 could be explored more extensively, for example the scoring could be repeated using an H-Score to assign a continuous rather than categorical value for intensity. Using an H-score helps to weight a score if more than one staining intensity is observed in each case/core. The proportion of each intensity observed as a percentage is weighted by its strength i.e. the percentage observed as "high" would be multiplied by 3, "medium" by 2 and "low" by 1. This system is vastly more time intensive and is vulnerable to operator bias. Though appealing to apply to this study, repeating scoring using H-scoring may be of limited use. It may benefit the overall research strategy better it the potential time and resources were spent staining a TMA with survival data available, or, immunostaining for a different GOI, namely Decorin.

Proteins coded by genes with multiple probes of interest that qualified for investigation as protein validation were: DCN, EDNRA IGF2, NAV3, TNFAIP6, WTAP and PPFIBP1. IGF2 and WTAP were discounted from further IHC investigation as they are well reported in cancer and ovarian cancer (Sayer *et al.,* (2005), Little *et al.,* (2000)) thus not likely to generate novel information, however philosophically, outside of the remit of this project are still worth research endeavours. Just because a biomarker has already been reported it cannot be concluded from this that there is nothing more that can be learned from further investigation with a different focus. Comprehensive collaborative review articles often comment on the abundance of publications which implicate a gene or protein expression to ovarian cancer yet the distinct lack of follow-up, validation and weaving these findings into a larger biological story/ scene (Braem *et al.,* (2011), Bowtell *et al.,* (2015)). Bowtell *et al.,* (2015) calls for describes how the ovarian cancer research area is now at the stage to move from a parts list to focus resources to a more comprehensive, integrated approach combining acquired knowledge resources. This project has highlighted the speed mass spectrometric protein analysis technology has advanced in recent years and minute protein concentration changes that were previously undetectable could now be possible, but reanalysis of the appropriate patient cohort with the correct detection system is key.

For example Sayer *et al.,* (2005) found IGF2 expression to associate with cancer progression however it was not significantly helpful in detecting early stage ovarian cancer – a core clinical question. It could be that a different detection method such as an MRM would help For a biomarker to be applicable in a clinical setting routine screening needs to be possible in a hospital lab at low cost. IGF2 overexpression has been shown in relation to the cancer progression (Sayer *et al.,* 2005), thus is less likely to be a good biomarker of the early stage disease. However this depends on the sensitivity of the detection method applied. It is possible that there are currently unmeasurably small fluctuations in IGF2 expression in earlier stage but for now current investigations may be more informative if focused on its potential to guide treatment by predicting poor responders who develop a resistance to platinum therapy. Evidence from cell line studies have shown IGF2 as a possible therapeutic target (Sayer *et al.,* 2005).

Ganzfried *et al.,* (2013) is an example of how mining databases of curated microarray data can be teamed with protein expression information to provide supporting evidence. In this example they match a trend of significantly lower survival time in ovarian cancer patients whose tumours have a higher expression of CXCL12, their hypotheses was from published data (Popple *et al.,* 2012 as referenced in Ganzfried *et al.,* (2013)). However, on close inspection, out of a database of 21 cohorts mentioned the significant trend is reported only in three. It is not clear whether this is deliberate (i.e. the remaining 18 were not able to be assessed) or whether in these the result was not significant.

Additionally, when KM Plotter is searched, none of the probes for CXCL12 were found to be significantly associated with survival time. Eight out of the ten cohorts on KM plotter are the same as in Ganzfried *et al., (*2013) leaving these finding ambiguous as they are not obviously reproducible.

Researchers must take care not to bias results by 'cherry-picking' significant values, yet conversely not discount a potential finding as it does not meet a stringency criteria that has been arbitrarily set.

If mass spectrometric verification were to be attempted, there is no guarantee that if the protein product of the gene of interest were to be present, even then, it still may not ionise or be detectable by mass spectrometry. The absence of detection could not count as a confirmed failure to verify genomic data. For example, as different peptides ionise with different efficiencies on different sources (Greaves and Roboz 2014), researchers facing an absence of detection would be inclined to try a mass spectrometric method with a different source amongst other method development before accepting a null hypothesis. Additionally, said hypothetical biomarker may be present, but at a level below the limit of detection (Chapman *et al.,* 2014), or be lost in source decay.

There are several ways the verification stage of this study could be re-examined. If considered a worthwhile endeavour this could yield further insight or value to genes of interest and at low financial cost.

For example KM plotter could be further explored by systematically cycling through the sample filters available in its interface. Using its filters to select subsets of samples an examination of a GOI could be investigated and focus the search to one attribute, namely drug treatments. For example, it may be possible to implicate a GOI to be linked to a platinum Chemoresistance pathway. If a high significance between long and short term survival were observed in a platinum treated subset and not in a subset treated by another drug. However, the feasibility of this has not been explored. KM plotter is a valuable tool, however its potential use is still confined by the limitations explored above (section 6.2.3.).

6.5. Conclusion of Verification Strategies

It was concluded that 7 of the 56 genes (DCN, EDNRA IGF2, NAV3, TNFAIP6, WTAP and PPFIBP1) of interest were verified to significantly associate with ovarian cancer when investigated using KM plotter. Only one of these genes were attempted to be verified by immunohistochemistry, EDNRA was shown to significantly associate with stage, grade and histology of ovarian cancer.

7. Discussions for Future Work and Overall Conclusions

This thesis sought to identify markers of ovarian cancer using the available technology *at that point in time*. Through developments of MALDI, LC-MALDI and expression array platforms, a number of challenges were encountered. In some cases alternative method strategies were designed and implemented, in other cases recommendations for future work were made, or, reanalysis when upgraded technological platforms capable of accurate quantitation better placed to deliver this become available.

The body of work presented here has taken a holistic/bottom-up approach avoiding the use of animal models, which do not best represent human disease (Vaughan *et al.,* 2012) or cell lines, which do not reflect the vast diversity of the disease (Fleury *et al.,* 2015).

In summary, after a lack of assurance of the validity of results from protein biomarker investigations using MALDI-TOF-MS, a change of direction of the project was made to focus on the data mining of gene array experiments. A meta-analysis of two data sets using three analytical methods generated a focus list of 56 genes of interest.

7.1. Ongoing Unmet Clinical Need and Future Work Required

This project has explored and evaluated potential biomarker strategies for diagnostic and prognostic biomarkers of ovarian cancer. Several transcriptomic markers were verified in a larger cohort of samples, and one was verified at a protein level. Feasibility has been a common limitation throughout all approaches, thus the success and order of future work will be dependent on availability of resources. It is foreseeable to overcome some obstacles, namely, patient material appropriate to answer the designed clinical question, and some limitations are beyond the remit of any one individual researcher to tackle i.e. the availability and integration of gene and protein nomenclature, interaction or signalling pathway information.

There are numerous possibilities for future work and a clearly defined and urgent unmet clinical demand (Bowtell *et al.,* 2015). However, the obstacles hindering this research remain the same, namely, lack of sample cohorts relevant to the clinical question especially in the case of early biomarker detection and the heterogeneity of the disease hindering classification and understanding.

Recently Bowtell *et al.,* (2015) reviewed and defined an international collaborative agreement to prioritise and steer future research in ovarian cancer, these include;

- To move from a ‘parts list’ to a more integrated view. Meaning integration of existing and emerging evidence to further understanding.
- Improving the currently used experimental models.
- Understanding drug response as much as chemo-resistance.
- Gain a deeper understanding of the tumour microenvironment.
- Harnessing and exploiting the immune response and interaction within the tumour microenvironment.
- Understanding clonal diversity, recurrent disease and exceptional responders.
- Move to stratified trials of high grade serous ovarian cancer (HGSOC) subsets.
- Implement strategies that could make a rapid impact on prevention and clinical care.

Arguably the most important of these is to act on anything that could make a rapid impact on prevention or clinical care. Most poignant to the current body of work is the aim of achieving a wider integrated understanding of the disease as opposed to a ‘parts list’, this statement is an echo of Braem *et al.,* (2011) 4 years prior, urging a move from biomarker discovery to investigation and verification. As this body of work has also shown, the discovery platforms and investigative tools are fast advancing, namely global protein expression and quantitation in different samples (Section 4.6.1., Law *et al.,* 2013). The ability to compute and combine large, complex data sets is also advancing fast (Section 2.1.2., SCIEX 2016), the ability to marry significant genomic/transcriptomic biomarker observations to proteomic observations from the same samples would be highly insightful to the understanding on cellular processes in subgroups of ovarian cancer.

The chance of attaining any future publication or funding to further the above body of research would be more successful if based around these collaboratively agreed guidelines of Bowtell *et al* (2015). The sentiment of some of the above points were bore in mind when undertaking this work, as discussed in section 4.6 where the call to re-examine existing data (Braem *et al.,* 2011) influenced the work towards the analysis of microarray data, however sample availability or lack of, inadvertently steers research. Bowtell *et al.,* (2015) emphasises the need to direct research to answer the specific clinical questions, however, this is not possible without the appropriate data set. A well devised huge-scale prospective databank collection, such as UKCTOCS, is the only way to attain this.

It is possible that much of the information needed to understand the complexities of ovarian cancer already is available as Braem *et al.,* (2010) suggests, and that re-examining all existing information by an enormous-scale cross platform data mining of databases and integration of genomic, proteomic, transcriptomic, cell line or aetiological studies could unearth additional and relevant information. However, this is not likely to attract funding/support due to the large amount of resources needed to conduct, for an undefined potential final measurable outcome.

7.1.1. Targeted Protein Mass Spectrometry Based on Transcriptomic Discovery as a Strategy for Future Work.

An MRM/SRM experiment could be designed for the coded proteins of the verified genes of interest generated from transcriptomic data in Chapter 5. A seemingly appropriate potential data set that could be re-examined this way does now exist (Russell *et al.,* 2016), this study incorporates patient cohorts which are sourced from the UKCTOCS investigation conducted in 2001 to 2005 (Jacobs *et al.,* 2015), however their raw data is not yet publicly available for re-analysis.

The biomarker discovery strategies explored in Chapter 3 and 4 were found to be of insufficient quality to be used as a biomarker discovery tool. However, other mass spectrometric approaches could be more strategically employed should more time, funding, resources become available. With current, more advanced technology, the protein transcripts of the gene markers of interest generated in the above chapters could be explored. The genomic validation step from KM plotter (Chapter 6.2) filtered the interest list down to seven genes, a number suitable for a targeted mass spectrometry strategy as recommended by Marx (2013) (Chapter 4). As evidenced by Morcavello *et al.,* (2014) DCN, one of the seven GOIs, is able to exist in the extracellular space, and its ligand role to IGF2 makes it a particularly interesting focus for potential targeted mass spectrometry experiments such as MRM/SRM or SWATH (see section 4.6). In such a method the expected fragment *m/z* values from each peptide fragment could be calculated using the known amino acid sequence the protein of the genes of interest (available via Uniprot) and an appropriate software such as Skyline, which is also freely available. From this, a selective reaction monitoring experiment could be set up with an appropriate mass spectrometer to selectively measure the quantity, if any, of each of the fragments in the sample with both accuracy and high sensitivity.

Using the targeted MRM/SRM/SWATH mass spectrometric approach for protein verification is advantageous over immuno-body based techniques (Marx 2013, Baker 2015) it is arguably cost effective, and accurate both qualitative and quantitative accuracy of measurement, increasing the size of experiment possible: If the capital equipment is in place (as antibodies, or designed kits do not need to be purchased) it is possible to assay more target molecules for the same cost. Additionally the accuracy of antibodies compared to MRM experiments for confirmation of target protein presence has been discussed (Marx 2013, Baker 2015), it would be possible to confirm and quantify all of the coded proteins of the genes of interest. For an example the reader is referred to Russell *et al.,* (2016), who, utilise the latest data independent tandem mass spectrometry technology for ovarian cancer biomarker verification. The data independent SWATH method applied in this study negated designing a specific MRM experiment. All peptide fragments from the digested serum were quantified, then algorithms applied to assign which fragment data matched to the values of their genes of interest. A higher number of potential biomarkers can be investigated using a SWATH data independent approach as all peptide fragments in the analyte are directed through all quadrupoles in the mass spectrometer to the detector where they are measured. This moderately decreases the sensitivity when compared directly to a specific MRM experiment as signal at the detector is distributed across all the other fragments hitting it at any given time, but does still provide high quality quantitative data.

As discussed in Chapter 4.6.2, the serologic proteomic analysis of the samples available to this study (see chapter 4.6) was found to be a suboptimal use of resources in a research landscape including such studies as Russell *et al.,* (2016), who, both utilise the next generation of MS technology and have greater sample numbers. Instead/ moreover, should the opportunity arise the raw data acquired in the Russell *et al.,* (2016) could be utilised to add to the body of evidence built in the current study: the raw/SWATH data could be mined to quantitate the calculated transition values of DCN, EDNRA, IGF2, NAV3, TNFAIP6, PPFIBP1 and WTAP proteins providing translational verification. For example if the raw/SWATH data was made available via a resource such as PRIDE. As patient survival information is available and the sample number is comparative or larger, this would be a highly relevant and appropriate translational validation of the genes of interest arising from the present study (Chapter 6.1).

7.1.2. Clarity and Cohesiveness of Current Resources Challenge Future Work

As discussed in section 5.4.5. and demonstrated above, re-exploration of publically available genomic or transcriptomic large data sets is an endeavour with potential to generate both novel results, and, a novel perspective on existing data. Moreover, meta-analyses to increase both sample number and the scope of type of data analysed, adds rigor to any results generated. Hence this should be continued to be incorporated into future study design. Data sets found to be comparable/akin to those used in the Chapter 5 and those of utilised in (Ganzfried *et al.,* 2013) would be most relevant to continue research with a goal of subcategorising subpopulation of ovarian cancers.

Encouragingly, more resources are constantly becoming available as discussed in section 5.4.5. However, if this work were to be conducted care must be taken to clarify the source of data before is used as discussed in section 2.2.3. It is important to note that in the spirit of data sharing data from the same sample may be represented by numerous databases for example the ICGC sources data from TCGA and CPTAC. Before utilising one we need to untangle where each source their data. For example the Cancer Genome Project (CGP) originated from the Human Genome Project at the Sanger Institute and, The International Cancer Genome Consortium (ICGC), the Cancer Genome Atlas (CGA) and the Clinical Proteomic Tumour Analysis Consortium (CPTAC). It must be considered that the/this multiple representation of the same data by a different title/name could transpire to misinterpreted or misrepresented findings and even further bias the emphasis on that data. Misrepresentation double reporting of data is a potential problem: for example, a researcher conducts a novel experiment to generate a list of genes or proteins of interest in relation to a cancer, a logical progression of this research would be to compare to already published results then proceed to validation. They may find two pieces of evidence on data repositories linking a gene to the cancer where: Firstly, it may not be obvious without a significant amount of secondary research that the two pieces of information came from the same sample. Secondly, after reporting their results, the role of the already associated genes to the cancer may be exaggerated by over reporting and the genes that are not or ambiguously linked via literature search, though potentially relevant, may remain unmentioned or under reported.

7.1.3. Integration of Cross-Platform Data Challenges Future Work

As discussed above, the concept of using one or a few biomarkers to detect and monitor a complex system, such as a tumour surviving within a host disease has been argued to be a reductionist, or overly simplistic (Strimbu *et al.,* 2011) measurement, however, a measurable factor from an easily attainable patient sample is key to improving patient prognosis. As outlined by Hanahan and Weinburg (2011), Strimbu *et al.,* (2011) suggests biomarkers will only be able to be implemented as accurate measurements of clinical endpoints once normal cellular function and physiology has been fully mapped out, which it has not yet been. Gil *et al.,* (2015) points out, the technical platform or computing power to attain this does not currently exist.

This study contributes to a growing body of evidence or each of the thousands of genes investigated, not just those that were shown to have a significant relation to survival time. The relevance of each gene in relation to survival time from ovarian cancer contributes to an ever expanding body of evidence connected through online resources mapping out cellular function under diseased and normal conditions such as STRING, Reactome, KEGG and more. In chapter 6 the three of the finalised GOIs are placed in two cellular pathways already known to manifest some hallmark phenotypes of cancer, this offered insight into their relevance and whether they may be of use as future biomarkers or targets of therapy. However, the KEGG database from which these were created, though extensive, is not yet comprehensive. An integration of data acquired at all levels of cellular function is called for.

As Hanahan and Weinburg (2011) and Strimbu *et al.,* (2011) elude to, an ideal-scenario-tool for an integrated research model, that is not currently logistically fathomable/foreseeable would enable the analysis to be expanded across acquisition platforms. More specifically encompass genomic transcriptional and proteomic translational data, and as software becomes available, to integrate measurements from all experimental platforms: genomic, proteomic, metabolomics lipidomic etc. The overall goal being a quantified ‘ome’ from multiple derivatives of each sample. For example, cancer tissue, sera, plasma, saliva, urine, white blood cells and more from each patient of a cohort, ideally the cohort would contain over 1000 cases.

A realistic best-case scenario using currently available platforms would include experimental measurements of the genome as measured by *de-novo* next generation sequencing (Koboldt *et* *al.,* 2013), and these measurements easily matched to and compared against the translated proteome as measured by a data independent *de-novo* sequenced platform data which included all post translational modifications, and any future platforms of measurement based in place of probing of known or matching to known factors.

Other considerations for future work:

- To reiterate what is discussed in section 6.5, some genes were found to associate with survival time from ovarian cancer, IGF2 and WTAP were discounted from further verification for this work as the previous reporting was counted as verification (Sayer *et al.,* (2005), Little *et al.,* 2000). However these genes and proteins should not be discounted from future validation or investigation. Although their association has already been reported their role each tumour microenvironment still remains to be expanded, and made part of a larger biological story/ scene (Braem *et al.,* 2011, Bowtell *et al.,* 2015).
- If a biomarker is recognised to be a suitable drug target this may still be of limited use. A drug designed to target one component in the molecular pathway, it is likely that inhibiting/blocking one pathway component the other components of the pathway compensate. To further complicate the matter, these are not necessarily of tumour origin and are more likely in or related to the tumour microenvironment. (Vaughan *et al.,* 2012). A full system of up or down regulated genes or proteins would need to be isolated and targeted.
- Though limited cell lines of well characterised histopathological origins are reported to be needed. New technologies such as three-dimensional growth platforms are available to better recapitulate the micro-tumours growth pattern and signalling (Vaughan *et al.,* 201, Hickman *et al.,* 2016).
- Immunofluorescence to view subcellular localisation of genes of interest would be insightful.

7.2. Overall Summary and Conclusions

The purpose of this study was to meet an ongoing unmet clinical need to detect and diagnose subpopulations to effectively treat ovarian cancer (Wilson and Junger (1968) in Nossov *et al.,* (2008), Vaughan *et al.,* 2012, Menon *et al.,* 2014, Bowtell *et al.,* 2015). As the stage of an ovarian cancer at diagnosis is the key prognostic indicator (Erickson *et al.,* 2014). Thus, the strategic clinical question to address first was to discover novel serologic protein biomarkers for early detection of ovarian cancer. Thus, the study began with this aim.

Chapter 3 and 4 investigated the merit and reliability of MALDI TOF MS/MS as a biomarker generation platform. It was concluded that the instrumentation used had limited applicability in the biomarker discovery process. In Chapter 5 a meta-analysis of serous Stage 3 ovarian cancer samples generated a list of 56 genes paired with evidence from three analyses, Cox univariate, ANN and T-test, implying their significant association with survival time of patients with ovarian cancer. Finally in Chapter 6 the results were validated on a larger cohort of microarray data narrowing the interest list down to 7 genes. One of the genes was further verified at a protein level and evidence supporting its differential protein suppression between cancer stage and grade and histopathology were shown to be significant.

In Chapter 3 a preliminary, bottom-up MALDI-TOF-MS based analysis of serum protein was conducted on a sample cohort collected in a manner addressing criticisms of peer research at the time. Unique patterns in protein expression in the sera were detected by MALD-TOF-MS and ANNs and were used in an attempt to distinguish cancer from control on a blinded validation set. Unfortunately insufficient data or evidence linked the *m/z* of the peptide peaks expressed differentially in the tested serum to protein identities by matching them to data from linked LC-MALDI-MS with the MS profiles. The full panel of peptide peaks could not be identified so full validation of the biomarker panel could not be attempted.

In Chapter 4 the next generation of MALDI-MS analysis platform was evaluated for use as an ovarian serum protein biomarker detection tool. It was concluded that the intensity values of MALDI-TOF-MS data can be used to indicate relative protein quantity within a sample, however it was noted that the accuracy of this type of analysis was low and did not allow a robust and meaningful comparison to be made between samples. One sample preparation technique resulted in generating more reproducible peptide identities than others. The differences in the LC-MALDI-MS profiles of serum samples produced under different conditions was shown using a qualitative and semi-quantitative method and neither yielded list of proteins large or consistent enough to contribute with novelty to serum protein biomarker discovery.

Evidence was also provided to suggest that, both sample processing workflows tested reduced the number of reproducible identities attained from samples and clinical samples should be run in at least triplicate to reduce the number of false identities attained. Of the two methods of data export and analysis conducted, different conclusions can be drawn from the raw data collected. Most importantly from this evaluation it was concluded that this sample platform was not reproducible enough to warrant the use of limited, rare sample cohort and resulted in modifying the research focus to include different data sources and the application of bioinformatic data mining of publicly available data sets in an attempt to discover valid ovarian cancer biomarkers. A meta-analysis of mRNA microarray data from two cohorts of serous Stage 3 ovarian cancer samples generated a list of 56 genes paired with evidence from three analyses; Cox univariate, ANN and T-test implying their significant association with survival time from ovarian cancer.

Within the listed caveats of the material available for study, 56 genes were found to significantly and consistently associate with survival from ovarian cancer. The meta-analysis tactic added confidence to the 56 inferred associations. Verification of the 56 genes was conducted on a wider sample cohort using a freely available web tool. Evidence confirmed the association of the following genes expression with survival time from Stage 3 serous ovarian cancer: *DCN*, *EDNRA, IGF2, NAV3, TNFAIP6, WTAP* and *PPFIBP1*. Literature review was used to select one of these genes, *EDNRA*, for preliminary protein verification, using immunohistochemical staining of a TMA of ovarian tumour tissue. The results correlated EDNRA expression with ovarian stage, grade and cancer histology.

This study generated evidence supporting the role of several genes effect on survival time from ovarian cancer at a transcriptomic level and one at both gene and protein. These results contribute to current knowledge and provide substantial evidence to base future investigation upon.

References

Abdel-Fatah T.M.A, Mcardle, S.E.B, Johnson C., Moseley P.M., Ball G.R., Pockley A.G. & Ellis I.O. (2014). "HAGE (DDX43) Is a Biomarker for Poor Prognosis and a Predictor of Chemotherapy Response in Breast Cancer." *British Journal of Cancer* 110 (April): 2450–61. doi:10.1038/bjc.2014.168.

Abdel-fatah, T. M. A., Agarwal, D., Liu, D., Russell, R., Rueda, O. M., Liu, K., Xu, B., Moseley, P. M., Green, A. R., Pockley, A. G., Rees, R. C., Caldas, C., Ellis, I. O., Ball, G. R. & Chan, S. Y. T. (2016). "SPAG5 as a Prognostic Biomarker and Chemotherapy Sensitivity Predictor in Breast Cancer : A Retrospective , Integrated Genomic , Transcriptomic , and Protein Analysis." *Lancet Oncology* 17 (7). Elsevier Ltd: 1–15. doi:10.1016/S1470-2045(16)00174-1.

Adam, M., Saller, S., Ströbl, S., Hennebold, J. D., Dissen, G. A., Ojeda, S. R., Mayerhofer, A. (2012). "Decorin is a part of the ovarian extracellular matrix in primates and may act as a signaling molecule". *Human Reproduction* (Oxford, England), 27(11), 3249–58. doi.org/10.1093/humrep/des297

Agilent Technologies. (2010). "Agilent 3100 OFFGEL Fractionator Achieve Unprecedented Sensitivity in LC / MS-Based Proteomics Experiments." *User Guide Publication Number 5990-5596EN*, 1–7.

Ahmad, A. S., Ormiston-Smith, N. & Sasieni, P. D. (2015). "Trends in the lifetime risk of developing cancer in Great Britain: comparison of risk for those born from 1930 to 1960". *British Journal of Cancer*, *112*(5), 943–7. doi.org/10.1038/bjc.2014.606.

Albrethsen, J. 2011. "The First Decade of MALDI Protein Profiling : A Lesson in Translational Biomarker Research." *Journal of Proteomics* 74 (6). Elsevier B.V.: 765–73. doi:10.1016/j.jprot.2011.02.027.

Allison, D. B., Cui, X., Page, G. P. & Sabripour, M. (2006). "Microarray data analysis: from disarray to consolidation and consensus". *Nature Reviews. Genetics*, *7*(1), 55–65. doi:10.1038/nrg1749.

An, H. J., Miyamoto, S., Lancaster, K. S., Kirmiz, C., Li, B., Lam, K. S., Leiserowitz, G. S. & Lebrilla, C. B. (2006). "Profiling of Glycans in Serum for the Discovery of Potential Biomarkers for Ovarian Cancer" *research articles*, 1626–1635.

Anderson, N. L. (2010). "The Clinical Plasma Proteome: A Survey of Clinical Assays for Proteins in Plasma and Serum." *Clinical Chemistry* 56 (2): 177–85. doi:10.1373/clinchem.2009.126706.

Anderson, L. & Anderson N.G. 1977. "Plasma Proteins." Biochemistry 74 (12): 5421–25. http://www.pnas.org/content/74/12/5421.full.pdf.

Anderson, N.L. & Anderson, N.G. (2002) "The Human Plasma Proteome Historical, Character and Diagnostic Prospects". *Molecular and Cellular Proteomics.*1.11 845-867

ArrayExpress (2011). Available at http://www.ebi.ac.uk/arrayexpress/ Accessed December 2011

ArrayExpress (2013). Available at http://www.ebi.ac.uk/arrayexpress/ Accessed May 2013

Atlas Antibodies HPA014087, (2012). Anti-EDNRA Product Datasheet, https://atlasantibodies.com/#!/products/EDNRA-antibody-HPA014087 / Accessed May 2014

Baba, A. & Câtoi, C. (2007). "Comparative Oncology". *Chapter 3, TUMOR CELL MORPHOLOGY*. Bucharest: The Publishing House of the Romanian Academy. Retrieved from http://preview.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=componc&part=ch3#ch3.s2

Baggerly, K. A., Morris J. S. & Coombes K. R. (2003). "Cautions about Reproducibility in Mass Spectrometry Patterns : Joint Analysis of Several Proteomic Data Sets."

Baggerly, K. A., Morris, J. S., Edmonson, S. R. & Coombes, K. R. (2005). "Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer". *Journal of the National Cancer Institut*e, 97(4), 307–9. doi.org/10.1093/jnci/dji008

Baker, M. (2015). "Blame It on the Antibodies." *Nature* 521: 274–75. doi:10.1038/521274a.

Balkwill, F. (2004). "Cancer and the chemokine network". *Nature reviews. Cancer*, 4(7), 540–50. doi:10.1038/nrc1388

Balog, J., Sasi-Szabó, L., Kinross, J., Lewis, M. R., Muirhead, L. J., Veselkov, K., Mirnezami, R., Dezső, B., Damjanovich, L., Darzi, A., Nicholson, J. K. & Takáts, Z. (2013). "Intraoperative tissue identification using rapid evaporative ionization mass spectrometry". *Science Translational Medicine*, 5(194), 194ra93. doi.org/10.1126/scitranslmed.3005623

Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K. K., Carter, S. L., Frederick, A. M., Meyerson, M. (2012). "Sequence analysis of mutations and translocations across breast cancer subtypes". *Nature*, 486(7403), 405–9. doi.org/10.1038/nature11154

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Garraway, L. A. (2012). "The Cancer Cell Line Encyclopaedia enables predictive modelling of anticancer drug sensitivity". *Nature*, 483(7391), 603–607. doi.org/10.1038/nature11003

Beck, V., Herold, H., Benge, A., Luber, B., Hutzler, P., Tschesche, H., Kessler, H., Schmitt, M., Geppert, HG and Reuning, U. (2005). ADAM15 decreases integrin αvβ3/vitronectin-mediated ovarian cancer cell adhesion and motility in an RGD-dependent fashion. *International Journal of Biochemistry and Cell Biology*. 590-603(37) doi.org/10.1016/j.biocel.2004.08.005

Benk, A. & Roesli, C. (2012). "Label-free quantification using MALDI mass spectrometry: considerations and perspectives". *Analytical and Bioanalytical Chemistry*. Retrieved from http://www.springerlink.com/index/155125V755U620Q0.pdf

Benzinger, A. (2005). "Targeted Proteomic Analysis of 14-3-3 , a p53 Effector Commonly Silenced in Cancer". *Molecular & Cellular Proteomics*, 4(6), 785–795. doi.org/10.1074/mcp.M500021-MCP200

Bi, X.L. & Yang, W (2013). “Review Decorin in Tumor Development and Progression Decorin in Tumor Angiogenesis Decorin as a Potential Anticancer Agent Decorin in Tumor Metastasis”, *Chin J Cancer*; Vol 32 Issue 5.

Bignotti, E., Tassi, R. A., Calza, S., Ravaggi, A., Romani, C., Rossi, E., Santin, A. D. (2006). “Differential gene expression profiles between tumor biopsies and short-term primary cultures of ovarian serous carcinomas: identification of novel molecular biomarkers for early diagnosis and therapy”. *Gynecologic Oncology*, 103(2), 405–16. doi.org/10.1016/j.ygyno.2006.03.056

Biognosys AG. (2014). “DIA / SWATH Market Assessment,” no. December. http://www.biognosys.ch/fileadmin/Uploads/Media/2014_Biognosys_DIA_Market_Assessment_Report.pdf.

Biomax, (2014). Biomax. Available at: http://www.biomax.us/tissue-arrays/Ovary/OV6161 (Accessed November 7, 2014).

Bland, J.M. & Altman, D.G. (2004) “The Logrank test”. *BMJ* p.1073. ;328:1073

Bleeker, F. E., Lamba, S., Rodolfo, M., Scarpa, A., Leenstra, S., Vandertop, W. P. & Bardelli, A. (2009). “Mutational profiling of cancer candidate genes in glioblastoma, melanoma and pancreatic carcinoma reveals a snapshot of their genomic landscapes”. *Human Mutation*, 30(2), E451–E459. doi.org/10.1002/humu.20927

Bonome, T. L., D., Shih, J., Randonovich, M., Pise-Masison, C.A., Bogomolniy, F., Ozbun, L., Brady, J., Barrett, J.C., Boyd, J. & Birrer, M.J. (2008). “A Gene Signature Predicting for Survival in Suboptimally Debulked Patients with Ovarian Cancer.” *Cancer Research* 68 (13): 5478–86. doi:10.1158/0008-5472.CAN-07-6595.

Bowtell, D.D., Böhm, S., Ahmed, A. A., Aspuria, P.-J., Bast, R.C., Beral, V., Berek, J.S., Birrer, M.J., Blagden, S., Bookman, M. A., Brenton, J.D., Chiappinelli, K.B., Martins, F.C., Coukos, G., Drapkin, R., Edmondson, R., Fotopoulou, C., Gabra, H., Galon, J., Gourley, C., Heong, V., Huntsman, D.G., Iwanicki, M., Karlan, B.Y., Kaye, A., Lengyel, E., Levine, D. a., Lu, K.H., McNeish, I. A., Menon, U., Narod, S. A., Nelson, B.H., Nephew, K.P., Pharoah, P., Powell, D.J., Ramos, P., Romero, I.L., Scott, C.L., Sood, A.K., Stronach, E. A., Balkwill, F.R.: (2015). “Rethinking Ovarian Cancer II: Reducing Mortality from High-Grade Serous Ovarian Cancer.” *Nature Reviews Cancer*. doi:10.1038/nrc4019.

Braem, M. G. M., Schouten, L. J., Peeters, P. H. M., Den Brandt, P. a Van, & Onland-Moret, N. C. (2011). “Genetic susceptibility to sporadic ovarian cancer: A systematic review”. *Biochimica et biophysica acta*, *1816*(2), 132–46. doi:10.1016/j.bbcan.2011.05.002

Bridges, P. J., Jo M., Alem L. A., Na G., Su W., Gong M. C., Jeoung M. & Ko C. (2010). “Production and Binding of Endothelin-2 (EDN2) in the Rat Ovary: Endothelin Receptor Subtype A (EDNRA)-Mediated Contraction.” *Reproduction, Fertility and Development* 22 (5): 780–87. doi:10.1071/RD09194.

Brown, K. R. & Jurisica I. (2007). “Unequal Evolutionary Conservation of Human Protein Interactions in Interologous Networks.” *Genome Biology* 8: 1–11. doi:10.1186/gb-2007-8-5-r95.

Brown, M. R., Blanchette, J. O. & Kohn, E. C. (2000). “Angiogenesis in ovarian cancer. Baillière’s Best Practice & Research”. *Clinical Obstetrics & Gynaecology*, 14(6), 901–18. doi:10.1053/beog.2000.0134.

Buchan, B. W., Riebe K. M. & Ledeboer N. A. (2012). “Comparison of the MALDI Biotyper System Using Sepsityper Specimen Processing to Routine Microbiological Methods for Identification of Bacteria from Positive Blood Culture Bottles.” *Journal of Clinical Microbiology* 50: 346–52. doi:10.1128/JCM.05021-11.

Buys, S. S. (2011). “Effect of Screening on Ovarian Cancer Mortality. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Randomized Controlled Trial.” JAMA: *The Journal of the American Medical Association* 305 (22): 2295. doi:10.1001/jama.2011.766.

Cadron, I., Van Gorp, T., Timmerman, D., Amant, F., Waelkens*, E. &* Vergote, I. (2009). “Application of proteomics in ovarian cancer: which sample should be used?” *Gynecologic oncology*, *115*(3), 497–503. doi:10.1016/j.ygyno.2009.09.005

Caldas, C. & Stingl, J. (2007). “Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis”. *Nature Reviews. Cancer,* 7: 791-99. Retrieved from www.ncbi.nlm.nih.gov/pubmed/17851544 doi:| doi:10.1038/nrc2212

Cancer Research UK (2012) Available at: http://info.cancerresearchuk.org (Accessed May 2012)

Cancer Research UK (2015) Available at: http://info.cancerresearchuk.org (Accessed April 2015

Cancer Research UK (2016) Available at: http://www.cancerresearchuk.org/content/cancer-statistics-for-the-uk (accessed Jan 2016)

Carey, V. J., Gentry, J., Sarkar, D., Gentleman, R. & Ramaswamy, S. (2008). “SGDI: SYSTEM FOR GENOMIC DATA INTEGRATION.” Pac Symp BioComput 83 (13): 141–52. doi:10.1021/ac200812d.

Carlsson, E., Ranki, A., Sipilä, L., Karenko, L., Abdel-Rahman, W M., Ovaska, K., Siggberg, L., Aapola, U., Ässämäki, R., Häyry, V., Niiranen, K., Helle, M., Knuutila, S., Hautaniemi, S., Peltomäki, P and Krohn, K. (2012). “Potential Role of a Navigator Gene NAV3 in Colorectal Cancer.” *British Journal of Cancer* 106 (3): 517–24. doi:10.1038/bjc.2011.553.

Carmen D.M., Vega F, Moreno-Bueno G., Artiga M.J., Sanchez L., Palacios J., Ridley A. & Timms J.F. (2008). “Characterisation of Tumoral Markers Correlated with ErbB2 (HER2/Neu) Overexpression and Metastasis in Breast Cancer.” Proteomics. Clinical Applications 2 (9): 1313–26. doi:10.1002/prca.200780020.

Catalona, W. J. (1996). “Clinical utility of measurement free and total prostate-specific antigen (PSA): A review”. *The Prostate*. 2927,64-69

Caudil, M. (1987). “Neural networks primer part 1”. *AI Expert*, 2(12), 46–52.

Chang, L.-C., Sheu, H.-M., Huang, Y.-S., Tsai, T.-R. & Kuo, K.-W. (1999). “A novel function of emodin”. *Biochemical Pharmacology, 58*(1), 49–57. doi.org/10.1016/S0006-2952(99)00075-1

Chapman, J. D, Goodlett D. R. & Masselon C. D. (2014). “Multiplexed and Data Independent Tandem Mass Spectrometry for Global Proteome Profiling.” *Mass Spectrometry Reviews*, no. 33: 452–70. doi:10.1002/mas.

Chen, L., Wang H., Zeng Q., Xu Y., Sun L., Xu H. & Ding. L. (2009). “On-Line Coupling of Solid-Phase Extraction to Liquid Chromatography--A Review.” *Journal of Chromatographic Science* 47 (8): 614–23. doi:10.1093/chromsci/47.8.614.

Christgen, M., Ballmaier, M., Bruchhardt, H., Wasielewski, R., Kreipe, H. & Lehmann, U. (2007). ”Identification of a distinct side population of cancer cells in the Cal-51 human breast carcinoma cell line.” *Molecular and Cellular Biochemistry*, 306(1-2), 201–212. doi.org/10.1007/s11010-007-9570-y

Christoforou, A. & Lilley K. S. (2011). “Taming the Isobaric Tagging Elephant in the Room in Quantitative Proteomics.” *Nature Methods* 8 (11): 911–13. doi:10.1038/nmeth.1736.

Conrads, T P., Fusaro, V A., Ross, S., Johann, D., Rajapakse, V., Hitt, B A., Steinberg, S M., Kohn, E C., Fishman, D A., Whitely, G., Barrett, J C., Liotta, L A., Petricoin, E F & Veenstra, T D. (2004). “High-Resolution Serum Proteomic Features for Ovarian Cancer Detection.” *Endocrine-Related Cancer* 11 (2): 163–78. http://www.ncbi.nlm.nih.gov/pubmed/15163296.

Cooper, GM. 2000. “The Development and Causes of Cancer.” In *The Cell: A Molecular Approach. 2nd Edition.*, 2nd ed. Sunderland: Sinauer Associates. https://www.ncbi.nlm.nih.gov/books/NBK9963/.

Coveney, C., Boocock D., Rees R., Deen S. & Ball G. (2015). “Data Mining of Gene Arrays for Biomarkers of Survival in Ovarian Cancer.” *Microarrays* 4 (3): 324–38. doi:10.3390/microarrays4030324

Cox, D . R . (1972). “Regression Models and Life-Tables.” *Journal of the Royal Statistical Society* 34 (2): 187–220.

Crijns, A. P. G., Fehrmann, R. S. N., de Jong, S., Gerbens, F., Meersma, G. J., Klip, H. G., Hollema, H., Hofstra, R. M. W., te Meerman, G. J., de Vries, E. G. E. & van der Zee, A. G. J. (2009). “Survival-related profile, pathways, and transcription factors in ovarian cancer.” *PLoS medicine*, *6*(2), e24. doi:10.1371/journal.pmed.1000024

Croft, D., Kelly G. O., Wu, G., Haw R., Gillespie M., Matthews L., Caudy M., Garapati, P., Gopinath G., Jassal B., Jupe S., Kataskaya I., Mahajan S., May B., Ndegwa N., Schmidt E., Shamovsky V., Yung C., Birney E., Hermjakob H., Eustachio P.D. & Stein L. (2010). “Reactome : A Database of Reactions , Pathways and Biological Processes.” *Nucleic Acids Research*, 1–7. doi:10.1093/nar/gkq1018.

DAVID. (2016). DAVID. Retrieved May 1, 2016, from https://david.ncifcrf.gov/home.jsp

Dayhoff, O. M. (1969). “National Biomedical Research Foundation,. Atlas of Protein Sequence and Structure (1st ed., Vol. 1)”. Maryland: *Silver Spring*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21986951

de Gramont, A., Watson, S., Ellis, L. M., Rodón, J., Tabernero, J., de Gramont, A. & Hamilton, S. R. (2014). “Pragmatic issues in biomarker evaluation for targeted therapies in cancer.” Nature Reviews Clinical Oncology, 12 (April 2015), 197–212. doi.org/10.1038/nrclinonc.2014.202

Dent, R., Trudeau, M., Pritchard, K. I., Hanna, W. M., Kahn, H. K., Sawka, C. A, Lickley L.A , Rawlinson E., Sun P. & Narod S.A (2007). “Triple-negative breast cancer: Clinical features and patterns of recurrence”. *Clinical Cancer Research*, *13*(15), 4429–4434. doi.org/10.1158/1078-0432.CCR-06-3045

Devlin, B., Roeder, K. & Wasserman, L. (2003). “False discovery or missed discovery?” *Heredity*, *91*(6), 537–8. doi:10.1038/sj.hdy.6800370

Dhondalay, G. K.,. Tong D. L. & Ball G. R. (2011). “Estrogen Receptor Status Prediction for Breast Cancer Using Artificial Neural Network.” 2011 *International Conference on Machine Learning and Cybernetics*, July. Ieee, 727–31. doi:10.1109/ICMLC.2011.6016771.

Diamandis, E. P. (2004). “Analysis of Serum Proteomic Patterns for Early Cancer Diagnosis: Drawing Attention to Potential Problems.” JNCI Journal of the National Cancer Institute, 96(5), 353–356. doi:10.1093/jnci/djh056

Distler, U; Kuharev, J; Navarro, P; Levin, Y; Schild, H & Tenzer, S (2014) “Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics”. *Nature Methods*; Feb2014, Vol. 11 Issue 2, p167

Dreiseitl, S., Ohno-machado L., Kittler H., Vinterbo S. & Binder M. (2001). “A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions.” *Journal of Biomedical Informatics* 36: 28–36. doi:10.1006/jbin.2001.1004.

Drug Bank (2016) Available at http://www.drugbank.ca/drugs/DB06268 Accessed January 2016

Dunn, WB., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., Brown, M., Knowles, JD., Halsall, A., Haselden, JN., Nicholls, AW., Wilson, ID., Kell, DB & Goodacre, R. (2011). “Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry”. *Nature Protocols*. 1060-1083

Eckstein, N. (2011). “Platinum resistance in breast and ovarian cancer cell lines”. *Journal of Experimental & Clinical Cancer Research : CR, 30*(1), 91. doi.org/10.1186/1756-9966-30-91

Eisen, M. B., Spellman P.T., Brown P. O. & Botstein D. (1999). “Proc. Natl. Acad. Sci. USA.” *Proc. Natl. Acad. Sci.*, no. 22: 12930–33.

ELRIG. (2016). “European Laboratory Research & Innovation Group.” In *European Laboratory Research & Innovation Group*, Lab book 007 pages114 115. Nottingham 23March2106. doi:10.1017/CBO9781107415324.004.

EMBL-EBI Online Training Available at: http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-you-will-learn/what-next-generation-dna- (Accessed Feb 2015)

Engwegen, J. Y. M. N., Gast, M.-C. W., Schellens, J. H. M. & Beijnen, J. H. (2006). ”Clinical proteomics: searching for better tumour markers with SELDI-TOF mass spectrometry”. *Trends in pharmacological sciences*, *27*(5), 251–9. doi:10.1016/j.tips.2006.03.003

Erickson, B. K., Martin, J. Y., Shah, M. M., Straughn, J. M. & Leath, C. a. (2014). “Reasons for failure to deliver National Comprehensive Cancer Network (NCCN)-adherent care in the treatment of epithelial ovarian cancer at an NCCN cancer centre”. *Gynecologic Oncology*, 133(2), 142–6. doi:10.1016/j.ygyno.2014.02.006

Escher, C., Reiter L., Maclean B., Ossola R., Herzog F., Chilton J., Maccoss M. J. & Rinner O. (2012). “Using iRT , a Normalized Retention Time for More Targeted Measurement of Peptides.” *Proteomics*, 1111–21. doi:10.1002/pmic.201100463.

Ettre, L S. (1993). “Nomenclature for Chromatography (IUPAC Recommendations 1993).” *Pure and Applied Chemistry* 65 (4): 819–72. doi:.org/10.1351/pac19936504819.

Faca, V. Pitteri, S. J., Newcomb, L., Glukhova, V., Phanstiel, D., Krasnoselsky, A., Zhang, Q., Struthers, J., Wang, H., Eng, J., Fitzgibbon, M., McIntosh, M. & Hanash, S. (2007) “Contribution of protein fractionation to depth of analysis of the serum and plasma proteomes.” *J. Proteome Res.* 6, 3558–3565.

Farley, J., Smith L. M., Darcy K. M., Sobel E., O’Connor D., Henderson B., Morrison L. E. & Birrer M. J. (2003). “Cyclin E Expression Is a Significant Predictor of Survival in Advanced, Suboptimally Debulked Ovarian Epithelial Cancers: A Gynaecologic Oncology Group Study.” *Cancer Research* 63 (6): 1235–41.

Fenn, J. B. (2002). “Electrospray Ionization Mass Spectrometry: How It All Began.” *Journal of Biomolecular Techniques* 13 (3): 101–18.

Fleury, H., Communal L., Carmona E., Portelance L., Arcand S. L., Rahimi K., Tonin P. N., Provencher D. & Mes-Masson A. (2015). “Novel High-Grade Serous Epithelial Ovarian Cancer Cell Lines That Reflect the Molecular Diversity of Both the Sporadic and Hereditary Disease.” *Genes & Cancer* 6 (9-10): 378–98. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4633166&tool=pmcentrez&rendertype=abstract.

Fu, P., Thompson J. A. & Bach L. A. (2007). “Promotion of Cancer Cell Migration: An Insulin-like Growth Factor (IGF)-Independent Action of IGF-Binding Protein-6.” *Journal of Biological Chemistry* 282 (31): 22298–306. doi:10.1074/jbc.M703066200.

Galazis, N., Olaleye, O., Haoula, Z., Layfield, R. & Atiomo, W. (2012). “Proteomic biomarkers for ovarian cancer risk in women with polycystic ovary syndrome: A systematic review and biomarker database integration.” *Fertility and Sterility*, 98(6). doi.org/10.1016/j.fertnstert.2012.08.002

Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X. V., Ahmadifar, M., Birrer, M.J., Parmigiani, G., Huttenhower, C., Waldron, L. (2013). "curatedOvarianData: Clinically Annotated Data for the Ovarian Cancer Transcriptome." Database : *The Journal of Biological Databases and Curat*ion 2013: bat013. doi:10.1093/database/bat013.

Garg, M., Kanojia, D., Khosla, A., Dudha, N., Sati, S., Chaurasiya, D., Jagadish, N., Seth, A. Kumar, R., Gupta, S., Gupta, A., Lohiya, N. K. & Suri, A. (2008). "Sperm-Associated Antigen 9 Is Associated with Tumor Growth, Migration, and Invasion in Renal Cell Carcinoma." *Cancer Research* 68 (20): 8240–48. doi:10.1158/0008-5472.CAN-08-1708.

Gates, P. 2016. "High Performance Liquid Chromatography Mass Spectrometry HPLC MS." *The University of Bristol, Life Sciences Mass Spectrometry Facility Natural Environment Research Council*. Accessed October 19. http://www.bris.ac.uk/nerclsmsf/techniques/hplcms.html.

GeneCards (2013) Available at http://www.genecards.org Accessed October 2013

GeneCards (2014) Available at http://www.genecards.org/cgi-bin/carddisp.pl?gene=EDNRA&search=68d7ab3090aceaa545e123e269be507a Accessed June 2014

GeneCards (2015) Available at http://www.genecards.org Accessed April 2015

Gene Expression Omnibus (2015) Available at https://www.ncbi.nlm.nih.gov/geo/ (Accessed March 2015)

Gene Ontology Consortium. (2004). "The Gene Ontology (GO) database and informatics resource." *Nucleic Acids Research*, 32 (Database issue), 258D–261. doi.org/10.1093/nar/gkh036

Ghiselli, G. (2006). "SMC3 Knockdown Triggers Genomic Instability and p53-Dependent Apoptosis in Human and Zebrafish Cells." *Molecular Cance*r 5: 52. doi:10.1186/1476-4598-5-52.

Gil, C., Calvete, J. J. & Corrales, F. J. (2015). "The proteome quest to understand biology and disease (HUPO 2014)." *Journal of Proteomics*, 127, 223–224. doi.org/10.1016/j.jprot.2015.09.025

Gillet, L. C, Navarro P., Tate S., Ro H., Selevsek N., Reiter L., Bonner R. & Aebersold R. (2012). "Targeted Data Extraction of the MS / MS Spectra Generated by Data-Independent Acquisition : A New Concept for Consistent and Accurate Proteome Analysis ." *Technological Innovation and Resources* 11 (6): 1–17. doi:10.1074/mcp.O111.016717.

Goossens, N., Nakagawa, S., Sun, X. & Hoshida, Y. (2015). HHS Public Access, 73(4), 389–400. doi.org/10.1530/ERC-14-0411.Persistent

Greaves, J & Roboz, J. (2014) "Mass Spectrometry for the Novice." *CRC Press Taylor and Francis Group*, New York pp

Griffiths, J. (2008). “A brief history of mass spectrometry.” *Analytical Chemistry*, 80, pp.5678–5683.3

Griffiths, J. R., Chicooree, N., Connolly, Y., Neffling, M., Lane, C. S., Knapman, T. & Smith, D. L (2014). “Mass spectral enhanced detection of ubls using SWATH acquisition: MEDUSA - Simultaneous quantification of SUMO and ubiquitin-derived isopeptides.” *Journal of the American Society for Mass Spectrometry*, 25, pp.767–777.

Gundry, R. L. & Cotter, R. J. (2007) “The Albuminome as a Tool for Biomarker Discovery, in Clinical Proteomics: From Diagnosis to Therapy” *Proteomics - Clinical Applications* (eds J. E. Van Eyk and M. J. Dunn), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany. doi: 10.1002/9783527622153.ch18

Gundry, R. L., Fu Q., Jelinek C. A., Van Eyk J. E. & Cotter R. J. (2007). “Investigation of an Albumin-Enriched Fraction of Human Serum and Its Albuminome.” *Proteomics - Clinical Applications*. doi:10.1002/prca.200600276.

Gyorffy, B., Surowiak P., Budczies J. & Lanczky A. (2013). “Online Survival Analysis Software to Assess the Prognostic Value of Biomarkers Using Transcriptomic Data in Non-Small-Cell Lung Cancer.” *PLoS ONE* 8 (12). doi:10.1371/journal.pone.0082241.

Hale, J. E., Butler J. P., Gelfanova V., You J. & Knierman M. D. (2004). “A Simplified Procedure for the Reduction and Alkylation of Cysteine Residues in Proteins prior to Proteolytic Digestion and Mass Spectral Analysis” *Analytical Biochemistry* 333: 174–81. doi:10.1016/j.ab.2004.04.013.

Hager, Y. (2008). “Going with the Flow.” Chemistry World, 70. http://www.rsc.org/images/ Water Company profile_tcm18-131651.pdf

Halling, K. C., Schrijver I. & Persons D. L. (2012). “Test Verification and Validation for Molecular Diagnostic Assays.” *Archives of Pathology & Laboratory Medicine* 136 (1): 11–13. doi:10.5858/arpa.2011-0212-ED.

Hanahan, D., Weinberg, R. A. & Francisco, S. (2000). “The Hallmarks of Cancer Review.” *Cell* University of California at San Francisco, 100, 57–70.

Hanahan, D. & Weinberg, R. A. (2011). “Hallmarks of cancer: the next generation.” *Cell*, 144(5), 646–74. doi.org/10.1016/j.cell.2011.02.013

Hanash, S.M., Pitteri, S.J., Faca, V.m. (2008) “Mining the plasma proteome for cancer biomarkers.” *Nature* 452 (7187) p. 571-9

Hao, Z., Zhang, Y., Eliuk, S. & Blethrow, J. (2012). “A Quadrupole-Orbitrap Hybrid Mass Spectrometer Offers Highest Benchtop Performance for In-Depth Analysis of Complex Proteomes.” *Thermofisher*. Cn. Available at: http://thermofisher.cn/Resources/201301/16164727718.pdf.

Harris, E. K., Kanofsky, P., Shakarji, G. & Cotlove, E. (1970). “Biological and analytic components of variation in long-term studies of serum constituents in normal subjects. II. Estimating biological components of variation.” *Clinical Chemistry*, 16(12), 1022–1027.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Hubbard, T. J. (2012). “GENCODE: The reference human genome annotation for the ENCODE project.” *Genome Research*, 22(9), 1760–1774. doi.org/10.1101/gr.135350.111

Hattan, S. J., Marchese, J., Khainovski, N., Martin, S. & Juhasz, P. (2005). “Comparative Study of [Three] LC-MALDI Workflows for the Analysis of Complex Proteomic Samples research articles.” *Journal of proteome research*1 931–1941.

Hawkins, D.M. (2004). “The Problem of Overfitting.” *J. Chem. Inf. Comp. Sci* 44: 1–12. doi:10.1021/ci0342472.

Hays, J. L., Kim, G., Giuroiu, I. & Kohn, E. C. (2010). “Proteomics and ovarian cancer: integrating proteomics information into clinical care.” *Journal of proteomics*, *73*(10), 1864–72. doi:10.1016/j.jprot.2010.05.013

Helleman, J., Jansen, M. P. H. M., Span, P. N., Van Staveren, I. L., Massuger, L. F. a G., Meijer-Van Gelder, M. E., Berns, E. M. J. J. (2006). “Molecular profiling of platinum resistant ovarian cancer.” *International Journal of Cancer*. doi.org/10.1002/ijc.21599

Heller, M. J. (2002). “DNA Microarray Technology: Devices , Systems , and Applications INTRODUCTION.” *Annu. Rev. Biomed. Eng* 4: 129–53. doi:10.1146/annurev.bioeng.4.020702.153438.

Henry, N. L. & Hayes, D. F. (2012). “Cancer biomarkers.” Molecular Oncology, 6(2), 140–146. doi.org/10.1016/j.molonc.2012.01.010

Hickman, G. J, Boocock D. J., Pockley A. G. & Perry C. C. (2016). “The Importance and Clinical Relevance of Surfaces in Tissue Culture.” ACS Biomaterial Science and Engineering 2 (2): 152–64. doi:10.1021/acsbiomaterials.5b00403.

Hillenkamp, F. & Karas M. (2000). “Matrix-Assisted Laser Desorption/ionisation, an Experience.” *International Journal of Mass Spectrometry* 200 (1-3): 71–77. doi:10.1016/S1387-3806(00)00300-6

Ho, J., Tan M. K., Go D. B., Yeo L. Y., Friend J. R. & Chang H-c. (2011). “Paper-Based Microfluidic Surface Acoustic Wave Sample Delivery and Ionization Source for Rapid and Sensitive Ambient Mass Spectrometry.” *Analytical Chemistry* 83: 3260–66. doi:org/10.1021/ac200380q.

Hoffman, E. de. (1996). “Tandem Mass Spectrometry a Primer.” *Journal of Mass Spectrometry* 31: 129–37. doi:10.1002/(SICI)1096-9888(199602)31:2<129::AID-JMS305>3.0.CO;2-T.

Hsieh, E. J., Bereman, M. S., Durand, S. & Valaskovic, G. A. (2014). “Effects of Column and Gradient Lengths on Peak Capacity and Peptide Identification in nanoflow LC-MS/MS of Complex Proteomic Samples.” *J Am Mass Spetrom.* 24(1), 148–153. doi.org/10.1007/s13361-012-0508-6.Effects

Huang, D. W., Sherman B. T. & Lempicki R. A. (2008). “Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources.” *Nature Protocols*, no. 2: 44–57. doi:10.1038/nprot.2008.211.

Huang, D. W., Sherman B.T., Tan Q., Kir J., Liu D.,Bryant D., Guo Y., Stephens R., Baseler M. W., Lane H. C. & Lempicki R A . (2007). “DAVID Bioinformatics Resources : Expanded Annotation Database and Novel Algorithms to Better Extract Biology from Large Gene Lists.” *Nucleic Acids Research* 35: 169–75. doi:10.1093/nar/gkm415.

Human Protein Atlas. (2014). “The Human Protein Atlas.” Accessed January 2014. Available at: http://www.proteinatlas.org/.

Illumina. (2016). BaseSpace 2016. Retrieved May 20, 2012, from https://basespace.illumina.com/apps/

Illustrated Health. (2014). The difference between NORMAL and CANCER cells. [Online Video]. 1 November 2016. Available from: https://www.youtube.com/watch?v=1MuWqQiqWnc. [Accessed: 4 January 2016].
IMEx (2015) Available at http://www.imexconsortium.org/home Accessed Feb2015

IntAct EMBL-EBI (2013) Available at http://www.ebi.ac.uk/intact/ Accessed December 2013

Irving-Rodgers, H. F. & Rodgers, R. J. (2006). “Extracellular matrix of the developing ovarian follicle.” *Semin Reprod Med*, 24(4), 195–203. doi.org/10.1055/s-2006-948549

Jacobs, I. J. & Menon, U. (2004). “Progress and challenges in screening for early detection of ovarian cancer.” *Molecular & Cellular Proteomics : MCP*, *3*(4), 355–66. doi:10.1074/mcp.R400006-MCP200

Jacobs, I. J., Menon, U., Ryan, A., Gentry-Maharaj, A., Burnell, M., Kalsi, J. K., Skates, S. J. (2015). “Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial.” *The Lancet, 6736*(15), 1–12. doi.org/10.1016/S0140-6736(15)01224-6

Jennings, K. R. (2012). “A History of European Mass Spectrometry”. Edited by Keith R Jennings. Journal of The American Society for Mass Spectrometry. Chichester, England: IM Publications LLP. doi:10.1007/s13361-013-0731-9.

Jensen, O. N. (2004). “Modification-Specific Proteomics: Characterization of Post-Translational Modifications by Mass Spectrometry.” *Current Opinion in Chemical Biology* 8 (1): 33–41. doi:10.1016/j.cbpa.2003.12.009.

Jin, D.-I., Lee, S. W., Han, M.-E., Kim, H.-J., Seo, S.-A., Hur, G.-Y., Oh, S.-O. (2012). “Expression and roles of Wilms’ tumor 1-associating protein in glioblastoma.” *Cancer Science*. doi.org/10.1111/cas.12022

Jones, S., Wang TL., Shih IM., Mao TL., Nakayama K., Glas R., Slamon D., Diaz Jr L.A., Vogelstein B., Kenneth W., Velculescu V. E. & Papadopoulos N. (2011). “Frequent Mutations of Chromatin Remodelling Gene ARID1A in Ovarian Clear Cell Carcinoma.” *Science* 330 (6001): 228–31. doi:10.1126/science.1196333.Frequent.

Kafetzopoulou, L. E., Boocock D. J., Dhondalay G. K. R., Powe D. G. & Ball G. R. (2013). “Biomarker Identification in Breast Cancer : Beta-Adrenergic Receptor Signalling and Pathways to Therapeutic Response.” *Computational and Structural Biotechnology Journal* 6 (7). doi:org/10.5936/csbj.201303003.

Kanehisa, M. & Goto S. (2000). “KEGG : Kyoto Encyclopaedia of Genes and Genomes” *Nucleic Acids Research* 28 (1): 27–30.

Kanety, H., Kattan, M., Goldberg, I., Kopolovic, J., Ravia, J., Menczer, J., Karasik, A. “Increased insulin-like growth factor binding protein-2 (IGFBP-2) gene expression and protein production lead to high IGFBP-2 content in malignant ovarian cyst fluid.” *Br. J. Cancer*. 73, 1069–73 (1996).

Kaplan, E L, & Meier P. (1958). “Nonparametric Estimation from Incomplete Observations A.” *Journal of the American Statistical Association* 53 (282): 457–81.

KEGG Kyoto Encycopedia of Genes and Genomes (2015) Available at: http://www.genome.jp/kegg/ Accessed November 2015

KEGG Kyoto Encycopedia of Genes and Genomes (2016) Available at: http://www.genome.jp/kegg/kegg2.html Accessed January 2016

Kelland, L. R. (2000). Preclinical perspectives on platinum resistance. *Drugs, 59 Suppl 4,* 1–8; discussion 37–38. doi.org/10.2165/00003495-200059004-00001

Kerrien, Samuel, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-carter, Carol Chen, Margaret Duesbury, Dumousseau M., Feuermann M., Hinz U., Jandrasits C., Jimenez R. C., Khadake J., Mahadevan U., Masson P., Pedruzzi I., Pfeiffenberger E., Porras P., Raghunath A., Roechert B., Orchard S. & Hermjakob H. (2012). “The IntAct Molecular Interaction Database in 2012.” *Nucleic Acids Research* 40 (November 2011): 841–46. doi:10.1093/nar/gkr1088.

Kim, A., Ueda Y., Naka T. & Enomoto T. 2012. “Therapeutic Strategies in Epithelial Ovarian Cancer.” *Journal of Experimental & Clinical Cancer Research* : CR 31 (1). BioMed Central Ltd: 14. doi:10.1186/1756-9966-31-14.

KMPlotter. (2014) “KMPlotter.”
http://kmplot.com/analysis/index.php?p=service&cancer=ovar Accessed June 2014

Koboldt, D. C., Steinberg K. M., Larson D. E., Wilson R. K. & Mardis E.R. (2013). “Review The Next-Generation Sequencing Revolution and Its Impact on Genomics.” *Cell* 155 (1). Elsevier Inc.: 27–38. doi:10.1016/j.cell.2013.09.006.

Kong, F., White C. N., Xiao X., Feng Y., Xu C. & He Da. (2006). “Using Proteomic Approaches to Identify New Biomarkers for Detection and Monitoring of Ovarian Cancer.” *Gynaecologic Oncology* 100: 247–53. doi:10.1016/j.ygyno.2005.08.051.

Kozak, K.R., Amneus, M.W. Pusey, S.M., Su, F., Luong, M.N., Luong, S.A. Reddy, S.T., Farias-Eisner, R. (2003). "Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: potential use in diagnosis and prognosis." *Proceedings of the National Academy of Sciences of the United States of America*, *100*(21), 12343–8. doi:10.1073/pnas.2033602100

Kuo, KT., Mao TL., Jones S., Veras E., Ayhan A., Wang TL, Glas R., Slamon D., Velculescu V. E., Kuman R. J. & Shih IM. (2009). "Frequent Activating Mutations of PIK3CA in Ovarian Clear Cell Carcinoma." *The American Journal of Pathology* 174 (5): 1597–1601. doi:10.2353/ajpath.2009.081000.

Laatio, L., Myllynen, P., Serpi, R., Rysä, J., Ilves, M., Lappi-Blanco, E., Ruskoaho, H., Vähäkangas, K., Puistola, U (2011). "BMP-4 expression has prognostic significance in advanced serous ovarian carcinoma and is affected by cisplatin in OVCAR-3 cells." *Tumor Biol*. 32, 985–995.

Lancashire, L. J., Rees, R. C. & Ball, G, R,. (2008). "Identification of Gene Transcript Signatures Predictive for Estrogen Receptor and Lymph Node Status Using a Stepwise Forward Selection Artificial Neural Network Modelling Approach." *Artificial Intelligence in Medicine* 43 (2): 99–111. doi:10.1016/j.artmed.2008.03.001.

Lancashire, L. J., Lemetre, C. & Ball, G. R. (2009). "An introduction to artificial neural networks in bioinformatics--application to complex microarray and mass spectrometry datasets in cancer studies." *Briefings in bioinformatics*, *10*(3), 315–29. doi:10.1093/bib/bbp012

Lancashire, L.J., Lemetre, C., Ball, G.R. (2009) "An introduction to artificial neural networks in bioinformatics application to complex microarray and mass spectrometry datasets in cancer studies." *Artificial neural networks in bioinformatics* 10 (3). doi:10.1093/bib/bbp012.

Lancashire, L. J., Powe, D. G., Reis-Filho, J. S., Rakha, E., Lemetre, C., Weigelt, B., Ball, G. R. (2010). "A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks." *Breast Cancer Research and Treatment*, *120*(1), 83–93. doi.org/10.1007/s10549-009-0378-1

Law, K. P., and Yoon Pin Lim. (2013). "Recent Advances in Mass Spectrometry : Data Independent Analysis and Hyper Reaction Monitoring." *Expert Review of Proteomics* 10 (6): 551–66. doi:org/10.1586/14789450.2013.858022.

Lec, S. & Guĕgan, J.-F. (2000). "Artificial Neuronal Networks Application in Ecology and Evoloution." (A. R, U.Forstner, & S. W, Eds.). Springer-Verlag Berlin Heidelberg New York.

Lee, E.-J., Mircean, C., Shmulevich, I., Wang, H., Liu, J., Niemistö, A., Kavanagh, J.J., Lee, J.-H., Zhang, W. (2005). "Insulin-like growth factor binding protein 2 promotes ovarian cancer cell invasion." *Mol. Cance*r. 4, 7

Licata, L., Briganti L., Peluso D., Perfetto L., Iannuccelli M., Galeota E., Sacco F.Sacco F., Palma ., Nardozza A. P., Santonico E., Castagnoli L. & Cesareni, G. (2012). "MINT , the Molecular Interaction Database : 2012 Update." *Nucleic Acids Research* 40 (November 2011): 857–61. doi:10.1093/nar/gkr930

Life Technologies (2013) Available at http://www.invitrogen.com/site/us/en/home/References/Ambion-Tech-Support/rna-gene-expression/general-articles/the-basics-what-is-a-gene-array.html Accessed May 2013

Little, N. a, Hastie, N.D., Davies, R.C. (2000) “Identification of WTAP, a novel Wilms’ tumour 1-associating protein.” *Hum. Mol. Genet*. 9, 2231–2239.

Machin, D., Cheung Y. B. & Parmar M. K.B. (2006). *Survival Analysis: A Practical Approach*. Chapter 9. Second. John Wiley & Sons, Ltd. doi:10.1002/0470034572.

Majors, R. E. (2012). “Current Trends in HPLC Column Usage.” *LC-GC North America*. 30 (1): 20,22,24,26,28,30–32,34.

Maliniemi, P., Carlsson E., Kaukola A., Ovaska K., Niiranen K., Saksela O., Jeskanen L., Hautaniemi S. & Ranki A. (2011). “NAV3 Copy Number Changes and Target Genes in Basal and Squamous Cell Cancers.” *Experimental Dermatology* 20: 926–31. doi:10.1111/j.1600-0625.2011.01358.x.

Mallick, P. & Kuster B. (2010). “Proteomics: A Pragmatic Perspective.” Nature Biotechnology 28 (7). *Nature Publishing Group*: 695–709. doi:10.1038/nbt.1658.

Marchini, S., Fruscio, R., Clivio, L., Beltrame, L., Porcu, L., Fuso Nerini, I., D’Incalci, M. (2013). “Resistance to platinum-based chemotherapy is associated with epithelial to mesenchymal transition in epithelial ovarian cancer.” *European Journal of Cancer* (Oxford, England : 1990), 49(2), 520–30. doi:10.1016/j.ejca.2012.06.026

Marcus, C. S., Maxwell, G. L., Darcy, K. M., Hamilton, C. A, & McGuire, W. P. (2014). “Current approaches and challenges in managing and monitoring treatment response in ovarian cancer.” *Journal of Cancer*, 5(1), 25–30. doi:10.7150/jca.7810

Margulies, D. H. & Shevack E. M. (1996). “Removal of Albumin from Multiple Human Serum Samples.” *BioTechniques* 20 (1): 1–2.

Martin, L. P., Hamilton, T. C. & Schilder, R. J. (2008). “Platinum resistance: the role of DNA repair pathways.” *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *14*(5), 1291–5. doi:10.1158/1078-0432.CCR-07-2238

Marx, V. (2013). “Targeted Proteomics.” *Nature Methods*, *10*(1), 19–22. doi:10.1038/nmeth.2285

Mathieu, M G, Knights A.J., Pawelec G., Riley C.L., Wernet D., Lemonnier F.A., Thor P., Ludmila S., Rees R.C. & Mcardle S.E.B. 2007. “HAGE, a Cancer / Testis Antigen with Potential for Melanoma Immunotherapy : Identification of Several MHC Class I / II HAGE-derived Immunogenic Peptides.” *Cancer Immunol Immunother* (2007) 56:1885–1895 1885–95. doi:10.1007/s00262-007-0331-2.

Matrix DB (2015) Available at http://matrixdb.ibcp.fr/?conversationContext=1 Acessed Feb 2015

Matsuzaki, S., Yoshino, K., Ueda, Y., Matsuzaki, S., Kakuda, M., Okazawa, A., Kimura, T. (2015). "Potential targets for ovarian clear cell carcinoma: a review of updates and future perspectives." *Cancer Cell International, 15*(1), 117. doi.org/10.1186/s12935-015-0267-0

Mclafferty, Fred W. 1981. "Spectrometry." *Science* 214 (4518): 280–87. doi:10.1126/science.7280693.

Menon, U., Griffin, M. & Gentry-Maharaj, A. (2014). "Ovarian cancer screening--current status, future directions." *Gynaecologic Oncology, 132*(2), 490–5. doi.org/10.1016/j.ygyno.2013.11.030

Miller, D. S., Blessing, J. a, Krasner, C. N., Mannel, R. S., Hanjani, P., Pearl, M. L., Boardman, C. H. (2009). "Phase II evaluation of pemetrexed in the treatment of recurrent or persistent platinum-resistant ovarian or primary peritoneal carcinoma: a study of the Gynaecologic Oncology Group.*" Journal of Clinical Oncology*, 27(16), 2686–91. doi:10.1200/JCO.2008.19.2963

Millipore Corporation (2005). "User guide for Reversed Phase ZipTip®, Pipette tips for sample preparation." P36110, Rev G, 05/05.

Miow, Q. H., Tan, T. Z., Ye, J., Lau, J. a, Yokomizo, T., Thiery, J.-P. & Mori, S. (2014) "Epithelial-mesenchymal status renders differential responses to cisplatin in ovarian cancer." *Oncogene*. 1-9, 2014, May 2. doi:10.1038/onc.2014.136

Morcavallo, A., Buraschi, S., Xu, S.-Q., Belfiore, A., Schaefer, L., Iozzo, R. V, & Morrione, A. (2014). "Decorin differentially modulates the activity of insulin receptor isoform A ligands." *Matrix Biology : Journal of the International Society for Matrix Biology*, 35, 82–90. doi.org/10.1016/j.matbio.2013.12.010

Naora, H. & Montell, D. J. (2005). "Ovarian cancer metastasis: integrating insights from disparate model organisms." *Nature Reviews. Cancer*, 5(5), 355–66. doi:10.1038/nrc1611

Nash, M. A, Loercher, A. E. & Freedman, R. S. (1999). "In vitro growth inhibition of ovarian cancer cells by decorin: synergism of action between decorin and carboplatin." *Cancer Research*, 59(24), 6192–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10626812

National Cancer Institute, Offfice of Cancer Clinical Proteomics Research CPTAC Data Portal 2015. https://cptac-data-portal.georgetown.edu/cptac/s/S013 Accessed December 2015

National Caner Institute, Office of Cancer Clinical Proteomics Research 2016, http://proteomics.cancer.gov/programs/cptacnetwork/background Accessed March 2016

Nature Editorial. (2013). "Method of the Year 2012." *Nature Methods* 10 (1): 2013. doi:10.1038/nmeth.2329.

NBDA. (2016). "NBDA National Biomarker Development Aliance." http://nbdabiomarkers.org/nbda-events/workshops/workshop-i/workshop-i-reports. Accessed May 2016.

NCBI Gene ID: 8496, updated on 6-Dec- (2015) http://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=ShowDetailView&TermToSearch=8496 (accessed December 2015)

Nelson, J., Bagnato, A., Battistini, B. & Nisen, P. (2003). "The endothelin axis: emerging role in cancer." *Nature Reviews. Cancer*, 3(2), 110–6. doi.org/10.1038/nrc99

Nesvizhskii, A. (2007). "Protein Identification by Tandem Mass Spectrometry and Sequence Database Searching." In *Methods in Molecular Biology*, edited by R Mattiesen, 367:87–119. Totowa: Humana. doi:10.1385@1-59745-275-0@87.

Neubert, H., Bonnert, T. P., Rumpel, K., Hunt, B. T., Henle, E. S. & James, I. T. (2008). „Label-free detection of differential protein expression by LC/MALDI mass spectrometry." *Journal of proteome research*, *7*(6), 2270–9. doi:10.1021/pr700705u

NICE. (2016). "Ovarian Cancer: Recognition and Initial Management Clinical Guideline [CG122] Published Date: April 2011." Available at: https://www.nice.org.uk/guidance/cg122/chapter/ftn.footnote_4. Accessed October 2016

Nier, A O. (1991). "The Development of a High Resolution Mass Spectrometer: A Reminiscence." *American Society for Mass Spectrometry* 2: 447–52.

Nossov, V., Amneus, M., Su, F., Lang, J., Janco, J. M. T., Reddy, S. T. & Farias-Eisner, R. (2008). "The early detection of ovarian cancer: from traditional methods to proteomics. Can we really do better than serum CA-125?" *American Journal of Obstetrics and Gynaecology*, 199(3), 215–23. doi:10.1016/j.ajog.2008.04.009

Nosov, V., Su, F., Amneus, M., Birrer, M., Robins, T., Kotlerman, J., Reddy, S. & Farias-Eisner, R. (2009). "Validation of serum biomarkers for detection of early-stage ovarian cancer." *American journal of obstetrics and gynaecology*, *200*(6), 639.e1–5. doi:10.1016/j.ajog.2008.12.042

Obermajer, N., Muthuswamy, R., Odunsi, K., Edwards, R. P. & Kalinski, P. (2011). "PGE(2)-induced CXCL12 production and CXCR4 expression controls the accumulation of human MDSCs in ovarian cancer environment." *Cancer Research*, 71(24), 7463–70. doi:10.1158/0008-5472.CAN-11-2449

Oesterling, J. E. (1991). "Prostate specific antigen: a critical assessment of the most useful tumor marker for adenocarcinoma of the prostate." *The Journal of Urology*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1707989

Ong, K. R., Sims, A. H., Harvie, M., Chapman, M., Dunn, W. B., Broadhurst, D., Howell, A. (2009). "Biomarkers of dietary energy restriction in women at increased risk of breast cancer." *Cancer Prevention Research* (Philadelphia, Pa.), 2(8), 720–31. doi.org/10.1158/1940-6207.CAPR-09-0008

Ontario Cancer Biomarker Network (2012). Available at: http://www.ocbn.ca/insolution.html Accessed January 2012

Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., Bridge, A., Briganti, L., Brinkman, F. S. L., Cesareni, G., Chatr-aryamontri, A., Chautard, E., Chen, C.l, Dumousseau, M., Goll, J., Hancock, R. E. W., Hannick, L. I, Jurisica, I., Khadake, J., Lynn, D. J., Mahadevan, U., Perfetto, L., Raghunath, A. & Ricard-blum, S. (2012). “Protein Interaction Data Curation : The International Molecular Exchange (IMEx) Consortium.” *Nature Methods* 9 (4): 345–50. doi:10.1038/nmeth.1931.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., Van Roey, K., Cesareni, G. & Hermjakob, H. (2014). “The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases.” *Nucleic Acids Research*, 42 (November 2013), pp.358–363.

Oxford Dictionaries Language matters (2015). Available at: http://www.oxforddictionaries/definition/english/biomarker Accessed January 2015

Pan S., Cheng l., White J.T., Lu W., Utleg A.G.,, Yan X., Urban N.D., Drescher C.W., Hood L. & Lin B. (2009) “Quantitative Proteomics Analysis Integrated with Microarray Data Reveals That Extracellular Matrix Proteins, Catenins, and P53 Binding Protein 1 Are Important for Chemotherapy Response in Ovarian Cancers.” *OMICS: A Journal of Integrative Biology*. August 2009, Vol. 13, No. 4: 345-354

PANTHER. (2016). PANTHER. Retrieved May 1, 2016, from http://www.pantherdb.org/about.jsp

Park, SH., Lydia Cheung W.T., Wong A.S.T, & Leung P. C. K. (2008). “Estrogen Regulates Snail and Slug in the down-Regulation of E-Cadherin and Induces Metastatic Potential of Ovarian Cancer Cells through Estrogen Receptor Alpha.” *Molecular Endocrinology* (Baltimore, Md.) 22 (9): 2085–98. doi:10.1210/me.2007-0512.

Pavlopoulou, A., Spandidos, D. & Michalopoulos, I. (2014). ”Human cancer databases (Review).” *Oncology Reports*, 3–18. doi.org/10.3892/or.2014.3579

Peckham, M., Knibbs, A and Paxton, S. The Histology Guide Universuty of Leeds (2004) at http://www.histology.leeds.ac.uk/female/FRS_ovarian_fol.php Accessed April 2016

Perkel, J. (2012). “Choosing the Optimal Ionization Source for Your Mass Spectrometry Needs.” Retrieved from http://www.biocompare.com/Editorial-Articles/41599-Choosing-the-Optimal-Ionization-Source-for-Your-Mass-Spectrometry-Needs/

Perkins, D. N., Pappin D. J. C., Creasy D. M. & Cottrell J. S. (1999). “Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data Proteomics and 2-DE.” *Electrophoresis* 20: 3551–67. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2

Peto, R. & Peto J. (1972). “Asymptotically Efficient Rank Invariant Test Procedures.” *Journal of the Royal Statistical Society* 135 (2): 185–207. doi:10.2307/2344317.

Petricoin III, E.F., Ardekani, A.M., Hitt, B.A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G.B., Simone, C., Fishman, D. A., Kohn, E. C., Liotta, L.A. (2002). "Mechanisms of disease Use of proteomic patterns in serum to identify ovarian cancer." *Mechanisms of Disease* 359, 572–577.

Picotti, P., Bodenmiller, B. & Aebersold, R. (2012). "Proteomics meets the scientific method." *Nature Methods*, *10*(1), 24–27. doi:10.1038/nmeth.2291

Podwojski, K., Fritsch, A., Chamrad, D. C., Paul, W., Sitek, B., Stühler, K., Mutzel, P., Stephan, C., Meyer, H. E., Urfer, W., Ickstadt, K. & Rahnenführer, J. (2009). "Retention time alignment algorithms for LC/MS data must consider non-linear shifts." *Bioinformatics (Oxford, England)*, *25*(6), 758–64. doi:10.1093/bioinformatics/btp052

Pontén, F, Jirström K. & Uhlen M. (2008). "The Human Protein Atlas a Tool for Pathology." *The Journal of Pathology* 216 (September): 387–93. doi:10.1002/path.2440.

Popple, a, L G Durrant, I Spendlove, P Rolland, I V Scott, S Deen, and J M Ramage. (2012). "The Chemokine, CXCL12, Is an Independent Predictor of Poor Survival in Ovarian Cancer." *British Journal of Cancer* 106 (7). Nature Publishing Group: 1306–13. doi:10.1038/bjc.2012.49.

Reactome A curated Pathway Database (2013) Available at http://www.reactome.org/ Accessed December 2013

Ringnér, M. (2008). "What Is Principal Component Analysis ?" *Nature Biotechnology* 26 (3): 303–4. doi:doi:10.1038/nbt0308-303.

Roboz, J. (2005). "Mass Spectrometry in Diagnostic Oncoproteomics." Cancer Investigation 23: 465–78. doi:10.1081/CNV-200067182.

Rosanò, L., Spinella F., Salani D., Di Castr V., Venuti A., Nicotra M. R., Natali P. G. & Bagnato A. (2003). "Therapeutic Targeting of the Endothelin a Receptor in Human Nasopharyngeal Carcinoma." *Cancer Science* 97 (28): 1388–95. doi:10.1111/j.1349-7006.2006.00333.x

Rosanò, L., Di Castro V., Spinella F., Decandia S., Natali, P. G. & Bagnato A. (2006). "ZD4054, a Potent Endothelin Receptor A Antagonist, Inhibits Ovarian Carcinoma Cell Proliferation." *Experimental Biology and Medicine* (Maywood, N.J.). http://www.ncbi.nlm.nih.gov/pubmed/16741063.

Rosanò, L., Cianfrocca, R., Spinella, F., Di Castro, V., Nicotra, M. R., Lucidi, A., Bagnato, A. (2011). "Acquisition of chemoresistance and EMT phenotype is linked with activation of the endothelin A receptor pathway in ovarian carcinoma cells." *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, 17(8), 2350–60. doi:10.1158/1078-0432.CCR-10-2325

Rumelhart, D. E., Hinton, G. E. & Williams, R. (1986). "Learning representations by back-propagating errors." *Nature*, (323), 533–536.

Russell, M. R., Walker, M. J., AU - Williamson, A. J. K., AU - Gentry-Maharaj, A., Ryan, A., Kalsi, J., Graham, R. L. J. (2016). "Protein Z: A putative novel biomarker for early detection of ovarian cancer." *International Journal of Cancer*, 1–9. doi.org/10.1002/ijc.30020

Sayer, R. A, Lancaster, J. M., Pittman, J., Gray, J., Whitaker, R., Marks, J. R. & Berchuck, A. (2005) "High insulin-like growth factor-2 (IGF-2) gene expression is an independent predictor of poor survival for patients with advanced stage serous epithelial ovarian cancer." *Gynaecologic Oncology*, 96(2), 355–61. (2005). doi:10.1016/j.ygyno.2004.10.012

Sciex. (2016). Sciex. http://sciex.com/applications/life-science-research/multi-omics-bioinformatics. Accessed May 2016

Scotting, P. (2011). "Cancer A Begginers Guide (Illustrate)." Oneworld Publications, Oxford, England 2011.

Scotton, C. J., Wilson, J. L., Milliken, D., Stamp, G. & Balkwill, F. R. (2001) "Epithelial Cancer Cell Migration : A Role for Chemokine Receptors?" *Cancer Research*, 4961–4965.

Sechi, S. & Chait B. T. (1998). "Modification of Cysteine Residues by Alkylation . A Tool in Peptide Mapping and Protein Identification Although Mass Spectrometric Peptide Mapping Has Be- of Proteins Isolated by Polyacrylamide Gel Electrophoresis." *Anal. Chem. 1* 70 (24): 5150–58. doi:10.1021/ac9806005.

Shepherd, T. G., Thériault B. L. & Nachtigal M. W. (2008). "Autocrine BMP4 Signalling Regulates ID3 Proto-Oncogene Expression in Human Ovarian Cancer Cells." *Gene* 414 (1-2): 95–105. doi:10.1016/j.gene.2008.02.015.

Sherman-Baust, C. A., Weeraratna, A. T., Rangel, L. B. a, Pizer, E. S., Cho, K. R., Schwartz, D. R., Morin, P. J. (2003). "Remodelling of the extracellular matrix through overexpression of collagen VI contributes to cisplatin resistance in ovarian cancer cells." Ca*ncer Cell,* 3(4), 377–386. doi.org/10.1016/S1535-6108(03)00058-8

Shin, H., Sheu, B., Joseph, M. & Markey, M. K. (2008). "Guilt-by-association feature selection: identifying biomarkers from proteomic profiles." *Journal of biomedical informatics*, *41*(1), 124–36. doi:10.1016/j.jbi.2007.04.003

Siegel, R., Naishadham, D. & Jemal, A. (2013) "Cancer Statistics 2013." *Ca Cancer J Clin.* 63(1), 11–30. doi:10.3322/caac.21166.

Sigma-Aldrich. (2012). "ProteoPrep ® 20 Plasma Immunodepletion Kit Cat.Nos. PROT20 PROT20S." *User Guide*. 1 1-7

Singh, R. & Mukhopadhyay, K., (2011). "Survival analysis in clinical trials: Basics and must know areas." *Perspectives in clinical research*, 2(4), pp.145–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3227332&tool=pmcentrez&rendertype=abstract (Accessed January 23, 2015).

Snel, B, Lehmann G., Bork P. & Huynen M. A. (2000). "STRING : A Web-Server to Retrieve and Display the Repeatedly Occurring Neighbourhood of a Gene." *Nucleic Acids Research* 28 (18): 3442–44. doi:10.1093/nar/gkv1145.

Snyder, L R., Kirkland, J, J., Dolan, J, W. "Introduction to Modern Liquid Chromatography." Third Edition.Jon Wiley and Sons, Inc., Hoboken New Jersey. 2010

Sorace, J. M. & Zhan, M. (2003) "A data review and re-assessment of ovarian cancer serum proteomic profiling." *BMC bioinformatics*, *4*, 24. doi:10.1186/1471-2105-4-24

Stein, D. R., Xiaojie H., Stuart J. M., Garrett R. W., Plummer F. A., Ball. T. B. & Carpenter M. S., (2013). "High pH Reversed-Phase Chromatography as a Superior Fractionation Scheme Compared to off-Gel Isoelectric Focusing for Complex Proteome Analysis." *Proteomics* 13 (20): 2956–66. doi:10.1002/pmic.201300079.

Strimbu, K. & Tavel, J. A. (2011). "What are Biomarkers?" *Curr Opin HIV AIDS*, 5(6), 463–466. doi.org/10.1097/COH.0b013e32833ed177.

STRING 9.05 (2013) Available at http://string-db.org/ Accessed May 2013.

Suckau, D., Resemann, A., Schuerenberg, M., Hufnagel, P., Franzen, J. & Holle, A. (2003). A novel MALDI LIFT-TOF/TOF mass spectrometer for proteomics. *Analytical and Bioanalytical Chemistry*, 376(7), 952–965. doi.org/10.1007/s00216-003-2057-0

Suresh, K., Thomas, S. V. & Suresh, G. (2011). "Design, Data Analysis and Sampling Techniques for Clinical Research." *Annals of Indian Academy of Neurology*. doi:10.4103/0972-2327.91951.

Takáts, Z., Wiseman J. W., Golgan B. & Cooks G. (2014). "Mass Spectrometry Sampling Under Ambient Conditions with Desorption Electrospray Ionization." *Science* 471 (2004): 10–13. doi:10.1126/science.1104404.

Tanaka, K., Waki ., Ido Y., Akita S. & Yoshida Y. 1988. "Protein and Polymer Analyses up to Mlz 100 000 by Laser Ionization Time-of-Flight Mass Spectrometry*." Rapid Communications in Mass Spectrometry* 2 (8): 151–53. doi:10.1002/rcm.1290020802.

Tang, N., Tornatore P. & Weinberger. S.R. 2004. "Current Developments in SELDI Affinity Technology." *Mass Spectrometry Reviews* 23 (March 2003): 34–44. doi:10.1002/mas.10066.

Tang, H.-Y., Beer, L. A, Tanyi, J. L., Zhang, R., Liu, Q. & Speicher, D. W. (2013). "Protein isoform-specific validation defines multiple chloride intracellular channel and tropomyosin isoforms as serological biomarkers of ovarian cancer." *Journal of proteomics*, *89*, 165–78. doi:10.1016/j.jprot.2013.06.016

Taub, E. F., Deleo J. M. & Thompson B. (1983). "Sequential Comparative Hybridizations Analyzed by" *DNA* 2 (4): 309–27. doi:10.1089/dna.1983.2.309.

Tavani, A., Gallus, S., Lavecchia, C., Conti, E., Montella, M. & Franceschi, S. (2000). "Aspirin and Ovarian Cancer : An Italian Case-Control Study." *Annals of Oncology* 11: 11: 1171–73.

Thériault, B. L., Shepherd, T. G., Mujoomdar M, L. & Nachtigal M. W. (2007). "BMP4 Induces EMT and Rho GTPase Activation in Human Ovarian Cancer Cells." *Carcinogenesis* 28 (6): 1153–62. doi:10.1093/carcin/bgm015.

Thermo Fisher Scientific. (2016). “Thermo Fisher Scientifiic.” https://www.thermofisher.com/uk/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-post-translational-modification.html# Accessed May 2106.

Thomas, P. D., Kejariwal A., Campbell M. J., Mi H., Diemer K., Guo N., Ladunga I., Ulitsky-lazareva B., Muruganujan A., Rabkin S., Vandergriff J. A. & Doremieux O. (2003). “PANTHER : A Browsable Database of Gene Products Organized by Biological Function , Using Curated Protein Family and Subfamily Classification.” *Nucleic Acids Research* 31 (1): 334–41. doi:10.1093/nar/gkg115.

Timms, J, F., Menon, U., Deventyarov, D., Tiss, A., Camuzeaux, K, M., Nourentidnov, I., Burford, B., Smith, C., Gentry-Maharaj, A., Hallett, R., Ford, J., Luo, Z., Vovk, V., Gammerman, A., Cramer, R. & Jacobs Ian. (2011) “Early Detection of OvCa in Samples PreDiagnosis Using CA125 and MALDI-MS Peaks.” *Cancer Genomics and Proteomics* 8(289-306)

Tipping, M. E. (2001). “Sparse Bayesian Learning and the Relevance Vector Machine.” *Journal of Machine Learning Research*, no. 1: 211–44.

Tong, D. L., Coveney, C. & Ball, G. R. (2012). “MS-Labeller: Bioinformatics support for quality assessment on high resolution mass spectrometry sample,” *25*(*Bhi*), 964–967.

Uhlen, M., Per O., Linn F., Lundberg E., Jonasson K., Forsberg M., Zwahlen M., Kampf C., Wester K., Hober S., Wernerus H., Björling L. & Ponten F. (2010). “Towards a Knowledge-Based Human Protein Atlas.” *Nature Biotechnology* 28 (12): 1248–50. doi:10.1038/nbt1210-1248.

Vafadar-Isfahani, B., Laversin, S. A., Ahmad, M., Ball, G., Coveney, C., Lemetre, C., Katheen Miles, A. K. , Van Schalkwyk, G., Rees, R. & Matharoo-Ball, B. (2010). “Serum biomarkers which correlate with failure to respond to immunotherapy and tumor progression in a murine colorectal cancer model.” *Proteomics Clinical applications*, *4*(8-9), 682–696. doi:10.1002/prca.200900218

van Deemter, J J, Zuiderwef F. J. & Klinkenberg A. 1956. “Diffusion and Resistance to Mass Transfer Nonideality in Chromatography.” *Chemical Engineering Science* 5: 271–89. doi.10.1016/0009-2509(56)80003-1

Van den Berg, R. A, Hoefsloot, H. C. J., Westerhuis, J. A, Smilde, A. K. & Van der Werf, M. J. (2006). “Centering, scaling, and transformations: improving the biological information content of metabolomics data”. *BMC genomics*, *7*, 142. doi:10.1186/1471-2164-7-142

Vaughan, S., Road, C., Ka, L., Centre, S., Way, R. & Coukos, G. (2012) .”Rethinking Ovarian Cancer: Recommendations for Improving Outcomes.” *NIH Public Access* 11(10), 719–725. doi:10.1038/nrc3144

Vizcano, J. A., Csordas A., del-Toro N., Dianes, J., Griss J., Lavidas I., Mayer G., Perez-riverol Y., Reisinger, F., Ternent T., Xu QW., Wang R. & Hermjakob.(2016). “2016 Update of the PRIDE Database and Its Related Tools.” *Nucleic Acids Research* 44 (November 2015): 447–56. doi:10.1093/nar/gkv1145.

Vlahou, A., Schorge, J. O., Gregory, B. W. & Coleman, R. L. (2003). “Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data.” *Journal of biomedicine & biotechnology*, *2003*(5), 308–314. doi:10.1155/S1110724303210032

Wagner, E.F. & Nebreda, A.R. (2009) “Signal integration by JNK and p38 MAPK pathways in cancer development.” *Nature Reviews. Cancer*. 9, 537–549.

Waugh, C. E., Shing, E. & Avery, B. (2015). “The neuroendocrine phenotype of gastric myofibroblasts and its loss with cancer progression.” *Emotion Review*. doi.org/10.1093/biostatistics/manuscript-acf-v5

Wedemeyer, W. J., Welker E., Narayan M. & Scheraga H. A. (2000). “Current Topics Disulphide Bonds and Protein Folding.” *Biochemistry* 39 (15).

Williams, G. Z., Widdowson, G. M. & Penton, J. (1978). “Individual character of variation in time-series studies of healthy people. II. Differences in values for clinical chemical analyses in serum among demographic groups, by age and sex.” *Clinical Chemistry*, 24(2), 313–320.

Wilson, J. M. & Jungner, Y. G. (1968). “Principles and practice of mass screening for disease.” *Boletín de La Oficina Sanitaria Panamericana. Pan American Sanitary Bureau*, 65(4), 281–393. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1875838”

Wu, T.-J., Shamsaddini, A., Pan, Y., Smith, K., Crichton, D. J., Simonyan, V. & Mazumder, R. (2014). “A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE).” Database : *The Journal of Biological Databases and Curation*, 2014, bau022. doi.org/10.1093/database/bau022

Yang, J., Zhu, Y., Guo, H., Wang, X., Gao, R., Zhang, L., Zhao, Y. & Zhang, X. 2013). “Identifying Serum Biomarkers for Ovarian Cancer by Screening With Surface-Enhanced Laser Desorption / Ionization Mass Spectrometry and the Artificial Neural Network.” *International Journal of Gynaecological Cancer 00*(00), 1–6. doi:10.1097/IGC.0b013e31827e1989

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N. & Liang, H. (2014). “Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types.” *Nature Communications*, 5, 3231. doi:10.1038/ncomms4231

Yu, J., Zheng, S., Tang, Y. & Li, L. (2005). “An integrated approach utilizing proteomics and bioinformatics to detect ovarian cancer.” *Journal of Zhejiang University. Science. B*, *6*(4), 227–31. doi:10.1631/jzus.2005.B0227

Zhang, L., Nuo Yang, N., Jin-wan Park, J., Katsaros, D., Fracchioli, S., Cao, G.,Brien-jenkins, A. O., Randall, T, C., Rubin, S. C. & Coukos. G. (2003). “Tumor-Derived Vascular Endothelial Growth Factor Up-Regulates Angiopoietin-2 in Host Endothelium and Destabilizes Host Vasculature , Supporting Angiogenesis in Ovarian Cancer Tumor-Derived Vascular Endothelial Growth Factor Up-Regulates Angiopoietin-2 in.” *Cancer Research* 63: 3403–12.

Zhang, Z., Bast, R. C., Yu, Y., Li, J., Sokoll, L. J., Rai, A. J, Rosenzweig, J. M., Cameron, B., Wang, Y. Y., Meng, X., Berchuck, A., Van Haaften-Day, C., Hacker, N. F., de Bruijn, H. W. A., van der Zee, A. G. J., Jacobs, I. J., Fung, E. T. & Chan, D. W. (2004). “Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer.” *Cancer research*, *64*(16), 5882–90. doi:10.1158/0008-5472.CAN-04-0746

Zhang, H., Kong, B., Qu, X., Jia, L., Deng, B. & Yang, Q. (2006). “Biomarker discovery for ovarian cancer using SELDI-TOF-MS.” *Gynaecologic oncology*, *102*(1), 61–6. doi:10.1016/j.ygyno.2005.11.029

Zhang, X., George, J., Deb, S., Degoutin, J. L., Takano, E. A, Fox, S. B., Harvey, K. F. (2011). “The Hippo pathway transcriptional co-activator, YAP, is an ovarian cancer oncogene.” *Oncogene*, 30(25), 2810–22. doi.org/10.1038/onc.2011.8

Zhang, W., Ota, T., Shridhar, V., Chien, J., Wu, B. & Kuang, R. (2013). “Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment.” PLoS *Computational Biology*, 9(3), e1002975. doi.org/10.1371/journal.pcbi.1002975

Zhu, W., Wang, X., Ma, Y., Rao, M., Glimm, J. & Kovach, J. S. (2003). “Detection of cancer-specific markers amid massive mass spectral data.” *Proceedings of the National Academy of Sciences of the United States of America*, *100*(25), 14666–71. doi:10.1073/pnas.2532248100

Ziqi, Y. A. N., Yuan, Z., Yichu, S., Qi, W. U., Shen, Z., Zhen, L., Lihua, Z. & Yukui, Z. (2014). “Label-Free Quantification of Differentially Expressed Proteins in Mouse Liver Cancer Cells with High and Low Metastasis Rates by a SWATH Acquisition Method” *Science China* 57 (5): 718–22. doi:10.1007/s11426-014-5093-z.

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Grczak, K., Klamecka, K., Figlerowicz, M. & Siatkowski, I. (2015). “The Impact of Normalization Methods on RNA-Seq Data Analysis.” *BioMed Research International* 2015. doi:10.1155/2015/621690.