

BL ✓

Ernos ✓ 12/14

41 0675835 4



ProQuest Number: 10183131

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10183131

Published by ProQuest LLC (2017). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

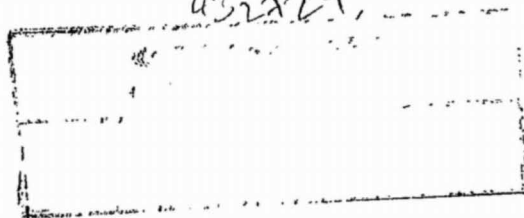
An Integrated Proteomic and Bioinformatic
Analysis for the Diagnosis and Prognosis of
Cancer

Erika Parkinson

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

November 2006

432729



Acknowledgements

Firstly I would like to thank my Director of Studies, Professor Bob Rees, and Dr Shahid Mian for giving me the opportunity to carry out this piece of research and for their guidance during this time. I would also like to give many thanks to Dr Balwir Matharoo-Ball whose supervision, guidance and support for the last 18 months of my PhD was invaluable, many thanks also go to Dr Graham Ball for his support and supervision over the past 3 ½ years, especially from the bioinformatics side of my work. I would also like to thank Lee Lancashire who, along with Graham, was willing to answer questions and help with server problems at the drop of a hat. Thanks must also go to Steve Reeder who helped me with the mammoth task of growing up over 100 melanoma cell lines and to Rob Davy for helping me to cryostat and H and E stain samples from 132 pieces of melanoma tissue, no mean feat. I would like to thank the EU for its part in funding this PhD as part of the 5th framework EU project, OISTER, and all of the partners of this project for their help and friendship, especially to Dr Selma Ugurel and Professor Dirk Schadendorf for providing all of the melanoma cell line, serum and tissue material and clinical information, and to Dr Per Guldberg for providing the genetic mutation information. Thanks go to Ian Ellis for providing the breast cancer tissue samples and tissue microarray data; and to Xenova, Cambridge, and the Health Protection Agency, London, for allowing me access to their SELDI machines during the course of my research. I would also like to thank everyone, past and present, who I have worked with in labs 008, 009 and 015 for their help and friendship during my time at NTU.

Finally my support system; my friends, especially Roger, Amanda, Nikki, Dilek, Morgan, Dan and Nadine; my family, especially my parents, my sisters, my Nan and Grandad and my mother and brother-in-law; and the most important people of all, my husband, Gary, and my dog, Max. I cannot begin to explain how much the love and support you have all given to me over the past 3 ½ years has meant to me, especially during the 'dark days' when I was ready to pack it all in, there is absolutely no way I could have gone through all of the ups and downs this PhD has put me through if you weren't all behind me, supporting me every step of the way, for this I am ever grateful to you all.

Contents

Acknowledgements.....	1
Contents.....	2
Figures.....	8
Tables.....	13
Abbreviations.....	15
Abstract.....	17
Chapter 1 Introduction.....	19
1.1 Cancer development, progression and treatment.....	19
1.1.1 Carcinogenesis.....	19
1.1.2 Oncogenes and Tumour Suppressor Genes.....	20
1.1.3 Cancer: A Multistep Disease.....	28
1.1.4 Tumour Progression and Survival.....	28
1.1.5 Cancer Therapy.....	31
1.2 Melanoma.....	33
1.2.1 Melanoma Tumorigenesis.....	33
1.2.2 Genetic involvement in Melanoma.....	34
1.2.3 Melanoma Staging.....	37
1.2.4 Melanoma Treatment.....	37
1.2.5 The Future for Melanoma Detection and Treatment.....	38
1.3 Breast Cancer.....	40
1.3.1 Breast Cancer Tumorigenesis.....	40
1.3.2 Genetic Involvement in Breast Cancer.....	41
1.3.2.1 Her-2(Her-2/neu or erbB-2).....	41
1.3.2.2 BRCA1.....	42
1.3.2.3 BRCA2.....	42
1.3.3 Therapy of Breast Cancer.....	44
1.3.4 The Future for Breast Cancer Detection and Treatment.....	45
1.4 Biomarker Identification.....	46
1.5 The Application of Proteomics in Cancer Research.....	48

1.5.1	2-D Polyacrylamide Gel Electrophoresis (2D PAGE)	48
1.5.2	Protein Microarrays	50
1.5.3	Mass Spectrometry	51
1.6	MALDI MS.....	53
1.7	SELDI MS	58
1.8	Bioinformatics approaches for data analysis	61
1.8.1	Artificial Neural Networks (ANNs)	62
1.8.2	Support Vector Machines (SVM)	65
1.8.3	Decision Trees	66
1.8.4	Principal Components Analysis (PCA)	66
1.8.5	Cluster Analysis.....	67
1.9	Cancer Therapy of the Future – Personalised Medicine.....	68
1.10	Aims and Objectives.....	70
1.11	Collaborations.....	71
Chapter 2 Methodology.....		72
2.1	Materials	72
2.1.1	Reagents and list of suppliers	72
2.1.2	Equipment.....	73
2.1.2.1	Cell lines and Media	73
2.1.2.2	Tumour Tissue	73
2.1.2.3	Serum	74
2.1.2.4	Buffers	74
2.1.2.5	Consumables.....	75
2.1.2.6	Hardware.....	75
2.1.2.7	Software.....	76
2.1.2.8	Company Addresses	76
2.2	Methods	78
2.2.1	Sample Preparation.....	78
2.2.1.1	Melanoma cell lines.....	78
2.2.1.2	Melanoma Tissue.....	78
2.2.1.3	Breast Cancer Tissue preparation	79
2.2.1.4	Serum preparation.....	79

2.2.2	Protein Microassay	79
2.2.3	1-Dimensional Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis (1-D SDS PAGE).....	80
2.2.4	SELDI MS	81
2.2.4.1	SELDI MS of melanoma cell lines and melanoma and breast cancer tissue 81	
2.2.4.2	SELDI MS of melanoma serum samples.....	82
2.2.4.3	SELDI MS of melanoma cell line samples analysed for the SELDI vs MALDI study.....	82
2.2.5	MALDI MS.....	83
2.2.5.1	ZipTip Method.....	83
2.2.5.2	Mass Spectrometric Analysis.....	83
2.2.6	Bioinformatic analysis	85
2.2.6.1	ANN analysis with parameterisation	86
2.2.6.2	Step Wise Approach	88
Chapter 3 Proteome profiling and bioinformatic analysis of melanoma cell lines carrying specific genetic mutations.....		90
3.1	Introduction.....	90
3.1.1	Combined genetic and proteomic analyses to identify the next generation of melanoma-specific markers	90
3.1.2	Melanoma-specific genetic mutations	91
3.1.2.1	BRAF	91
3.1.2.2	TP53.....	92
3.1.2.3	N-RAS	93
3.1.2.4	CDKN2A (P16 ^{INK4A})	94
3.1.2.5	CDK4.....	95
3.1.2.6	PTEN	96
3.1.2.7	CTNNB1 (β -catenin)	96
3.2	Methods	97
3.3	Results.....	100
3.3.1	Protein quantification and sample quality control.	100
3.3.2	SELDI MS	105

3.3.3	Bioinformatic analysis	113
3.3.3.1	Braf analysis	115
3.3.3.2	p16 ^{INK4A} analysis.....	117
3.3.3.3	Tp53 mutation analysis.....	119
3.3.3.4	Mycoplasma Infection – Validation of results.....	123
3.4	Discussion.....	125
Chapter 4 Proteome profiling and bioinformatics analysis of paired melanoma cell lines and tissues.....		132
4.1	Introduction.....	132
4.2	Methods	134
4.3	Results.....	136
4.3.1	Protein quantification and sample quality control.....	136
4.3.2	SELDI MS	137
4.3.3	Bioinformatic analysis	144
4.3.3.1	Prediction of the site of growth of the tumour for both cell lines and tissues (Skin vs Lymph Node).....	146
4.3.3.2	Prediction of patient survival for both cell lines and tissues	150
4.4	Discussion.....	156
Chapter 5 Breast cancer tissue proteome profiles associating with a basal phenotype and p53 expression.....		161
5.1	Introduction.....	161
5.1.1	Gene expression profiling of Breast Cancer.....	162
5.1.2	Proteomic profiling of Breast Cancer	162
5.2	Methods	164
5.3	Results.....	167
5.3.1	Protein quantification and sample quality control.....	167
5.3.2	SELDI MS	168
5.3.3	Bioinformatics Analysis	173
5.3.3.1	Basal Phenotype.....	173
5.3.3.2	P53 expression	176
5.4	Discussion.....	178

Chapter 6	SELDI MS and Bioinformatic approaches for the evaluation of serum biomarkers in stage I-IV melanoma	182
6.1	Introduction.....	182
6.1.1	Human blood proteome based studies	183
6.1.2	Serum-based studies using SELDI MS.....	184
6.1.3	Current problems associated with serum proteome analysis	185
6.1.4	Current serum biomarkers in melanoma.....	186
6.2	Methods	187
6.3	Results.....	189
6.3.1	SELDI MS analysis	189
6.3.2	Bioinformatics Analysis	194
6.3.2.1	Stage I vs Stage IV serum analysis.....	195
6.3.2.2	Stage III progressors vs non-progressors.....	197
6.4	Discussion.....	200
Chapter 7	Comparison of mass spectrometry platforms for protein biomarker identification in melanoma cell lines carrying specific genetic mutations	203
7.1	Introduction.....	203
7.2	Methods	206
7.3	Results.....	208
7.3.1	Mass spectrometry	208
7.3.2	Bioinformatic analysis	210
7.3.2.1	Braf analysis	210
7.3.2.2	P16 ^{INK4A} analysis	213
7.3.2.3	Tp53 analysis	217
7.4	Discussion.....	223
Chapter 8	Conclusions and Further Study	228
8.1	The potential use of cancer cell lines in proteomic analysis.....	228
8.2	Analysis of the Cancer Tissue Proteome	230
8.3	The Cancer Serum Proteome	231
8.4	Bioinformatic studies.....	232

8.5 Identification of cancer biomarkers using proteomic profiling: A question of reproducibility and validity.....	234
8.6 Cancer diagnosis and treatment – personalised medicine	237
References	240
Communications resulting from study.....	273
Appendix I	274
Appendix II.....	278
Appendix III.....	280
Appendix IV	284
Appendix V.....	288

Figures

Figure 1-1 Ras activated signalling pathways, Downward 2003.....	21
Figure 1-2 Retinoblastoma pathway.....	24
Figure 1-3 The p53 pathway.....	26
Figure 1-4 Figure showing melanoma progression from normal skin (A), into RGP melanoma (B), and then into VGP melanoma (C).	35
Figure 1-5 Figure showing the genes and proteins known to be involved in specific stages of melanoma progression.....	35
Figure 1-6 The mechanism of desorption and ionization of an analyte via MALDI.....	55
Figure 1-7 Diagram of a typical linear TOF mass spectrometer.	56
Figure 1-8 Schematic showing the basic layout of a neural network.	63
Figure 2-1 Representation of reproducible SELDI spectra obtained from the aliquots of normal serum.	82
Figure 2-2 Representation of reproducible MALDI spectra obtained from aliquots of normal serum.	84
Figure 3-1 A schematic representation showing the flow of sample preparation and analysis methods used in this study	99
Figure 3-2 Figure showing the process of protein estimation in all samples analysed.	102
Figure 3-3 1-D SDS PAGE of three cell line lysates, MEWO, ESTDAB 005 and Ma Mel 39a.....	103
Figure 3-4 Photographs of three cell lines in culture.....	104
Figure 3-5 Figure showing 4 SELDI MS profiles in the 5000-20000 Da mass range representing the reproducibility of the SELDI MS technique.	105
Figure 3-6 SELDI MS profiles in the 2000-10000 Da range of cell lines with or without <i>tp53</i> mutations.	107
Figure 3-7 SELDI MS profiles in the 2000-10000 Da range of cell lines with or without <i>Braf</i> mutations.	110
Figure 3-8 SELDI MS profiles in the 2000-10000 Da range of cell lines with or without <i>p16^{INK4A}</i> mutations.	112
Figure 3-9 Classification by ANN analysis of 85 cell lines with or without <i>braf</i> mutations.	116

Figure 3-10 Sensitivity analysis of the top 1000 ions used to classify cell lines with or without *braf* mutations..... 117

Figure 3-11 Classification by ANN analysis of 71 cell lines with or without *p16^{INK4a}* mutations..... 118

Figure 3-12 Sensitivity analysis of the top 1000 ions used to classify cell lines with or without *p16^{INK4a}* mutations. 119

Figure 3-13 Classification by ANN analysis of 85 cell lines with or without *tp53* mutations. 120

Figure 3-14 Sensitivity analysis of the top 1000 ions used to classify cell lines with or without *tp53* mutations. 121

Figure 3-15 Graph showing the model performance with each input addition for the stepwise ANN analysis of the *tp53* mutation data..... 122

Figure 3-16 Classification of 68 cell lines with or without *tp53* mutations after mycoplasma infected cell lines were removed from analysis..... 124

Figure 4-1 A schematic demonstrating how the cell lines and tissues were processed and analysed for this study. 135

Figure 4-2 1-D PAGE of cell line and tissue lysates representing sample quality of all samples analysed as part of the cell lines vs tissue study. 136

Figure 4-3 SELDI MS profiles in the 5000-15000 Da mass range showing the profiles obtained from a cell line are overlaid on the profile of the parent tissue. 138

Figure 4-4 SELDI spectra showing the differences and similarities between melanoma cell lines and tissues obtained from skin or lymph node..... 140

Figure 4-5 SELDI spectra showing the differences and similarities of cell lines and tissues taken from patients with a survival, after sample collection, of less than or more than 8 months..... 143

Figure 4-6 Classification by ANN analysis of 38 cell lines as either originating from skin or from lymph node..... 147

Figure 4-7 Sensitivity analysis of the top 1000 ions used to classify cell lines that originated from melanomas taken from either the skin of lymph nodes. 148

Figure 4-8 Classification by ANN analysis of 38 melanoma tissues as either originating from skin or lymph nodes. 149

Figure 4-9 Sensitivity analysis of the top 1000 ions used to classify melanoma tissues that originated from either the skin or the lymph nodes..... 150

Figure 4-10 Classification by ANN analysis of 43 melanoma cell lines originating from patients with a survival, after sample collection, of less than or more than 8 months. 151

Figure 4-11 Sensitivity analysis of the top 1000 ions used to classify cell lines originating from patients with a survival, after sample collection, of less than or more than 8 months..... 152

Figure 4-12 Classification by ANN analysis of 43 tissues originating from patients that survived less than or more than 8 months after tissue extraction. 153

Figure 4-13 Sensitivity analysis of the top 1000 ions used to classify tissue samples that originated from patients that survived less than or more than 8 months after tissue extraction. 154

Figure 5-1 A schematic showing the methods for sample processing and analysis used in this study..... 166

Figure 5-2 1-D PAGE of breast cancer tissue lysates representing sample quality of all samples analysed as part of the breast cancer study. 167

Figure 5-3 SELDI spectra obtained from breast cancer tissue lysates highlighting the differences and similarities between samples with different basal phenotypes, in the 2500 – 10000 Da range..... 169

Figure 5-4 SELDI spectra obtained from breast cancer tissue lysates highlighting differences and similarities between samples that do or do not express p53, in the 5000-15000 Da mass range. 171

Figure 5-5 Classification by ANN analysis of 140 breast cancer tissues displaying a basal-like or non-basal-like phenotype. 174

Figure 5-6 Graph showing the results of a sensitivity analysis of the top 1000 ions used to classify samples as those with or those without a basal-like phenotype. 175

Figure 5-7 Classification by ANN analysis of 135 breast cancer tissue samples according to whether or not they express p53. 176

Figure 5-8 Sensitivity analysis of the top 1000 ions used by ANNs to classify breast cancer samples according to whether or not they express p53. 177

Figure 6-1 Scheme showing the methods used in this study for the analysis of melanoma serum samples..... 188

Figure 6-2 SELDI spectra of serum samples highlighting the differences and similarities between samples taken from patients with either stage I or stage IV melanoma. 190

Figure 6-3 SELDI spectra of sera taken from stage III melanoma patients comparing the peak patterns of spectra generated from patients that progress to stage IV melanoma and patients that do not. 192

Figure 6-4 Classification by ANN analysis of 205 stage I and IV melanoma serum samples. 196

Figure 6-5 Sensitivity analysis of the top 1000 ions used to classify samples as stage I or stage IV melanoma serum samples..... 197

Figure 6-6 Classification by ANN analysis of 56 stage III patients according to whether the patients progressed to stage IV melanoma or not. 198

Figure 6-7 Sensitivity analysis of the top 1000 ions used to classify stage III sera taken from progressing or non-progressing melanoma patients. 199

Figure 7-1 Figure showing a flow diagram of the method followed in order to compare SELDI MS and MALDI MS technologies.207

Figure 7-2 SELDI and MALDI spectra generated from melanoma cell line FM 93/2.208

Figure 7-3 SELDI and MALDI spectra generated from analysis of melanoma cell line ESTDAB 027.....209

Figure 7-4 Classification by ANN analysis and sensitivity analysis of SELDI data generated from cell lines with or without a *braf* mutation.211

Figure 7-5 Classification by ANN analysis and sensitivity analysis of MALDI data generated from cell lines with or without a *braf* mutation.212

Figure 7-6 Classification by ANN analysis and sensitivity analysis of SELDI data generated from cell lines with or without a *p16^{INK4a}* mutation.215

Figure 7-7 Classification by ANN analysis and sensitivity analysis of MALDI data generated from cell lines with or without a *p16^{INK4a}* mutation.216

Figure 7-8 Classification by ANN analysis and sensitivity analysis of SELDI data generated from cell lines with or without a *tp53* mutation.218

Figure 7-9 Classification by ANN analysis and sensitivity analysis of MALDI data generated from cell lines with or without a *tp53* mutation.219

Figure 7-10 Stepwise model performance after addition of each input during the analysis of <i>tp53</i> mutation data generated by SELDI MS.....	221
Figure 7-11 Stepwise model performance after addition of each input during the analysis of <i>tp53</i> mutation data generated using MALDI MS.	223

Tables

Table 1-1 Oncogenes.	22
Table 1-2 Known tumour Suppressor Genes.....	27
Table 1-3 A summary of genes involved in malignant melanoma.	36
Table 1-4 A summary of the different treatments currently used for the treatment of melanoma, adapted from Lugovic <i>et al.</i> 2005; Kim <i>et al.</i> 2002a.	38
Table 1-5 Breast cancer associated genes.....	43
Table 2-1 Table showing the coefficients of variation (%) for mass and intensities of most prominent peaks in normal serum analysed by SELDI and MALDI MS, n=18.	85
Table 3-1 Values of the main peaks observed in the spectra shown in figure 3-6.	108
Table 3-2 Values of the main peaks observed in the spectra shown in figure 3-7.	111
Table 3-3 Values of the main peaks observed in the spectra shown in figure 3-8.	113
Table 3-4 Summary table of the data shown in figure 3-15	123
Table 4-1 Values of the main peaks observed in the spectra shown in figure 4-3.	139
Table 4-2 Values of main peaks observed in spectra shown in figure 4-4	141
Table 4-3 Values of the main peaks observed in the spectra shown in figure 4-5	144
Table 4-4 Table showing the paired cell lines and tissues used for the bioinformatics part of this study.....	145
Table 4-5 Survival data (in months) of the patients from which the cell lines and tissues analysed in this study were taken.	155
Table 5-1 Values of the main peaks observed in the spectra presented in figure 5-3.	170
Table 5-2 Values of the main peaks observed in the spectra presented in figure 5-4.	172
Table 6-1 Values of the main peaks observed in the spectra shown in figure 6-2.	191
Table 6-2 Values of the main peaks observed in the spectra of serum samples taken from stage III progressing or non-pogressing melanoma patients.	193
Table 7-1 Table comparing MALDI MS and SELDI MS technologies, adapted from Engwegen <i>et al.</i> 2006.	204
Table 7-2 Stepwise analysis of SELDI data generated from cell lines with or without <i>tp53</i> mutations.....	220
Table 7-3 Stepwise analysis of MALDI data generated from cell lines with or without <i>tp53</i> mutations.....	222

Table 7-4 ANN data generated by the analysis of SELDI and MALDI spectra of cell lines with or without <i>braf</i> mutations.	225
Table 7-5 ANN data generated by the analysis of SELDI and MALDI spectra of cell lines with or without <i>p16^{INK4a}</i> mutations.	226
Table 7-6 ANN data generated by the analysis of SELDI and MALDI spectra of cell lines with or without <i>tp53</i> mutations.	226
Table 8-1 Some of the identified biomarker proteins discovered using SELDI MS, adapted from Engwegen <i>et al.</i> 2006.	235

Abbreviations

1-D PAGE	1-Dimensional Polyacrylamide Gel Electrophoresis
ACN	Acetonitrile
ANNs	Artificial Neural Networks
BN	Benign naevus
BP	Back Propagation
BPA	Benign Prostatic Hyperplasia
BSA	Bovine serum albumin
CAMs	Cell adhesion molecules
CDK	Cyclin dependent kinases
CGD	Conjugate Gradient Descent
CV	Coefficient of Variation
Da	Daltons
DCIS	Ductal Carcinoma <i>in situ</i>
DGGE	Denaturing Gradient Gel Electrophoresis
DTT	Dithiothreitol
ECM	Extracellular matrix
ER	Estrogen receptor
ESI	Electrospray Ionisation
FGF	Fibroblast growth factor
GAPs	GTPase activating proteins
HDACs	Histone deacetylators
HLA	Human Leukocyte Antigen
HPLC	High performance liquid chromatography
IDCA	Invasive ductal carcinoma
IR	Infra-red
kDa	Kilodaltons
LCM	Laser Capture Microdissection
LDH	Lactate dehydrogenase
LOH	Loss of heterozygosity

<i>m/z</i>	Mass-to-charge ratio
MALDI MS	Matrix-Assisted Laser Desorption Ionisation Mass Spectrometry
MS	Mass Spectrometry
OGP	Octyl- β -D-glucopyranoside
PCa	Prostatic cancer
PCA	Principal components analysis
PCR	Polymerase Chain Reaction
PMF	Peptide mass fingerprinting
PSD	Post Source Decay
RGP	Radial growth phase
SELDI MS	Surface-Enhanced Laser Desorption Ionisation Mass Spectrometry
SPA	Sinapinic Acid
SSCP	Single-stranded conformation polymorphism
SVM	Support vector machines
TFA	Tri-Fluoro Acetic Acid
TMA	Tissue microarray
TOF	Time-of-flight
UV	ultraviolet
VEGF	Vascular endothelial growth factor
VGP	Vertical growth phase

Abstract

The advent of proteomics and high-throughput technologies has allowed scientists to derive protein expression patterns of potential use in predictive medicine. The application of bioinformatics to analyse complex data makes it possible to identify important protein biomarkers. These biomarkers may have predictive capability to determine, for example, the presence and progression of disease and how an individual patient might respond to therapy.

Mass spectrometry (MS) has increasingly become the method of choice for the analysis of complex samples and new MS systems have been developed that can rapidly profile and generate proteomic 'fingerprints' from tissue and body fluids. In particular, MALDI mass spectrometry coupled with Ciphergen® chip technology (SELDI MS) has been widely used to identify discriminatory patterns to distinguish patients at different clinical stages of disease, for example, in ovarian, prostate, colon and breast cancer. All of these studies incorporate the use of computer algorithms to mine the proteomic data obtained from the mass spectra, allowing large cohorts of samples to be included into the analysis.

The aim of this study was to introduce the use of MS and bioinformatics to analyse the cancer proteome, in particular melanoma and breast cancer and to investigate the information obtained from profiling cell lines, tissue and serum samples, as well as evaluating the type of analytical methods currently available.

The methods used in this study for sample preparation and analysis demonstrate that good quality proteomic data from cell lines, tissue and serum can be obtained and that it is possible to generate discriminatory protein profiles that correlate with clinical outcomes when analysed using Artificial Neural Networks (ANNs). Through the analysis of the proteome of melanoma cell lines, it is possible to classify samples according to the presence of specific genetic mutations, the site of the tumour sample from which the cell line was derived, as well as the overall survival of a patient. Comparison of melanoma cell line proteomes and their tumour tissue of origin revealed that both sample types were able to provide discriminating patterns that correlated to clinical outcomes. This finding has significance for future proteomic-based biomarker discovery research where it is possible to use cell lines in place of "precious" tumour tissue for the identification of clinically relevant biomarkers. The presence of a basal phenotype, which signifies the aggressive

nature of breast cancer, can be identified from the proteomic profiling of patients' breast cancer tissue. The analysis of melanoma patient serum was investigated and patterns that predicted the stage of disease, as well as disease progression, were identified, using SELDI MS and ANNs. These results demonstrate that it is possible to obtain clinically valid information from the proteome of samples derived from melanoma and breast cancer patients through the use of SELDI MS and ANN analysis.

Although SELDI MS has proven useful in generating protein profiles that can be used for identifying patients with different clinical outcomes, this technology has limitations. One aspect of the study was to determine if similar, or more accurate, discriminatory analysis could be achieved using higher resolution and higher sensitivity MALDI instrumentation. A set of melanoma cell line samples were subjected to SELDI MS and MALDI MS analysis and the data from both methods were analysed in the same way by ANNs. Slightly different sample preparation methods were used prior to MS analysis, thus the spectra obtained by SELDI MS and MALDI MS was dissimilar; the data revealed that MALDI MS did not improve upon the accuracy of classifying samples.

The work presented demonstrates a proof-of-principle of the different types of information that can be obtained from samples derived from melanoma and breast cancer patients. It has also been revealed that the analysis of MS spectra by ANNs can be used for predicting blind datasets which is not necessarily dependent on the MS method used; however, this is likely to have significant implications for biomarker identification as the different methods used will reveal different disease-associated proteins.

Chapter 1 Introduction

1.1 *Cancer development, progression and treatment*

Cancer is a disease that involves dynamic changes in the genome and is defined as any malignant growth or tumour caused by abnormal and uncontrolled cell division. (Hanahan *et al.* 2000). There are over 100 different types of cancer and they can occur in almost any tissue in the body. Cancer occurs as the result of the disrupted regulation of biomolecular processes that determine whether a cell remains quiescent, proliferates, differentiates or dies via apoptosis (Li *et al.* 2000). The following sub-sections detail some of the events necessary for cell transformation.

1.1.1 Carcinogenesis

Random mutations in the genes which control apoptosis or proliferation are responsible for promoting uncontrolled cell growth leading to cancer. A large proportion of the mutations are not inherited but are in fact spontaneous and occur in response to chemical damage of the DNA, resulting in the altered functions of crucial genes. In order for chemical damage to cause an inheritable change in the DNA, it is necessary for DNA replication and cell division to occur. The frequency of mutations is at a rate of 1 cell in 1 million but with a large number of proliferating stem cells there is a high probability of unrepaired DNA damage causing a single mutation, leading to the formation of an initiated cell. Mutations in at least 5 genes are required for a cell to become fully malignant, however the probability of an “initiated cell” obtaining the additional 4 mutations is low, due to the number of repair mechanisms or death pathways that become activated as a result of cellular irregularity (Loeb *et al.* 2003). The mutations have 2 consequences, they either allow the inappropriate expression or activation of genes, oncogenes, or they result in the functional inactivation of tumour suppressor genes.

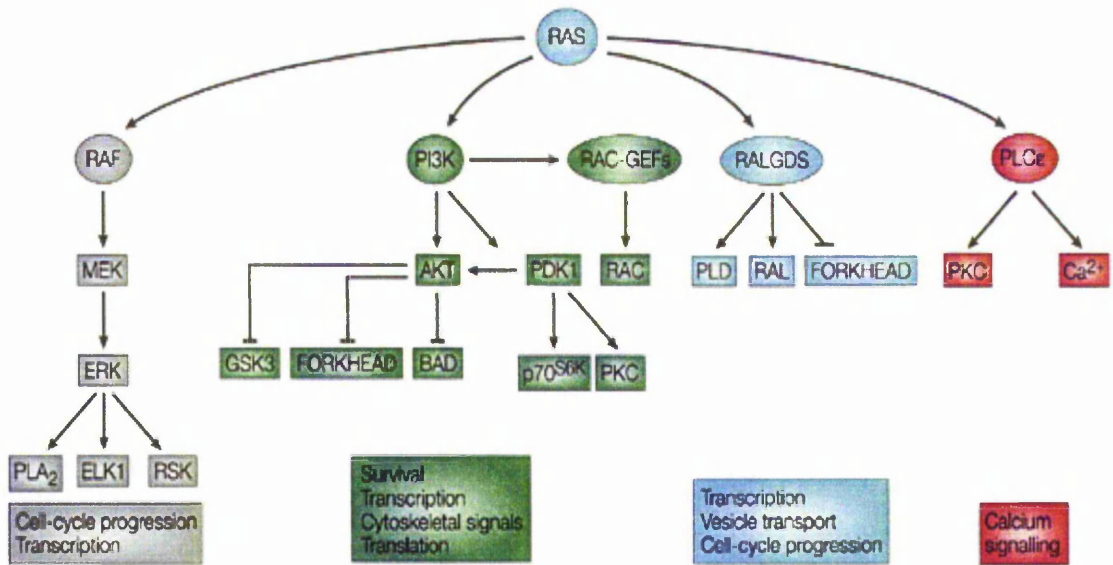
Six essential pathways must be altered in the normal cell physiology to dictate the formation of a malignant tumour and its capability of spreading to secondary sites; the development of independent growth stimulatory signals; resistance to growth inhibitory signals; the development of resistance to apoptosis; the acquisition of infinite proliferation;

the capacity to form new blood vessels and capillaries; and the ability to undergo tissue invasion and metastasis (Bertram 2001; Hanahan & Weinberg 2000).

1.1.2 Oncogenes and Tumour Suppressor Genes

Under normal homeostasis oncogenes and tumour suppressor genes are involved in the control of the cell cycle. Proto-oncogenes encourage cell growth and when mutated can become carcinogenic oncogenes that drive excessive multiplication. The mechanisms of action to transform proto-oncogenes into carcinogenic oncogenes are via mutations, gene amplification or chromosome rearrangement and the activation of proto-oncogenes leads to an unrestrained progression of the cell cycle and cell growth (Rieger 2004). The proteins encoded by proto-oncogenes fall into one of four categories, growth factor receptors, growth factors, transcription factors and signal transducers. Since oncogenes are dominant genes, only 1 allele needs to carry the mutation for it to be expressed and cause a gain-of-function (activating) mutation in cancer (Fearon *et al.* 1999). Many oncogenes have been discovered that are involved in tumour initiation, progression, angiogenesis and metastasis and more than 50 oncogenes have been identified in human cancers including *K-ras*, *N-ras*, *H-ras*, *cdk4*, *bcr-abl* and *neu (erb-b2)*, table 1-1 summarises some of the important known oncogenes (Haber *et al.* 1998; Michor *et al.* 2004).

Ras proteins have essential roles in controlling the activity of several signaling pathways that regulate normal cellular proliferation. Ras proteins activated by point mutations in the ras gene are expressed in 20% of all tumours (Bos 1989). When activated in human tumours, ras proteins contribute to the deregulation of cell growth, programmed cell death, invasiveness and the induction of angiogenesis (Shields *et al.* 2000). Ras proteins can be bound to GTP in their active state or GDP in their inactive state. In normal cells, the activity of ras proteins is controlled by the ratio of bound GTP to GDP (Campbell *et al.* 1998). GTP-bound ras can bind and activate effector enzymes and thus control cell proliferation and survival. Mutations in the ras gene compromise the GTPase activity of ras proteins, preventing the GTPase activating proteins (GAPs) from promoting the hydrolysis of GTP on ras. This causes ras to accumulate in the GTP bound active form allowing it to interact with several effector proteins resulting in the stimulation of their catalytic activity and triggering downstream signaling pathways. The best characterized effector pathways activated by ras are shown in the figure 1-1 (Downward 2003).



Nature Reviews | Cancer

Figure 1-1 Ras activated signalling pathways, Downward 2003.

Ras controls several signaling pathways that regulate normal cellular proliferation. Ras binds GTP allowing it to bind effector enzymes and triggering downstream pathways involved in cell cycle progression, transcription, and cell survival.

Gene	Mechanism of Activation of gene	Function of gene products	Human tumours associated with oncogene mutation/activation
<i>Myc</i>	Amplification, chromosomal translocation	Nuclear transcription factor	Lymphomas, carcinomas
<i>Bcl 2</i>	Chromosomal translocation	Cytoplasmic perhaps mitochondrial	Follicular and undifferentiated lymphomas
<i>Erb B1</i>	Amplification	Growth factor receptor	Mammary carcinoma, glioblastoma
<i>Erb B2</i>	Amplification	Cell surface growth factor receptor	Mammary, ovarian and stomach cancers
<i>Ras</i>	Point mutation	GDP/GTP binding	Bladder cancer (<i>H-ras</i>), lung and colon cancer (<i>K-ras</i>), lymphomas and carcinomas (<i>N-ras</i>)
<i>Raf</i>	Rearrangement	Cytoplasmic serine/threonine kinase	Stomach cancer
<i>Hst</i>	Rearrangement	Growth factor	Stomach cancer
<i>Ret</i>	Rearrangement	Growth factor receptor	Thyroid cancer
<i>Cdk4</i>	Amplification, point mutation	Cyclin dependent kinase	Sarcoma, familial melanoma

Table 1-1 Oncogenes.

Examples of oncogenes their mechanism of activation, and function of the gene products for main cancer types.

The mechanisms of action of tumour suppressor genes to initiate cancer cells result from the loss of large portions of their genetic sequence, methylation of promoter regions or complete loss of one allele plus a mutation in the other (Rieger 2004). This causes a loss of functional suppressor proteins which deprive the cell of the crucial brakes that prevent inappropriate growth. Tumour suppressor genes are recessive as in most cases the normal tumour suppressor allele can function in the presence of the damaged allele. Mutations in both alleles are required for the loss of function (inactivation) in that particular gene. Loss of only one allele of a tumour suppressor gene allows that gene to become silent and so permits the germ-line inheritance of the damaged gene (Weinberg 1991). *Rb1*, *tp53*, *p16^{INK4A}*, *Pten* and *APC* are all examples of tumour suppressor genes associated with cancer, a summary of some of these genes is shown in table 1-2 (Haber & Fearon 1998). The retinoblastoma gene (*Rb1*) and the *tp53* gene are the best characterized tumour suppressor genes.

Retinoblastoma is a rare malignant tumour of the developing retina that occurs in children. The *Rb* gene has been mapped to chromosome 13q14 by linkage studies and detection analysis (Godbout *et al.* 1983). *Rb1* encodes the retinoblastoma protein, Rb, which functions as a transcription factor and is regulated by phosphorylation through the cell cycle, playing a critical role in the control of proliferation (DiCiommo *et al.* 2000). Many of the amino acid substitutions and in-frame deletions that occur in *Rb* mutations affect the A/B pocket in Rb, which is necessary for biological functions including regulating growth and differentiation and biological activities including transcriptional regulation (Kouzarides 1995). E2F proteins activate genes that are required for DNA synthesis by recruiting general transcription factors such as TBP and TFIID. When Rb binds to E2F proteins it prevents E2F interacting with factors like TBP and so represses transcription (Pearson *et al.* 1997). Hypophosphorylated Rb binds target proteins such as E2F and histone deacetylators (HDACs) and so can arrest cells in the G1 phase of the cell cycle. CDK mediated phosphorylation of Rb occurs to reverse this block and so drive cells through G1 into S phase (Mittnacht 1998). Hyperphosphorylated Rb causes the release of free E2Fs and the dissociation of transcriptional repression complexes. The freed E2F proteins then activate the transcription genes necessary for S phase entry and progression of the cell cycle (Lipinski *et al.* 1999). A summary of the Rb pathway is shown in figure 1-2.

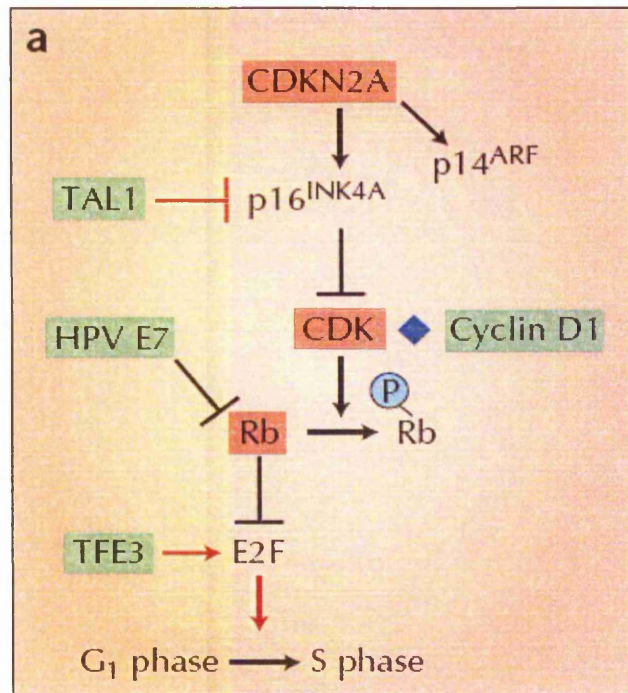


Figure 1-2 Retinoblastoma pathway.

Rb functions as a transcription factor, playing a critical role in cell proliferation. It is regulated through its phosphorylation by cyclin dependent kinases. Rb binds E2F, stopping E2F from activating transcription, thereby arresting cells in the G₁ phase of the cell cycle. When Rb is phosphorylated, E2F is freed and the cell cycle progresses from the G₁ to S phase of the cell cycle (Vogelstein *et al.* 2004)

The *tp53* gene has been shown to be lost or contain mutations that inactivate the p53 protein in approximately 50% of human cancers (Hollstein *et al.* 1996). Most of the mutations are missense mutations that cause single residue changes in the binding core domain of the protein (Sigal *et al.* 2000). Mutations in codons 175, 249, 273 and 282 accounts for 40% of total missense mutations reported in *tp53* positive human cancers (May *et al.* 1999). Mutations in the *tp53* gene can result in loss of function of the p53 protein or acquisition of new functions that contribute to the transformation and tumourigenic potential (Dittmer *et al.* 1993). P53 can regulate a wide range of cellular processes including cell cycle control, DNA repair, genome stability, programmed cell death, differentiation, senescence, angiogenesis and is a major component of the DNA damage response pathway (Rotter *et al.* 1994). A summary of the pathways p53 affects is shown on figure 1-3. The expression of p53 is up-regulated in response to hypoxia, nucleotide deprivation and DNA damage caused by ionizing radiation, UV light and

chemicals (Lakin *et al.* 1999). It is also subjected to extensive post-translational modifications including phosphorylation and acetylation which modulate its stability and activities. Some of the normal cellular functions can be modulated and sometimes inhibited by the interaction with the cellular protein MDM-2, a product of the *mdm-2* oncogene which can be amplified in some types of tumour. This alternative mechanism of p53 activation is common in many tumours including soft tissue sarcomas, bladder, cervical and breast carcinomas and leukemia (May & May 1999). Li-Fraumeni syndrome is an inherited disease where *tp53* is mutated in one allele in the germ line. This syndrome causes the onset of cancers at an early age, most commonly these are sarcomas but breast, colon and several other types of cancers can occur (Frebourg *et al.* 1992). One major consideration when developing therapies that try to restore wild-type p53 to tumour cells containing mutated *tp53* is that p53 acts as a tetramer and the presence of mutated *tp53* acts in a dominant manner and inhibits the actions of the wild-type protein (Bertram 2001). Approximately 50% of all cancers have *tp53* somatic mutations in both alleles, teratocarcinomas, however, are one of the few types of cancer that do not usually have *tp53* mutations. In this particular cancer, the p53 protein is not functional and so the p53 protein can be activated to induce apoptosis, therefore teratocarcinomas respond well to chemotherapy and can be cured in most cases (Lutzker *et al.* 1996).

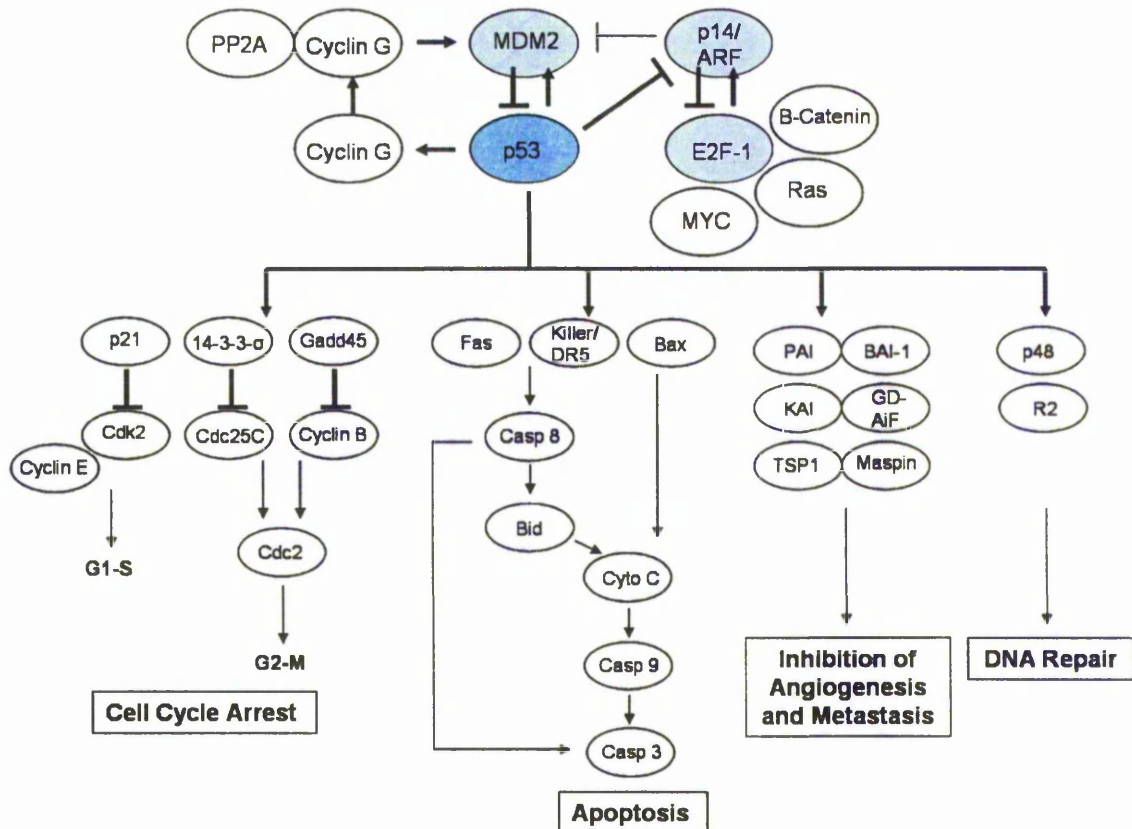


Figure 1-3 The p53 pathway.

P53 regulates a wide range of cellular processes including cell cycle control, DNA repair, genome stability, programmed cell death, differentiation, senescence, angiogenesis and is a major component of the DNA damage response pathway. It is regulated by *p14/arf* and has been shown to bind MDM-2 which blocks the ability of p53 to act as a transcription factor. MDM-2 has also been shown to be transcriptionally regulated by p53 forming an autoregulatory loop where increased p53 activity increases MDM-2 levels which in turn decreases p53 activity resulting in declining MDM-2 levels (Bertram 2001; Vogelstein & Kinzler 2004).

Gene	Chromosomal location	Function	Human tumours and cancer syndromes associated with mutation
<i>Rb1</i>	13q14	Transcriptional regulator of the cell cycle	Retinoblastoma, osteosarcoma
<i>WT1</i>	11p13	Transcriptional regulator	Nephroblastoma, Wilms tumour
<i>tp53</i>	17q11	Transcriptional regulator, growth arrest, apoptosis	Sarcomas, breast, brain tumours, Li-Fraumeni syndrome
<i>NF1</i>	17q11	Ras-GAP activity	Neurofibromas, sarcomas, gliomas
<i>APC</i>	5q21	Binds/regulates β -catenin activity	Colon cancer, familial adenomatous polyposis
<i>ink4a</i>	9p21	<i>p16^{INK4A}</i> cdki for cyclin D/cdk 4/6	Melanoma, pancreatic cancer, familial melanoma
<i>BRCA1</i>	17q21	Transcriptional regulator, DNA repair	Breast and ovarian tumours, familial breast cancer
<i>BRCA2</i>	13q12	Transcriptional regulator, DNA repair	Breast and ovarian tumours, familial breast cancer
<i>Pten</i>	10q23	Dual specificity phosphatase	Glioblastoma, prostate and breast cancer

Table 1-2 Known tumour Suppressor Genes

Outlining the main tumour suppressor genes identified to date, their chromosomal location, function and the cancers or cancer syndromes reported to be associated with gene inactivation.

1.1.3 Cancer: A Multistep Disease

Each cell in the body has the potential to change its genotype and phenotype on the pathway to malignancy. Each tissue can be classified at various points on the progression to malignancy including fully normal, hyperplastic, metaplastic, neoplastic and metastatic. These different stages suggest that a single genetic change only pushes a cell part of the way towards a malignant state, and that multiple changes need to occur for the process to complete (Weinberg 1994). Fearon *et al.* (1990) have illustrated this multistep carcinogenesis in the colon cancer model, as follows: the mutation of the tumour suppressor gene *APC* allows the growth of polyps in the colon, subsequently a somatic mutation in the *ras* oncogenes may occur to create a more advanced, benign polyp which may then eventually gain mutations in its *DCC* and *tp53* tumour suppressor genes causing the uncontrolled growth of the colon carcinoma cell (Fearon *et al.* 1990).

1.1.4 Tumour Progression and Survival

As previously mentioned, cancer cells must obtain the ability to generate their own mitogenic signals, to resist exogenous growth inhibitory signals, to evade apoptosis, to infinitely proliferate, to acquire vasculature and to invade and metastasize in order to survive and progress (Eccles 2005; Rieger 2004).

For a normal cell to move from a quiescent into a proliferative state, growth signals are transmitted into the cell via transmembrane receptors. Many known oncogenes mimic normal growth signals and so disrupt an important mechanism which usually ensures the proper behaviour of cells within tissue (Hanahan & Weinberg 2000). Growth factor receptors are also dysregulated during carcinogenesis and their over expression allows the cancer cell to become hyper-responsive to levels of growth factors that would not normally trigger proliferation (Fedi *et al.* 1997). Extracellular matrix (ECM) receptors (integrins) link cells to the ECM and transduce signals into the cytoplasm (from the ECM) that influences cellular quiescence in normal tissue, cell motility, resistance to apoptosis and entrance into the cell cycle. Cancer cells can alter integrin expression, favouring ones that transmit pro-growth signals.

Soluble growth factor inhibitors and immobilized inhibitors are molecules that are embedded in the ECM and on the surfaces of nearby cells act as anti-proliferative signal

that function to maintain normal cellular quiescence and tissue homeostasis. These signals can block proliferation in two ways; either by forcing cells into quiescence (G_0), from which they may start proliferating at a later stage when signaled to do so, or by being induced into a post-mitotic state (Hanahan & Weinberg 2000). Most anti-proliferative signals are regulated by the retinoblastoma protein (pRb) as well as related proteins such as p107 and p130. When hypophosphorylated, pRb sequesters and alters the function of E2F transcription factors. These transcription factors control the expression of genes that are essential for the progression of cells from G_1 into the S phase and so when altered, proliferation is blocked (Weinberg 1995). The disruption of the pRb pathway activates the E2F transcription factors causing expression of the genes that allow cell proliferation to occur. This also means that these cells are resistant to the anti-growth factors that would normally act on the pRb pathway blocking the cell cycle progression.

The ability of a cell to enter into the apoptotic pathway is a feature of nearly all types of cell found in the body. When this pathway is triggered, the cellular membranes become disrupted, the cytoplasmic and nuclear skeletons are broken down, the cytosol is extruded, the chromosomes become degraded and the nucleus fragments. Finally, the degenerating cell is engulfed by nearby cells in the tissue and disappears (Wyllie *et al.* 1980). The mitochondria receive the majority of the proapoptotic signals and respond by releasing cytochrome C. The Bcl-2 family have either pro-apoptotic (Bax, Bak, Bid, Bim) or anti-apoptotic (Bcl-2, Bcl-XL, Bcl-W) functions and part of their mechanism of action is to direct the mitochondrial signals via the release of Cytochrome C. P53 can induce apoptosis by upregulating the expression of Bax in response to DNA damage (Green *et al.* 1998). The most common mechanism of resistance to apoptosis in cancer cells is via the loss of the *tp53* tumour suppressor gene and is seen in over 50% of human cancers (Harris 1996).

Many cells have an intrinsic program that limits the number of times they can replicate and it operates independently of the signaling pathways mentioned previously; this pathway must be disrupted in order for a clone of cells to expand to such a size that it becomes a life threatening tumour. This program involves telomeres, located at the ends of chromosomes, which are made up of several thousand repeats of a 6 base pair sequence. At each replication, 50-100 base pairs are lost from the telomere and eventually this shortening of the telomeres causes the ends of the chromosomes to become unprotected. The chromosomes fuse end-to-end and cause the death of the affected cell (Shay *et al.* 2005).

Maintenance of the telomere occurs in virtually all types of malignant cell and 85-90% of these cells achieve this by upregulating telomerase, which adds 6 base pair repeats onto the ends of the telomere (Ahmed *et al.* 2003; Shay *et al.* 1997). By maintaining the telomeres at a length above the critical threshold, the cells are permitted to replicate limitlessly.

All cells within a tissue must reside within 100 μ m of a capillary or blood vessel in order to receive the oxygen and nutrients required for cell function and survival. The induction of angiogenesis is an early to mid-stage event in many cancers and is necessary if a solid tumour is to grow beyond 1-2mm. Angiogenesis is activated by changing the balance of angiogenic inducers and inhibitors seen in normal tissues. This can be achieved by either upregulating the expression of VEGF and/or FGF (angiogenic activators) or by downregulating the expression of thrombospondin-1 or β -interferon (angiogenic inhibitors) (Hanahan *et al.* 1996; Volpert *et al.* 1997).

During the development of most cancers, cells from the primary tumour develop the ability to invade into adjacent tissues and travel to and colonise distant sites. A tumour cell detaches from the primary tumour and actively infiltrates the surrounding stromal tissue. It then enters the circulatory system, travelling to specific sites to establish a secondary tumour. The altered binding of cell adhesion molecules (CAMs) and the activation of extracellular proteases, contributes to the invasive and metastatic ability of cancer cells. CAMs are divided into 2 main groups the immunoglobulin superfamily and some members of the cadherin family (Harlozinska 2005). E-cadherin has been widely studied as being an important factor in tumour cell invasion and metastasis and decreased expression or loss of function occurs in a majority of epithelial cancers. E-cadherin transmits anti-growth signals via β -catenin which in turn activates intracellular signaling pathways and transcription factors. The loss of E-cadherin function, therefore, occurs via inactivation of the *E-cadherin* and *β -catenin* genes and transcriptional repression (Christofori *et al.* 1999). As well as altering the physical relationship of cells with their microenvironment, increased extracellular protease activity can occur. This can be achieved by the upregulation of the proteases, the downregulation of protease inhibitors and activation of inactive forms of proteases (Werb 1997). Cancer cells can invade into the nearby stroma, across blood vessel walls, and through normal epithelial cell layers via the docking of proteases onto the cell surface.

Recently it has been suggested that there is a small subset of cells within leukaemias and solid tumours that have the ability to proliferate extensively and form new tumours and also have the ability to reproduce the variety of cell types that comprise the tumour, these cells have been identified as 'cancer stem cells' (Al-Hajj *et al.* 2004). The existence of these cancer stem cells was first suggested in 1963 and provided an explanation of the heterogenous nature of tumours but the best evidence of the existence of these cells has come from studies of haematological malignancies (Bhatia *et al.* 1997; Bonnet *et al.* 1997; Bruce *et al.* 1963; Holyoake *et al.* 1999). A cancer stem cell is defined as a cell that has the ability to self-renew giving rise to another malignant stem cell as well as being able to differentiate and produce the diverse number of non-tumourigenic cancer cell types. The origin of cancer stem cells is unclear: they may have derived from a transformed normal stem cell, where only its proliferative pathways have been altered, or from a progenitor cell that has acquired oncogenic mutations allowing it to self-renew and so acting as a stem cell. Evidence suggests that it is most likely that the former hypothesis is true and has been supported by functional assays (Al-Hajj *et al.* 2004; Hope *et al.* 2004; Wang *et al.* 2005; Zhang *et al.* 2006a).

1.1.5 Cancer Therapy

There are a number of therapies currently available for the treatment of cancer, including surgery, radiotherapy, chemotherapy, hormone therapy and immunotherapy. All of these treatments are used in different ways to treat different types of cancers, but all have their limitations. Surgery can be effective as long as the tumour is localised and has not invaded surrounding tissues. Chemotherapy and radiotherapy both target rapidly dividing cells, including but not exclusively tumour cells and therefore significant toxicity usually occurs. Hormone therapy is only effective on certain types of cancer and immunotherapy is still not developed enough for widespread use, but it has potential useage as a treatment that only has low toxicity as the immune system is highly specific and could be directed to target only tumour cells.

The choice of treatment usually depends on the type and the stage of the cancer but to date no therapy has been shown to substantially prolong the survival in patients with advanced

disease (Wagner *et al.* 2005). It is desirable, therefore, to diagnose the cancer early in the disease process when treatment is more likely to be effective.

The presence of a cancer stem cell population within a tumour has implications for the diagnosis and treatment of cancer. If tumour formation is driven by these cells, then the goal of therapy should be to identify and target this population. The failure to eliminate these cells could be responsible for the re-growth seen in many tumours once the chemotherapy treatment has been stopped (Clarke *et al.* 2006). The identification and characterisation of this group of cells as well as the identification of the differences and similarities between these and normal stem cells could provide novel targets for future cancer therapies (Perez-Caro *et al.* 2006).

1.2 Melanoma

Melanoma is an aggressive malignant cancer of melanocytes that occurs most frequently in the skin but also less commonly in the eye and mucosal surfaces as primary sites. Despite the increase in public awareness of the dangers of excessive sun exposure, one of the risk factors associated with melanoma, the incidences of this cancer are still rising in the UK (Kalkman *et al.* 2004). Malignant melanoma is the least common form of skin cancer but it is the most aggressive, claiming 1,700 lives each year in the UK compared with only 514 deaths for non-melanoma skin cancer. The incidence rates of melanoma increase with age, the highest rate being in the over 75s. It is the second most commonly diagnosed cancer in the 15-34 age group, however (Melanoma Statistics and Prognosis 2006). The cure rates in patients with early stage disease is up to 90%, however the survival rate for patients with locoregional disease is 24 months and for patients with metastatic disease the median survival is 6 months, therefore prognosis remains poor for patients with advanced disease (Balch *et al.* 1997).

1.2.1 Melanoma Tumourigenesis

Six steps have been identified during melanoma tumourigenesis; 1) common acquired melanocytic naevus (BN); 2) melanocytic naevus with lentiginous melanocytic hyperplasia (abberant differentiation); 3) melanocytic naevus with aberrant differentiation and melanocytic nuclear atypia – melanocytic dysplasia; 4) radial growth phase of primary melanoma (RGP); 5) vertical growth phase of primary melanoma (VGP); and 6) metastatic melanoma (Clark *et al.* 1984). Most melanomas progress through a slow RGP that is restricted to just the epidermis and into a more rapid VGP, see figure 1-4. Alongside this progression, the treatment options, cure rates and survival rates decrease dramatically. In the RGP the melanoma cells lack the ability to invade and metastasise and therefore can be cured, in general, by surgical excision. The VGP, however, is more aggressive, invades the dermis and possesses the ability to metastasise (Chudnovsky *et al.* 2005). These metastases can spread locally through the lymphatic system or to distant sites via the bloodstream to any organ but lung and liver metastases are the most common (Souhami *et al.* 1998). Each of the six steps in melanoma tumourigenesis involves genetic alterations including allelic loss, microsatellite instability and alterations of tumour suppressor genes, mismatch

repair proteins, oncogenes and some growth factors (Hussein 2004). The earliest genetic changes appear to involve mutation of the melanocytes of the melanocytic dysplastic naevi (Hussein *et al.* 2002). The transition from a normal melanocyte to a BN or dysplastic naevi involves the loss of genes at 1p or 9p and/or 10q chromosomal regions. Progression of tumourigenesis from the dysplastic naevi into the RGP involves further loss of genes at the 9p, 10q and 6q chromosomal regions. Finally the progression of the VGP into metastatic disease involves the loss of genes at the 1p, 11q and 17q as well as other undetermined chromosomal regions (Hussein *et al.* 2003; Park *et al.* 1998).

1.2.2 Genetic involvement in Melanoma

Several genes have been shown to be involved in malignant melanoma including germline mutations in *cdkn2A*, *Arf*, *cdk4* and somatic mutations in *Pten* and *Braf*. These melanoma associated genes have been summarised in table 1-3, but this is by no means an exhaustive list as the genetic events leading to the transformation of a melanocyte is still being investigated (Castellano *et al.* 1999). A summary of the six steps that take place during tumourigenesis along with some known mutations at each stage is illustrated in figure 1-5. The expression of cyclin D1, cyclin A, CDK1, CDK2 and STAT1 have been identified in the radial growth phase; Ki 67, survivin and PKC-B have been shown to be expressed in the vertical growth phase; and expression of cyclin D1, cyclin D3, loss of *p16^{INK4A}*, loss of *Bcl-2* and loss of MUM1 have been shown in metastatic melanoma. In general, the increase of expression of cyclins and CDKs in conjunction with a loss of CDK inhibitors facilitates the progression to advanced stages of melanoma (Alonso *et al.* 2004). The main genes involved in melanoma development and progression are discussed more fully in chapter 3, section 3.1.2.

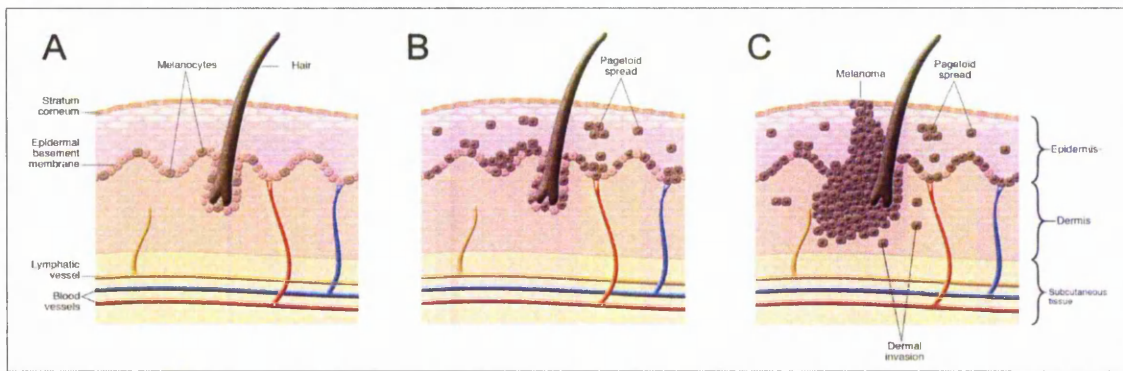


Figure 1-4 Figure showing melanoma progression from normal skin (A), into RGP melanoma (B), and then into VGP melanoma (C).

Typically in melanoma, the cancer cells progress through a slow radial growth phase (RGP) that is restricted to just the epidermis and into a more rapid vertical growth phase (VGP). In the RGP the melanoma cells lack the ability to invade and metastasise, however the VGP is more aggressive. In the VGP the cancer cells invade the dermis which allows them access to the lymphatic and blood vessels, thereby allowing metastasis to distant sites (Chudnovsky *et al.* 2005).

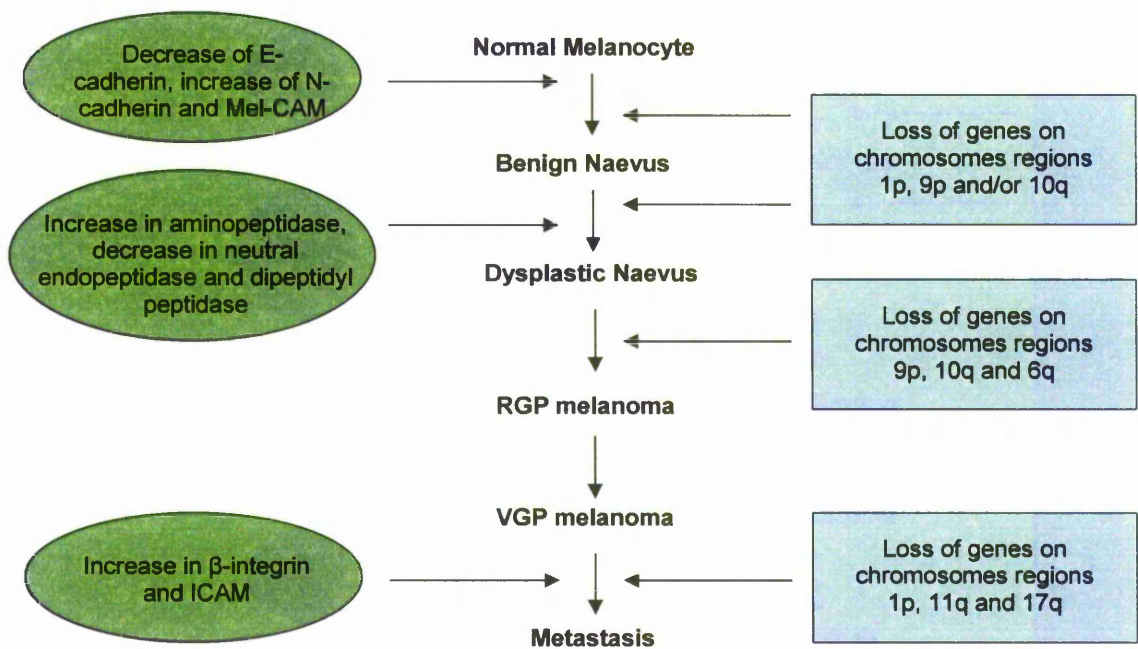


Figure 1-5 Figure showing the genes and proteins known to be involved in specific stages of melanoma progression.

There are six stages of melanoma progression, from normal melanocytes to metastasis and each stage involves the loss or mutation of a set of genes that facilitate the progression of melanoma. Although the identity of specific genes involved at each stage has not been fully characterised, losses of genes at chromosomal regions of 1p, 9p and 10q are thought to have a role. It is thought that changes in the expression of various ECM related proteins lead to the aggressive nature of melanoma (Bogenrieder *et al.* 2002)

Gene or chromosomal region	Mutation present in familial or sporadic melanoma?	Nature of mutation/genetic alteration
<i>p16^{INK4A}</i>	Familial and Sporadic	Point mutation, homozygous deletion, promoter methylation
<i>cdk4</i>	Familial and Sporadic	Point mutation
1p36	Familial	-
6p	Familial	-
<i>Rb1</i>	Sporadic	Point mutation
<i>tp53</i>	Sporadic	Point mutation
6q	Sporadic	Loss of heterozygosity and cytogenetic alterations
10q23; <i>Pten</i>	Sporadic	Loss of heterozygosity and point mutation
11q22-23	Sporadic	Loss of heterozygosity
<i>N-ras</i>	Sporadic	Point mutation
<i>B-catenin</i>	Sporadic	Point mutations
<i>c-myc</i>	Sporadic	Overexpression
MC1R melanocortin receptor	Sporadic	Point mutations

Table 1-3 A summary of genes involved in malignant melanoma.

There are a number of genes which are thought to be involved in familial and/or sporadic melanoma, however characterisation of melanoma related genes and their mutations is still being investigated.

1.2.3 Melanoma Staging

The prognosis of melanoma is determined according to the thickness (Breslow scale) and/or level of invasion (Clark level) of the melanoma, mitotic index, the presence of tumour infiltrating lymphocytes, number of regional lymph nodes involved, ulceration or bleeding at the primary site and lactate dehydrogenase (LDH) levels in stage IV disease. Clark *et al.* 1969 have described 5 separate levels of invasiveness and demonstrated that prognosis correlated well with the depth of invasion. Work by Breslow 1975 has suggested that the vertical tumour thickness in millimetres may be a better guide to prognosis, tumours less than 0.75mm in thickness rarely metastasise. In 2002, a new AJCC staging system for melanoma was formed providing more accurate and precise information regarding patient prognosis (Kim *et al.* 2002b). Patients are classified into primary tumour (T), regional lymph nodes (N) and distant metastases (M). T is now calculated by tumour thickness (Breslow) and the presence of ulceration in the primary tumour. The Clarks level still carries prognostic information in patients with 'thin' (<1.0mm) melanoma and so pathologists still include the Clark level in the histology report of the primary tumour. The presence of lymph node metastases (N) is one of the most important predictors of survival in melanoma patients. These are identified by haematoxylin and eosin (H and E) staining and immunohistochemistry with S-100 and HMB-45 stains of the sentinel lymph node. M identifies patients with distant and systemic metastatic melanoma. Three factors are important in determining prognosis for advanced stage melanoma: the site of the metastasis, the number of metastatic sites and the level of LDH in the patient's serum. M1 disease includes distant skin, subcutaneous or lymph node metastases, M2 disease includes lung metastases and M3 includes all other visceral organ or distant metastases and/or elevated serum LDH levels.

1.2.4 Melanoma Treatment

As well as surgical therapy, which is the main treatment for melanoma, adjuvant chemotherapy, immunotherapy, radiotherapy and biologic therapy are also important. A summary of melanoma treatments is shown in table 1-4.

Different treatments available for melanoma.
--

Therapy for primary melanoma

- | |
|--|
| <ul style="list-style-type: none"> • Complete surgical excision of the primary tumour • Elective lymph node dissection • Sentinel lymph node biopsy |
|--|

Therapy for regional metastases

- | |
|---|
| <ul style="list-style-type: none"> • Surgery • Isolated limb perfusion • Adjuvant therapy (radiotherapy, chemotherapy, regional limb perfusion, IFN-alpha) |
|---|

Therapy for distant metastases

- | |
|---|
| <ul style="list-style-type: none"> • Surgery • Radiotherapy • Chemotherapy (single-agent or combination including Dacarbazine (DTIC), Carmustine (BCNU), Lomustine (CCNU), Vindesine, taxanes, platinum compounds and combinations of DTIC, BCNU and Cisplatin) • Chemoimmunotherapy (IFN-alpha + chemotherapy) • Biologic therapy (IFN-alpha, IL-2, monoclonal antibodies, melanoma vaccines) |
|---|

Table 1-4 A summary of the different treatments currently used for the treatment of melanoma, adapted from Lugovic *et al.* 2005; Kim *et al.* 2002a.

1.2.5 The Future for Melanoma Detection and Treatment

Advanced melanoma is relatively resistant to therapy and so new and innovative approaches for melanoma detection and treatment are needed (Kim *et al.* 2002a). The identification of tumour markers could aid detection and treatment but as yet none have been identified that are associated with the early detection of melanoma. Most of the serological markers detected so far are detected in advanced disease, these include S100, melanoma inhibitory activity, cytokines, cytokine receptors, neurone specific enolase and melanin metabolites (Brochez *et al.* 2000). S100 levels have no clinical value in detecting early disease in melanoma patients but it has been associated with aggressive tumour behaviour (Hansson *et al.* 1997). There has also been the identification of melanoma associated antigens, for example MART-1 (also known as Melanin-A) which is widely

expressed in primary and metastatic melanoma and so could be used as a histological marker of melanoma or as a potential target in peptide vaccinations for immunotherapy of patients (Pittet *et al.* 2002). Other potential peptide targets for immunotherapy include tyrosinase, gp100/pMEL17, gp75/TRP-1, TRP-2, MAGE-1, MAGE-2, MAGE-3, BAGE, GAGE-1,2, NY-ESO-1, mutated β -catenin and p15 (Kim *et al.* 2002a). However, despite advances in our understanding of the disease the incidence of melanoma has continued to rise, with no decline in death rates and no progression into its treatment of disease (Herlyn 2002). In the absence of advances in the non-surgical treatment of advanced melanoma, diagnosing patients with early disease, where the invasion of the tumour is much shallower, will result in much better prognosis and the possibility of complete cure with surgical excision of the tumour (Brenner *et al.* 2002). With this in mind it is important to take advantage of new technologies and embark upon new strategies for identification of genetic and proteomic markers that associate with disease progression, diagnosis, prognosis and more significantly response to therapy which will hopefully lead to the early diagnosis and more successful treatment of the disease.

1.3 Breast Cancer

Breast cancer is the most common cancer in women with an incidence rate of 1 in 10. The term breast cancer describes a breast malignancy that has invaded beyond the basement membrane of the duct and therefore has the potential to spread to the lymph nodes and distant sites. It is viewed as a clonal disease that depends on genetic mutation and can originate from either of the glands (lobular) or the ducts of the breast; caused by the irregular growth of cells found in either of these areas. Virtually all invasive breast cancers are adenocarcinomas with 85% ductal adenocarcinomas and 15% lobular adenocarcinomas. The risk of developing breast cancer increases with advancing age; the risk is 5.8 times higher after the age of 65 although this is not the only risk factor for developing the disease. The other risks include age at pregnancy, age at menopause, hormone replacement therapy, alcohol consumption and exposure to ionizing radiation but overall they can be assigned to two categories: excessive exposure to oestrogens or deficiency in the maintenance of genomic integrity (Singletary 2003). Mortality from breast cancer has fallen by 31% in the UK since 1989 but still accounts for 17% of female deaths from cancer and for women in the 35-54 yr age group, breast cancer accounts for 17% of all deaths. The reason for the decline in mortality rates includes widespread mammographic screening, more precise diagnosis and better treatment (Anon 2006)

1.3.1 Breast Cancer Tumourigenesis

The development of breast cancer involves many genes but the sequential steps involved in the progression of the tumour are not clear. Inconclusive evidence suggests that it may begin with the hyperproliferation of the epithelial cells progressing through a preneoplastic phase called ductal carcinoma *in situ* (DCIS), which is contained within the basement membrane of the duct, to invasive breast cancer, where the cancer has breached the basement membrane (Lakhani 1999). LOH in the chromosomal regions of 8p, 16q and 17q have been detected in the precursor stages, specifically in ductal hyperplasia and atypical hyperplasia, suggesting that these are early genetic changes. In the same study it was also determined that atypical, ductal hyperplasia is a premalignant lesion for the development of breast cancer (Amari *et al.* 1999). The ovarian hormone, oestrogen, has been found to be involved in stimulating breast cancer growth, by stimulating proliferation in malignant

cells, leading to the development of therapies that inhibit the synthesis of this hormone or block its receptor. Approximately 15-25% of the epithelial cells within the normal breast are oestrogen receptor (ER) positive and these have been found to be mainly non-dividing cells. Proliferation that is stimulated by oestrogen occurs mainly in ER-negative cells that surround the ER-positive cells, probably due to the secretion of paracrine factors. Proliferation of ER-positive cells is mainly regulated by oestrogen although the conversion of these cells from a non-dividing state remains unclear (Ali *et al.* 2002). Two-thirds of breast cancers are ER-positive and most of these respond to endocrine therapy (Khan *et al.* 1998).

1.3.2 Genetic Involvement in Breast Cancer

Most cases of breast cancer are sporadic not familial and are caused by genetic damage to the breast cells that are acquired during a woman's lifetime. Chromosomal instability caused by enormous chromosomal abnormalities tends to be the most characteristic feature of the breast cancer genome along with the widespread hypermethylation of regulatory regions of the genome (Lerebours *et al.* 2002; Widschwendter *et al.* 2002). Single nucleotide instability is less common in breast cancer but it has been shown that alterations in mitochondrial DNA occurs (Bianchi *et al.* 2001). A wide variety of genes have been implicated in the development and progression of the disease including genes encoding growth factors and receptors, intracellular signaling molecules, cell cycle regulators, apoptosis regulators, adhesion molecules and the high risk breast cancer susceptibility genes *BRCA1* and *BRCA2*. Three of the more breast cancer specific genes are discussed briefly below.

1.3.2.1 *Her-2*(*Her-2/neu* or *erbB-2*)

This gene is located on chromosome 17q and encodes a tyrosine kinase growth factor receptor, which is initiated by binding to specific ligands leading to the activation of multiple signaling cascades including the MAP kinase and PI3K/Akt pathways causing proliferation, angiogenesis, increased cell motility and resistance to apoptosis (Osborne *et al.* 2004). This gene is rarely amplified in benign breast disease and is found to be either

amplified or over-expressed in 10-34% of invasive breast cancers and in the majority of high-grade DCIS cases (Ross *et al.* 1999; van de Vijver *et al.* 1988).

1.3.2.2 *BRCA1*

BRCA1 is a tumour suppressor gene that is located on chromosome 17q21 and it has been estimated that approximately 0.12% of the population carries a mutation in this gene. *BRCA1* mutations account for approximately 5% of all breast cancer cases in women under the age of 40 but this figure rises to more than 90% if there is a family history of more than four cases of breast cancer and more than 1 case of ovarian cancer (Ford *et al.* 1995). *BRCA1* encodes a 208 kDa protein that has been shown to have an involvement in transcription, replication and DNA repair (Liu *et al.* 2002; Scully *et al.* 1997). Over 200 individual mutations of this gene have been described including deletions, substitutions and insertions and these can be found along the whole length of the gene (Osborne *et al.* 2004). The location of the mutation can suggest the severity of the disease with mutations at either the amino or carboxy terminus associating with tumours that have a high proliferation rate (Sobol *et al.* 1996).

1.3.2.3 *BRCA2*

The *BRCA2* gene is located at chromosome 13q12-13 and is also a tumour suppressor gene, sharing many of the same features as *BRCA1*. The structures of *BRCA1* and *BRCA2*, however, are different, for example this gene also localises to areas of damaged DNA. Over 100 mutations of this gene have been described most of which cause premature truncation of the protein. The incidence of someone in the general population carrying a mutation in this gene is the same as in *BRCA1*. The *BRCA2* gene encodes a 384 kDa protein that has many of the same functions as *BRCA1*, although evidence suggests that *BRCA2* has more of a direct role in DNA repair than *BRCA1* (Yoshida *et al.* 2004). Mutations in *BRCA2* are also associated with increased risk of developing other cancers including melanoma, prostate cancer and gastric cancer (Consortium 1997).

Gene	Chromosomal Location	Normal Function	Type of Genetic Event	% of tumours
<i>c-myc</i>	8q24	Transcription factor	Amplification	15-25%
<i>Her-2</i>	17q12-21	Tyrosine kinase receptor	Amplification	30-70%
<i>tp53</i>	17p	Induces cell cycle arrest, triggers apoptosis	Deletion	41-73%
<i>BRCA1</i>	17q21	Regulates DNA transcription	Deletion	30-70%
<i>BRCA2</i>	13q12-13	Repairs damaged DNA	Deletion	33-75%
<i>Pten</i>	10q	Phosphatase	Deletion	Increased risk of breast cancer
<i>E-cadherin</i>	16q22	Cell adhesion molecule	Loss of homozygosity	>85%
<i>Cyclin D1</i>	11q13	Cell cycle regulator	Over-expressed	40-50%
<i>Cyclin E</i>	19q12	Cell cycle regulator	Amplification	2%

Table 1-5 Breast cancer associated genes.

This table shows the most well-known breast cancer associated genes, their chromosomal locations, normal function, type of genetic alteration and the frequency of their mutations in breast cancer (Ingvarsson 1999; Osborne *et al.* 2004).

1.3.3 Therapy of Breast Cancer

Treatment of breast cancer involves 3 main areas: treatment of the breast with either breast conservation surgery and radiotherapy or mastectomy and/or radiotherapy and/or breast reconstruction; treatment of the regional lymph nodes with axillary clearance or sentinel node biopsy or a combination of both; and systemic treatment which could include hormone therapy (tamoxifen, aromatase inhibitors or ovarian suppression), chemotherapy or new treatments such as monoclonal antibodies (Brennan *et al.* 2005).

A number of chemotherapy combinations have been shown to be effective in the treatment of breast cancer. A regimen of cyclophosphamide, methotrexate and fluorouracil (CMF) is commonly used as a therapeutic treatment of breast cancer and has been shown to improve the 10 year survival in patients under the age of 50 by 7-11% but regimens containing modern anthracyclines, such as adriamycin and epirubicin, have proven more effective although causing more side effects (Howell *et al.* 2005).

Endocrine therapy is considered for patients that have breast tumours expressing oestrogen or progesterone receptors as this therapy has been found to be reasonably non-toxic. This type of therapy is aimed at reducing the oestrogen that is available to the tumour cells by blocking the oestrogen receptor or by reducing the levels of oestrogen in the blood. Tamoxifen is a non-steroidal partial anti-oestrogen which was first used in clinical trials for the treatment of breast cancer in 1971 (Jordan 1988). It has been a highly successful drug for the treatment of ER-positive breast cancer and works by blocking the oestrogen receptor, although it is now clear that extended use of tamoxifen for more than five years has a detrimental effect on patients, causing endometrial cancer and thromboembolism with long term use (Fisher *et al.* 2001). Aromatase inhibitors work by blocking aromatase, which is an enzyme that converts testosterone to oestrogen in peripheral tissues, but are only effective in post-menopausal women. Anastrozole (Arimidex®), letrozole (Femara®) and exemestane (Aromasin®) are third generation aromatase inhibitors and recent trials have shown that these are more effective than tamoxifen in preventing breast cancer relapse and have fewer side effects (Winer *et al.* 2005). Trastuzumab (Herceptin) is a monoclonal antibody that targets the cell surface receptor Her-2 and so is only effective in those patients whose tumours have an amplification or overexpression of the gene that encodes for this receptor. In trials where trastuzumab was given to patients after completion of chemotherapy, the disease free survival at 2 years was 85.8% compared with 77.4% for

patients who received no further treatment although deterioration in left ventricular function was noted. Data from clinical trials has shown that trastuzumab should be considered as an adjuvant therapy for patients with Her-2 positive tumours but long-term safety data stills needs to be obtained before it can become routinely used (Houssami *et al.* 2006).

1.3.4 The Future for Breast Cancer Detection and Treatment

The discovery that the endocrine system has an important role in the development of breast cancers has lead to the development of effective and safe drugs, such as tamoxifen, that have contributed to the decrease in mortality seen in the past few years. However, a large proportion of patients with localized disease and all patients with metastatic disease become resistant to endocrine therapies, which have fuelled the search for alternative strategies for overcoming or bypassing this resistance (Ali & Coombes 2002). Traditional methods for the treatment of metastatic breast cancer have included the use of multiple lines of hormone therapies and the increased duration, dose and intensity of chemotherapy but despite prolonging the time to progression of the cancer, there has been no improvement in the overall survival of patients. Another problem also occurs in patients that develop anthracyclin and taxane resistant tumours as there are few alternative therapies available for effective treatment (Awada *et al.* 2003). New strategies are needed to provide prognostic and predictive information for established treatments as well as for the discovery of new therapeutic targets. It may be possible, with the advancement of DNA microarrays and proteomics, to tailor treatment for each patient as well as gain a better understanding into the progression of breast cancer and the mechanism by which endocrine therapy and chemotherapy resistance occurs.

1.4 Biomarker Identification

The combination of tumour size or depth (T), lymph node spread (N) and presence or absence of metastases (M), the TNM staging system, has provided the basis for diagnosis of cancer progression, predicting survival and deciding the choice of treatment in cancer patients since 1958. Individual or groups of molecular markers are now becoming more widely used to classify tumours into subsets that behave differently from each other (Ludwig *et al.* 2005). Biomarkers can indicate the physiological state of the cell at a specific point in time and so the identification of these markers could prove important in gaining an understanding of the physiology and pathology of cancers and therefore improve patient diagnosis and treatment. Biomarkers that are currently in use to assess cancers include chorionic gonadotropin and α -fetoprotein for germ cell tumours, monoclonal immunoglobulin and urine electrophoretic peaks in myeloma, prostate serum antigen (PSA) for prostate cancer (Pritzker 2002), VEGF mRNA expression in lung cancer (Yano *et al.* 2000), tyrosine kinases in leukaemia (Druker *et al.* 2001) and HER-2 in breast cancer (Molina *et al.* 2001).

Potential important biomarkers are those genes and proteins that are involved in the growth and maintenance of cancer cells and can arise due to over-expression, mutation, chromosomal rearrangement, altered pre-mRNA processing, post-translational modifications and viral antigens (Miles *et al.* 2006). Relevant biomarkers have yet to be identified that have a high enough sensitivity and specificity to be accepted in the assessment of early-stage cancers. It is also important to attempt to identify new biomarkers than can be detected in patients via non-invasive tests, for example in urine, saliva and sputum (Negm *et al.* 2002). For example, molecular assays have shown mutations in *tp53* in urine taken from bladder cancer patients and ras mutations in stool samples taken from colorectal cancer patients (Sidransky *et al.* 1991; Sidransky *et al.* 1992).

There are a number of ways tumour antigens or biomarkers can be identified. cDNA expression cloning is a technique used for the isolation of tumour antigens recognized by T cells. This method has provided a catalogue of melanoma antigens including mutated peptides derived from genetic alterations in the tumour cells, self-peptides derived from

tissue-specific proteins, over-expressed proteins in tumour cells and cancer testis antigens. Direct identification of tumour antigens bound to tumour cell human leukocyte antigens (HLA) has proven successful in isolating naturally presented antigens including those with post-translational modifications using HPLC fractionation and mass spectrometry. These two methods have identified antigens, for example peptides, for use, more specifically, as targets for vaccine therapy (Clark *et al.* 2001; Kawakami *et al.* 2004).

cDNA expression cloning using serum IgG antibodies from cancer patients, known as SEREX, has led to the identification of antigens which have proven useful for the diagnosis, prognosis and therapy of different cancers (Heubeck *et al.* 2006; Li *et al.* 2004; Okada *et al.* 2006).

Expressed sequence tags (ESTs) can be used to predict the expression patterns of specific genes in different tissues. By data mining these ESTs it may be possible to identify genes that are expressed in cancer tissues and therefore identify new biomarkers. It is then possible to validate these findings using real-time RT-PCR.

Protein-based technologies have recently been applied to the discovery of new biomarkers. These technologies include low-throughput methods such as western blotting, *in situ* hybridisation, immuno-histochemistry and 2-D gel electrophoresis. More high throughput methods have recently lead to biomarker identification including protein microarrays, MALDI MS and ProteinChip technology, also known as SELDI MS (Miles *et al.* 2006).

Each of these methods has yet to provide antigens that have significantly facilitated the diagnosis or treatment of cancer patients. An approach that incorporates both genomic and proteomic technologies along with sophisticated data mining methods may prove to be the key to identifying biomarkers that have real clinical significance. Success must rely on using these approaches to validate identified markers against disease endpoints, to establish quantitative criteria for the presence or absence of the biomarkers and to confirm the predictive value of the marker in prospective population trials (Tockman *et al.* 1992).

1.5 The Application of Proteomics in Cancer Research

Cancer is known as a disease caused by defects or mutations in certain genes but genes, however, only contain information that has yet to become functional. The functional information, in the form of proteins, occurs as a result of the translation of mRNA (Posadas *et al.* 2005). The gene-encoded information can also change by the time the proteins are translated. This is due to gene amplification, alternative RNA splicing, co-translational modifications, post-translational modifications, differential stability and secretion of proteins (Verma *et al.* 2001).

To date the majority of studies have analysed single genetic mutations as potential biomarkers but correlating the presence of these mutations with disease progression has proven difficult. Other studies have investigated RNA expression levels in cancer, but this does not necessarily correlate with protein expression and does not take into account any post-translational modifications of these proteins. The advent of proteomics and high-throughput technology allows scientists to determine protein expression patterns that may allow the discrimination between disease states and with further interrogation the identity of the proteins. These proteins may then become the “next generation” of clinically significant biomarkers important in diagnosis, disease progression and response to therapy. Development of sophisticated, high-throughput and sensitive technologies has enabled researchers to begin to dissect the genomic changes, expression events and the differential expression, activation and signalling of a wide variety of proteins isolated from tumour samples. These technologies include 2-dimensional polyacrylamide gel electrophoresis (2D PAGE), protein microarrays, laser capture microdissection and mass spectrometry (Michener *et al.* 2002). The application of proteomics to cancer marker identification is becoming increasingly popular. These approaches include the analysis of protein expression in normal and tumour tissue to detect over or under-expressed proteins that may serve as novel markers, analysis of secreted proteins in cell lines and direct serum protein profiling to identify potential new markers (Shin *et al.* 2002).

1.5.1 2-D Polyacrylamide Gel Electrophoresis (2D PAGE)

2D PAGE has the capability to separate thousands of proteins in a single analysis according to isoelectric point (pI) and then according to molecular weight. In practice, high resolution

gels can resolve approximately 3000 proteins depending on the sample used and the sensitivity of the staining technique (Lopez 1999). For the last 25 years 2D PAGE has been the technique of choice for analyzing the protein content of human samples and over this time there have been many alternatives in the method to try to improve the resolving power. These improvements include the use of immobilised pH gradient (IPG) strips, (Righetti 1990), the introduction of new reducing agents, (Herbert *et al.* 1998), and new surfactants, (Chevallet *et al.* 1998), to improve sample solubilisation. Traditional 2D PAGE required large amounts of material and complex sample preparation, however with recent advances in technology this is no longer necessary (Liotta *et al.* 2001b). To improve the detection of low abundant proteins in whole cell lysates or body fluids, the use of chromatography as a prefractionation step has been applied. This is not for the purification of individual proteins but instead for the enrichment of certain classes of proteins using group specific resins, for example the use of lectin affinity resins to enrich for glycoproteins (Lopez 1999).

The methods for the identification of proteins present in single 2D PAGE spots have been improved with the development of mass spectrometric techniques including matrix assisted laser desorption ionisation (MALDI) and electrospray ionisation (ESI) mass spectrometry. Proteins from gel spots can be characterised using mass spectrometry via peptide mass fingerprinting (PMF). The masses of peptides derived from a protease (usually trypsin) digest of 1 gel spot are determined using mass spectrometry and the masses are then used to search various databases such as Genbank and SwissProt. Depending on the mass accuracy of the instrument used, an unknown protein from a gel spot can be identified from as few as 5 peptides (Lopez 1999).

A recent study has shown that only the most abundant proteins are seen on a 2D gel and identified by mass spectrometry (Gygi *et al.* 2000). Despite the limitations of 2D PAGE it is still widely used for the analysis of protein mixtures and one of its strengths is the reliable and rapid comparison of healthy and diseased protein samples (Issaq *et al.* 2002a). 2D PAGE has been used to analyse a wide range of samples including plasma taken from Alzheimer's patients, colon carcinoma cell lines, hepatocellular carcinoma samples, membrane proteins and cerebrospinal fluid, to name a few (Seow *et al.* 2000; Simpson *et al.* 2000; Ueno *et al.* 2000). It has also facilitated the discovery of biomarkers in ovarian, oesophageal and bladder cancer, amongst others (Celis *et al.* 2000; Jones *et al.* 2002; Petricoin *et al.* 2002b; Zhou *et al.* 2002).

It has been suggested that a mammalian cell could produce on average 50,000 unique proteins. This means that even though it is possible to visualise between 3000 and 10,000 protein spots on a 2D gel, these only represent between 7-24% of the most abundant proteins in a cell (Vuong *et al.* 2000). This then suggests that 76% of proteins expressed in a cell are below the detection limits of standard gel staining methods. Improvements to this resolution problem could include the use of radioactive labels, or increasing the separation by using ultrazoom gels.

As well as its ability to resolve thousands on proteins and provide rapid comparison of proteins between samples of different types, it is also economical when compared with other separation methods such as HPLC and capillary electrophoresis (CE), which both need special equipment. The limitations far outweigh the advantages, however, and include the poor solubility of hydrophobic and membrane proteins, narrow dynamic range, difficulty in focusing highly basic and acidic proteins, inadequate sensitivity, poor quantification and the fact that it is very labour intensive, and that post-translationally modified proteins will increase the repertoire of the proteome (Lilley *et al.* 2002; Timperman *et al.* 2000).

1.5.2 Protein Microarrays

Protein microarrays allow the systematic analysis of thousands of proteins simultaneously and can be used to determine the post-translational modifications of the proteins analysed, their levels in biological samples and their selective interaction with other proteins and antibodies. Most protein microarrays are affinity-based, being chemically robust, stable and compact with a high binding efficiency and specificity. Protein-to-protein based arrays include antibody, phage displayed antibodies or polypeptide recognition moieties (Jenkins *et al.* 2001). One of the main problems in using this approach, however, is the lack of available antibodies. Also, post-translational modifications are not captured using recombinant proteins or antibodies that do not distinctly recognize specific forms of a protein (Misek *et al.* 2004). Finally, there is also a problem with cross-reactivity of the proteins with affinity agents (Baak *et al.* 2003).

1.5.3 Mass Spectrometry

The principle of mass spectrometry is to generate ions from either inorganic or organic compounds by a suitable method, separate those ions by their mass-to-charge ratio (m/z) and detect them either qualitatively or quantitatively. Mass spectrometers are capable of generating 2 types of information, firstly an accurate measurement of molecular weight and secondly a partial amino acid sequence, which can be achieved using tandem mass spectrometry.

A mass spectrometer consists of three essential components, an ion source, a mass analyser and a detector which are operated under high vacuum. The analyte can be ionised thermally, by electric fields or by impacting energetic electrons, ions, photons, neutral atoms or heavy cluster ions causing the production of gas-phase ions. The mass analyser then separates the ions by static or dynamic electric or magnetic fields or by time-of-flight (TOF) in a field free region according to their m/z value before they strike the detector. When ions hit the detector they are represented in mass spectrum; this is a 2-D representation of signal intensity versus m/z . The signal intensity of a peak reflects the abundance of ions at that particular m/z value that have been generated via ionisation of the analyte. Because the charge (z) is often equal to one (more specifically in the case of MALDI-MS) then the m/z value directly reflects the mass of the ion (Gross 2004a). The development of both MALDI and electrospray ionisation (ESI) to create ions from large molecules have been important breakthroughs for proteomic studies. MALDI creates ions by firing a laser at a sample which is combined with an energy absorbing matrix. The energy from the laser causes the excitation of the matrix and subsequent ejection of the matrix and analyte ions into the gas-phase. ESI creates ions by the application of a potential to a flowing liquid, usually coupled to LC, causing the liquid to become charged and then spray. The spray creates small droplets of solvent containing analyte. The solvent is removed by heat or energetic collisions with a gas, as the droplets enter the mass spectrometer, whilst multiply charged ions are formed (Yates 2000).

There are four basic types of mass analysers that are currently used in proteomic studies, each with their own design and performance; Quadrupole mass analysers, which resolve m/z by applying radio frequency and DF voltages. This allows only a narrow mass range to reach the detector, and so have limited mass range and low resolution; Time-of-flight analysers are most commonly used with MALDI. These accelerate ions by using a short

voltage gradient and measure the time it takes for the ions to travel through the flight tube, the flight time is proportional to the square root of the m/z . The resolution of this analyser is quite good, it is able to separate ions to 0.1Da; Quadrupole ion traps focus the ions into a small volume where ions are captured (or trapped), activated and ejected by the electronic manipulation of an oscillating electric field. This concentration of ions means that these analysers are very sensitive but have low mass accuracy. This is due to the limited number of ions that can be accumulated before space-charging distorts their distribution and therefore the accuracy. Recently, linear ion traps have been developed where the ions are stored in a larger area allowing the increase in sensitivity, resolution and mass accuracy. Ion traps can also rapidly shift between scanning for the masses of the analyte (MS scan) and generating fragmentation spectra of the ions detected in the MS scan (MS/MS scan); Fourier Transform Ion Cyclotron Resonance (FT-ICR) uses high magnetic fields to trap the ions and cyclotron resonance is used to detect the ions. This analyser has extremely high resolution as it can separate ions to 1mDa, although the expense, operational complexity and the low peptide fragmentation efficiency has limited their widespread use in proteomics research (Aebersold *et al.* 2003; Wysocki *et al.* 2005).

New mass spectrometry-based systems have been developed that can rapidly profile and generate proteomic “fingerprints” from tissues and body fluids. SELDI MS has been applied to the analysis of patient sera to identify protein patterns characteristic of different tumour types and stages where only microlitre quantities of crude serum are required for MS analysis. Protein patterns, distinct for different cancer patient populations, have been reported for a number of different cancers including prostate, colon, breast and ovarian cancer (Oehr 2003; Petricoin *et al.* 2002a). A relatively new approach is to use multidimensional liquid chromatography directly coupled to mass spectrometry, either on-line with ESI-MS or off-line with MALDI-MS. This technology is under development and needs further work for detecting potential biomarker candidates, but it is very promising (Shin *et al.* 2002; Wang *et al.* 2003).

1.6 MALDIMS

MALDI-MS was first introduced by Tanaka *et al* and Karas *et al* in 1988 and has now become a widespread analytical tool for the study of peptides, proteins and other biomolecules (Tanaka *et al.* 1988; Karas *et al.* 1988). The analyte of interest is cocrystallised with a large molar excess of an organic matrix that strongly absorbs the energy from a laser to desorb and rapidly ionize intact high mass compounds (Chaurand *et al.* 1999a), see figure 1-6.

MALDI matrices are crystalline solids which absorb light at the wavelength which is intended to be used for the experiment. They can serve as protonating or deprotonating agent or as an electron-donating or -accepting agent so acting to analyse the analyte in both positive and negative ion modes. Common matrices for use in UV-MALDI include 2,5-dihydroxybenzoic acid, DHB, for oligosaccharides; α -cyano-4-hydroxycinnamic acid, CHCA, for peptides; and 3,5-dimethoxy-4-hydroxycinnamic acid, SPA or sinapinic acid, for proteins. Other matrices are also available with their structures based around an aromatic core. In IR-MALDI the wavelengths are absorbed by O-H and N-H vibrations and so malonic acid, succinic acid, malic acid, urea and glycerol can all be used as effective matrices.

Lasers at wavelengths ranging from ultraviolet (UV) to infra-red (IR) can be used but UV nitrogen lasers (337nm) are the most common. The pulse of laser light is focused onto a small spot which is usually 0.05-0.2mm in diameter. An attenuator can be used in the laser optical path to adjust the laser irradiance. For example, a rotating UV filter can allow the transmission of 1-100% of the laser beam energy and this can be adjusted for each measurement. UV lasers emit pulses between 3-10ns long whilst IR laser pulses are usually between 6-200ns long. Short laser pulses are needed to cause a sudden ablation of the sample layer, to avoid thermal degradation during the short time interval of ion generation and to give a better starting pulse for the TOF measurement, although this latter advantage is also improved with the use of delayed extraction.

Several theories have been developed to explain desorption of molecules from the target. Firstly the Thermal Spike Model suggests that the matrix molecules sublime from the surface after local heating above a certain laser intensity but a low laser fluence. The ejection of the intact analyte molecules is attributed to poor vibrational coupling between

the matrix and analyte (Vertes *et al.* 1990). Conversely the Pressure Pulse Theory suggests that a pressure gradient is created at the surface and desorption of molecules is enhanced by momentum transfer from collisions with fast-moving matrix molecules (Johnson *et al.* 1991). Many chemical and physical pathways have been suggested as ionization mechanisms for MALDI including gas-phase photoionization, excited-state proton transfer, ion-molecule reactions and desorption of preformed ions. It has been suggested that photoionization occurs when using UV-absorbing matrices as positive and negative radical ions ($M^{+\bullet}$ and $M^{\bullet-}$) have been observed and these can only be generated by the addition or removal of an electron. The most widely accepted mechanism involves gas-phase proton transfer in the expanding plume and depends critically on the matrix-analyte combination (Gross 2004b).

There are 3 types of mass analyzers typically used with the MALDI ionization source, a linear time-of-flight (linear TOF), a reflectron TOF, and a Fourier Transform Mass Analyzer. The Linear TOF is the simplest of the three and is based on accelerating a set of ions to a detector where all of the ions are given the same amount of energy, a diagram is shown in figure 1-7. The principle of this method is that if ions are accelerated with the same potential at a fixed point and a fixed initial time and are allowed to drift, the ions will separate according to their mass to charge ratios. A reflectron TOF combines TOF technology with an electrostatic analyser, the reflectron. The main drawback of linear TOF analysers is their poor mass resolution. This is affected by factors that cause a distribution in the flight time amongst ions with the same m/z value. These factors include length of the ion formation pulse (time distribution), the size of space where the ions are formed (space distribution), the variation of the initial kinetic energy of the ions (kinetic energy distribution). The reflectron increases the amount of time the ions need to reach the detector while reducing their kinetic energy distribution therefore increasing the mass resolution but at the expense of sensitivity, also it has a relatively low mass range, < 10,000 Da. Another way to reduce the kinetic energy distribution among ions of the same m/z value, and therefore improve mass resolution, is to use delayed extraction, also known as delayed pulsed extraction, pulsed ion extraction, pulsed extraction or dynamic extraction. This introduces a time lag or delay between the formation of the ions and extraction into the TOF analyser. The ions are allowed to expand in the field free region of the source and after a delay of ~150ns a voltage pulse is applied to extract the ions into the analyser. The

benefits of delayed extraction are diminished, however, with proteins of mass $> 30,000$ Da. (Lewis *et al.* 2000).

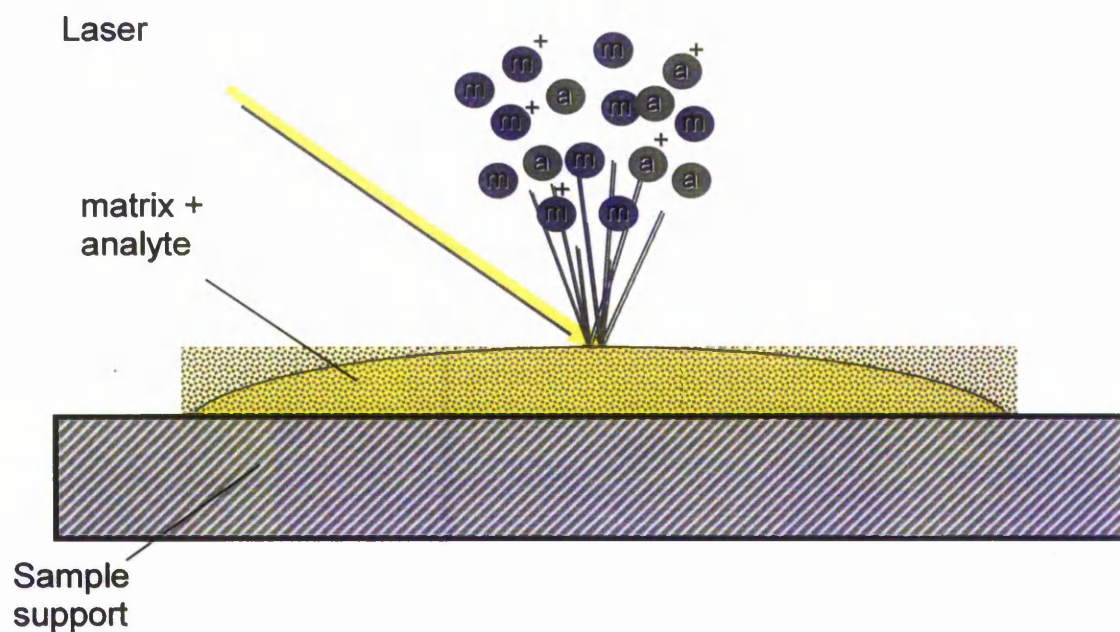


Figure 1-6 The mechanism of desorption and ionization of an analyte via MALDI.

A laser pulse is fired on to the sample where the matrix molecules (m) absorb the energy from the laser causing the desorption of the matrix and analyte (a) from the MALDI target surface. Ionisation of the analyte occurs in the gas-phase before the ions enter into the mass analyzer and detector where the m/z values of the ions can be determined.

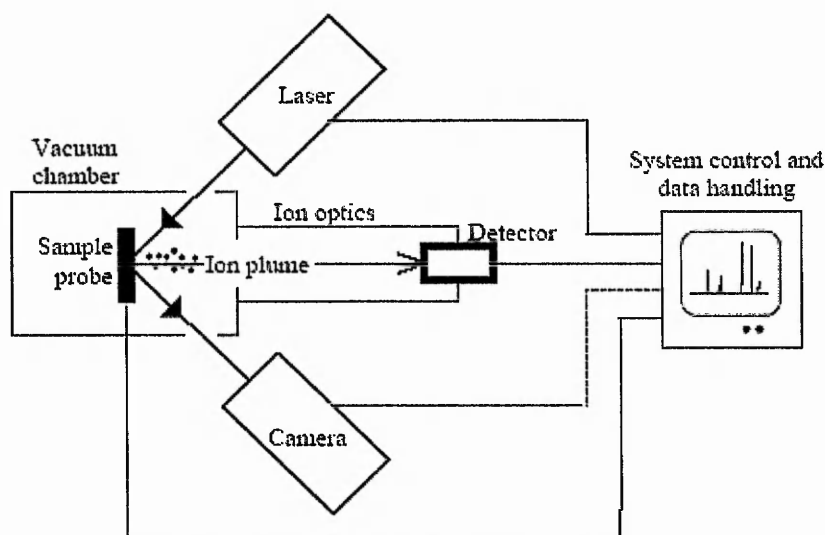


Figure 1-7 Diagram of a typical linear TOF mass spectrometer.

The ions are given the same amount of energy accelerating them towards the detector. As they are accelerated with the same potential at a fixed point and a fixed initial time, they drift towards the detector, separating according to their mass to charge ratios.

Peptide Mass Fingerprinting (PMF) can be performed using MALDI-TOF for the identification of specific proteins as it is relatively easier to use than ESI-MS and is more tolerant to salts and other contaminants that may be present after the tryptic digestion of a sample. Some pre-mass spectrometry clean-up can be carried out with the use of ZipTips (Millipore). Delayed extraction is used within the reflectron MALDI-TOF to improve the mass accuracy allowing for a more accurate PMF (Nyman 2001).

Post Source Decay (PSD) can also be performed using MALDI-TOF in which peptides are partially sequenced. The ions are fragmented in the TOF mass analyzer by collisions between the analyte ions and neutral matrix ions or residual gas molecules during the desorption and acceleration stage. As these fragments have the same velocities as their precursor ions, they will reach the detector at the same time in a linear TOF and so cannot be identified in the mass spectrum. In a reflectron TOF these fragments enter the reflectron at different depths and are spatially differentiated causing the individual fragments to hit the detector at different times and so are represented in the mass spectrum (Lewis *et al.* 2000).

Due to the complexity of biological samples a pre-fractionation and/or purification is required before MALDI-MS analysis. As a result the majority of biomarker identification

studies in cancer have involved separation of proteins by 2-D PAGE before MALDI-MS analysis although there are other methods available. One particular study used serum taken from patients with non small cell lung cancer and control patients. The serum was fractionated using isoelectric focusing before MALDI-TOF mass spectrometry. Potential biomarkers were detected in the spectra using computer-based algorithms and then identified (Howard *et al.* 2003). A new technique involving MALDI MS has emerged in recent years, Imaging Mass Spectrometry, where tissue sections can be directly placed upon a MALDI target, a matrix applied on top and spectra obtained for different regions of the tissue. From the systematic analysis of a tissue section, protein maps that are directly correlated with the tissue architecture or morphology can be simultaneously obtained from hundreds of different protein species (Chaurand *et al.* 2004; Stoeckli *et al.* 2001). It is possible that this new approach will have an impact three major areas: tissue-based diagnosis, prognosis determination, and prediction of response to specific modes of therapy (Chaurand *et al.* 2006).

1.7 SELDI MS

Surface-enhanced laser desorption ionization mass spectrometry (SELDI MS) is a novel, analytical high throughput, array-based technology, developed by Hutchens and Yip (Hutchens *et al.* 1993) and is currently utilized in the ProteinChip system developed by Ciphergen® Biosystems Inc (Freemont, CA, USA). SELDI technology tries to overcome the requirement for purification and separation of proteins prior to mass spectrometry analysis by directly analysing proteins retained on the chromatographic surfaces of aluminium or stainless steel chips. Spectra are produced of the complex protein mixtures and are based on the mass-to-charge ratio of the proteins and on their binding affinity to the chip surface. The ProteinChip system is similar to DNA chip technology except that the ProteinChip system is much more complex due to the varying nature of the proteins compared to DNA.

A variety of chips can be used on which complex samples, for example body fluids and cell extracts, can be spotted. The different types of chip surface available include hydrophobic, ionic, hydrophilic, receptor, ligand, enzyme, antibody, DNA and protein A/G and these will retain a subset of proteins (Bischoff *et al.* 2004). Biochemically treated surfaces, such as antibody treated, are designed to interact with a specific protein, whilst the chemically treated surfaces retain whole classes of proteins (Issaq *et al.* 2002b). A variety of wash steps can be applied to the chips to remove any contaminants as well as enhance the selectivity of the chips used. Once washed, a matrix is applied to the chip and then it can be placed directly into the linear-TOF mass spectrometer and the retained proteins analysed in the same way as MALDI mass spectrometry.

Although SELDI MS has not improved on the resolving power of 2D PAGE, this approach does overcome a number of limitations presented by MALDI-MS, most importantly the direct analysis of crude samples such as body fluids and cell extracts (Vlahou *et al.* 2005). SELDI MS has proven its worth in recent years in pattern recognition and biomarker identification in the field of cancer, through the analysis of body fluids such as plasma, serum and urine as well as tissue samples. SELDI MS screening of plasma samples taken from 11 men with prostate cancer and 12 men with benign prostatic hyperplasia identified 3 protein peaks which were present in the plasma of the cancer patients and not in the controls (Lehrer *et al.* 2003). Another study identified a panel of 7 protein biomarkers that

were unique to the plasma samples taken from ovarian cancer patients (Rai *et al.* 2002). Although CA125 is already used as a prognostic tool for detecting and diagnosing ovarian cancer, 4 of the 7 markers identified improved the sensitivity of the screening method when used in combination with CA125. Zhukov *et al.* (2003) identified protein profiles that were unique to malignant lung tumours and pre- malignant epithelium. Three peaks in particular were increased in the tumour samples when compared to normal cells, and 1 peak was found to be not present in the controls (Zhukov *et al.* 2003). Assessment of urine from patients with TCC of the bladder led to the identification of 5 potential biomarkers that were preferentially expressed in the cancer samples (Vlahou *et al.* 2001). SELDI MS has also been used to identify 4 well characterized prostate cancer associated biomarkers in prostate cancer specific cell lysates, serum and seminal plasma (Wright Jr *et al.* 1999). Li *et al.* (2002) screened serum samples taken from breast cancer patients with stage 0, I, II, and III disease, healthy women and patients with benign breast disease. A panel consisting of 3 biomarkers was found to separate breast cancer and non-cancer samples (Li *et al.* 2002a). SELDI mass spectra were generated from ovarian cancer serum and identified patterns that discriminated cancer from non-cancer with a sensitivity of 100% and a specificity of 95% (Petricoin *et al.* 2002a), however this data has been proven by other groups to be misleading due to the manner in which the experiment was performed (Baggerly *et al.* 2004).

One of the main advantages of this technique is that contaminants such as salts and detergents can be removed prior to analysis and only proteins that interact with the chip spots are analysed allowing for the physicochemical property of identified protein peaks to be determined. The analysis of the chip is fully automated so up to 100 samples a day can be prepared and analysed. There are limitations to this technique, however. Firstly, the resolution and the sensitivity are poor for proteins with molecular weights greater than 30kDa. Secondly, this technology was developed initially to investigate the differential expression of proteins between samples of different types. The identification of some of these differentially expressed proteins is not possible using this system. Instead, samples would need to be pre-fractionated, proteins of interest isolated and digested and proteins identified by their PMF or partial sequence obtained by tandem mass spectrometry. Therefore the identification of differentially expressed proteins is still a difficult and low throughput task. Also, due to the difficulty in protein quantification the expression levels

of identified protein peaks is not possible. Despite these limitations the SELDI ProteinChip's strengths lie in the systematic identification and characterization of proteins for diagnostic and prognostic markers in tissues and body fluids and the speed at which potential targets for therapy can be identified is increased (Seibert *et al.* 2004). Improvements in the instrumentation and bioinformatics methods associated with SELDI MS will be needed, however, before this method could be considered as a clinical tool (Vlahou & Fountoulakis 2005).

1.8 Bioinformatics approaches for data analysis

To identify differential biomarkers, a large cohort of patient samples has to be analysed for the consistency of the presence of potential markers associated with disease classes of interest. Mass spectrometry in conjunction with robotics now allows a large number of samples to be analysed in just a few hours. The difficulty then arises when trying to analyse such complex data: for a study that may identify the presence of significant biomarkers it is necessary to analyse large sample numbers. If the mass spectrum between the 2kDa and 30kDa range is analysed then that would equate to approximately 18,000 data points per sample and therefore several million data points in the whole study. Much of the spectra contain background noise and low amplitude peaks so identifying true protein peaks requires a system that can rapidly search through all of the data to distinguish between noise and true protein peaks. To determine the presence of any consistent biomarkers amongst all this data would be virtually impossible without data interrogation using computer algorithms, which would allow a computer to sort through all of the data in a much faster and efficient manner. Many data mining systems are being developed but most fall into one of 2 main approaches; supervised learning systems that require data where the outcome or classification is already known; or unsupervised learning systems where the data is clustered without any previous knowledge of the outcome or classification. Supervised systems include linear regression models, non-linear feed forward neural networks, genetic algorithms and support vector machines. Unsupervised systems include K-means clustering, principal components analysis, hierarchal clustering, Euclidean distance based non-linear methods, fuzzy logic and self organized maps (Petricoin & Liotta 2002b). With the vast array of learning systems available, there is no ideal single method for the analysis of complex proteomic data. Once the data mining technique has been selected, feature selection and data transformation must be performed, where appropriate, to obtain the best generalised solution to the question being asked. Dimensionality is a result of the asymmetry between the number of inputs (for example, peaks) and the number of samples. Any classification algorithm that is given too many features will be able to find a solution to the problem that has been presented but these may not be representative biomarkers. A given mass spectrum can have tens of thousands of data points per sample used and so a process of feature selection must be undertaken before data mining can begin.

One method is to select the m/z values that represent peaks, peak detection. Other methods include use of t-tests, unsupervised learning approaches or supervised learning approaches that have feature selection incorporated into them. After selecting the features, data transformation is often performed where the values of the features are constrained within a defined range reducing the impact that a high variance of a given input feature may have on altering the distribution of the original data. Data transformation can be achieved by log transformation, square root transformation and linear and logarithmic scaling, but not all data mining approaches require data transformation (Fung *et al.* 2005). Below is a brief explanation of a few of the data mining techniques that can be used for the analysis of proteomics data.

1.8.1 Artificial Neural Networks (ANNs)

Artificial neural networks (ANNs) are an example of a computer-based algorithm used in this study to analyse complex proteomic data to identify key prognostic biomarkers. They are based on biological neurons and so can organise and process information allowing them to handle data which contains high levels of noise and redundancy (Ball *et al.* 2002). ANNs are non-linear which allow better fit to the data, are robust, have the ability to learn and adapt, allow the system to modify its internal structures, and have the capability to generalise, which enables application of the model to unseen data. ANNs have been used in a variety of applications including modelling, classification, pattern recognition, and multivariate analysis and are made up of an input layer, a hidden layer and an output layer with a series of weighted links connecting each layer, see figure 1.8.

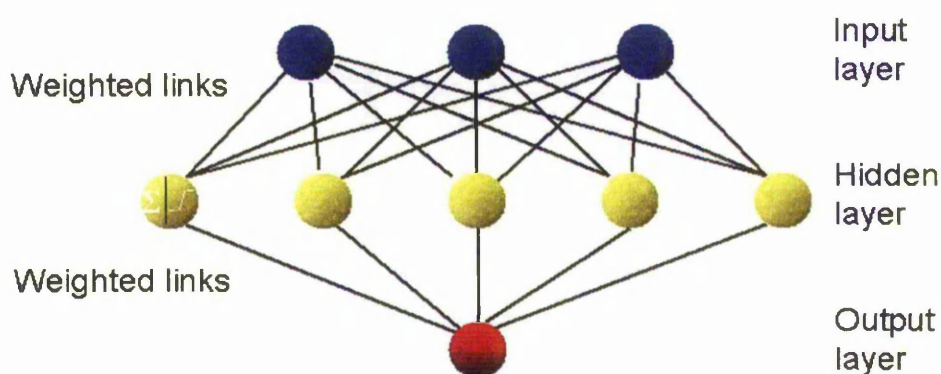


Figure 1-8 Schematic showing the basic layout of a neural network.

A basic neural network contains one input layer, which represent independent variables used for training such as mass and intensity data from a mass spectral profile; 1 hidden layer, which contains the mathematical workings of the model; and 1 output layer which is calculated by the network and can then be compared to the actual output. The weighted links are adjusted in proportion to the difference between the actual output and the predicted output, the training algorithm is applied here (Ball *et al.* 2002).

The ANNs are trained using the input layer (representing independent variables), which could represent the m/z value and intensity from a set of mass spectral profile data. The hidden layer represents the mathematical workings of the model and does not interact with the external environment but merely processes the information received from the input layer and passes it to the output layer via weighted links. The output layer is calculated by the network based on the data inputted, and is then compared with the actual known output. For example, in a classification model an output of 1 would represent a control sample and an output of 2 would represent a cancer sample, and then the error between the predicted and actual outputs is calculated. The training algorithm is applied to the weighted links which are adjusted in proportion to the difference between the actual output and the predicted output. When using the back-propagation algorithm, the weights are updated in proportion to the error calculated and are governed by the learning rate and momentum parameters. Other algorithms that could be used are conjugate gradient descent, Levenberg-Marquardt, quick propagation and Delta-Bar-Delta (StatSoft 2006). The learning process continues repeatedly and attempts to obtain a set of weights that correspond to a predetermined parameter, for example a minimum target error or the failure of the network to improve after a certain amount of training cycles (epochs).

Data inputted for neural network analysis can be split into training and test data where the ANNs train on the training set and validates its outputs on the test set whilst the training occurs. After the training has been completed it is validated by a further set of cases that are completely blind to the model, they are applied and an output is calculated based upon the new data (Basheer *et al.* 2000; Lancashire *et al.* 2005). The performance on unseen data indicates whether a generalised model has been obtained or not. By using multiple models with different unseen datasets (random sample cross validation) a more generalised model may be obtained with only a small sample number. Random sample cross validation randomly selects different sample sets into training and test sets for every model run. All of the inputs can then be treated as unseen data and by using different data splits, different models can be created allowing confidence intervals to be calculated and outliers to be identified. Parameterisation is a process that allows the identification of the importance of all of the inputs to the model and this can then lead to the elimination of inputs that are of little or no importance thereby reducing the complexity of the data and increasing the predictive performance. Parameterisation can be achieved by performing a weightings analysis, sensitivity analysis or by stepwise approaches. As already described, the ANNs work by weighting the links from the inputs to the outputs; the stronger the weight leading from a particular output, the greater the influence it has on the model. A sensitivity analysis can also be applied to the trained ANN model where the error in the performance is monitored when an input is removed from the system; the greater the change in the error after removal of the input, the more influential that particular input is. The predictive performance when the input is present is compared to the predictive performance when the input has been removed and this gives a sensitivity ratio. Therefore, a sensitivity ratio of 1 or close to 1 indicates there is no influence of that particular input on the system, a value greater than 1 indicates that the input has a positive effect on the predictive performance as its loss causes degradation in the model indicating its importance.

A stepwise approach is a detailed method of parameterisation whereby the stepwise addition of inputs is undertaken allowing the determination of the best subset of ions to predict a particular outcome. Interactions between the ions may also be identified in this way. The analysis is based on the hypothesis that the change in performance when an input is added indicates its influence on the model. Multiple sub models are run with input 1,

then with input 2, and then with input 3 and so on until all of the inputs are modelled separately. The error is determined for each sub model and the input that gives the lowest error is selected and put with all of the remaining inputs sequentially in a number of sub models once more. This process continues until there is no improvement in the error and so a subset of ions is finally identified.

The applications of ANNs are divided into two main classes; supervised and unsupervised methods. For supervised learning, a dataset consisting of both input and output data is presented to the ANN during a training phase. The ANN tries to find a link between the two with the least error, by adjusting the weighted links until the error falls below a certain threshold. Once the ANN has established a connection between the inputs and outputs, the model can then be used for unseen data. When using unsupervised learning, only an input dataset is presented and the ANN is free to search for hidden relationships amongst these data (Tafeit *et al.* 1999). Recent studies have shown that the application of ANN-based approaches can be used to identify patterns strongly associated with specific disease stages (Ball *et al.* 2002; Petricoin *et al.* 2002a).

1.8.2 Support Vector Machines (SVM)

Support Vector Machines (SVMs) were invented by Vladimir Vapnik (1995) and comprise a set of supervised learning methods used for classification and regression. For the classification of samples, SVMs map a hyperplane in a multidimensional space of training inputs in an attempt to split the cases and the controls (Vapnik 1995). The simplest form of SVM plots a linear boundary between the two datasets but real-world problems are not usually this straightforward. More sophisticated SVMs can be used where polynomial or sigmoidal boundaries between the data points can be found but this has to be carefully controlled to ensure that a generalised model is found (Webb 2002). Recently, it has been shown that SVMs can outperform neural networks in learning to recognize patterns from a dataset with a low number of training inputs, in some instances. SVMs have been successfully applied in the analysis of proteomic data in a number of studies (Li *et al.* 2002a; Wu *et al.* 2003a; Zhang *et al.* 2004b).

1.8.3 Decision Trees

Decision trees use continual portioning of samples into sub-groups in order to achieve classification. They begin with the entire sample set, for example m/z values and intensities of mass spectral peaks, and create a rule that causes the sample set to divide into two groups. The decision rule looks at the input features and this leads to the creation of a second rule that states that if the intensity of a peak is less than x then a sample is partitioned to the left branch, if it is not then it is partitioned to the right. Each branch is studied in this way and further sub divided by creating another rule until the training is complete. New samples are classified by following all of the rules and observing which terminal node the sample ends up in (Fung *et al.* 2005). One advantage of this method is that it can create easily interpretable rules for the classification of samples and have therefore been used in the analysis of mass spectral data. Adam *et al.* (2002) used this method to distinguish between serum taken from prostate cancer patients from serum taken from those with benign prostate hyperplasia and healthy men. By using the intensity levels of the nine highest discriminatory peaks as features, classification was achieved with 96% accuracy (Adam *et al.* 2002). One disadvantage of this method, however, is that as the complexity of the tree increases there are fewer and fewer samples within each group but modifications such as bagging and boosting may help to overcome this problem (Fung *et al.* 2005).

1.8.4 Principal Components Analysis (PCA)

Principal Components Analysis (PCA) is a commonly used unsupervised learning technique that can reduce the number of variables and detect structure in the relationships between the variables, thereby having the ability to classify them (StatSoft 2006). It reorganizes information in a sample data set mathematically and is very useful when there are large numbers of variables. It maps highly-dimensional data into more manageable sets of dimensions by creating principal components, which accounts for the majority of the variability in the data, enabling the data to be explained with far fewer variables than were present in the original dataset. The first principal component is placed through the data in the direction that explains the most variability. The second principal component is then placed at right angles to this to describe the maximum amount of the remaining variability;

subsequent principal components are placed at right angles to the principal component before it. Once the directions of the principal components are plotted, the values of the individual samples can be expressed as linear summations of the original data multiplied by the coefficient that best describes the principal components, these new values are known as Eigenvalues and each sample will have a score for each principal component (Davies *et al.* 2005; Fung *et al.* 2005).

1.8.5 Cluster Analysis

Cluster analysis encompasses a number of different algorithms and methods for grouping objects into categories. It is an exploratory tool which aims at sorting different objects into groups in a way that if two objects belong to the same group then they must be highly associated and if they are not placed in the same group then the reverse is true. It can be used to discover structures in data without providing an explanation/interpretation as to why they exist (StatSoft 2006). Clustering is done hierarchically to form a dendrogram, or relational tree, where the leaves, or nodes, of the tree represent the individual objects, and the branches of the tree group the nodes according to the similarity or difference of the objects and this is represented by the branch length. Cluster analysis is used extensively in the mining of data generated by cDNA microarray experiments but its application to proteomic studies is still fairly limited (Harris *et al.* 2002). Schmidt *et al.* (2003) clustered peak lists extracted from the mass spectra of spots on 2-DE gels. They used clustering to purify peak lists by removing peaks that came from neighboring spots and so improving upon protein identification (Schmidt *et al.* 2003). Tibshirani *et al.* (2004) used the clustering of peaks across many mass spectra in order to classify samples from patients according to disease status from protein MS data and Beer *et al.* (2004) used clustering of LC-MS/MS spectra to reduce the large amounts of data generated into a more manageable size (Tibshirani *et al.* 2004; Beer *et al.* 2004).

1.9 Cancer Therapy of the Future – Personalised Medicine

One major goal in cancer research is the development of diagnostics and therapeutics that will decrease morbidity and mortality of cancer patients within the next decade. The aim is to identify diagnostic biomarkers that can detect disease at earlier stages with high specificity, and to develop therapies that effectively reduce the tumour burden in patients without inducing any side effects (Conrads *et al.* 2005). As cancer is known to be a heterogenous disease it is no surprise that cancers vary genetically and phenotypically between patients that may have an identical type and stage of cancer and so explains the unpredictable responses of patients to existing cancer therapies (Jain 2004). Advances in genetics and proteomics have led to the concept of personalised medicine for improvement in the healthcare of patients. This type of medicine involves the direction of specific treatments that are best suited to the individual patient's genotype and various other factors that can influence the outcome of the disease and patient's response to the treatment (Jain 2002). There are a number of ways proteomics can be of use in the personalised management of cancer; it can help in the progress in understanding the pathophysiology of cancer thereby helping to search for safer and more effective treatments; global changes in the protein expression patterns due to transcriptional and post-transcriptional control and post-translational modifications can be monitored; the molecular diagnosis of cancer can be improved; glycoproteins that are found on the cell membrane of cancer cells, distinguishing them from normal cells, can be identified aiding in the targeted killing of cancer cells; the combination of proteomic approaches with biochemical assays for the identification of target antigens could lead to antibody-based therapies for cancer; and finally, proteomics can also be of use in the identification of biomarkers that will improve the diagnosis, prognosis and therapy of patients (Jain 2004).

The use of gene arrays for the molecular profiling of diseases has shown potential in classifying patients according to their disease stage or survival (Golub *et al.* 1999). Microarray-based profiling techniques have also provided an opportunity investigate the relationship between thousands of genes and the clinical phenotypes of the samples. Several groups have used this approach and have been able to identify biologically and prognostically distinct tumour subgroups in a number of cancers (Agarwal *et al.* 2006). In a study of early breast cancer, a signature associated with metastasis-free survival has been

identified, validated in an independent dataset, and is currently the subject of a multi-centre randomized clinical trial to establish its use in selecting patients for adjuvant chemotherapy (van 't Veer *et al.* 2002; van de Vijver *et al.* 2002). Due to the success of these gene profiling studies it is likely that protein profiling studies would also give information on the relationship between proteins and clinical phenotypes of samples. Protein profiling studies also have an advantage over gene profiling studies as these only provide information at the genetic level and do not necessarily reflect protein expression or modifications that occur during transcription and translation. Also, the expression of certain genes does not provide details of protein-protein interactions and so proteomic analysis should be included with molecular profiling when selecting an appropriate treatment regimen for a patient (Petricoin *et al.* 2002c). The integration of multiple disciplines such as genetics, proteomics, transcriptomics and metabolomics can be used to develop truly personalized therapies by directing the selection of appropriate drugs or drug combinations to achieve increased benefits to patients and to reduce toxicity and to also predict the response of an individual patient to a particular therapy.

1.10 Aims and Objectives

The identification of patterns within proteomic profiles of patient samples to predict diagnosis and prognosis of patients and identify suitable therapeutic regimens is the highlight of one facet of research into personalised medicine. The use of SELDI MS and bioinformatics to provide this patient information has been widely researched in the fields of ovarian, prostate and breast cancer, but as yet these methods have still not been validated for their clinical use.

The aim of this research is to apply the principles of SELDI MS and artificial neural networks as tools to identify patterns that correlate to certain clinical outcomes in both melanoma and breast cancer. Also by identifying patterns, the identification of possible biomarkers that can reflect these clinical outcomes will be investigated. As research into personalised medicine seeks to identify patterns that can indicate certain clinical outcomes, patient material is usually utilised, but this is sometimes limited and difficult to obtain in sufficient quantity and numbers to make the findings statistically significant. The first part of this study aims to identify whether melanoma cell lines can be used for proteomic studies to identify biological traits correlating with clinical information. This will be useful for biomarker identification studies, where large amounts of material may be required but cannot be obtained. In parallel, tissue samples taken from melanoma patients with known clinical history will be investigated in order to determine whether the proteomic patterns relate to clinical outcome in melanoma and breast cancer.

The majority of SELDI MS studies to date have utilised patient serum to investigate defined clinical outcomes, however research in melanoma has been limited. This aspect of the research will investigate 205 melanoma serum samples to determine if stage of melanoma can be predicted from the patient serum proteome. Proteomic profiles will also be correlated with progression of patients from stage III to stage IV. Disease progression has not been widely investigated using pattern recognition and the results of such studies may be important for clinical management.

The final part of this study will compare SELDI MS with MALDI MS in order to determine if a better predictive capability can be determined using a MS method with improved sensitivity and resolution. This will be conducted using melanoma cell line lysates which will be subjected to SELDI MS analysis using C₁₆ (H4) chips and to MALDI

MS analysis after C₁₈ ZipTip sample clean up. The spectra obtained from the SELDI MS analysis will then be analysed using ANNs to determine if patterns can be found within the data that can classify blind samples according to their genetic mutation. The MALDI spectra will be analysed using the same ANN parameters and the results of the analysis with the SELDI and MALDI methods compared.

1.11 Collaborations

The results of the present study form part of a 3 year 5th framework European Union project, OISTER (Outcome and Impact of Specific Treatment in European Research in melanoma). This study involved 7 partners across Europe and was designed to collect tissue, serum samples and cultured cell lines from melanoma patients with known clinical outcome, for analysis of gene expression patterns, expression of proteins, HLA type and other immunological markers. Clinical information on those patients was available in order to correlate the scientific findings to specific criteria such as stage of disease or response to treatment, thereby allowing clinically significant biomarkers to be identified.

Chapter 2 Methodology

2.1 Materials

2.1.1 Reagents and list of suppliers

All of the reagents were stored as per the manufacturer's instructions and used before the expiry date.

<u>Sample Preparation reagents</u>	<u>Company</u>
Acrylamide/Bis	National Diagnostics
BSA	Sigma
BSA standard	Bio-Rad
Coomasie Blue	Sigma
Dithiothreitol	Apollo Scientific
Ethanol	Sigma
Glycine	ICN
Methanol	Sigma
Octyl glucopyranoside	Apollo Scientific
PBS	Bio Whittaker Europe
Protein Assay Dye Reagent	Bio-Rad
Sodium Azide	Sigma
Sodium Dodecyl Sulphate	Sigma
Sodium Tetraborate	Sigma
Trizma Base	Sigma
Urea	Sigma
Water	Sigma
<u>Tissue culture reagents</u>	<u>Company</u>
DMEM	Bio Whittaker Europe
DMSO	Acros
Ethanol	BDH
Foetal calf serum	Bio Whittaker Europe

Glutamine	Cambrex
RPMI	Bio Whittaker Europe
Trypan blue	Sigma
Trypsin	Gibco

Proteomics Reagents**Company**

Acetonitrile	Fisher Scientific
Dichloromethane	Fisher Scientific
Hexane	Fisher Scientific
Propanaol	Fisher Scientific
Protein calibrant Mix 2	Laser Bio Labs
Sinapinic Acid (3,5 Dimethoxy 4-Hydroxycinnamic Acid)	Sigma
Trifluoroacetic acid	Fisher Scientific

2.1.2 Equipment*2.1.2.1 Cell lines and Media*

All of the cell lines used were adherent cell lines. These were obtained from the ESTDAB tumour cell line bank, <http://www.ebi.ac.uk/ipd/estdab/> and therefore were of unknown origin, or were established in the laboratory at DKFZ, Mannheim from patients' metastases before the frozen cells were sent to Nottingham Trent University. A comprehensive list of the cell lines used and their origin can be found in appendix I.

All of the cell lines used were cultured in RPMI + 10% FCS + 2mM L-glutamine.

2.1.2.2 Tumour Tissue

All of the melanoma tumour tissue used in this study was obtained from Professor Dr Dirk Schadendorf, DKFZ, Mannheim, Germany, as part of the 5th framework OISTER project. A comprehensive list of the samples used can be found in appendix II but briefly, 62 melanoma tissue samples from stage III and IV patients were analysed.

All of the breast cancer tumour tissue used in this study was obtained from Professor Ian Ellis, City Hospital, Nottingham. A comprehensive list of the samples used can be found in appendix III but briefly, 165 breast cancer tissue samples from patients of all stages of the disease were analysed.

2.1.2.3 Serum

A list of the serum samples used in this study can be found in appendix IV. The samples were taken from patients attending the German Cancer Research Centre (DKFZ) in Mannheim, Germany; n=101 stage I and n=104 stage IV melanoma sera as well as n=28 progressor and n=28 non-progressor stage III sera.

2.1.2.4 Buffers

Buffers were prepared as indicated below:

<u>Name</u>	<u>Composition</u>
PBS Freshly prepared each day	1 tablet dissolved into 100ml dH ₂ O (Oxoid)
Lysis buffer Aliquoted and stored at -80°C	9.5M Urea 2% w/v DTT 1% w/v OGP
Sample reducing buffer Stored at room temperature	2.5ml 0.5M Tris HCl buffer (pH 6.8) 400mg SDS 2ml Glycerol 200mg DTT A few grains of bromophenol blue make up to 20ml with option 4 H ₂ O
1.5M Tris HCl buffer (pH 8.8) (resolving gel buffer) (1-D PAGE) Stored at room temperature	18.16g Trizma base 0.4g SDS make up to 100ml with option 4 H ₂ O adjust pH to 8.8 with HCl
0.5M Tris HCl buffer (pH 6.8) (stacking gel buffer) (1-D PAGE) Stored at room temperature	6g Trizma base 0.4g SDS make up to 100ml with option 4 H ₂ O adjust pH to 6.8 with HCl
10x Running Buffer (1-D PAGE) Stored at room temperature	0.25M Trizma base (Sigma) 2M glycine (Fischer Scientific) 1% (w/v) SDS (Acros)

2.1.2.5 Consumables

All of the plastic ware used in this study were obtained from Sarstedt except for:

<u>Plastic Ware</u>	<u>Company</u>
10ml syringe	BD

<u>Miscellaneous Items</u>	<u>Company</u>
1ml cryovials	TPP
300µl glass vials	Chromacol
Glass slides	Menzel GmbH
H4 24 spot SELDI chips	Ciphergen®
H50 8/16 spot SELDI chips	Ciphergen®
384 MALDI target plate	Shimadzu
64 spot MALDI microscope slides	Shimadzu

2.1.2.6 Hardware

<u>Hardware</u>	<u>Model, Company</u>
Cryostore	Cryo 200, Forma Scientific
-80°C freezer	Ultima II, Revco U570 Premium, New Brunswick Scientific
Microcentrifuge	Microcentaur, MSE Mikro 22R, Hettich Zentrifugen
Centrifuge	Mistral 1000, MSE Mistral 2000R, MSE Falcon 6/300, MSE
Vortex	Whirlimixer, Fisher Brand
Safety cabinet	Microflow biological safety cabinet, Walker
Incubators	CO ₂ water jacketed incubator, Forma Scientific
Microscope and Camera	Model PIM, World Precision Instruments
Cryostat	CM 1900, Leica

Water bath	Y14, Grant
Sonicators	Precision Ultrasonic Cleaning, Ultawave Ultrasonic Cleaner, VWR
Electrophoresis gel tank	GeneFlow
Power supply for electrophoresis	Consort E122, GeneFlow
Absorbance plate reader	Model 680, Biorad
Mass Spectrometers	PBS II analyser, CIPHERGEN® Axima CFR+, Kratos
Robotic MALDI plate spotter	XCISE, Proteome Systems/Shimadzu

2.1.2.7 Software

<u>Software Product</u>	<u>Company</u>
CIPHERGEN® ProteinChip Software v3.2.0	CIPHERGEN®
Kompact Launchpad v2.4.1	Kratos Analytical Ltd
Statistica v6.1	StatSoft Inc

2.1.2.8 Company Addresses

<u>Company</u>	<u>Address</u>
Acros	Loughborough, UK
BD	Cowley, UK
BDH	Leicester, UK
Beckman Coulter	High Wycombe, Bucks, UK
Bio Whittaker Europe	Wokingham, UK
Biorad	Hemel Hempstead, UK
Cambrex	Nottingham, UK
Chromacol	Welwyn Garden City, Herts, UK
CIPHERGEN®	Guildford, Surrey, UK
Elkay	Basingstoke, UK
Fischer Scientific	Loughborough, UK
Forma Scientific (Thermo)	Basingstoke, UK

GeneFlow	Fradley, Staffs, UK
Gibco	Paisley, UK
Greiner Bio-One	Gloucestershire, UK
Hettich Zentrifugen	Tuttlingen, Germany
Kratos Analytical Ltd	Manchester, UK
Laser Bio Labs	Cedex, France
Leica	Milton Keynes, UK
Menzel GmBH	Braunschweig, Germany
MSE	London, UK
New Brunswick Scientific	Edison, New Jersey, USA
Proteome Systems	North Ryde, New South Wales, Australia
Revco	Asheville, North Carolina, USA
Sarstedt	Leicester, Leicestershire, UK
Scientific Laboratory Supplies (SLS)	Nottingham, Notts, UK
Shimadzu	Milton Keynes, UK
Sigma	Gillingham, Surrey, UK
StatSoft Inc	Bedford, UK
Stratec	Birkenfeld, Germany
TPP	Switzerland
Ultrawave	Cardiff, UK
VWR	Poole, Dorset, UK
Walker	Massachusetts, USA
Ward Systems Group, Inc	Frederick, Maryland, USA
World Precision Instruments	Stevenage, Herts, UK

2.2 *Methods*

2.2.1 **Sample Preparation**

2.2.1.1 *Melanoma cell lines*

The cell lines used in this study and related information can be found in appendix I.

One vial of each cell line was thawed, washed and seeded into a T25 flask that had been incubated with 10ml of complete media for at least 1 hour. Once confluent the cells were seeded into 2 x T75 flasks and placed into an incubator which was maintained at 37°C in a humidified atmosphere of 5% CO₂. Once the cells were 70% confluent, photographs were taken for entry into the ESTDAB database (<http://www.ebi.ac.uk/ipd/estdab/>) before the cells were scraped from the flasks and washed in PBS. An 100µl of cell lysis buffer was added to the cell pellet, giving two lysates for each cell line; each lysate was then aliquoted into three vials and stored at -80°C for protein quantification and mass spectrometric analysis.

2.2.1.2 *Melanoma Tissue*

The tissue samples used and related information can be found appendix II.

Samples were snap frozen in liquid nitrogen within 30 minutes of surgical resection and stored at -80°C.

The tissues were then sectioned using a Leica 1900 cryostat to allow the percentage cellularity of the samples to be determined whilst conserving the melanoma tissues. The specimen holder was cooled down to -12°C and the cryochamber cooled down to -25°C prior to use. A piece of melanoma tissue was fixed directly onto the metal chuck using a small drop of freezing medium (OCT) and freezing spray. This was left for a few of minutes in the cryochamber to ensure the tissue was properly fixed and then the chuck was placed onto the specimen holder. The blade and the anti-roll plate were wiped with a small amount of acetone to prevent static and the specimen holder was set to cut 9µm sections. Twenty to thirty sections were cut from each piece of tissue and placed onto a glass slide. The sections were then scraped off using a blade before being placed into 50µl of lysis buffer where the samples were pipetted up and down to shear the membranes, releasing the proteins and allowing them to solubilise in the buffer.

2.2.1.3 Breast Cancer Tissue preparation

The tissue used in this study along with the relevant clinical information can be found in appendix III.

The tissue samples were collected at Nottingham City Hospital and snap frozen in liquid nitrogen within 30 minutes of surgical resection prior to storage at -80°C.

The pieces of tissue were solubilised in 100µL of cell lysis buffer and macerated by pipetting up and down. 20µL of the protein lysate was aliquoted and stored at -80°C for SELDI analysis and the remaining 80µL was aliquoted and stored at -80°C for protein assay and 1-D PAGE.

2.2.1.4 Serum preparation

The serum samples used in this study along with their relevant clinical information can be found in appendix IV.

The serum was collected and stored at DKFZ, Mannheim, Germany.

Serum samples were selected from a frozen collection of sera from patients with histologically confirmed melanoma. All patient's gave informed consent and the study was approved by the IRB. The serum samples were processed using the following standardized protocol: blood was drawn from the patients' cubital vein into gel coated serum tubes and allowed to clot at room temperature for at least 30 min, but no longer than 60 min. Thereafter, the tubes were centrifuged at 2500 g for 10 min. The serum phase was harvested and subsequently frozen without any additives in 1 ml aliquots at -20 °C and then transferred to -80 °C, and not thawed until immediately prior to analysis.

2.2.2 Protein Microassay

A Bio-Rad protein microassay was used to determine the protein concentration of all of the samples used in this study, aside from the serum samples. This assay proved to be the most appropriate to use as most other kits available were not compatible with the 9.5M urea present in the lysis buffer.

A BSA protein standard was made up to concentrations of 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20µg in water to a total volume of 800µl. Two eppendorf tubes were set up for every

sample with 795µl of water and 5µl of sample, 5µl of cell lysis buffer and 795 µl of water was used as the negative control. 200µl of undiluted Bio-Rad protein assay dye concentrate was added to each eppendorf and left to incubate for 15 minutes at room temperature. 100µl of sample from each eppendorf was then aliquoted into a 96 well plate and the absorbance determined using a spectrophotometer at 595nm.

2.2.3 1-Dimensional Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis (1-D SDS PAGE)

All of the samples, apart from the serum samples, were resolved by 12.5% 1-D SDS PAGE to determine the sample integrity prior to mass spectrometric analysis.

A 30 ml of a 12.5% resolving gel was made up by placing 12.5 ml of 30% Acrylamide stock, 7.8ml 1.5M Tris HCl buffer (pH 8.8), 9.5 ml water, 150µl ammonium persulphate (100mg in 1ml water) and 15µl TEMED, this was enough for 4 gels. 6ml of this mix was then placed in each of the four protein gel casts and 100µl of butanol was placed on top to keep the gel even and to stop dehydration while it polymerised. Once polymerised, the butanol was removed using filter paper and the appropriate combs were put on top of the casts. A 4% stacking gel was made up by adding 2 ml of 30% Acrylamide, 2.6 ml of 0.5M Tris HCl buffer (pH 6.8), 5.3 ml water, 70µl ammonium persulphate and 7µl TEMED and then 2ml of this mix was placed on top of the resolving gel in each cast. Once polymerised, the combs were removed and the gels and casts were placed in to a tank containing 1 x running buffer.

The samples were prepared by mixing 40µg of sample with approximately 5µl of reducing sample buffer and denaturing the proteins for 5 minutes at 95°C.

The gels were completely covered in running buffer in the tanks and the samples were loaded into the wells in the gels, 5µl of a ProtoMetrics protein ladder was added to one well in each gel. Once loaded, the gels were run at 100V until the dye front reached the bottom of the stacking gel, approximately 30-40 minutes, the voltage was then increased to 150V until the dye front reached 1cm from the bottom of the gel, approximately 1 hour. The gels were stained with coomassie blue for 30-60 minutes and the excess stain was removed by placing the gels in destain (10% [v/v] acetic acid and 30% [v/v] methanol in HPLC grade water) over night.

2.2.4 SELDI MS

All of the samples were subjected to SELDI MS analysis in order to obtain protein fingerprints.

The spots of a 16 spot reverse phase C₆₋₁₂ coated H50 protein chip were wetted with 2µl of 20% [v/v] ACN and incubated for 2 minutes at room temperature in a humidity chamber. The samples were then placed onto the spots and incubated for 15 minutes at room temperature in a humidity chamber to allow proteins to bind to the chip surface. The spots were washed 5 times with 2µl of HPLC grade water and 0.8µl of SPA (saturated solution in 50% ACN + 0.5% TFA) was then added to each spot. The ProteinChip reader, a PBS II mass analyser from Ciphergen®, was calibrated using the singly and doubly charged peaks of known calibrants made up to 0.01nmol in 100µl of SPA. 0.5µl of this was added to each spot and enabled manual calibration of the mass spectrometer.

The SELDI MS was run in the positive ion mode with the time lag focusing switched on. The high mass was set at 25,000 Da with an optimization range of 1500 to 20,000 Da.

The pulse voltage was set at 3000V with a pulse lag time of 773ns and the detector voltage was at 1850V. The laser was fired 7 times at each position within a spot starting at position 20 and ending at position 80. The first two shots were warming shots, to ablate some of the matrix, and these were fired at a laser intensity of 155 and a detector sensitivity of 6, this data was not included in the final spectra. The final five shots were taken using a laser intensity of 270 and detector sensitivity of 6 and these shots were included into the spectra. The raw spectral data was exported as a .csv file and converted into a .xls file.

2.2.4.1 SELDI MS of melanoma cell lines and melanoma and breast cancer tissue

The SELDI MS parameters used were as described above in section 2.2.4 but with the following changes.

The cell line and tissue lysates were diluted down to 1µg/µl in cell lysis buffer 2µl of which was added to each spot on the chip. The calibrant used was a mix containing insulin B chain (3494.65 Da), bovine Cytochrome C (12362 Da), Apomyoglobin (16952.3) and Trysinogen (23981.1). After calibration the mass accuracy was 0.06% in automated mode.

2.2.4.2 SELDI MS of melanoma serum samples

The SELDI MS parameters used were as described above in section 2.2.4 but with the following changes. 1 μ l of neat serum was placed onto the spots of a 24 spot reverse phase C₁₆ coated H4 chips (CIPHERGEN®). The SELDI instrument was calibrated with the singly and doubly charged peaks of bovine superoxide dismutase (SOD) (15591.4 Da) and bovine ubiquitin (8564.8 Da) made up at 0.01nmol in 100 μ l of SPA. After calibration the mass accuracy of the instrument was 0.05% in automated mode.

2.2.4.3 SELDI MS of melanoma cell line samples analysed for the SELDI vs MALDI study

The method used for preparing the samples is as described in section 2.2.4.1 except that 24 spot reverse phase C₁₆ coated H4 chips (CIPHERGEN®) were used.

The reproducibility of the SELDI MS methods used is shown visually below where 18 aliquots of normal serum were spotted onto the surface of a H4 chip. The coefficients of variation (CV's) for the 5 most prominent peaks was calculated and shown in table 2-1. Reproducibility of SELDI MS using H50 chips is shown in appendix 5 where the CV's for the peak location and peak intensity of 5 calibrant peaks from a calibrant mix which was bound to 12 spots is calculated.

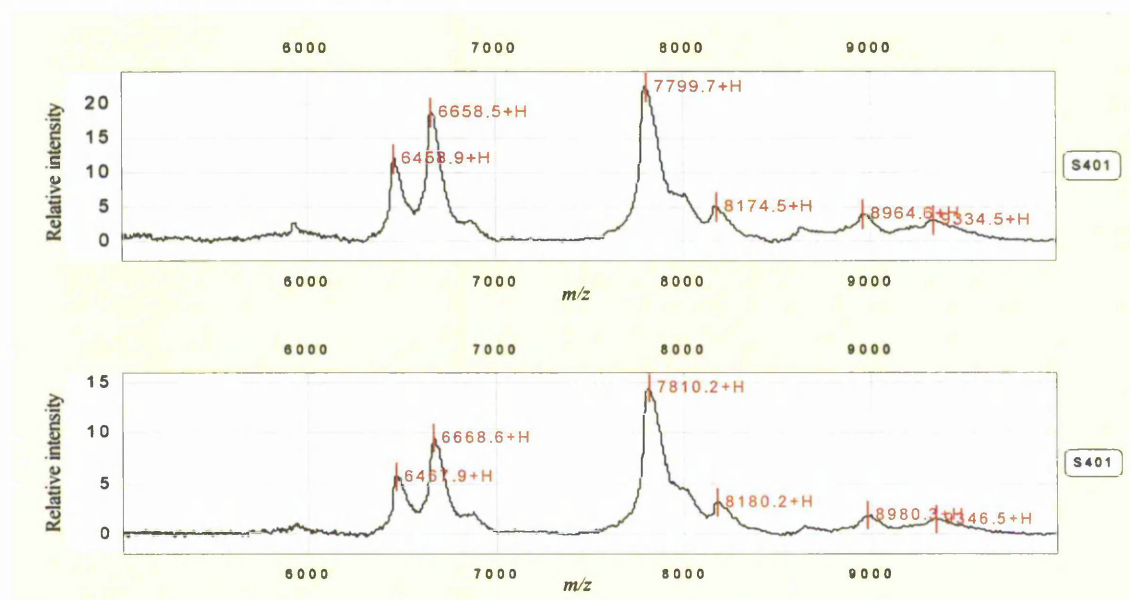


Figure 2-1 Representation of reproducible SELDI spectra obtained from the aliquots of normal serum. The spectra represent the reproducibility, in the 5-10 kDa mass range, of the SELDI technique, where the peak pattern remains consistent when a sample is spotted onto multiple spots on a H4 SELDI chip.

2.2.5 MALDI MS

The cell line lysates were prepared by diluting to $1\mu\text{g}/\mu\text{l}$ with 0.1 % trifluoroacetic acid (TFA) to give a total volume of $30\mu\text{l}$ and cleaned up using a ZipTip method prior to MS analysis.

2.2.5.1 ZipTip Method

The diluted cell line lysates were prepared by C_{18} ZipTip clean-up using a robotics machine (XCISE, Proteome Systems, Shimadzu) to ensure consistency of sample processing. The samples were bound to the ZipTip with 25 cycles of binding, followed by two washes in 0.1% TFA, where the washes were discarded. The samples were eluted off of the ZipTip in $4\mu\text{l}$ of 80% ACN + 0.1% TFA. $1\mu\text{l}$ of the sample was then spotted onto the MALDI target followed by $1\mu\text{l}$ of SPA matrix (10 mg/ml) and allowed to air dry. The samples were randomly spotted in duplicate on the MALDI target and then analysed by MALDI-TOF MS.

2.2.5.2 Mass Spectrometric Analysis

MALDI-TOF experiments were performed on an Axima CFR+ mass spectrometer (Shimadzu, Manchester, UK). Close external calibration, where every four spots were calibrated to 1 calibrant spot, was performed using protein calibration mix 2 (Laser Bio Labs). The protein calibrants were: Cytochrome C (horse heart) m/z 12361.12, Myoglobin, (horse) m/z 16181.06, Trypsinogen m/z 23981.98 and Insulin beta chain m/z 3494.65 ($3\mu\text{l}$ of 5mM). Mass spectral data from 1-25 kDa was collected in 'raster mode' using linear TOF-MS. The raw mass spectral data was then processed for bioinformatics analysis.

The reproducibility of the MALDI method used is shown visually below in figure 2-2, where 18 aliquots of normal serum ZipTipped using C_{18} ZipTips and then analysed by MALDI MS. The coefficients of variation (CV's) for the 5 most prominent peaks were calculated and are shown in table 2-1.

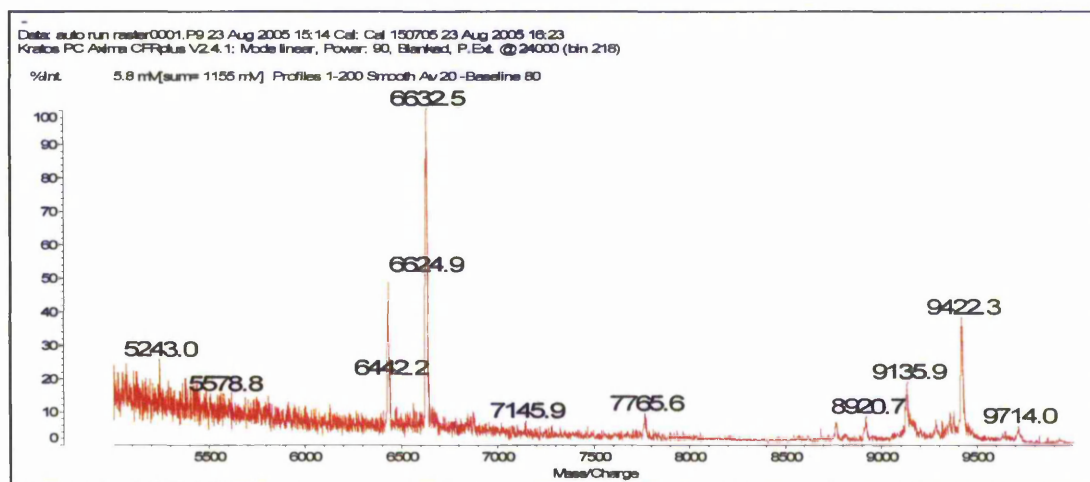
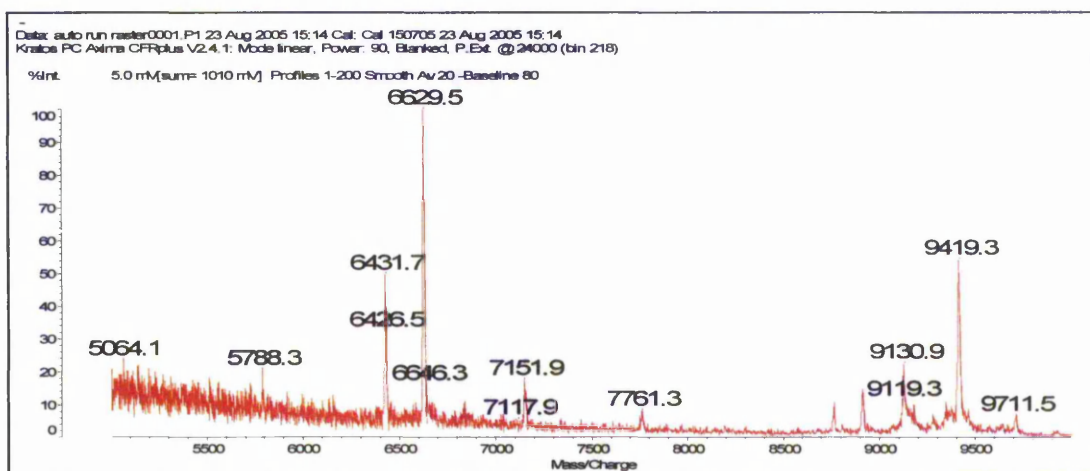


Figure 2-2 Representation of reproducible MALDI spectra obtained from aliquots of normal serum. Aliquots of normal human sera were cleaned up using C18 ZipTIPS and analysed using MALDI MS. Similar peak patterns shown for two of the spots represent the reproducibility of this technique.

Peak (<i>m/z</i>) (SELDI)	6468	6663	7805	8977	9339
Mass	0.13	0.10	0.10	0.11	0.09
Intensity Normalised to <i>m/z</i> 7805	18.2	18.4	-	20.6	14.5
Peak (<i>m/z</i>) (MALDI)	6433	6630	7765	8916	9421
Mass	0.09	0.09	0.07	0.07	0.07
Intensity Normalized to <i>m/z</i> 7765	29.3	25.7	-	47.7	42.5

Table 2-1 Table showing the coefficients of variation (%) for mass and intensities of most prominent peaks in normal serum analysed by SELDI and MALDI MS, n=18.

2.2.6 Bioinformatic analysis

Once all of the data was obtained by SELDI MS it was analysed by artificial neural networks (ANNs). The exported mass spectral data containing the mass and intensity information from 1-25 kDa for each sample was merged into 1 file in Excel. The intensity data was binned across the 1-25 kDa mass range where the median intensity value across a 3 Da range was determined to represent the intensity of a given mass at that mass value. The MALDI data was exported from the mass spectrometer as ASCII text files and then merged into 1 file in Excel. The intensity data was binned across the 1-25 kDa mass range where the median intensity value across a 1 Da range was determined to represent the intensity of a given mass at that mass value. The SELDI MS data was smoothed differently from the MALDI data due to the increased sensitivity and resolution of the MALDI instrument. Once the data had been merged and smoothed, the mass columns for every sample were deleted leaving the mass values in only the first column, before the samples were exported as a .txt file and imported into the Statistica software. The text file was transformed before being saved as a Statistica spreadsheet file, .sta, where the sample names were down the first column and the first row contained the smoothed *m/z* values, all of the remaining cells contained the relevant intensity values.

Once in this format the data was analysed by ANNs in 2 different ways. The first approach used the full data set to train the ANNs and then parameterization to identify the most important ions and the second approach, a step-wise approach, used each ion individually to train the ANNs.

A three layer multi-layer perceptron (MLP) ANN was used with a feed forward back propagation algorithm and a conjugate gradient descent algorithm. The data was scaled between 0 and 1 using minimums and maximums, where the raw values were scaled linearly so that the smallest value in the dataset is the minimum and the largest value is the maximum. This scaling method was used to ensure that all of the potential relationships amongst the variables were kept identical, therefore not introducing any bias into the data. The inputs used to train the ANNs represented the binned m/z values between the 1-25 kDa mass range together with their corresponding intensity values. Two hidden nodes were used in the hidden layer and the output later consisted of a single node, where one group of samples were coded as "1", and the other group of samples were coded as "2".

2.2.6.1 ANN analysis with parameterisation

The inputs into the ANNs consisted of 8000 variables specifying the intensity at given binned m/z values for every sample analysed by SELDI MS and 23001 variables specifying the intensity at given binned m/z values for every sample analysed by MALDI. The samples were split into training, test and blind data sets. The ANNs were trained using the training set, and the network error with regards to predictive performance was monitored with the test set, which was unseen during training. Once this error failed to improve for a pre-determined number of training events (epochs), training was terminated, and the model validated on the blind data set. The following parameters were set for all of the analyses performed in this study unless stated otherwise in each results chapter: A linear regression output function was used to map the output variables, where the sum-squared error function was used. The training was performed using a back-propagation (BP) algorithm until the error no longer improved for 10,000 epochs, followed by a conjugate gradient descent (CGD) algorithm for 5000 epochs at a learning rate of 0.1 and a momentum of 0.3. As well as setting the maximum number of epochs used for training it was also decided to specify the minimum level of error improvement over a given number of epochs. The 'window' was set to 2000 with the training and selection values left at 0, this means that if the

training or selection errors deteriorated over 2000 epochs then the training should stop. As the error fluctuates during training, sometimes increasing only to decrease soon afterwards, adjusting the 'window' variable can take this into account. This allows the training to stop when a clear trend of deterioration has set in and not stopping the training prematurely due an increase in error caused by random noise. Also weight decay regularisation was set so that overfitting could be avoided. Overfitting can occur when a network has a large number of weights in comparison to the number of training cases and achieves a low training error even though the underlying model in the data has not been found. An overfitted model has a high curvature as it tends to model noise as well as the underlying data. Regularisation penalises networks with a large curvature and so encourages the development of a smoother model. The models were trained using random sample cross validation, where the samples were randomly split into three groups for every model; 60% for the training set, 20% for the test set and 20% for the validation set. This whole process was repeated 100 times; so that each sample was treated as truly blind a number of times, enabling confidence intervals to be calculated for the network predictions on the blind data. The predicted outputs of the blind data for each sample over the 100 models were averaged and these values were plotted on a bar chart, with samples given an actual output value of 1 coloured in blue and samples with an actual output of 2 coloured in red. This population classification graph then allows the actual output of the samples to be easily compared to the output predicted by the ANNs. After the predicted outputs were averaged for each sample, the percentage of samples correctly classified, that is given a predicted output of less than 1.5 if its actual output is 1 and given a predicted output of greater than 1.5 if its actual output is 2, was calculated. The sensitivity (percentage of samples with outputs of 1 that were given predicted output values of less than 1.5) and specificity (percentage of samples with outputs of 2 that were given predicted output values of greater than 1.5) were also calculated.

Initially, all of the variables from the dataset were used as inputs in a network and trained over 100 randomly selected subsets. The network predictions and mean squared error values for these predictions were then calculated for each model with regards to the blind dataset set and an average prediction along with confidence intervals could be calculated over the 100 models trained to determine how well the ANNs could predict using the full dataset, this was illustrated in a population classification graph. After this, parameterisation was carried out, in the form of a sensitivity analysis, to identify the importance of inputs in

order to reduce the number of input variables so that the predictions could be improved upon. The sensitivity analysis removes each of the ions, in turn, from each of the models trained to determine if the error of that model increases or decreases with its loss. A ratio is calculated where the error once a particular ion has been removed is divided by the error with all of the input variables in. Therefore if an ion has a ratio value of less than one, it must have a negative influence on the training, if an ion has a ratio of greater than one it must have a positive influence; that is its removal causes the error to increase. Once the sensitivity ratios were calculated for each model, they were averaged over the 100 models and the top 1000 performing ions were selected as the input variables and the ANNs were retrained as before. The final results then show the average predictions and the average sensitivity ratio after using the top 1000 ions as the input variables to train 100 ANN models.

This approach is much quicker than the stepwise method and allows the user to determine if there are any predictive patterns in the proteomic data.

2.2.6.2 Step Wise Approach

The inputs into the ANNs consisted of 8000 variables specifying the intensity at given binned m/z values for every sample analysed by SELDI MS. The ANNs were trained with 1 hidden layer and 2 hidden nodes at a learning rate of 0.1 and a momentum of 0.5. The models were trained using random sample cross validation, as described above, using the training set, and the network error with regards to predictive performance was monitored with the test set, which was unseen during training. Once this error failed to improve for a pre-determined number of training events, training was terminated, and the model validated on the blind data set. This process was repeated 50 times; so that each sample was treated as truly blind a number of times, enabling confidence intervals to be calculated for the network predictions on blind data.

Initially, each variable from the dataset was used as an individual input in a network, thus creating 8000 individual models. These 8000 models were then trained over 50 randomly selected subsets and the network predictions and mean squared error values for these predictions were calculated for each model with regards to the separate validation set. The inputs were ranked in ascending order based on the mean squared error values for test data and the model input which performed with the lowest error was selected for inclusion into

the subsequent step. Thus, approximately 1 million models were trained and tested at each step of model development. Next, each of the remaining inputs were then sequentially added to the previous best input, creating $n-1$ models each containing two inputs. Training was repeated and performance evaluated. The model which showed the best capabilities to model the data was then selected and the process repeated, creating $n-2$ models each containing three inputs. This process was repeated until no significant improvement was gained from the addition of further inputs resulting in a final model containing the proteomic pattern which most accurately predicted between the two outcomes.

Although this approach is slower than using a sensitivity to select the best performing ions, it is possible to identify predictive patterns by using a small number of input variables, therefore allowing the user to suggest possible biomarker ions that could be used for diagnostic or prognostic purposes.

Chapter 3 Proteome profiling and bioinformatic analysis of melanoma cell lines carrying specific genetic mutations

3.1 Introduction

3.1.1 Combined genetic and proteomic analyses to identify the next generation of melanoma-specific markers

As previously described, cancer is caused by a complex set of genetic alterations occurring within a cell that leads to neoplastic transformation. Cancer cells continuously undergo genetic changes that cause the prolonged multistep sequence from early right through to the late clinical stages of the disease. The identification of genes that are involved in the development and progression of different cancers has been the focus of the majority of cancer related research to date, using techniques such as PCR, SEREX, Southern blotting, Northern blotting and DNA hybridization. More recently, the development of microarrays has provided a high-throughput expression profiling analysis with the aim of discovering genes and their pathways. With global analysis of gene expression it may be possible to identify previously unknown subtypes of cutaneous melanoma and predict phenotypic characteristics that may be important in identifying disease progression (Bittner *et al.* 2000). Molecular profiling of melanoma has lead to the identification of previously unknown candidate genes involved in melanoma progression, *Wnt-5A* and *B-raf* (Gray-Schopfer *et al.* 2005; Weeraratna 2005). A study comparing the gene expression profiles of melanocytes and melanomas has identified a number of novel genes and pathways that are up or down regulated in melanoma, including FGF13, a new autocrine factor, Twist, a clinically relevant prognostic marker and Necdin, a melanoma growth suppressor (Hoek *et al.* 2004).

As well as knowing which genes are involved in the progression of a cancer, it is also important to understand cellular behaviour and its response to external signals which will influence signalling pathways within a cell. Expression profiling can be used for the functional classification of dysregulated genes and a particular gene's involvement in a biological pathway can be determined. Understanding the biological pathways involved is

important to the development of reliable diagnostic markers and treatments (Nambiar *et al.* 2004).

The progression of melanoma can be associated with known genetic mutations which cause abnormal growth and dissemination of the cancer cells (Chin 2003). Genetic studies have revealed the existence of a number of subtypes of melanoma that differ in their clinical behaviour (Rodolfo *et al.* 2004). For example, acral lentiginous melanomas (ALM) show very early and multiple gene amplifications compared to superficial spreading melanomas (SSM) (Bastian *et al.* 2000). Cytogenetic changes in chromosomes 13q and 17p occur frequently in SSM and lentigo maligna (LM) and chromosomal imbalances such as gains in 1q, 6p and 8q have been shown in mucosal melanomas (Bastian *et al.* 2003; van Dijk *et al.* 2003). By combining both genomic and proteomic approaches, new and improved molecular diagnostics may be developed that are capable of classifying patients into subgroups based on the phenotype of the tumour or response of a patient to a particular type of treatment.

3.1.2 Melanoma-specific genetic mutations

The present study was undertaken to analyse cell line proteomes associated with the principal genetic characteristics (genes) associated with melanoma.

There are a number of genes that have been detected as having abnormal expression in melanoma including, *tp53*, *braf*, *n-ras*, *ctnnb1* (β -catenin), *c-myc*, *apc*, *wnt5a*, *pten*, *p16/ink4a*, *cdk4*, *p14/arf* and *cdkn2A*. Almost all of the genes that have been associated with melanoma so far encode products that play a role in one of the three main cellular processes; cell cycle regulation, DNA repair and receptor-mediated signal transduction (de Snoo *et al.* 2005). Those genes specific to this particular study are described below.

3.1.2.1 BRAF

The *BRAF* gene encodes a serine/threonine kinase which activates the ERK signalling pathway and is regulated by binding with the RAS protein. The ERK signalling pathway is involved in the regulation of gene expression, cytoskeletal rearrangements and co-ordination of responses to extracellular signals affecting proliferation, differentiation, senescence and apoptosis (Garnett *et al.* 2004). This pathway has been found to be

hyperactivated in approximately 30% of cancers (Hoshino *et al.* 1999) and *Braf* has been shown to be mutated in 7% of cancers. The V599E mutation of *Braf* lies within the kinase domain of the protein and mimics the phosphorylation of *Braf* inducing the activation of MEK and so causing the over activation of the RAS-ERK pathway (Mercer *et al.* 2003). The highest frequency of *Braf* mutations occurs in malignant melanoma where 70% of melanomas have an activating mutation in this gene. Over 40 different mis-sense mutations in the *Braf* gene have been identified across a number of different cancers, but one mutation in particular predominates. The substitution of a thymidine for an adenosine at nucleotide 1796 causes the conversion of the amino-acid valine (V599) for glutamine and this serves as a molecular marker for approximately 50% of melanomas (Davies *et al.* 2002). *Braf* mutations can be detected in the radial growth phase of melanoma so are an early event in melanoma development. They have also been associated with melanoma progression as a high frequency of mutations have been detected in invasive melanoma and also an association between *Braf* mutations and metastatic melanoma has been detected Shinozaki *et al.* 2004. The frequency of *Braf* mutations in melanoma varies according to the histological subtype and location of the primary tumour, for example in mucosal melanomas, which include sinonasal, laryngeal and anogenital melanomas, *Braf* mutations only occur in 6% of cases. In superficial spreading melanoma and nodular melanoma of the skin *Braf* mutations occur in 50% of cases, and in 11% of lentigo maligna melanoma cases (Sasaki *et al.* 2004). The overall frequency of *Braf* mutations in melanocytic nevi is similar to that in melanomas and also varies according to histological subtype, ranging from 0% in Spitz nevi to 90% in intradermal nevi (Yazdi *et al.* 2003).

3.1.2.2 TP53

TP53 is found on chromosome 17p13 and is known as a universal tumour suppressor gene that is regulated by *p14/arf*. It has many functions within the cell including transcriptional and cell cycle control and has been shown to bind to the protein product of the *Hdm-2* human oncogene (*Mdm-2* in mice) (Castellano & Parmiani 1999; Momand *et al.* 1992). The *Hdm-2* oncogene has been shown to be amplified and overexpressed in some human sarcomas. Its protein product binds to the p53 protein and blocks its ability to act as a transcription factor (Oliner *et al.* 1992). HDM-2 has also been shown to be transcriptionally regulated by p53 forming an autoregulatory loop where increased p53

activity increases HDM-2 levels which in turn decreases p53 activity resulting in declining HDM-2 levels (Picksley *et al.* 1993; Wu *et al.* 1993). This relationship can be disrupted in several ways; the *tp53* gene can be mutated so that the cell does not make HDM-2 proteins (occurring in 50-55% of cancers); the *Hdm-2* gene can be amplified so that it blocks p53 functions (found in 30% of sarcomas); phosphorylation of p53 can occur on or near the p53-HDM-2 binding sites and disrupt their interaction; HDM-2 can be inactivated by the ARF protein or by the interaction with some ribosomal proteins (Lohrum *et al.* 2003; Unger *et al.* 1999).

P21/Waf-1/Cip-1 is another gene regulated by the p53 transcription factor (el-Deiry *et al.* 1993). One function of the p21 protein is to bind to the cyclin E/cdk2 protein kinase that acts in the late G1 phase of the cell cycle and blocks its activity and is, therefore, one of the reasons why p53 activation can lead to arrest in the G1 phase of the cell cycle (Harper *et al.* 1995). There are also a set of genes regulated by p53 that promote apoptosis by helping to activate the release of cytochrome c from the mitochondria and contributing to the activation of caspase 9 and 3. P53 can also activate the increase in levels of Fas ligand and the KILLER DR receptor which then activates caspase 8 and eventually leads to apoptosis. P53 also regulates some genes that are involved in DNA repair as well as those that can alter the extracellular matrix and so affects the regulation of cell division, metastasis and angiogenesis.

82% of somatic mutations in the *tp53* gene are point mutations and the other 18% are made up of insertions, deletions or rearrangements. The most common point mutations have been localised in six hotspots within the *tp53* gene at codons 175, 245, 248, 249, 273 and 282 causing either the elimination of critical contacts with DNA or destabilisation of the protein structures required for DNA binding (Cho *et al.* 1994). Mutations in the *tp53* gene have been found to be an early genetic event in the development of UV-induced skin cancers (Ziegler *et al.* 1994).

3.1.2.3 *N-RAS*

Ras proteins have been briefly described in section 1.1.2 and as a group are comprised of 20-25 kDa proteins that are important in regulating growth and differentiation. The *N-ras* gene, along with *H-ras* and *K-ras*, is expressed in every tissue and cell type, however the type of organ and the stage of development denotes the variation in these gene's expression

patterns (Crespo *et al.* 2000). Ras gene products affect a number of downstream signalling pathways through a number of effectors that interact with Ras-GTP, these include Raf, PI3K and Ral-GDS (Campbell *et al.* 1998). Raf, as already described, stimulates the activation of MEK which then activates the ERK pathway. Phosphatidylinositol 3-kinase (PI3K) activates Akt/PKB and p70^{S6K} which take part in transcriptional control. The activation of Akt also generates an anti-apoptotic signal (Bos 1998). Mutated ras genes can be found in up to 35% of human tumours, more specifically, mutated *N-ras* can be found in 20% of melanoma lesions. *N-ras* mutations have also been found in naevi whereas *K-ras* and *H-ras* have been shown to be only occasionally involved in melanoma (Polsky *et al.* 2003). Activating mutations in the *N-ras* gene cause the reduction of intrinsic GTPase activity and the resistance against molecules that induce GTPase activity.

3.1.2.4 *CDKN2A* (*p16^{INK4A}*)

P16 is a cyclin dependant kinase inhibitor (CKI) that can bind to both cdk4 and cdk6 and inhibits the catalytic activity of the cdk4-6/cyclin D enzyme complex, which is required for the phosphorylation of Rb (Rocco *et al.* 2001). The suppression of the retinoblastoma protein caused by the inactivation of *p16^{INK4A}* is an early event commonly seen in melanomas (Kamb *et al.* 1994b). By inhibiting the phosphorylation of Rb the formation of the Rb-E2F repressive transcriptional complex is promoted resulting in the prevention of E2F dependant transcription and so blocking the progression of the cell cycle at the G1/S phase (Zhang *et al.* 1999). Many studies have identified a high frequency of *p16^{INK4A}* deletion in a number of human tumour cell lines including melanoma, oesophageal, lung, head and neck, breast and ovarian cancers (Kamb *et al.* 1994a; Nobori *et al.* 1994). Most tumour suppressor genes have been found to be inactivated by point mutations; however, small homozygous deletions are a major mechanism for the inactivation of *p16^{INK4A}*. Homozygous deletions of *p16^{INK4A}* have been detected in a number of primary tumours including bladder, prostate and breast cancers as well as melanoma, sarcoma and glioma (Cairns *et al.* 1995; Ohta *et al.* 1994). As well as homozygous deletions and rare inactivating mutations of *p16^{INK4A}*, inappropriate methylation of *p16^{INK4A}* promoter region has also been recognized as a mechanism of silencing this gene. The methylation of the 5' CpG island of *p16^{INK4A}* has been associated with a complete transcriptional block of this gene in a number of cancers including head and neck, lung, brain, colon and bladder cancer

(Gonzalez-Zulueta *et al.* 1995; Herman *et al.* 1995). The inactivation rate of $p16^{INK4A}$ in melanoma cell lines due to homozygous deletions, point mutations or promoter hypermethylation is high, at approximately 80%, whereas it is only at about 10% in primary melanomas (Castellano *et al.* 1997). It has been suggested that the absence of $p16^{INK4A}$ gives a growth advantage as the loss of expression correlates with the progression of invasive melanoma (Reed *et al.* 1995).

It is now known that all melanoma cases that have a homozygous deletion at exon 1 β of $p16^{INK4A}$ will also lose $p14^{ARF}$, as they share this exon. $P14^{ARF}$ has been shown to inhibit p53 degradation, therefore loss of the $p14^{ARF}$ tumour suppressor will result in the abnormal and uncontrolled degradation of p53 resulting in its inactivation (Pomerantz *et al.* 1998).

3.1.2.5 CDK4

It has been revealed that some of the molecules that are most often altered in cancer are those that are involved in the regulation of the G1/S phase of the cell cycle. It is at this phase that cells become committed to a new round of cell division. Cyclin dependent kinase-4 (CDK-4) functions in the G1/S phase of the cell cycle and it is driven by 3 D type cyclins: D1, D2 and D3. Together Cyclin D-CDK 4 targets pRb and its related proteins by phosphorylating them, causing their inactivation (Harbour *et al.* 1999). The inactivation of the Rb proteins allows the transcription of the 2F-controlled genes leading, eventually, to the initiation of the S phase (Bartek *et al.* 2001; Sherr 2000). The CDK-Cyclin D/INK4/pRb/E2F cascade has been found to be altered, by either mutations in the genes that encode these proteins or their upstream regulators, in more than 80% of human cancers (Ortega *et al.* 2002).

CDK 4/6-cyclin D also binds the Cip/Kip family of cell cycle inhibitors, preventing them from binding and inactivating CDK-2-cyclin E and thus allowing the further progression of G1 (Sherr *et al.* 1999). *Cdk4* has been found to be amplified or overexpressed in a wide variety of tumours including gliomas, sarcomas and breast tumours (An *et al.* 1999; Perry *et al.* 1999; Wei *et al.* 1999). Point mutations in *Cdk4* and *Cdk6* prevent the binding of the INK4 cell cycle inhibitors but still allow the binding of cyclin D1 causing the CDK4/6 kinase activity to remain 'switched on' at inappropriate times. In the absence of the negative regulation by INK4 of CDK 4/6, the phosphorylation of Rb proteins and the progression of the cell cycle through the G1/S phase is allowed (Ranade *et al.* 1995). *Cdk4*

has been recently identified as a melanoma susceptibility gene, although its overall role in familial melanoma is low (Soufir *et al.* 1998).

3.1.2.6 *PTEN*

Pten is a tumour suppressor gene that is located on chromosome 10q23 and encodes a phosphatase which can dephosphorylate proteins but primarily dephosphorylates the lipid phosphatidylinositol-3,4,5-tris-phosphate (PIP₃) (Myers *et al.* 1998). PIP₃ is undetectable in quiescent cells but increases in response to growth factors or extracellular matrix-dependant signalling, which recruit phosphatidylinositol-3-kinase (PI3K). PTEN negatively regulates the Akt/PI3K pathway which is important for cell growth, proliferation and survival. The indirect regulation of this pathway by PTEN occurs when PTEN dephosphorylates PIP₃. PIP₃ is, indirectly, a major activator of Akt and so by keeping levels of PIP₃ low via its dephosphorylation, phosphoinositide-dependant kinase-1 (PDK-1) is not activated which in turn prevents the phosphorylation and subsequent activation of Akt. By inhibiting Akt, apoptosis is stimulated and the cell cycle is halted at the G1/S phase (Wu *et al.* 1998). Genetic inactivation of *Pten* leads to the overactivation of Akt and results in the loss of proliferative and apoptotic control (Wu *et al.* 2003b).

Loss of homozygosity on regions of chromosome 10q have been observed in a number of cancer types including 30-50% of melanomas (Healy *et al.* 1995; Herbst *et al.* 1994). Studies have also shown mutations of *Pten* in 30-40% of melanoma cell lines and approximately 10% of primary melanomas (Guldberg *et al.* 1997; Tsao *et al.* 1998).

3.1.2.7 *CTNNB1* (β -catenin)

CTNNB1 encodes for the protein β -catenin which is the central and essential component in the Wnt pathway. This is one of the few pathways that interact to regulate proliferation, differentiation and cell-cell interactions (Brembeck *et al.* 2006). β -catenin has two functions, firstly in cell adhesion and secondly in transcriptional regulation and both of these functions are crucial for maintaining normal cellular function. The promotion of transcriptional regulation by β -catenin is caused by deregulation of normal cellular control leading to the deregulation in the phosphorylation of its tyrosine residues; this phosphorylation also causes β -catenin to lose its function in adhesion. This deregulation is

seen in the development and progression of many malignancies. In the Wnt signalling pathway, β -catenin acts as a transcriptional activator in conjunction with LEF/TCF binding proteins (Behrens *et al.* 1996). In cell adhesion, β -catenin functions at the plasma membrane by linking E-cadherin to α -catenin. The E-cadherin – β -catenin – α -catenin complex at adherens junctions forms a dynamic link to the cytoskeleton (Drees *et al.* 2005; Yamada *et al.* 2005). The loss of cell-cell adhesion and increased cell motility is a feature of tumour progression. It occurs, in part, due to the loss in function of E-cadherin causing the dissociation of the E-cadherin – β -catenin – α -catenin complex from the membrane (Perl *et al.* 1998). The dissociation of adherens junctions can also be induced by the phosphorylation of any of the components of this complex, including β -catenin (Roura *et al.* 1999). Most β -catenin mutations occur in or around exon 3 of its gene affecting its phosphorylation site and so making it resistant to degradation (Morin *et al.* 1997; Polakis 1999). Activating mutations in the β -catenin gene have been detected in approximately 10% of colorectal cancers and 40% of hepatocellular carcinomas (Legoix *et al.* 1999; Samowitz *et al.* 1999).

All of these mutations are important in the development and progression of melanoma at the genetic level; it is difficult, however, to determine what is actually happening at the functional level. As mentioned previously, most studies have reported gene mutations in melanoma but differences in protein fingerprints in relation to these mutations has not been reported; therefore the aim of this aspect of my research was to assess proteomic differences in cell lines according to their known genetic mutations, specifically *tp53*, *Braf* or *p16^{INK4A}*. Bioinformatics analysis was used to deconvolute the large amount of mass spectral data generated.

3.2 Methods

A more detailed account of the methods used for this study is given in Chapter 2 and a schematic representation, showing how the cell line samples were processed for mass spectrometry and ANN analysis, is shown in figure 3-1.

All of the cell lines processed by SELDI MS were analysed by Dr Guldberg and Dr Thor Straten for the presence of mutations of known melanoma associated genes, as part of the

OISTER project. The method used for obtaining the gene mutation information from each of the cell lines is described below.

Exons 11 and 15 of *Braf* and exons 4-10 of *tp53* were scanned for mutations by PCR in combination with DGGE, as previously described (Christensen *et al.* 2005; Guldborg *et al.* 1997). Exons 1 α , 2 and 3 of the *cdkn2A* gene were examined for deletions by PCR analysis and for mutations by single-stranded conformation polymorphism (SSCP) analysis, as previously described (Gronbaek *et al.* 2000). Abnormal DGGE and SSCP bands were excised from the gels, and the DNA was reamplified and sequenced using a ³³P-end-labeled primer and the ThermoPrime Cycle Sequencing Kit (Amersham Life Science, Cleveland, OH). The methylation status of the *cdkn2A* promoter CpG island was examined by methylation-specific PCR, using previously described primers (Herman *et al.* 1996).

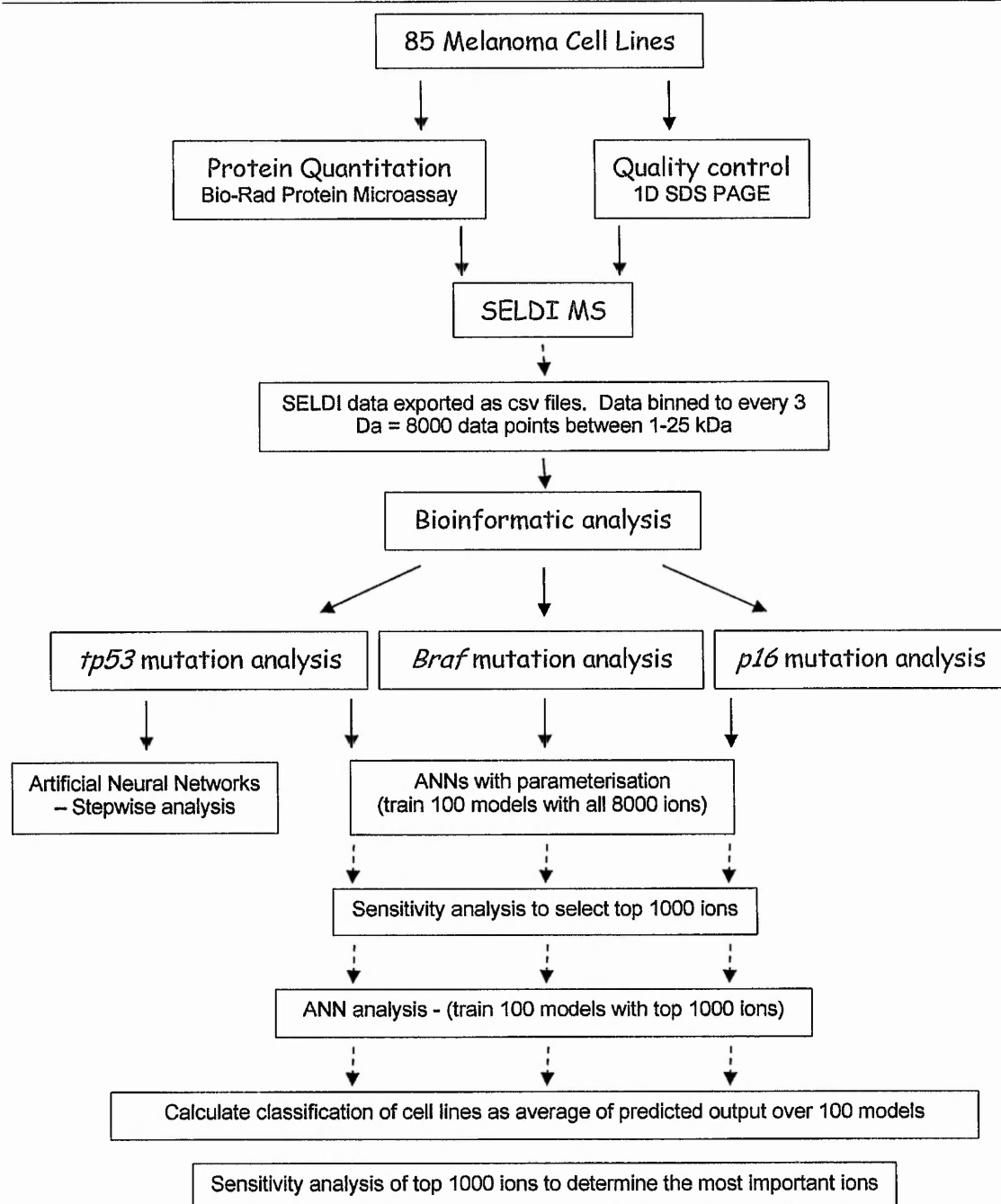


Figure 3-1 A schematic representation showing the flow of sample preparation and analysis methods used in this study

3.3 Results

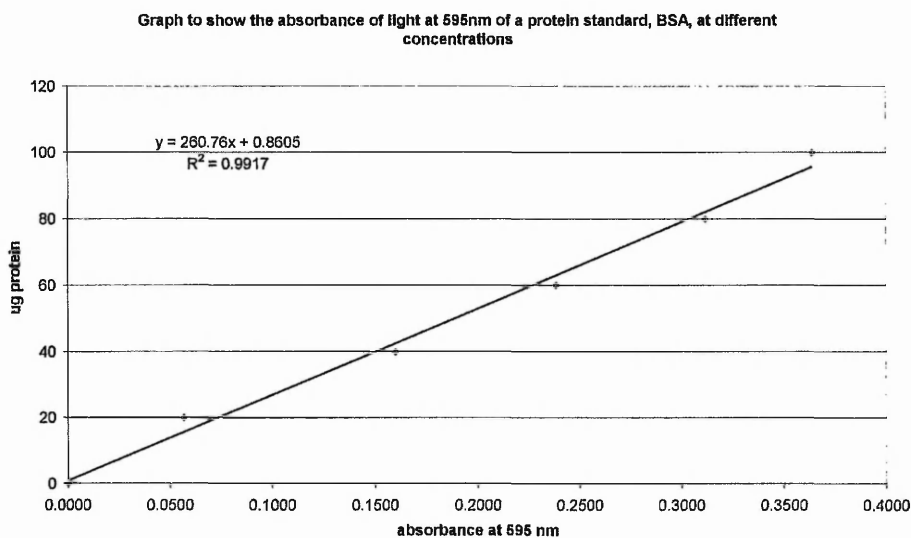
3.3.1 Protein quantification and sample quality control.

Protein quantification was performed using the Bio-Rad protein microassay as described in section 2.2.2. A standard curve was determined by plotting the absorbance at 595nm versus the total amount of BSA protein (μg). An equation of the line was determined which is calculated by the least squares fit for a line represented by the following equation $y = mx+b$, where m is the gradient of the line, b is the y -intercept, x is the absorbance value and y is the amount of protein in μg . All of the protein lysates were assayed in duplicate with each reading taken in duplicate, giving four absorbance readings per sample. The average absorbance was calculated and subtracted from the absorbance value of the reagent blank, giving a final absorbance value which was then substituted into the equation allowing the amount of protein, in μg , in the assay tube to be determined. As the sample was made up to an 800 μl volume, the amount of protein given in μg was divided by 0.8 to give the amount of protein in $\mu\text{g/ml}$. The amount of protein in $\mu\text{g/ml}$ needs to be multiplied by the dilution factor which, if 5 μl of neat sample was used and made up to 800 μl with water, would be $800 \div 5 = 160$, giving the protein in $\mu\text{g/ml}$, which is then divided by 1000 to give the amount of protein in $\mu\text{g}/\mu\text{l}$; these units were used throughout the study. Figure 3-2 shows an example of a standard curve calculated using BSA, an example of a calculation to determine the amount of protein in one of the cell line lysates and finally the protein concentrations of three cell line lysates. To determine whether any protein degradation had occurred during the production of the cell line lysates, the lysates were run through a 12.5% resolving gel to assess the banding pattern. The bands of three cell lines, MEWO, ESTDAB 005 and Ma Mel 39a, are shown in figure 3-3. The data shown in figures 3-2 and 3-3 are representative of the results obtained for all of the cell line lysates generated in this study. An adequate amount of protein was extracted from the cell lines as shown by the protein assay. The gels shown in figure 3-3 show the same bands for all three of the cell lines indicating that the lysates were of similar quality. Any non-specific degradation of a sample would cause the most prominent bands to fade or disappear affecting the high abundant as well as the low abundant proteins.

Figure 3-4 shows the morphology of three of the cell lines in culture, at x10 and x20 magnification, depicting the differences between established melanoma cell lines. It can be

assumed that each cell line will generate a slightly different proteomic profile based on the fact that the morphology of the cell lines are different in culture. Their morphologies are either octagonal, for example cell lines ESTDAB 005 and Ma Mel 39a shown in figure 3-4, or epithelial-like as shown by cell line MEWO. Although the shapes of the cells are either one of two types, the size of the cell, size of the nucleus and the granularity of the cells contribute to the different morphologies seen across all 85 cell lines grown in culture. Once the protein concentration was determined and the sample quality verified, the protein lysates were analysed by SELDI MS.

a)



b)

Calculation of the amount of protein in the MEWO cell line lysate.

Firstly the average of the 4 absorbance readings was calculated:

$$(0.4645 + 0.4426 + 0.4597 + 0.4503)/4 = \mathbf{0.4543}$$

Then the absorbance of the negative control was subtracted:

$$0.4543 - 0.2328 = \mathbf{0.2215}$$

This value was then put into the equation of the standard curve, as shown on the above graph:

$$y = 260.76x + 0.8605$$

$$y = (260.76 * 0.2215) + 0.8605$$

$$y = \mathbf{58.62\mu g}$$

The amount of protein in 1 ml was calculated:

$$58.62/0.8 = \mathbf{73.28\mu g/ml}$$

This was then multiplied by the dilution factor:

$$73.28 * 160 = \mathbf{11724.20\mu g/ml}$$

The value was then shown in $\mu g/\mu l$:

$$11724.20 / 1000 = \mathbf{11.72\mu g/\mu l}$$

c)

Protein concentrations of three of the cell line lysates:

$$\text{MEWO} = \mathbf{11.72\mu g/\mu l}$$

$$\text{ESTDAB 005} = \mathbf{10.43\mu g/\mu l}$$

$$\text{Ma Mel 39a} = \mathbf{5.34\mu g/\mu l}$$

Figure 3-2 Figure showing the process of protein estimation in all samples analysed.

A) shows the standard curve of the absorbance of light at 595nm of BSA standards at different concentrations, b) shows how the concentration of cell line lysates is calculated using the equation of the standard curve and c) shows the amount of protein estimated in three cell line lysates, MEWO, ESTDAB 005 and Ma Mel 39a.

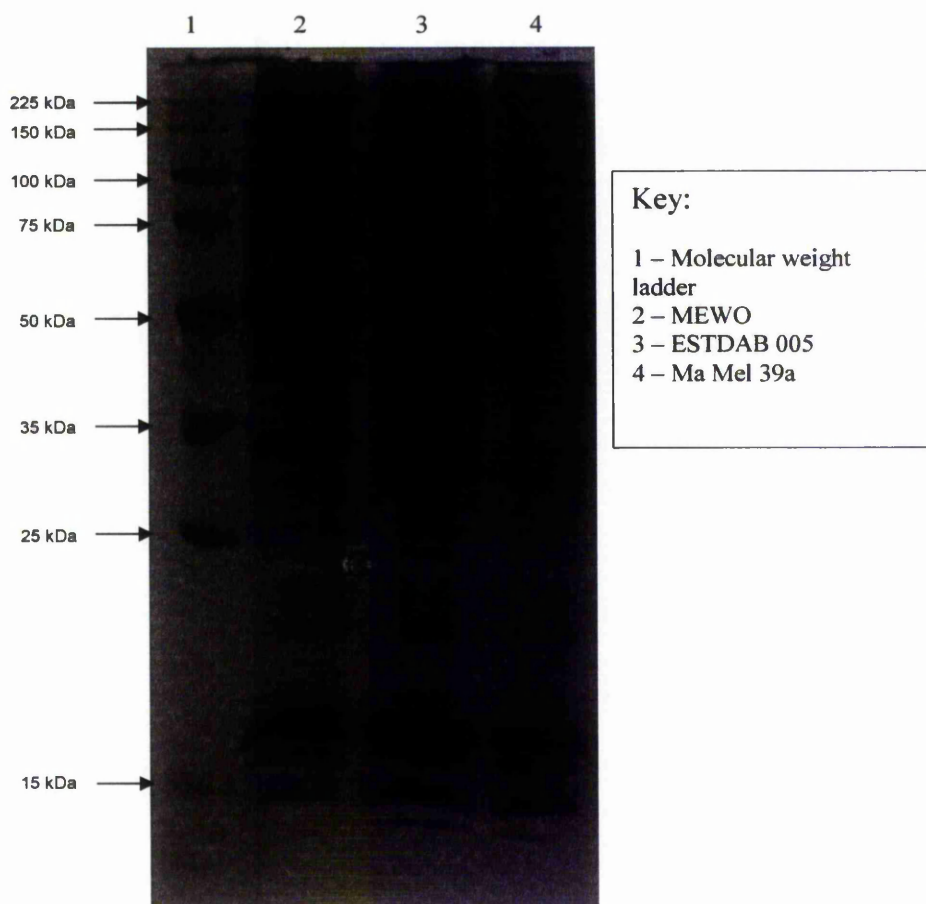


Figure 3-3 1-D SDS PAGE of three cell line lysates, MEWO, ESTDAB 005 and Ma Mel 39a.

The gel presented above represents an example of the data obtained from all of the cell line lysates used in this study. 40 μ g of each lysate sample was resolved in a 12.5% resolving gels and stained with coomassie blue to determine if degradation of the proteins occurred during sample preparation.

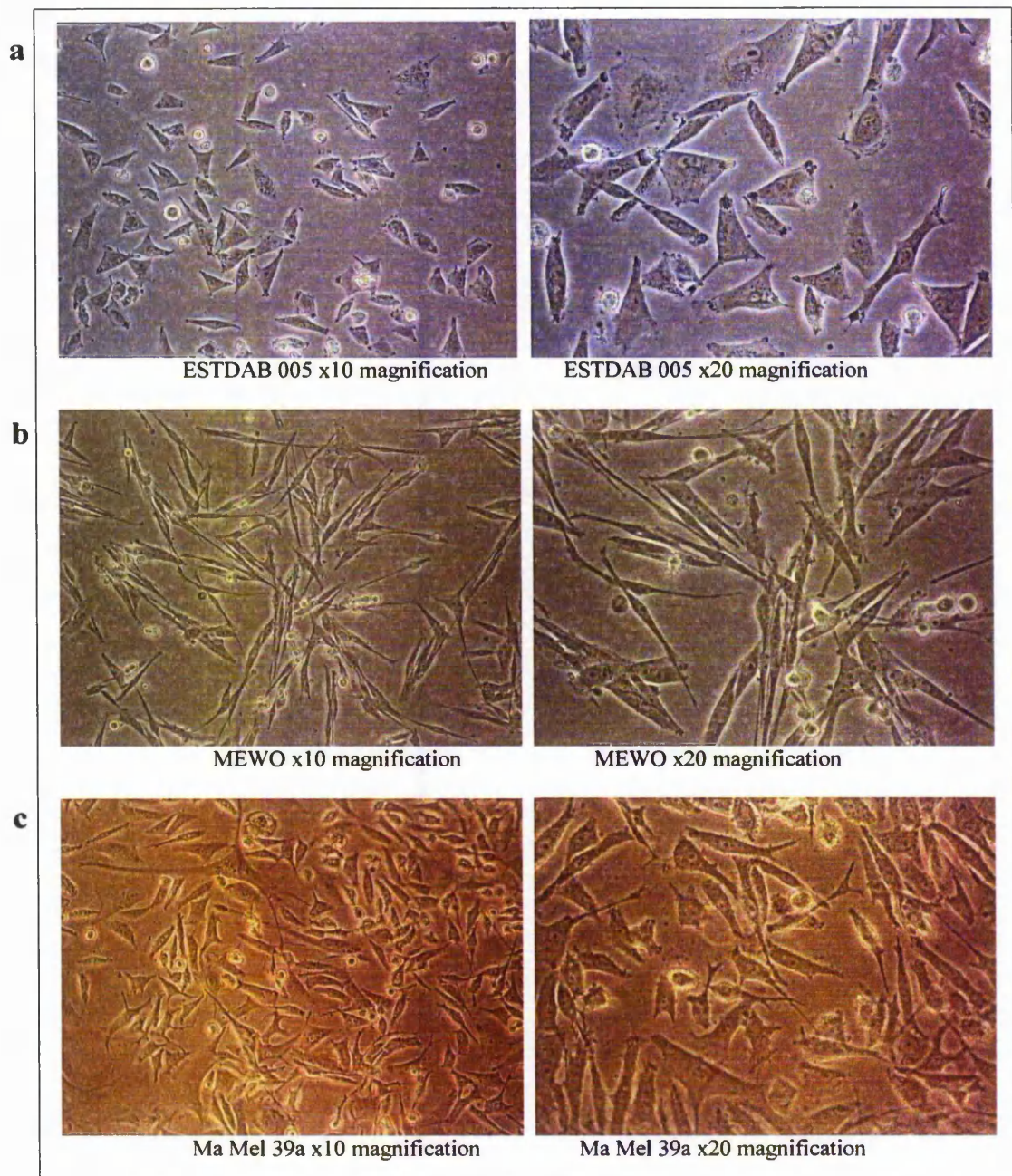


Figure 3-4 Photographs of three cell lines in culture.

This figure shows the morphology of three of the melanoma cell lines in culture at x10 and x20 magnification as a representation of the diverse morphology seen in all of the cell lines used in this study. a) shows ESTDAB 005 cell line, b) shows MEWO cell line and c) shows Ma Mel 39a cell line.

3.3.2 SELDI MS

The protocol used is described in section 2.2.4.1, but briefly the samples were bound to the hydrophobic surface on a H50 SELDI chip and analysed using a PBS II mass analyzer. Figure 3-5, shows 2 of the cell lines analysed in duplicate to represent the reproducibility of the SELDI MS method.

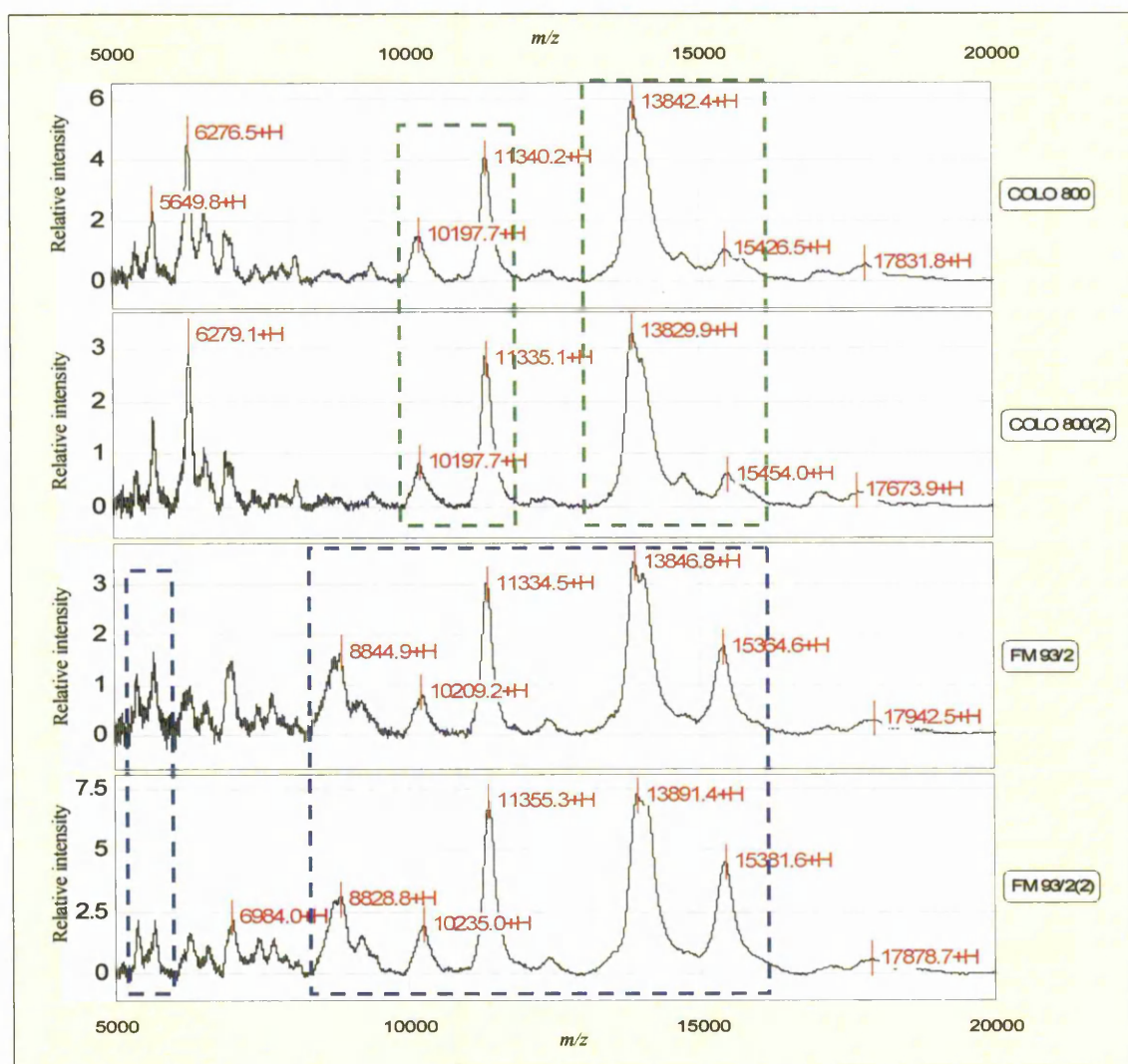


Figure 3-5 Figure showing 4 SELDI MS profiles in the 5000-20000 Da mass range representing the reproducibility of the SELDI MS technique.

This figure shows two cell line samples, COLO 800 and FM 93/2 spotted in duplicate on different SELDI chips to highlight the low chip-to-chip variation. The green boxes highlight the similarity of the profiles of COLO 800 cell line when run in duplicate and the blue boxes show the peak similarities in the duplicated FM 93/2 cell line.

This shows the reproducibility of the protein profiles when samples are spotted on different chips but analysed using the same automated protocols. This was also performed to check on variability between applying the sample onto different spots, chip to chip variability and to determine machine stability. The profile patterns are similar between the duplicates but the masses of the peaks and the peak intensities are visually different. The differences in the masses are due to the mass accuracy of the PBS II mass analyser and are taken into consideration when analysing the data. The variability in the intensities of the peaks is accounted for when the data is prepared for neural network analysis.

Figure 3.6 shows the SELDI MS profiles obtained from 4 cell lines, COLO 818, FM55 M2, GR-M and MEWO, containing mutations in their *tp53* gene overlaid on a profiled of a cell line without a *tp53* mutation, Mel SOE. The profiles show the 2000-10000 Da mass range, although data up to 30kDa was collected. The green boxes highlight peaks that are visibly similar in this mass range, whereas the blue boxes highlight where a peak is present in one of the profiles but not in the other. Table 3-1 shows all of the peaks that can be seen in all of the profiles shown in figure 3-6, with peaks that are similar between the cell lines containing a *tp53* mutation and Mel SOE highlighted in red. There are no overall similarities between cell lines with or without *tp53* mutations that can be observed visually except for one peak present in Mel SOE at 7450.6 Da that is not visible in any of the cell lines containing a *tp53* mutation.

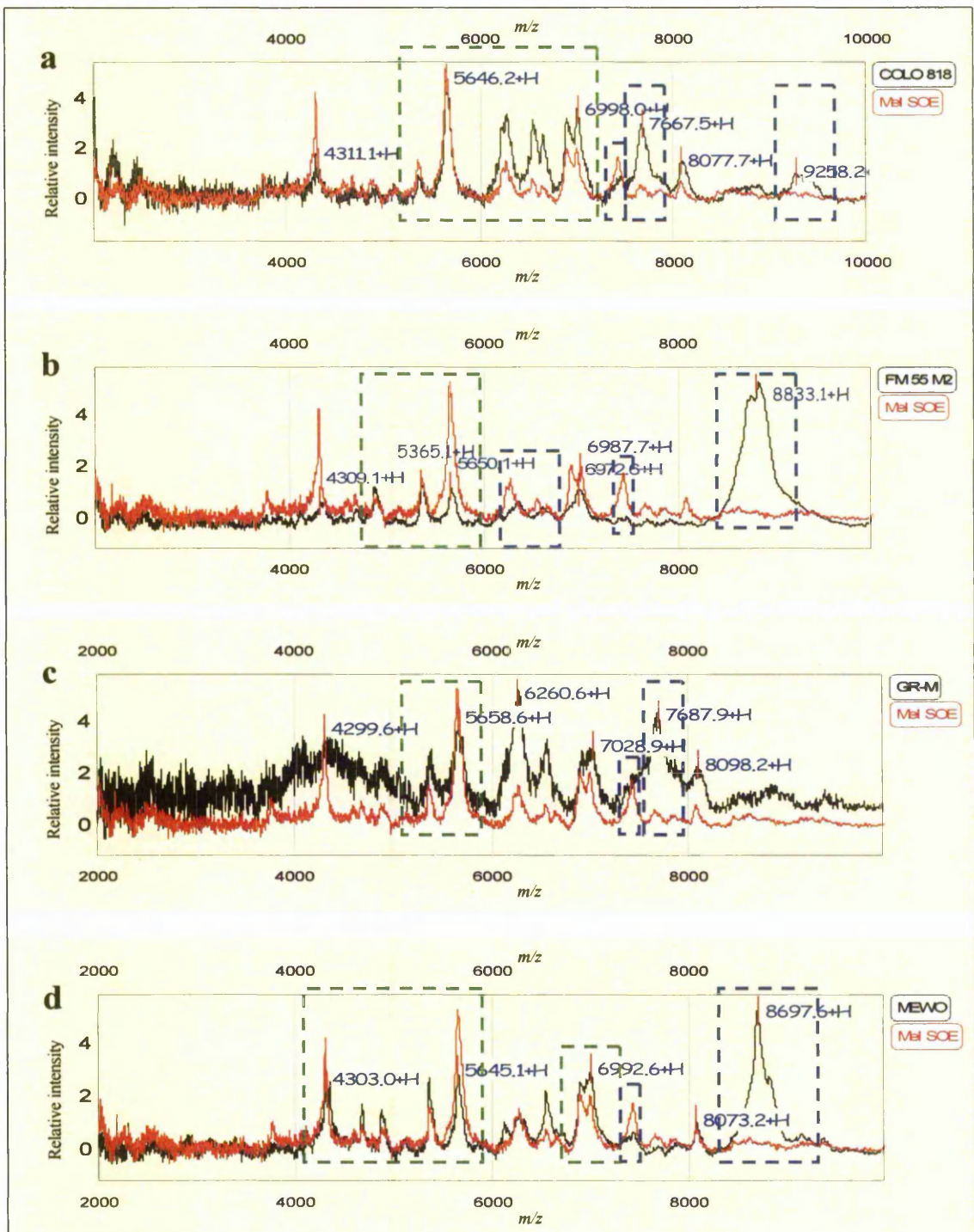


Figure 3-6 SELDI MS profiles in the 2000-10000 Da range of cell lines with or without *tp53* mutations. The 4 profiles show Mel SOE as an example of a profile taken from a cell line without a *tp53* mutation and this is overlaid with profiles taken from cell lines with a *tp53* mutation; a) shows the cell line COLO 818, b) shows cell line FM55 M2, c) shows cell line GR-M and d) shows cell line MEWO. The green boxes show the similarities between the *tp53* mutated cell lines and Mel SOE and the blue boxes show the differences between the *tp53* mutated cell lines and Mel SOE.

Mel SOE (wt <i>tp53</i>)	COLO 818 (mut <i>tp53</i>)	FM 55M2 (mut <i>tp53</i>)	GR-M (mut <i>tp53</i>)	MEWO (mut <i>tp53</i>)
3765.2	3770.5	-	-	-
4299.9	4311.1	-	-	4303.0
-	-	4364.8	-	-
-	-	4372.3	-	-
-	-	-	-	4669.9
-	-	-	-	4868.9
5351.2	5356.8	5360.5	5379.8	5351.6
5642.5	5646.2	5650.1	5658.6	5645.1
-	-	-	-	6110.5
6262.3	6266.2	-	6260.6	6260.5
-	-	6349.4	-	-
6530.3	6537.5	-	-	-
-	-	-	6560.3	-
-	-	6594.9	-	-
-	6639.0	-	-	-
6889.8	6888.2	-	-	6892.0
6987.7	6998.0	6972.6	7028.9	6992.6
7450.6	7422.3	-	-	7430.0
7666.6	7667.6	-	7687.9	-
8074.2	8077.7	-	8098.2	8073.2
-	9258.2	-	-	9363.3

Table 3-1 Values of the main peaks observed in the spectra shown in figure 3-6.

The table shows the main peaks (values given in Daltons) observed in the spectra obtained from a cell line with a *tp53* mutation and cell lines without *tp53* mutations, in the 2000-10000 Da mass range. Peaks that are similar between samples are shown in the same row, although the m/z values may be slightly different due to the mass accuracy of the instrument. Peaks that are observed in both the spectra of cell lines with wild-type *tp53* and mutated *tp53* are highlighted in red.

Figure 3.7 shows the SELDI MS profiles obtained from 4 cell lines, COLO 794, ESTDAB 005, WM 139 and WM 1205, containing *Braf* mutations overlaid on a profile of a cell line without a *Braf* mutation, where the green boxes highlight peaks that are similar between spectra and the blue boxes highlight the differences. Table 3-2 shows all of the peaks visible in the profiles shown in figure 3-7, with peaks that are similar between the cell lines containing a *Braf* mutation and FM-3 highlighted in red. There are no overall similarities between cell lines with or without *Braf* mutations that can be observed visually but there are two peaks present in FM-3 at 5129.1 and 5883.7 Da that are not visible in any of the cell lines containing a *Braf* mutation.

Figure 3.8 shows the SELDI MS profiles obtained from cell lines ESTDAB 107, Ma Mel 39a, WM 852 and NW 145 which contain mutations in their $p16^{INK4A}$ gene, overlaid on the cell line FM 79, which does not have any $p16^{INK4A}$ mutations, where similarities and differences are highlighted by green and blue boxes, respectively. Peaks visible in the spectra shown in figure 3-8 are listed in table 3-3, where peaks of similar m/z values from both types of cell line shows are highlighted in red. The results show that there are five peaks present in FM79 at 2923.9, 3398.1, 3899.9, 4521.1 and 5923.5 Da that are not visible in any of the cell lines that contain a $p16^{INK4A}$ mutation.

Due to the large amount of data generated in this study, the complexity of the spectra obtained from SELDI MS profiling and a high sample number, it was difficult to distinguish between the protein profiles of a cell line that has a particular genetic mutation and hence which ions are important and are associated with the mutation. Also, visual differences between cell lines with or without known mutations do not necessarily mean that they will be important predictive markers. Therefore, the mass spectrometry data was mined using artificial neural networks to derive and recognise patterns that associate with a given genotypic characteristic.

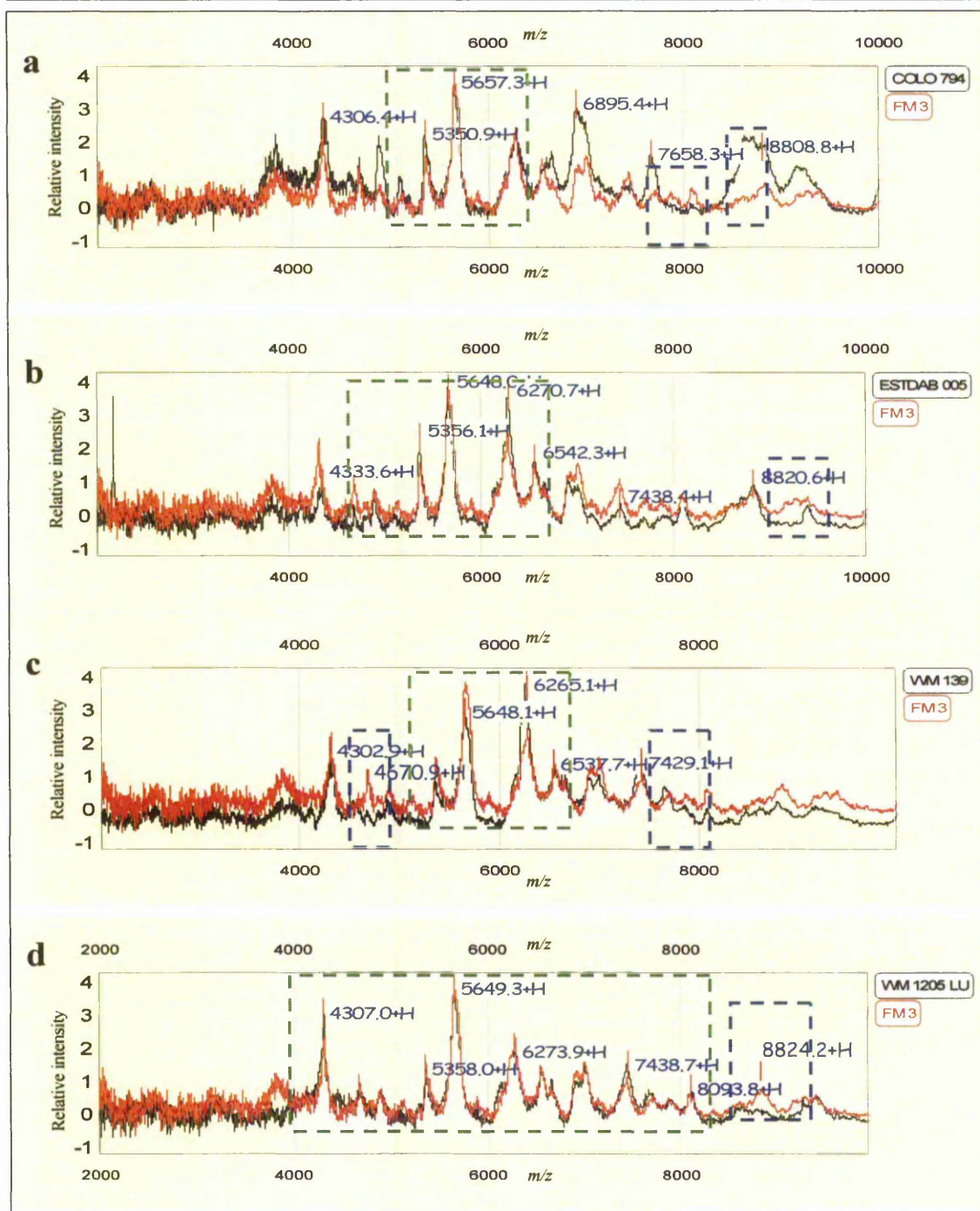


Figure 3-7 SELDI MS profiles in the 2000-10000 Da range of cell lines with or without *Braf* mutations. All 4 profiles show FM-3 as an example of a profile taken from a cell line without a *Braf* mutation and this is overlaid with profiles taken from cell lines with a *Braf* mutation; a) shows the cell line COLO 794, b) shows ESTDAB 005, c) shows cell line WM 139 and d) shows cell line WM 1205. The green boxes show the similarities between the cell lines containing the mutations and Mel SOE and the blue boxes show the differences between the cell lines containing the mutations and Mel SOE.

FM-3 (wt <i>braf</i>)	COLO 794 (mut <i>braf</i>)	ESTDAB 005 (mut <i>braf</i>)	WM 139 (mut <i>braf</i>)	WM 1205 (mut <i>braf</i>)
-	-	2173.4	-	-
3812.6	3830.0	-	-	-
-	-	4113.6	4117.2	-
4310.0	4306.4	4333.6	4302.9	4307.0
-	-	-	4497.3	-
-	4575.7	-	-	-
4671.6	4676.1	4675.6	-	4673.5
4880.0	4886.2	4885.2	4863.7	4900.5
5129.1	5091.6	-	-	-
5357.5	5350.9	5356.1	5354.5	5358.0
5652.4	5657.3	5648	5648.1	5649.3
5883.7	-	-	-	-
6266.5	6271.6	6270.7	6265.1	6273.9
6536.6	-	6542.3	6537.1	6546.7
-	6646.4	6645.7	-	6644.2
6906.6	6895.4	6900.6	-	6914.5
6994.0	-	7000.9	6994.3	7003.1
-	-	-	7229.1	-
7432.9	7458.9	7438.4	-	7438.7
7711.3	7658.3	7687.9	7654.2	7662.3
7860.9	-	-	-	7875.5
8079.4	-	8085.0	8081.1	8093.8
-	8622.1	-	-	-
-	-	8703.7	-	-
-	-	-	8790.7	-
8830.4	8809.4	8820.6	-	-
-	9160.7	-	9167.1	-
9266.0	-	-	-	9251.6
-	-	9383.2	-	9397.9

Table 3-2 Values of the main peaks observed in the spectra shown in figure 3-7.

The table shows the main peaks (values given in Daltons) observed in the spectra obtained from a cell line with a *braf* mutation and cell lines without *braf* mutations, in the 2000-10000 Da mass range. Peaks that are similar between samples are shown in the same row, although the m/z values may be slightly different due to the mass accuracy of the instrument. Peaks that are observed in both the spectra of cell lines with wild-type *braf* and mutated *braf* are highlighted in red.

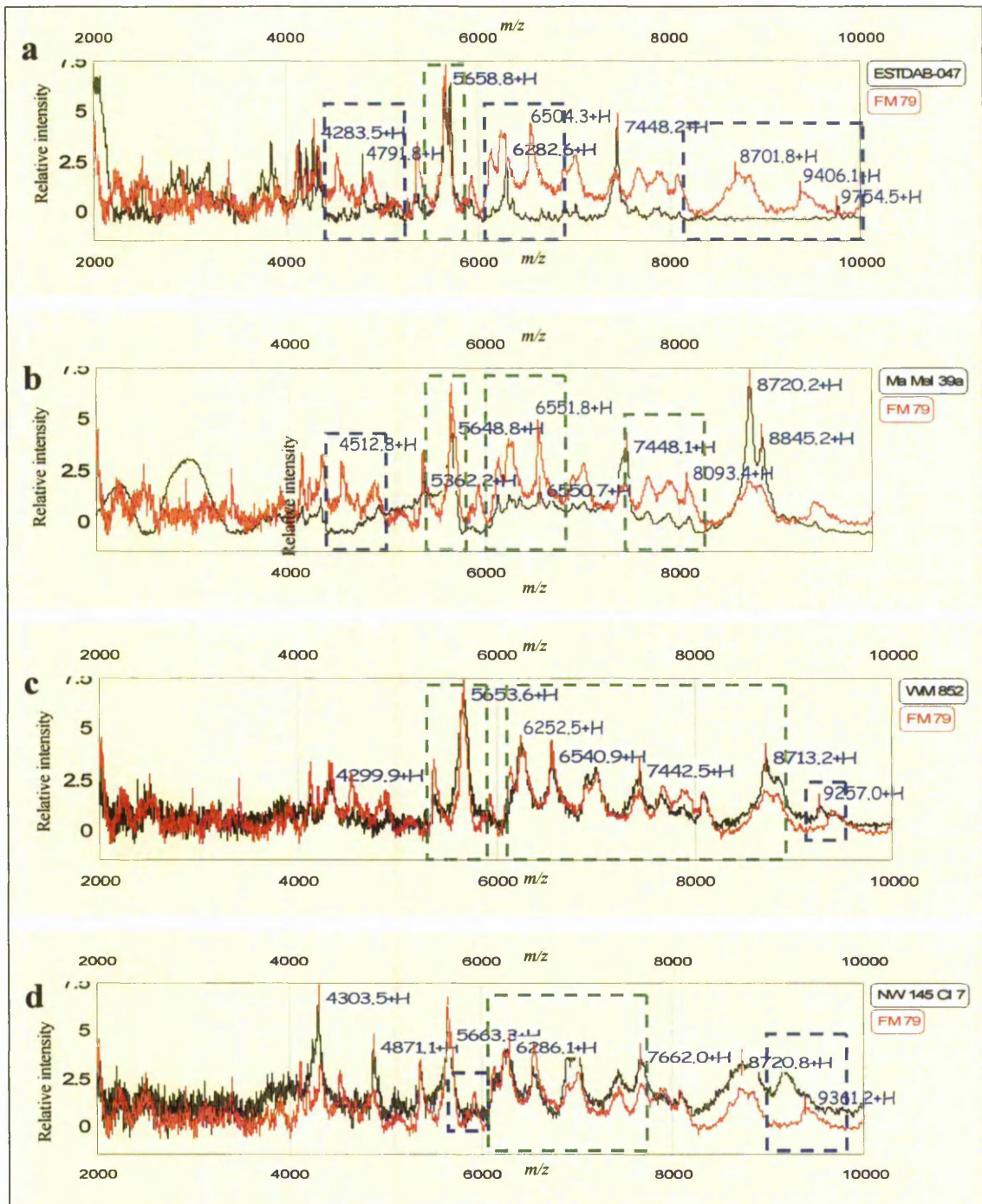


Figure 3-8 SELDI MS profiles in the 2000-10000 Da range of cell lines with or without $p16^{INK4A}$ mutations.

All 4 profiles show FM 79 as an example of a profile taken from a cell line without a $p16^{INK4A}$ mutation and this is overlaid with profiles taken from cell lines with a $p16^{INK4A}$ mutation; a) shows ESTDAB 047, b) shows Ma Mel 39a, c) shows cell line WM 852 and d) shows cell line NW 145. The green boxes show the similarities between the cell lines containing the mutations and Mel SOE and the blue boxes show the differences between the cell lines containing the mutations and Mel SOE.

FM-79 (wt $p16^{INK4a}$)	ESTDAB 047 (mut $p16^{INK4a}$)	Ma Mel 39a (mut $p16^{INK4a}$)	WM 852 (mut $p16^{INK4a}$)	NW 145 (mut $p16^{INK4a}$)
-	2059.0	-	-	-
2495.4	2498.6	-	-	-
-	2799.8	-	-	-
2923.9	2887.1	-	-	-
-	3190.6	-	-	-
3398.1	3369.2	-	-	-
-	3756.9	-	-	-
3899.9	3843.2	-	-	-
4115.3	4140.1	4118.7	-	-
4325.4	4326.6	4309.7	4299.9	4303.5
4521.1	-	-	-	-
4862.8	4791.8	-	-	4871.1
-	5045.3	4905.5	-	-
5358.0	5368.6	5362.2	5372.6	5365.2
5643.4	5658.8	5648.8	5653.6	5663.3
-	5707.6	-	-	-
5923.5	-	-	-	-
6127.5	-	6124.3	-	-
6230.4	6282.6	6239.5	6252.5	6286.1
-	-	6345.8	-	-
6540.4	-	6550.7	6540.9	6535.4
-	6886.8	6896.0	6910.2	-
7002.7	7008.2	7005.7	6995.5	7003.7
7431.4	7448.2	7448.1	7442.5	7419.5
7670.9	-	7666.6	7659.6	7662.0
7869.2	-	7883.3	7885.6	-
8076.4	-	8093.4	8073.7	8081.8
8708.8	-	8720.2	8713.2	-
-	-	-	8777.2	-
8826.8	-	8845.2	-	8832.9
-	-	-	-	9149.5
-	-	-	9257.0	-
9378.7	-	9385.2	-	-
-	9754.5	-	-	-

Table 3-3 Values of the main peaks observed in the spectra shown in figure 3-8.

The table shows the main peaks (values given in Daltons) observed in the spectra obtained from a cell line with a $p16^{INK4a}$ mutation and cell lines without $p16^{INK4a}$ mutations, in the 2000-10000 Da mass range. Peaks that are similar between samples are shown in the same row, although the m/z values may be slightly different due to the mass accuracy of the instrument. Peaks that are observed in both the spectra of cell lines with wild-type $p16^{INK4a}$ and mutated $p16^{INK4a}$ are highlighted in red.

3.3.3 Bioinformatic analysis

As genetic mutation information was made available for the cell lines, it was decided to use ANNs to data mine the SELDI spectral profiles to determine whether cell lines containing different known melanoma genetic mutations associated with a 'protein fingerprint' could be found. Eighty-five melanoma cell lines were analysed, and all had mutations in more

than one melanoma specific genes. Analysis was performed for the presence or absence of a specific genetic mutation rather than the presence or absence of multiple mutations, due to the low sample number. The following summarises the number of cell lines containing a mutation in a specific gene, although these cell lines would have also had mutations in other genes:

Braf mutations – 68 with mutations and 27 without
Nras mutations – 15 with mutations and 70 without
tp53 mutations – 24 with mutations and 61 without
cdk4 mutations – 8 with mutations and 77 without
ctnmb1 mutations – 4 with mutations and 81 without
Pten mutations/deletions – 23 with and 62 without
p16^{INK4A} deletions – 45 with deletions and 26 without

Due to time constraints it was decided that the SELDI MS data obtained from the profiling of cell line lysates would be mined for presence or absence of mutation in only three genes, *tp53*, *Braf* and *p16^{INK4A}*.

The SELDI MS data was exported into EXCEL as csv files and converted into .xls files. The data was merged into one file and smoothed by taking the median intensity value at every 3 Da, giving a total of 8000 ions between 1000-25000 Da. A multi-layer perceptron was trained with a back propagation algorithm and conjugate gradient descent, using Statistica software, on the smoothed mass/intensity data, using the ion intensity profile as the input variables. One hidden layer was used and the two groups were assigned either a 1 or 2 as the output for classification. It was then deemed beneficial to remove any ions, using a sensitivity analysis, that were less important to the ANNs for prediction and then retrain the ANNs to determine if the % correct prediction in all 100 models can be increased. To improve upon the predictions as well as speed up the training process it was decided that the top 1000 ions would be used for further neural network training. The protocol used to analyse the data is described in more detail in section 2.2.6. The following sections discuss in detail the results obtained from each of the neural network analyses performed.

3.3.3.1 *Braf* analysis

The m/z and intensity data from 1-25 kDa of 85 cell lines were used to train 100 ANN models using different training, test and blind data splits, where 60% of the cell lines were randomly selected for the training set, 20% were randomly selected for the test set and 20% were randomly selected for the blind data set for each model. For classification, the cell lines with a mutation in the *Braf* gene were assigned an output of 1 and the cell lines without a mutation in the *Braf* gene were assigned an output of 2. After the training had completed the predictions for the blind dataset of all 100 models were obtained, these were then averaged and the standard error of the mean calculated.

By training the ANNs with the top 1000 ions, correct classification of the cell lines as originating from either group 1 or group 2 was achieved for 100% of the samples in the training set and for 96.5% of the samples in the blind data set with a sensitivity of 98.3% and a specificity of 92.6%. The correct classification of the samples for the blind data set across the 100 models trained is illustrated in figure 3-9. For the ANNs to classify a cell line as one with or without a *Braf* mutation, it uses a cutoff of 1.5. If a cell line possesses a mutation in its *Braf* gene it should have an output of 1, therefore for the ANNs to correctly classify it, it must be assigned a value of below 1.5. Conversely if a cell line does not have a mutation in *Braf* then it should have an output of 2 and so for the ANNs to correctly classify this line, it should be assigned a value of greater than 1.5.

As can be seen for figure 3-9, 3 cell lines were misclassified; Ma Mel 39a and FM-3 were given values of 1.34 and 1.38 respectively when they should have had an output of 2, and FM 93_2 was given a value of 1.74 when it should have had an output of 1. A sensitivity analysis was performed on the top 1000 ions that were used to train the ANN models in order to determine the ions that were most important for the classification of the samples. The sensitivity ratios of the top 1000 ions were calculated for each of the 100 models trained, the average was determined and this is illustrated in figure 3-10.

Figure 3-10 shows that only 3 ions have ratio values that are significantly higher than the rest of the top 1000 ions for training the ANNs to classify between cell lines with or without *Braf* mutations. These ions are 6985, 13234 and 21187 with ratio values of 1.74, 1.30 and 1.21 respectively.

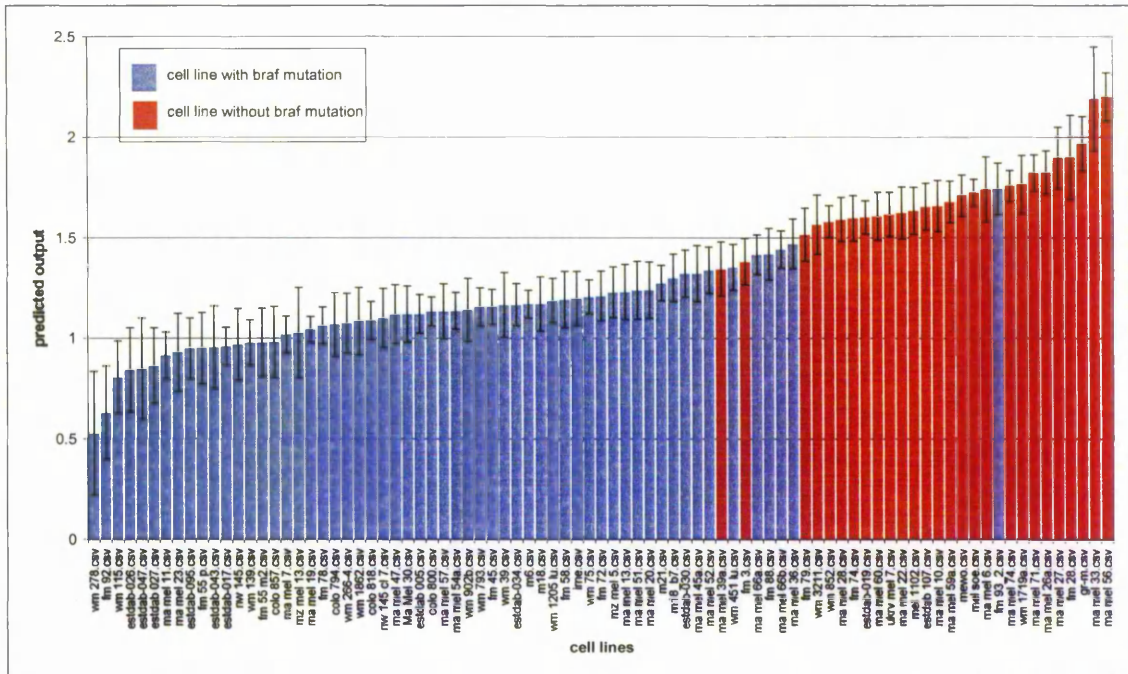


Figure 3-9 Classification by ANN analysis of 85 cell lines with or without *braf* mutations.

The figure shows the classification of 85 cell lines as predicted by ANN analysis, averaged over 100 models. The blue bars show the cell lines which have a *Braf* mutation and the red bars show the cell lines which do not have a *Braf* mutation. Cell lines with a predicted output of less than 1.5 are classified by the ANNs as having the mutation and cells lines with a predicted output of more than 1.5 are classified by the ANNs as not having the mutation. The error bars were determined by calculating the standard error of the mean with 95% confidence intervals. The figure shows that 3 cell lines were misclassified; Ma Mel 39a and FM-3 were given values of 1.34 and 1.38 respectively when they should have had an output of 2, and FM 93_2 was given a value of 1.74 when it should have had an output of 1.

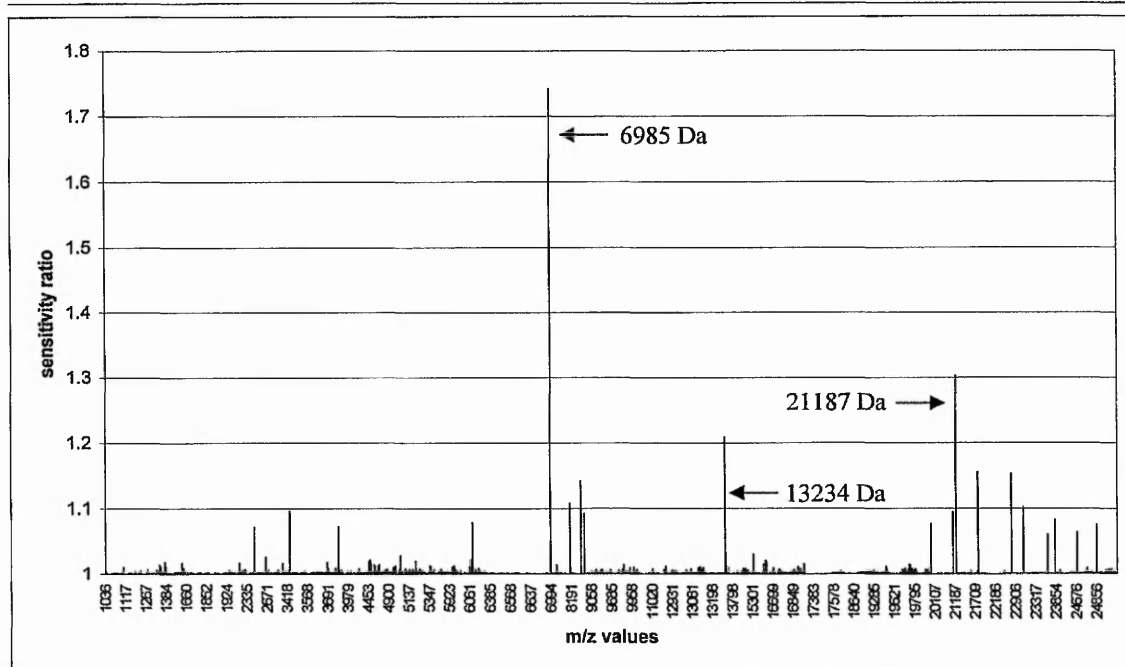


Figure 3-10 Sensitivity analysis of the top 1000 ions used to classify cell lines with or without *braf* mutations.

A list of the top 1000 ions used to classify cell lines with or without *braf* mutations in each ANN model is obtained using a sensitivity analysis. The data presented shows the sensitivity ratios of the top 1000 ions averaged over the 100 models trained.

3.3.3.2 *p16^{INK4A}* analysis

Only 71 cell lines were analysed for the presence or absence of mutations in the *p16^{INK4A}* gene as the mutational analysis (described in section 3.2) had not been fully completed at the time of bioinformatics analysis. The *m/z* and intensity data from the 71 cell lines was performed in the same way as described for the *Braf* analysis in section 3.3.3.1. For classification, the cell lines that contained a mutation in the *p16^{INK4A}* gene were assigned an output of 1 and cell lines without a mutation in the *p16^{INK4A}* gene were assigned an output of 2.

After training 100 ANN models using the top 1000 ions it was possible to achieve correct classification of the cell lines with 100% of the samples in the training set and with 97.2% of the samples in the blind data set, with 100% sensitivity and 92.3% specificity, illustrated in figure 3-11. This figure shows that 2 cell lines were misclassified; FM 92 and WM 39 were both given values of 1.38 when they should have had an output of 2.

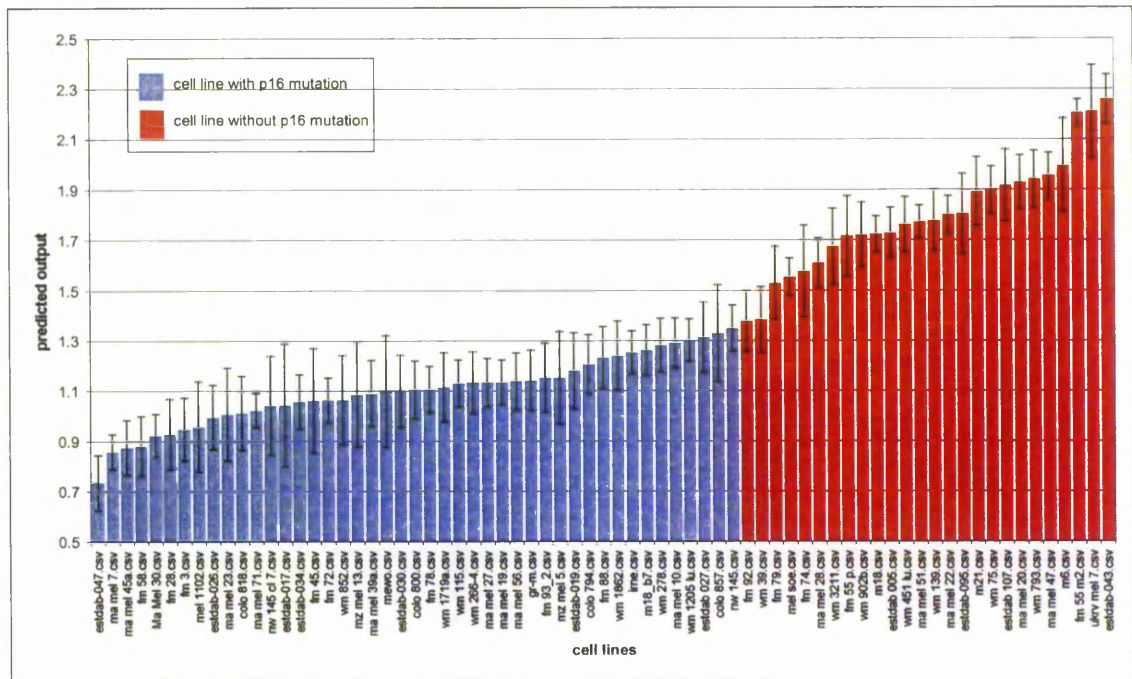


Figure 3-11 Classification by ANN analysis of 71 cell lines with or without $p16^{INK4a}$ mutations.

The figure shows the classification of 71 cell lines as predicted by ANN analysis, averaged over 100 models. The blue bars show the cell lines which have a $p16^{INK4a}$ mutation and the red bars show the cell lines which do not have a $p16^{INK4a}$ mutation. Cell lines with a predicted output of less than 1.5 are classified by the ANNs as having the mutation and cells lines with a predicted output of more than 1.5 are classified by the ANNs as not having the mutation. The error bars were determined by calculating the standard error of the mean with 95% confidence intervals. Two cell lines were misclassified; FM 92 and WM 39 were both given values of 1.38 when they should have had an output of 2.

Figure 3-12 shows the average sensitivity ratio of the top 1000 ions over the 100 models. Four ions have ratio values that are significantly higher than the rest of the top 1000 ions for training the ANNs to classify between cell lines with or without $p16^{INK4a}$ mutations. These ions are 1546, 7438, 8905 and 9013 with ratio values of 1.06, 1.05, 1.05 and 1.03 respectively.

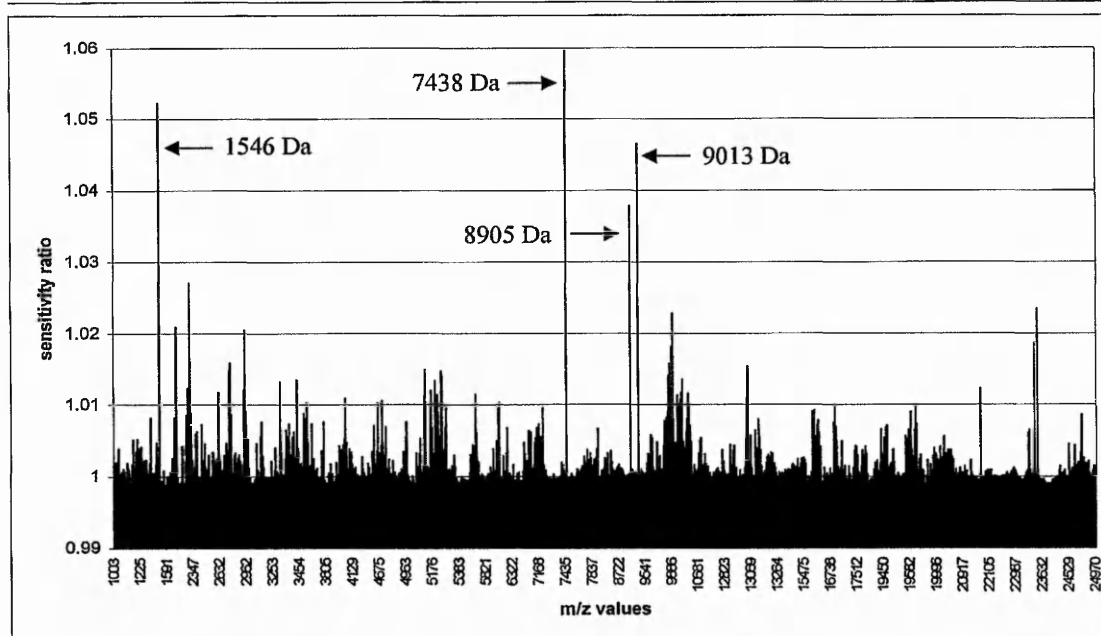


Figure 3-12 Sensitivity analysis of the top 1000 ions used to classify cell lines with or without $p16^{INK4a}$ mutations.

A list of the top 1000 ions used to classify cell lines with or without $p16^{INK4a}$ mutations in each ANN model is obtained using a sensitivity analysis. The data presented shows the sensitivity ratios of the top 1000 ions averaged over the 100 models trained.

3.3.3.3 *Tp53* mutation analysis

The analysis of 85 cell lines for the presence or absence of *tp53* mutations was performed in the same way as the *Braf* and $p16^{INK4A}$ mutation analyses, section 3.3.3.1 and 3.3.3.2. The cell lines that contained a mutation in *tp53* were assigned an output of 1 and cell lines without a mutation in *tp53* were assigned an output of 2.

It was possible to correctly classify 100% of the cell lines in the training set and 98.8% of the cell lines in the blind data set, with 100% sensitivity and 98.4% specificity when 100 ANN models were trained using the top 1000 ions; this is illustrated in figure 3-13. This figure shows that only 1 cell line was misclassified, FM92, as it was given a value of 1.38 when it should have an output of 2.

A sensitivity analysis was performed on the top 1000 ions that were used to train the ANN models in order to determine the ions that are most important for classification of the samples and illustrated as an average over the 100 models trained in figure 3-14.

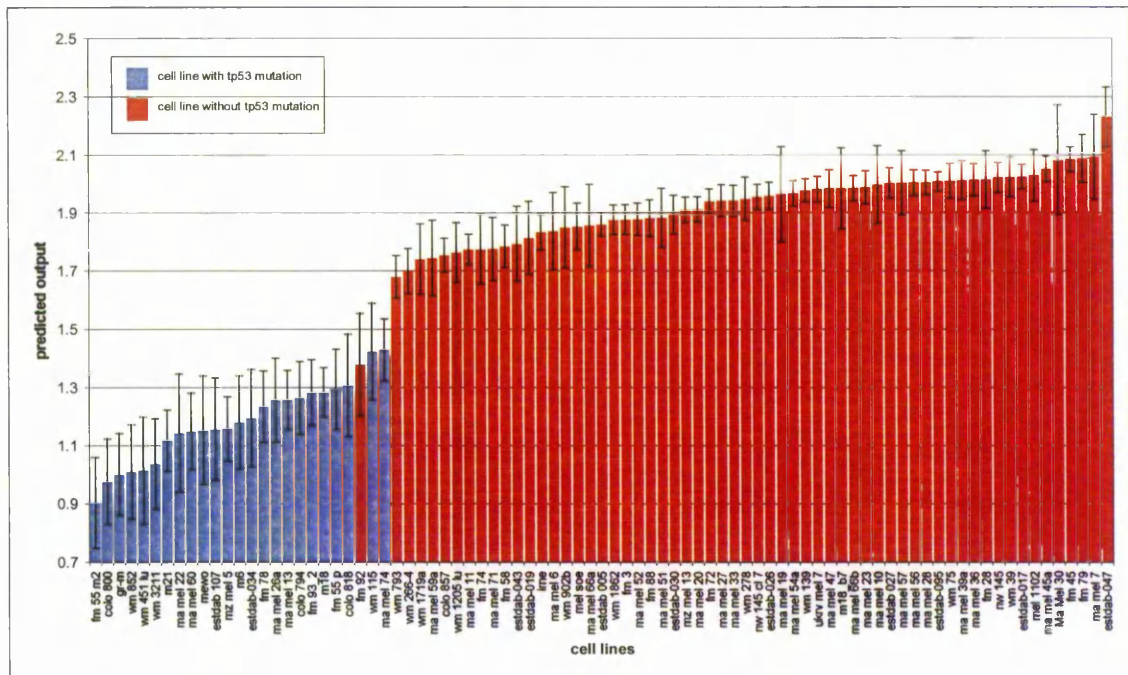


Figure 3-13 Classification by ANN analysis of 85 cell lines with or without *tp53* mutations.

The classification of 85 cell lines as predicted by ANN analysis, averaged over 100 models, is shown above. The blue bars show the cell lines which have a *tp53* mutation and the red bars show the cell lines which do not have a *tp53* mutation. Cell lines with a predicted output of less than 1.5 are classified by the ANNs as having a *tp53* mutation and cells lines with a predicted output of more than 1.5 are classified by the ANNs as not having a *tp53* mutation. The error bars were determined by calculating the standard error of the mean with 95% confidence intervals. Only 1 cell line was misclassified, FM92, it was given a value of 1.38 when it should have an output of 2.

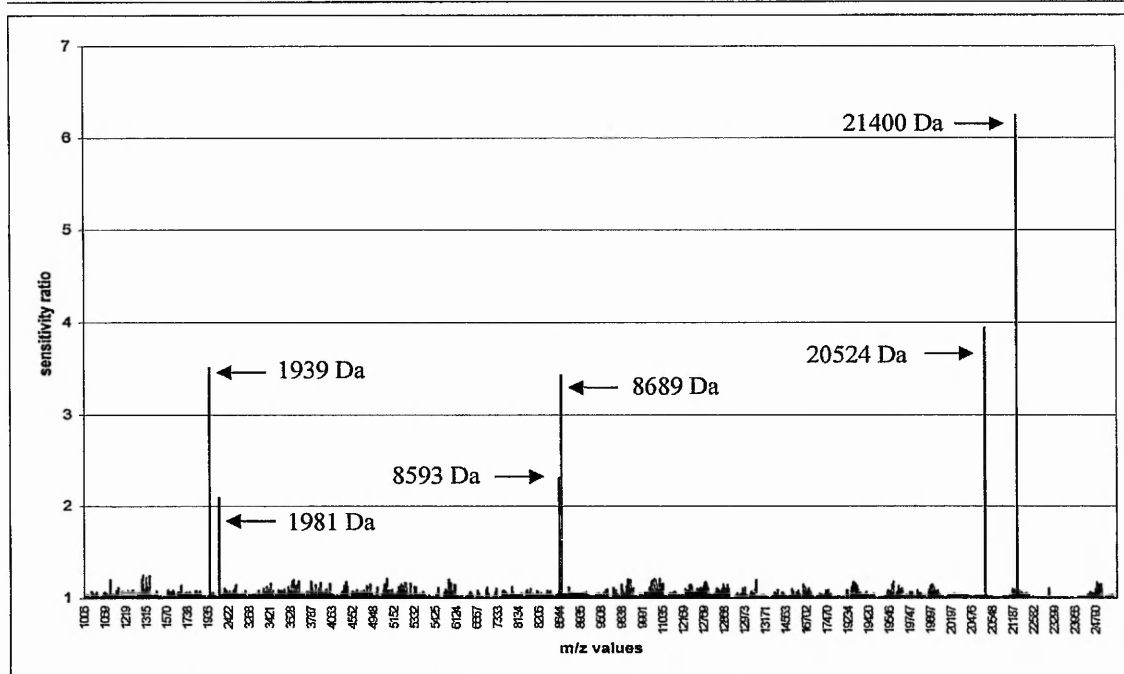


Figure 3-14 Sensitivity analysis of the top 1000 ions used to classify cell lines with or without *tp53* mutations.

A list of the top 1000 ions used to classify cell lines with or without *tp53* mutations in each ANN model is obtained using a sensitivity analysis. The data presented shows the sensitivity ratios of the top 1000 ions averaged over the 100 models trained, where 6 ions in particular have much higher ratios than the rest, therefore being the most important for the classification of these cell lines.

Figure 3-14 shows that 6 ions are particularly important for training the ANNs to classify between cell lines with or without *tp53* mutations, these are 1939, 1981, 8593, 8689, 20524 and 21400 Da. The ratio values of these six ions are quite high, between 2 and 7, showing that their removal from the models causes the error to increase significantly.

A stepwise analysis was also performed using the *tp53* mutation data to determine if a slightly different bioinformatic analysis on the same data set would give better predictive capability, and due to the nature of this analysis may also provide more accurate predictive biomarkers. The method used to perform this analysis is described fully in chapter 2, section 2.2.6.2.

Figure 3-15 illustrates the median accuracy and the mean squared error for the training, test and blind datasets as each input is added to the final model. The model that gave the best predictive performance contained 9 inputs, after this point the addition of further inputs did not result in any significant improvement in accuracies or error. This 9 input model was

able to classify cell lines having a *tp53* mutation and cell lines without a *tp53* mutation with a 92% accuracy; table 3-1 summarises the predictive performance after the addition of each input.

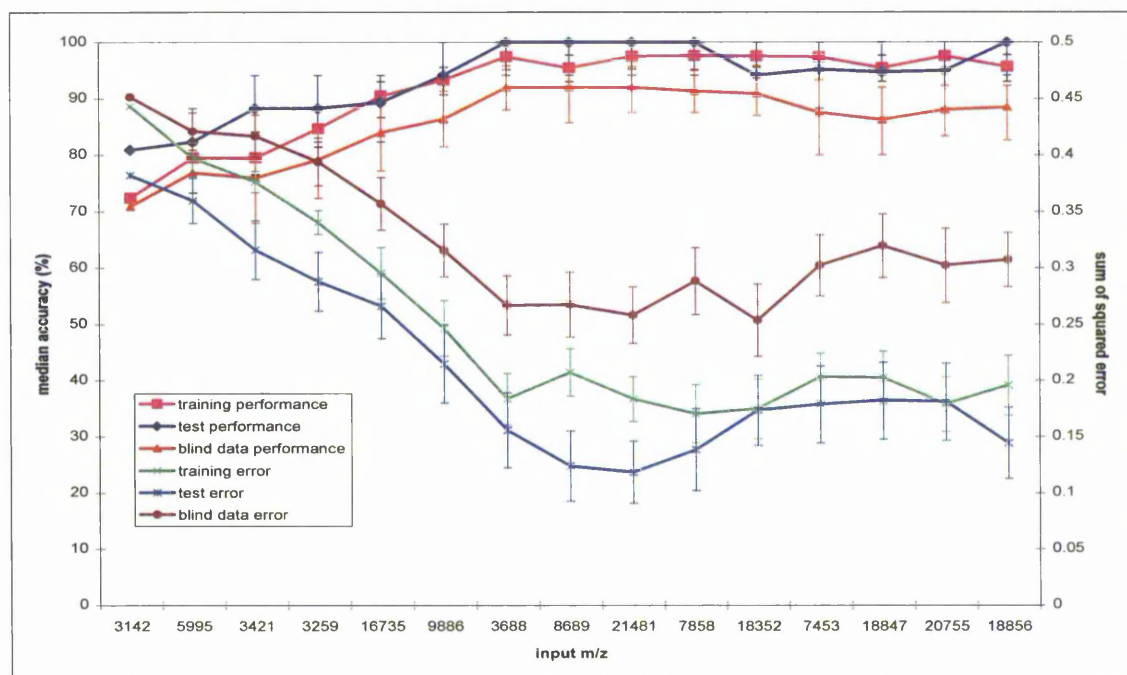


Figure 3-15 Graph showing the model performance with each input addition for the stepwise ANN analysis of the *tp53* mutation data.

Stepwise analysis of the *tp53* mutation data gave a 9 input model that could correctly classify the cell lines with 92% accuracy. The addition of further inputs did not significantly improve the error of the model. The pink line shows the median accuracy for the training data set, with the lower and upper quartile ranges represented as error bars. The dark blue line shows the median accuracy for the test data and the red line shows the median accuracy for the blind data. The green line shows the mean squared error for the predictions at each step with the error bars representing the 95% confidence intervals. The light blue line shows the mean squared error for the test and the purple line shows the mean squared error for the blind data.

Input number	m/z value	Training Accuracy (%)	Test Accuracy (%)	Blind Accuracy (%)	Training Error	Test Error	Blind Error
1	3142	72.4	80.9	70.9	0.4434	0.3825	0.4516
2	5995	79.5	82.4	76.9	0.3977	0.3600	0.4212
3	3421	79.5	88.2	76	0.3766	0.3160	0.4170
4	3259	84.7	88.2	79.2	0.3406	0.2880	0.3943
5	16735	90.5	89.3	84	0.2953	0.2663	0.3572
6	9886	93.3	94.1	86.4	0.2463	0.2147	0.3156
7	3688	97.4	100	92	0.1833	0.1557	0.2668
8	8689	95.5	100	92	0.2071	0.1239	0.2674
9	21481	97.5	100	92	0.1832	0.1184	0.2581
10	7858	97.6	100	91.3	0.1702	0.1383	0.2884
11	18352	97.6	94.1	90.9	0.1745	0.1732	0.2535
12	7453	97.4	95.2	87.5	0.2028	0.1785	0.3022
13	18847	95.5	94.7	86.2	0.2024	0.1819	0.3196
14	20755	97.6	94.9	88	0.1787	0.1808	0.3021
15	18856	95.7	100	88.5	0.1956	0.1442	0.3072

Table 3-4 Summary table of the data shown in figure 3-15

Table showing a summary of the median accuracies and the mean squared error for the training, test and blind data sets as each input is added to the model. The row highlighted in red shows at which point the accuracy and the error failed to improve with subsequent addition of inputs.

3.3.3.4 *Mycoplasma Infection – Validation of results*

A retrospective analysis of the cell line samples used in this study revealed that 17 of the cell lines tested positive for mycoplasma infection. It was necessary to reanalyse the ANN data to confirm that the patterns and top ions identified related to the presence or absence of certain genetic mutations and not to the presence or absence of mycoplasma infection. One way to validate the results was to take the top nine ions identified by the stepwise analysis of the *tp53* mutation data to retrain the ANNs. The mycoplasma infected cell lines were removed from the data set, leaving 68 cell lines, and ions 3142, 3259, 3421, 3688, 5995, 8689, 9866, 16735 and 21481 were the only ions used in the input layer. The following figure shows the classification of these 68 cell lines using only the top 9 ions to train the ANNs.

The correct classification of the cell lines as either originating from group 1, containing a *tp53* mutation, or group 2, without a *tp53* mutation, was achieved for 95.6% of the blind data set with a sensitivity of 89.5% and a specificity of 98% and this is illustrated in figure 3-16. This figure shows that 3 cell lines were misclassified; Ma Mel 74, ESTDAB 034 and

interestingly FM 92, which was also misclassified in the original *tp53* analysis containing all 85 cell lines, shown in figure 3-13.

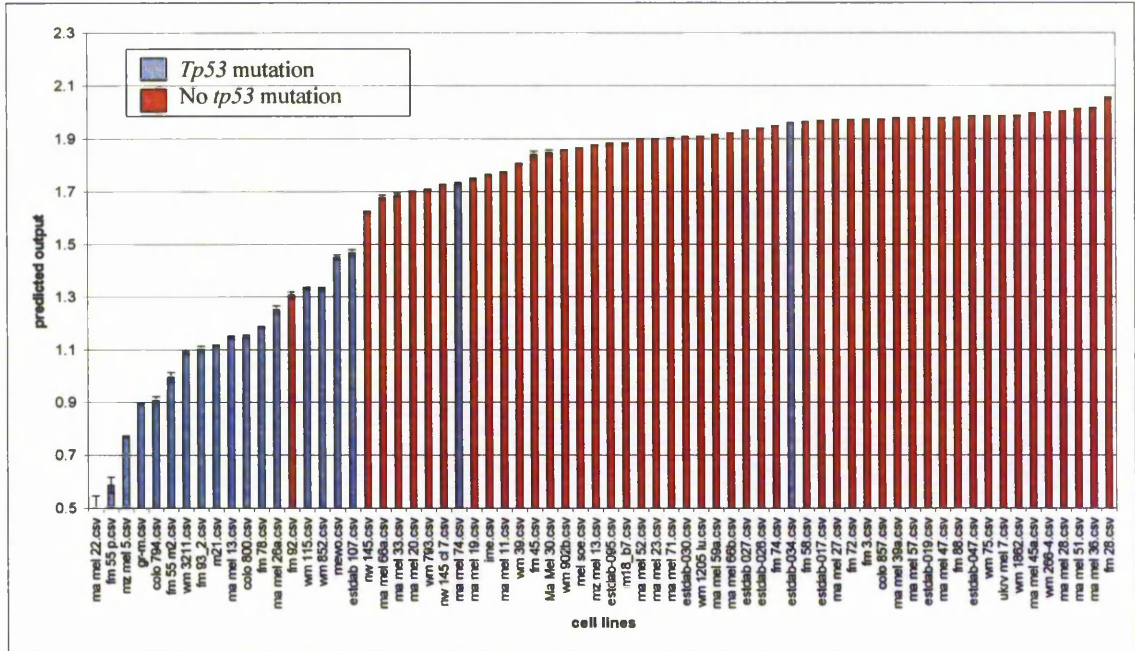


Figure 3-16 Classification of 68 cell lines with or without *tp53* mutations after mycoplasma infected cell lines were removed from analysis.

Figure showing the averaged classification of 68 cell lines with or without a *tp53* mutation as predicted by ANN analysis, using the top 9 ions, identified using the stepwise approach, to train 100 models. The blue bars show the cell lines which have a *tp53* mutation and the red bars show the cell lines which do not have a *tp53* mutation. Cell lines with a predicted output of less than 1.5 are classified by the ANNs as having a *tp53* mutation and cells lines with a predicted output of more than 1.5 are classified by the ANNs as not having a *tp53* mutation. The error bars were determined by calculating the standard error of the mean with 95% confidence intervals. Three cell lines were misclassified; Ma Mel 74, ESTDAB 034 and FM 92 which was also misclassified in the original *tp53* analysis described in section 3.3.3.3.

3.4 Discussion

Melanoma progression can be associated with the presence of genetic mutations; however it would prove beneficial to strengthen these findings with additional information relating to protein expression. The aim of this aspect of the research was to determine whether the proteomic profiles of melanoma cell lines could predict or be associated with the presence or absence of known genetic mutations and therefore further our understanding of events at the functional level once mutations have occurred. The three mutations, *tp53*, *Braf* and *p16^{INK4A}*, investigated have been shown to be upregulated at different stages of melanoma progression (Lomuto *et al.* 2004; Omholt *et al.* 2003).

All of the methods used in this study were optimized prior to sample preparation and analysis. When the cells were at 70% confluency, they were scraped into the media to avoid the effects of trypsin on cellular proteins, washed and then lysed using the lysis buffer. The buffer used for lysis had a high concentration of urea, 9.5M, to ensure the complete denaturation and solubilisation of the proteins, including membrane proteins, and also contained 2% Octyl- β -D-glucopyranoside (OGP) and 1% Dithiothreitol (DTT). OGP is a non-ionic gentle detergent that solubilises proteins whilst maintaining their native subunit structure; DTT is effective in sample buffers for reducing the protein disulphide bonds prior to SDS PAGE. The type of buffer used for lysis was optimized and it was decided that the use of protease inhibitors would be detrimental due to the fact that these enzymes could mask the biomarker profile of the cell lines generated by mass spectrometry. Since the aim of this study was to interrogate the proteome of melanoma cell lines via mass spectrometry, the addition of 'foreign' proteins could cause ion suppression upon mass spectrometry analyses. Protein quantification was performed using the Bio-Rad microassay to ensure that upon analysis, the amount of total protein added to the SELDI chip was similar for each sample. Although the proteomic methods used in this study are not quantitative, the aim was to identify differential intensities of peaks and establish patterns in the data; it was therefore necessary to ensure equal amounts of protein were analysed for each sample. Following quantification of protein within the samples, it was also important to determine sample quality using 1-D SDS PAGE. The samples were run through a 12.5% resolving gel with a molecular weight marker run alongside. The aim was to ensure a consistent banding pattern in all of the samples and to ensure that highly abundant proteins

were not degraded, thereby concluding that the samples were of good enough quality for mass spectrometry analysis.

Figure 3-4 shows the morphology of three of the cell lines in culture that were used in this study and it is clearly visible that there are morphological differences between the cell lines, although these differences could not be related to mutational events or proteomic profiles. 74 of the cell lines showed an octagonal morphology and the remaining 11 cell lines all showed epithelial-like growth and each of the cell lines within these two groups differed further in their morphologies as they varied in size and granularity.

The SELDI ProteinChip method (Ciphergen®) relies entirely on establishing a protein fingerprint to make a correct classification and diagnosis; it is not necessary to know the identity of the masses for this purpose. The SELDI MS analysis used in this study utilized H50 SELDI chips which bind proteins through a hydrophobic chromatography and have binding characteristics similar to that of a C6 to C12 alkyl chromatographic resin resulting in the binding of hydrophobic proteins. An important property of these affinity arrays, aside from their chromatography, is their specific binding of hydrophobic proteins from a complex biological sample. This means that only biomolecules whose properties match the binding characteristics of the surface, in this case hydrophobic, are retained and all other proteins are washed away. Another advantage of these chips is that contaminants, such as salts or detergents that can interfere with MS analysis, can be removed prior to processing simply by washing with deionised water. Also, due to the specificity of the chip surfaces, certain characteristics of the proteins identified are known. By using different washing conditions, the basic physico-chemical characteristics of the proteins can be determined. One disadvantage is that not all of the bound proteins can be “visualized” equally well. Although proteins below the 30000 Da mass range can be resolved, sensitivity for higher molecular weight proteins is low resulting in fewer signals at this higher range. Another disadvantage is that this technique is restricted towards the investigation of differential expression of proteins rather than their identification, which is important to establish protein identity and potentially new diagnostic and/or therapeutic targets (Seibert *et al.* 2004). Data was analysed from the 1000 – 25000 Da range due to the contamination of matrix peaks at lower mass ranges and the reduced sensitivity and resolution at the higher masses, as previously mentioned.

There has been much controversy recently concerning the use of SELDI MS, particularly the reproducibility of this method (Diamandis 2004b). There are a number of steps involved in SELDI MS analysis, including sample collection and preparation, protein chip selection and preparation, matrix selection and application, calibration, sample loading onto the chip, procedures for washing off unbound proteins, parameter setting of the mass spectrometer and data processing. This extensive and complex process can greatly reduce the reproducibility of SELDI MS and the introduction of chemical noise caused by matrix clusters can also be an important factor leading to variation of profiles. In quality control experiments it has been reported that the coefficients of variation (CV's) for absolute intensities of peaks can be as much as 50-60% (Yasui *et al.* 2003); thus quality control procedures are important to ensure reproducibility. Coombes *et al.* (2003) pooled samples of nipple aspirate fluid and applied this to 2 spots each on 3 IMAC SELDI chips on 4 successive days. Using a basic ANOVA to analyse the data from 356 peaks it was revealed that spot-to-spot variation was larger than day-to-day variation which was in turn much larger than chip-to-chip variation (Coombes *et al.* 2003). Qu *et al.* (2002) also investigated the reproducibility of SELDI MS by pooling serum from a healthy donor and spotting this onto 2 spots on each IMAC chip. By selecting 7 peaks from each spectra reproducibility was proven with the CV's for peak location of 0.05% and CV's of 15-20% for peak intensities quality controlled samples (Qu *et al.* 2002). Zhukov *et al.* (2003) investigated the reproducibility of SELDI MS by the analysis of multiple samples of normal and tumour cells from one patient with squamous cell carcinoma of the lung. These samples were run on IMAC chips 5 times using identical protocols and the mean mass and standard deviation for the detected protein peaks were calculated; the CV's ranged from 0.02 – 0.25% for peak location, showing acceptable reproducibility of protein detection (Coombes *et al.* 2003; Drake *et al.* 2004; Qu *et al.* 2002; Zhukov *et al.* 2003). More recently, Hong *et al.* (2005) investigated the reproducibility of SELDI MS by examining 144 pooled plasma samples. These were randomly placed on the spots of 12 IMAC chips placed in each of six bioprocessor plates. The average intensity and standard deviation of 5 peaks present in all 144 spectra was calculated and the CV's were found to be less than 20% (Hong *et al.* 2005). All of the quality assurance experiments described above were performed using the IMAC chips whereas the chips used in the present study were hydrophobic chips, where the quality assurance experiments performed involved the rigorous generation of samples from

all the melanoma cell lines used. The samples were cultured and lysates produced by a standard methodology. Protein quantification was performed using the Bio-Rad protein microassay and the sample quality determined by 1-D SDS PAGE using a 12.5% resolving gel. The buffer used in all experiments was from a single 500ml batch, frozen down in 1.5 ml aliquots. All of the samples analysed were thawed and diluted in buffer on the day of SELDI MS analysis, and all of the solvents and matrix used were prepared fresh on the day of analysis. Before analysis the SELDI MS was calibrated using Ubiquitin, Cytochrome C and Thioredoxin, made up separately in large batches giving over 100 aliquots of each calibrant, these were then stored at -80°C. The calibrants were diluted in matrix on the day of analysis, and any unused calibrant was discarded. Once calibrated, calibrant was placed onto 1 spot of every chip used, and the reproducibility from chip-to-chip within each experiment was determined by calculating the CV's for the peak location and peak intensities of 5 of the calibrant peaks. The CV's for peak location was between 0.05 – 0.12% and for the peak intensities the CV's were between 30-70%, see appendix 5 for data and were the same as those published by Zhukov *et al.* (2003) and the CV's for peak intensity were similar to those presented by Yasui *et al.* (2003), although the variation was greater. Within each experiment, every sample was spotted in duplicate on separate chips to give an indication of the reproducibility of the complex samples analysed. The profiles were assessed visually and the reproducibility of the protein patterns evaluated, (figure 3-5).

SELDI MS analysis was used to profile the protein expression of melanoma cell lines with different genetic mutations. Although visual differences were observed in the protein profiles, due to the dimensionality of the data and the number of samples analysed, a bioinformatics approach was used to mine the data using ANNs. The ANNs were able to correctly classify cell lines with or without particular mutations with over 96% accuracy for unseen data. The sensitivity analyses show that the majority of the top 1000 ions have sensitivity ratio, meaning that these ions have a positive influence on the neural network being able to classify between a cell line with or without a particular mutation. This shows that the presence of a genetic mutation not only affects the protein(s) transcribed from that gene but also proteins of other genes (figures 3-10, 3-12 and 3-14).

The ANNs correctly classified cell lines with or without *Braf* mutations with 96.5% accuracy for unseen data (n=85) with 3 cell lines (Ma Mel 39a, FM-3 and FM 93/2) that were misclassified. The sensitivity analysis revealed three ions that had ratio values much higher than the rest with *m/z* values of 6985, 13234 and 21187. In addition, the ANNs correctly classified cell lines with or without *p16^{INK4A}* mutations with 97.2% accuracy for unseen data (n=71) with 2 cell lines (FM 92 and WM 39) that were misclassified. The sensitivity analysis revealed four ions that had ratio values much higher than the rest with *m/z* values of 1546, 7438, 8905 and 9013. These analyses can also reveal outliers (those samples misclassified) which could reveal a previously unknown disease subtype. The ANNs correctly classified cell lines with or without *tp53* mutations with 98.8% accuracy for unseen data (n=85) with only one cell line (FM 92) misclassified. The sensitivity analysis revealed six ions that had ratio values higher than the rest that represent potentially important markers; these ions had *m/z* values of 1939, 1981, 8593, 8689, 20524 and 21400. The stepwise analysis is a much more accurate process for identifying potential biomarkers as each ion is modeled in turn and for cell lines with and without a *tp53* mutation revealed a 92% accuracy that could be achieved with only 9 ions. This demonstrates that the standard method of ANN analysis is only useful in determining whether the protein profiles reveal patterns that allow classification of unseen samples, whereas the stepwise analysis can also suggest important biomarker ions.

The presence of mycoplasma infection in some of the samples originally analysed led to the question of whether the correct classification of samples was due to patterns indicating presence of mycoplasma and not the presence of genetic mutations. The data from the infected samples was removed and the remaining data reanalysed using the ions identified as important for classifying samples according to the presence or absence of *tp53* mutations as inputs. This is also a good way to validate the original findings and this validation revealed that samples could be correctly classified with 95.6% accuracy and so the patterns and important ions identified in the original analysis were as a result of the presence or absence of mutations and not mycoplasma infection. Due to time constraints, it was not possible to reanalyse all of the data but this validation gives confidence to the original results obtained.

The high prediction rate of the *tp53* mutation data in both of the ANN approaches used has shown that there are sufficient patterns within the data to predict with a high degree of accuracy whether a cell line carries a mutation. The two approaches did, however, reveal different important ions for classifying the samples with only 1 ion (8689 Da) that was identified using both methods. It is important to note, therefore, that in future studies the use of different bioinformatics approaches may reveal different accuracies in the classifications of samples as well identifying different important ions. This has been shown using a prostate cancer dataset with 197 prostate cancer (PCA) patients, 92 benign prostate hyperplasia (BPH) patients and 96 age-matched unaffected healthy men, which was mined using 3 different bioinformatics approaches. These 3 approaches randomly split the data into 85% for the training set and 15% for the independent test set. Adam *et al* (2002) used “area under the ROC curve” to reduce the number of inputs and then used this data to develop a decision tree classification algorithm that used 9 masses to correctly classify 96% of the samples with a sensitivity of 83% and a specificity of 97%. Qu *et al* (2002) used a boosting tree algorithm to develop a classifier for separating prostate cancer samples from non-cancer. The first classifier, AdaBoost, managed to separate the samples with 100% sensitivity and specificity. The second classifier, the Boosted Decision Stump Feature Selection classifier, was easier to interpret and managed to separate samples with 97% sensitivity and specificity. Yasui *et al* (2003) combined boosting with linear discriminant analysis to analyse the data and managed to correctly classify PCA/BPH from controls with 98% sensitivity and 100% specificity.

The noisy and highly dimensional data obtained through proteomic analysis requires all aspects of data interrogation to be optimised including spectra pre-processing, quality control and dimensionality reduction. The biggest challenge comes from the development of algorithms that can generalise well enough for unseen data. There are many data mining methods that can be employed and, presently, no one method can provide the most accurate and reliable analysis.

The cell lines used in this study are mostly derived from metastases and these mutations are therefore associated with metastatic disease. It should be considered that changes in both the proteome and genome may have occurred during *in vitro* culture and hence there are likely to be differences between the original tissue and the cell line. However, due to the limited availability of tumour tissue, cell lines are important tools for many experimental

studies and proof of concept. Genetic mutations will lead to a number of changes in gene and protein expression through downstream effects. It has been documented that the differences in *Braf* mutation frequency between various subtypes of melanoma and benign nevi make it difficult to determine the impact on prognosis (de Snoo & Hayward 2005). Some studies have reported that there is no difference between mutation frequency and disease free survival whereas others show it to be a major predictor of survival (Deichmann *et al.* 2004; Kumar *et al.* 2003; Shinozaki *et al.* 2004; Thomas *et al.* 2004). The investigation into the downstream effects of *Braf* mutations such as the up or down regulation of proteins may provide further understanding of the biological effect of this mutation which potentially may lead to the identification of a reliable prognostic marker which could predict survival.

Cell lines have been used in several studies in an attempt to identify diagnostic and prognostic markers. One study investigated etoposide resistance in human neuroblastoma cell lines as a first step towards the development of novel prognostic markers of neuroblastoma chemotherapy. The study reported the overexpression of a number of proteins in the etoposide resistant cell line, paving the way for future *in vivo* studies (Urbani *et al.* 2005). A recent study using melanoma cell lines revealed several proteins that were induced in response to interferon α , some of which had not been previously reported to be type I interferon responsive (Craven *et al.* 2004). A large scale study into the investigation of markers that could be used to distinguish between ovarian and colon carcinoma was performed using 60 human cancer cell lines, which were at the protein, DNA and mRNA levels and revealed 2 candidate biomarkers, villin and moesin, that improved upon the accuracy of discriminating between the two cancers. Using tissue microarrays to validate the findings obtained with the cell lines, it was revealed that the anti-moesin antibody did not stain colon cancer cells but did stain the surrounding stromal cells and hence may have not been recognised as a marker originally if tumour tissue alone had been profiled (Nishizuka *et al.* 2003). These studies have revealed how cell lines systems can be used to provide useful information on the disease of interest, even at the protein level.

The present study provides an insight into protein markers identified by SELDI MS, that associate with important cancer-related gene mutations and has important implications for the discovery of proteins that are of potential diagnostic, prognostic and therapeutic benefit.

Chapter 4 Proteome profiling and bioinformatics analysis of paired melanoma cell lines and tissues

4.1 Introduction

There are several clinical and pathological factors that influence the progression and behaviour of melanoma including tumour thickness, ulceration, site of the primary tumour, age, gender, level of invasion, mitotic rate, tumour infiltrating lymphocytes (TILs), histological regression, microscopic satellites, histological subtype, vascular invasion and tumour cell type. The thickness of the tumour, ulceration and site of primary tumour are factors that are used to predict survival in patients with stage I and II melanoma but most studies to date have shown that the depth of tumour invasion is the most powerful independent prognostic factor, although even this is not always accurate (Li *et al.* 2002b). Therefore additional objective prognostic markers are needed with improved specificity and predictive value for melanoma patients.

Tumour cells are embedded in a matrix of structural extracellular proteins that are surrounded by other cells including endothelial cells, fibroblasts, inflammatory cells and immune cells (Bissell *et al.* 2001). These cells make up the tumour tissue and continuously interact with their neighbours. Collectively, these cells generate signals that can determine the growth and survival of the tumour and the tumour can send signals that can remodel the stroma to suit its changing needs (Ruiter *et al.* 2002). Signalling molecules generated by the tumour cells include, among others, basic fibroblast growth factor (bFGF), members of the vascular endothelial growth factor (VEGF) family, platelet derived growth factor (PDGF), epidermal growth factor receptor (EGFR) ligands, interleukins, colony stimulating factors, transforming growth factor- β (TGF β) among others. These factors can disrupt normal tissue homeostasis and can induce stromal reactions such as angiogenesis and the inflammatory response. These molecules also activate surrounding cells such as fibroblasts, smooth muscle cells and adipocytes causing the secretion of additional growth factors, promigratory extracellular matrix components as well as the upregulation of expression of serine proteases and matrix metalloproteinases that degrade and remodel the extracellular matrix. The induction of the inflammatory responses in the tumour stroma

also results in the production of factors that promote tumour progression (Mueller *et al.* 2004). A predominant stromal cell is the fibroblast which is responsible for the creation of most of the components of the connective tissue such as the different collagens, proteolytic enzymes and their inhibitors as well as growth factors. Other functions of the fibroblast include the deposition of the extracellular matrix, the regulation of epithelial differentiation, regulation of inflammation and involvement in wound healing (Parsonage *et al.* 2005). As each tissue has specialised requirements, fibroblasts from different organs produce different variations of the above mentioned molecules. Also, the fibroblasts in the stroma change according to different physiological signals. Evidence from a number of studies has revealed that the fibroblasts at the site of a tumour may have a role in the initiation and progression of cancer (Tlsty 2001) and are known as “carcinoma associated fibroblasts” (CAF); they are spindle like mesenchymal cells that share characteristics with smooth muscle cells and normal fibroblasts. The presence of CAFs in tumour stroma has been seen in many cancer types including breast, prostate and skin cancer (Chauhan *et al.* 2003; Olumi *et al.* 1999; Skobe *et al.* 1998).

It is questionable if cell lines are representative of the general cancer cell population found in tumour tissue and whether they can be used in proteomic studies to identify key biomarkers that can discriminate between different disease states. Wistuba *et al.* 1998 have reported that a majority of non small cell lung cancer cell lines retained the properties of their parental tumours; it has also been shown that cell lines derived from primary tumours or bone marrow metastases retain many of the features of the primary tumour (Ross *et al.* 2003). The RNA and protein signature of tumour cell lines can show similar patterns to those of the original tumour tissue (Myers *et al.* 1997; Ross *et al.* 2000; Scherf *et al.* 2000). An advantage of the use of cell lines in proteomic studies is that large numbers of cells are available for study. Clinical tumours contain endothelial cells, infiltrating lymphocytes, fibroblasts and other stromal cells as well as tumour cells, which may represent only a fraction of the overall tissue sample, and so the concentration of potential tumour specific proteins in the sample may be small. Nishizuka *et al.* (2003) used genomic, transcriptomic and proteomic profiling of colon and ovarian cancer cell lines to identify two new biomarkers that were present in tissue sections, thus showing their utility as relevant biomarkers. One of the biggest challenges in cancer proteomics is the complexity of the

tissue microenvironment. The protein networks within a cell are constantly changing and are dependent on the local microenvironment (Liotta *et al.* 2001a). One problem, therefore, in using cell lines is that the tumour microenvironment plays an important role in cancer and so the lack of cell heterogeneity in cell line studies could be a disadvantage.

The aim of this study was to compare the proteome of melanoma tissue and paired cell lines derived from 44 patients; SELDI MS profiles were derived and ANNs used to investigate the data.

4.2 Methods

Detailed methodology is given in Chapter 2 and the following scheme demonstrated how the cell line and tissue samples were processed for mass spectrometry and ANN analysis (figure 4-1). The clinical data obtained on the patients from which the cell lines and tissues were derived can be found in appendix 2.

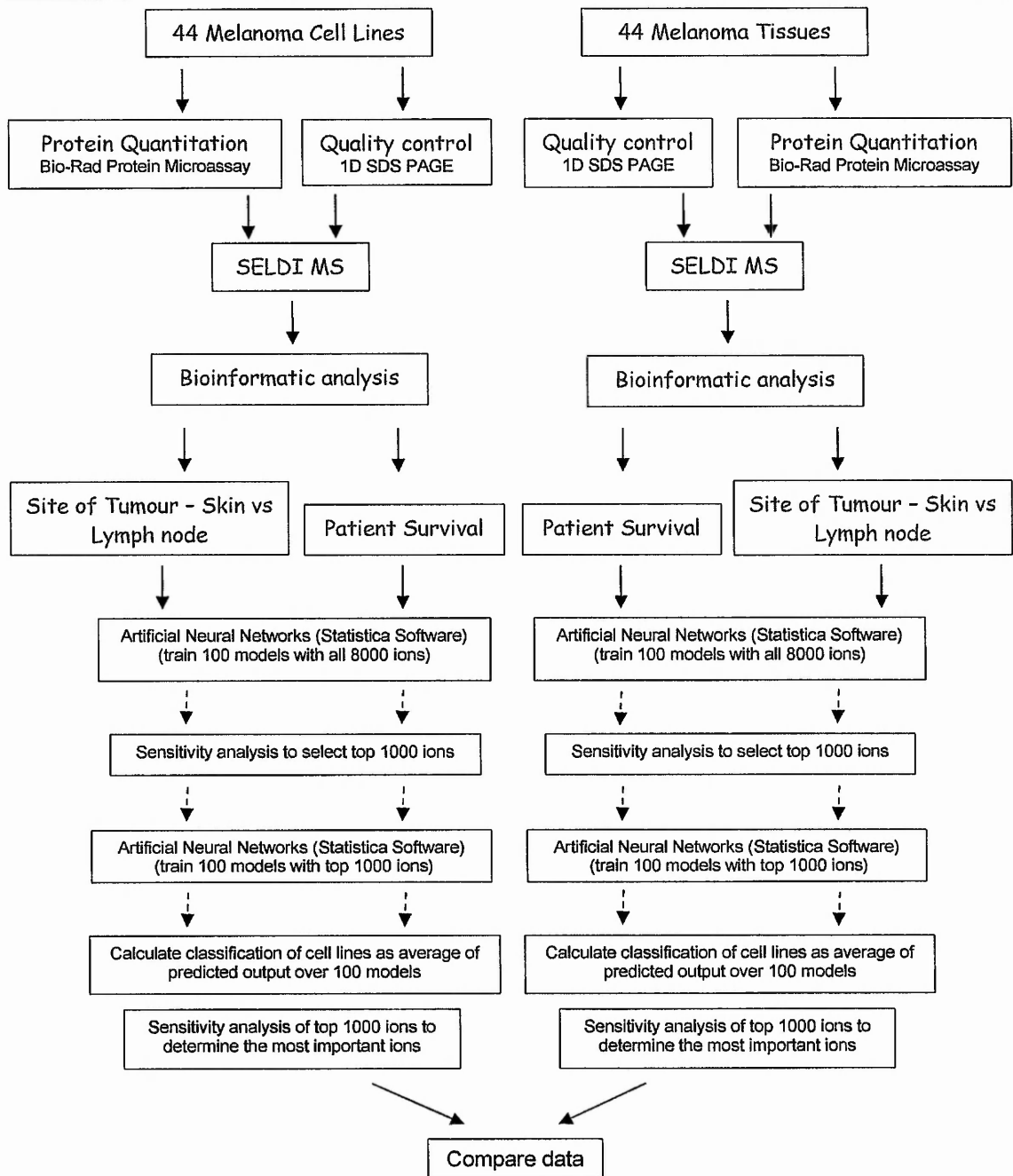


Figure 4-1 A schematic demonstrating how the cell lines and tissues were processed and analysed for this study.

4.3 Results

4.3.1 Protein quantification and sample quality control.

Protein quantification was performed using the Bio-Rad protein microassay as described in section 2.2.2. A standard curve was determined by plotting the absorbance at 595nm versus BSA protein standard concentration (μg). The calculation of the protein concentrations is explained in more detail in section 3.2.1. On average, concentrations of between 3 – 7 $\mu\text{g}/\mu\text{l}$ were obtained from the cell line lysates and concentrations of between 7 – 11 $\mu\text{g}/\mu\text{l}$ were obtained for the tissue lysates. The samples were then run through a 12.5% resolving gel to determine if sample integrity had been maintained, examples of the gels run are shown in figure 4-2. This figure shows a representation of the banding pattern observed with all of the samples that were run through the gel. As all of the samples showed very similar bands it was concluded that the samples were of good quality for further analysis and none of the samples were rejected at this stage.

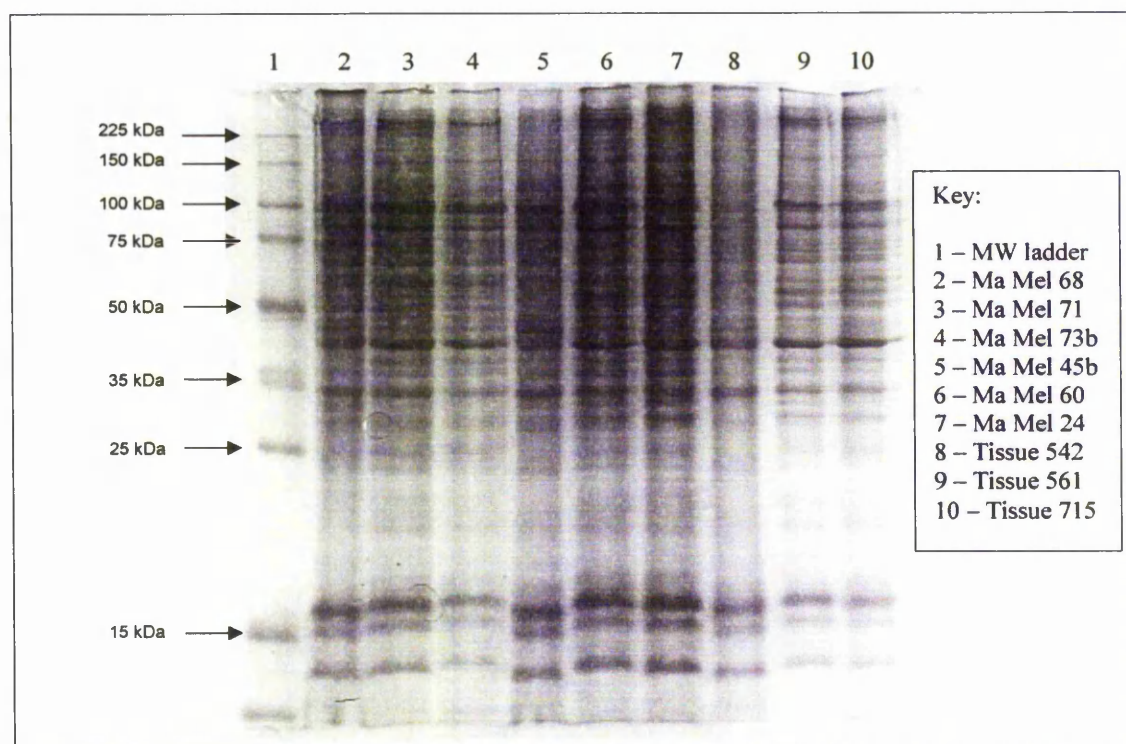


Figure 4-2 1-D PAGE of cell line and tissue lysates representing sample quality of all samples analysed as part of the cell lines vs tissue study.

The gel presented above represents an example of the data obtained from all of the cell line and tissue lysates used in this study. 40 μg of each lysate sample was resolved in a 12.5% resolving gels and stained with coomassie blue to determine if degradation of the proteins occurred during sample preparation.

4.3.2 SELDI MS

The protocol used is described in section 2.2.4.1, but briefly the samples were bound to the hydrophobic surface on a H50 SELDI chip and analysed using a PBS II mass analyzer. The following figures show some of the SELDI spectra obtained for this study. Figure 4-3 highlights some of the observed differences, shown by the blue boxes, and similarities, shown by green boxes, between the cell lines and their paired tissues: Profile a) for cell line Ma Mel 28 overlaid on its paired tissue (284), b) for cell line Ma Mel 36 overlaid on its paired tissue (304) and c) for cell line Ma Mel 48a overlaid on its paired tissue (414). Table 4-1 shows peak values of the profiles shown in figure 4-3, with peaks that are similar between the cell lines and their paired tissues highlighted in red; peak values at around 5650 and 6545 Da are seen in all of the spectra. Although cell line samples derived from the tissue samples show many similarities, the spectra obtained from the tissues is much more complex with a greater number of peaks observed in the 5000-15000 Da range. This may reflect protein ions derived from stromal as well as melanoma cells and the altered protein expression pattern of melanoma cells *in situ* compared with cultured cells.

The protein ions of melanoma cell lines derived from the skin or lymph node of different patients were compared (figure 4-4a and 4-4b). Similarly melanoma tissue samples taken from the skin or lymph node of different patients were profiled by SELDI MS in order to determine if melanoma tumours (and cell lines derived from those samples) present at different sites (skin or lymph node) display different protein spectra. Table 4-2 shows all of the peaks that can be seen in all of the profiles shown in figure 4-4, with peaks that are similar between the cell lines and their paired tissues highlighted in red.

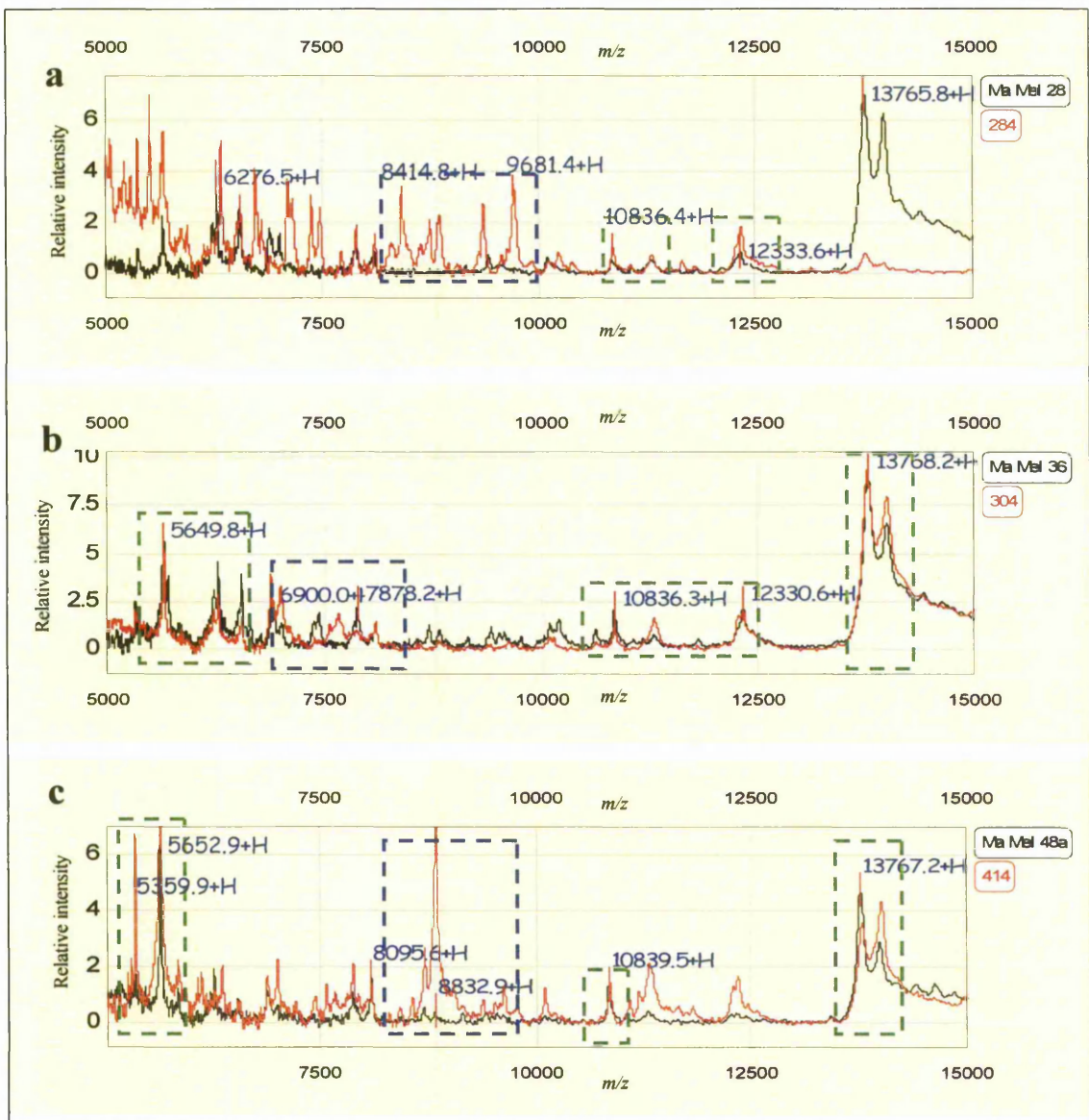


Figure 4-3 SELDI MS profiles in the 5000-15000 Da mass range showing the profiles obtained from a cell line are overlaid on the profile of the parent tissue.

This figure represents how the spectra of cell lines and their paired tissues compare with each other. a) shows cell line Ma Mel 28 overlaid on its paired tissue 284, b) shows cell line Ma Mel 36 overlaid on its paired tissue 304 and c) shows cell line Ma Mel 48a overlaid on its paired tissue 414. The green boxes show the similarities between the cell lines and their paired tissue and the blue boxes show the differences.

Pair 1		Pair 2		Pair 3	
Ma Mel 28 (cell line)	284 (tissue)	Ma Mel 36 (cell line)	304 (tissue)	Ma Mel 48a (cell line)	
-	5040.7	5316.5	-	-	5361.9
-	5202.1	-	5356.7	5652.9	5654.6
5359.1	5359	5649.8	5649.8	5862	5861
-	5499.4	6276.3	6274.8	-	6125
5648.3	5654.9	6544.5	6543.7	6276.7	-
-	5868.7	6900	6885.3	-	6367.9
6276.5	-	-	6997.7	6552	6550
-	6329.1	7441.9	-	6882.8	6890.6
6543.8	6550.3	-	7666.5	-	7011.4
-	6724.1	7878.2	-	-	7220.8
6886.8	-	8086.9	8087.7	-	7449.1
7003.7	-	8700.4	-	-	7576.1
-	7102.6	8824	-	7877.3	7888.5
-	7370.6	9150.8	-	8095.6	8100.1
-	7474	9412.2	-	-	8713.9
-	7717.5	-	10138.5	8765.2	-
7879.6	7888.9	10195.8	-	8832.9	8842.2
8091.8	8103.3	10617.5	-	-	9628.5
-	8408.1	10836.3	10834.4	-	10100.5
-	8735.7	11287.5	11298.1	10839.5	10851.7
-	8853.8	12330.6	12330.6	11043.6	11083.4
-	9351.3	13768.2	13764.8	11298.5	11314.6
9409.3	-	13987	13990	-	12346.5
-	9704.5	-	-	13767.2	13785.2
10091.9	-	-	-	13986.7	14008.3
-	10219.9	-	-	14633.8	-
10836.4	10848.2	-	-	-	-
11298.7	11310	-	-	-	-
-	11656.3	-	-	-	-
12333.6	12342.1	-	-	-	-
13765.8	13780.5	-	-	-	-
13988.7	-	-	-	-	-

Table 4-1 Values of the main peaks observed in the spectra shown in figure 4-3.

The table shows the main peaks (values given in Daltons) observed in the spectra obtained from paired cell lines and tissues in the 5000-15000 Da mass range. Peaks that are similar between each pair of samples are shown in the same row, although the m/z values may be slightly different due to the mass accuracy of the instrument. Peaks that are observed in both the spectra of cell lines and their paired tissue are highlighted in red.

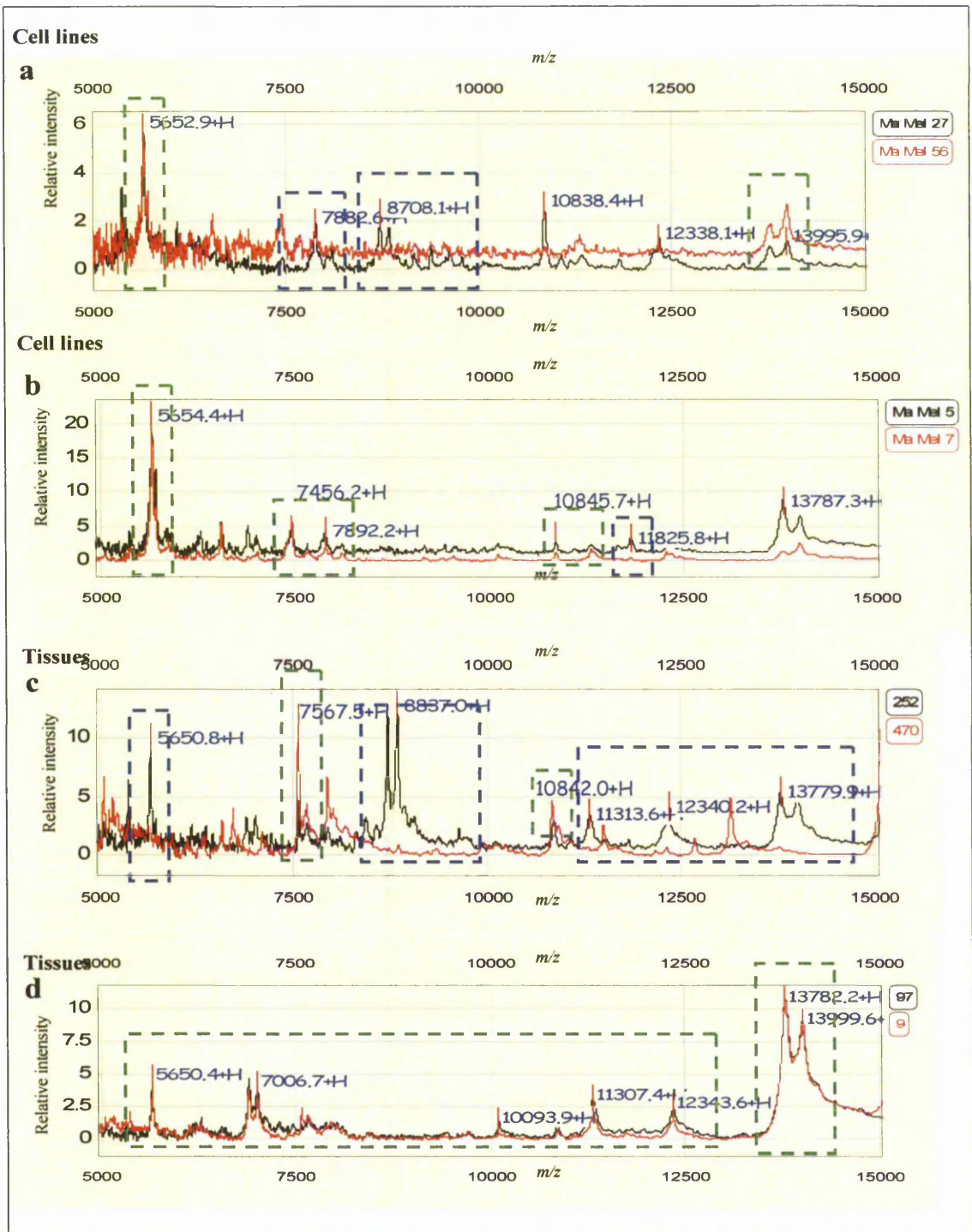


Figure 4-4 SELDI spectra showing the differences and similarities between melanoma cell lines and tissues obtained from skin or lymph node.

Spectra of samples originating from the skin are shown in black and spectra of samples originating from lymph nodes are shown in red. a) shows cell line Ma Mel 27 (black) overlaid on cell line Ma Mel 56 (red), b) shows cell line Ma Mel 5 (black) overlaid on Ma Mel 7 (red), c) shows tissue 252 (black) overlaid on tissue 470 (red) and d) shows tissue 97 (black) overlaid on tissue 9 (red). The green boxes show the similarities between the paired spectra and the blue boxes show the differences.

Cell lines				Tissues			
Ma Mel 27 (from skin)	Ma Mel 56 (from lymph node)	Ma Mel 5 (from skin)	Ma Mel 7 (from lymph node)	252 (from skin)	470 (from lymph node)	97 (from skin)	9 (from lymph node)
5362.5	-	4121	-	5360.1	5040.3	-	5359.9
5652.9	5651	4309.1	-	5650.8	5157.2	5650.4	5653.1
5857.7	-	4777.9	4781.4	-	6119.5	6280.6	6281.3
6082.5	-	5654.4	5655.1	6180	-	6896.3	6898.3
-	6549	5861.4	-	6347.4	6343.8	7006.7	7011.1
6724.7	-	6283	-	-	6573.5	-	7572.2
7457.1	7439.5	6551.1	6551.2	-	6713.1	7678.4	7651.5
-	7678.4	6896	-	6891.1	-	-	7944.3
7882.6	-	7009.2	7010.1	7007.4	-	-	8418.7
8093.6	-	7456.5	7452.1	-	7090.3	-	9695.2
8708.1	-	-	7664.9	7567.5	7559	10093.9	-
8828.1	-	7892.2	7891	7671.8	7663.2	10844.6	10849.3
9151.4	-	8108.5	8095.5	7882.5	-	11307.4	11312.6
9368.7	-	10845.7	10856.1	-	7930	12343.6	12344.8
9614.8	-	11308.6	11310.7	8434.7	-	13782.2	13783.1
10838.4	-	11825.8	-	8708.6	-	13999.6	14006.8
11063.1	-	12355.1	12347.1	8837	8847.6	-	-
11335.1	11304.3	13787.3	13786.4	9624.9	-	-	-
11815	-	14009.6	14005.4	10842	10824.6	-	-
12338.1	-	-	-	11069.9	-	-	-
13769.9	13784.1	-	-	11313.6	-	-	-
13995.9	13996.6	-	-	-	11488.5	-	-
-	-	-	-	-	12307.1	-	-
-	-	-	-	12340.2	-	-	-
-	-	-	-	-	12674.2	-	-
-	-	-	-	-	13137	-	-
-	-	-	-	13779.9	-	-	-
-	-	-	-	13998.1	-	-	-

Table 4-2 Values of main peaks observed in spectra shown in figure 4-4

Table showing the main peaks (in Daltons) observed in the spectra obtained from cell lines and tissues originating from either the skin or the lymph nodes, in the 5000-15000 Da mass range. Peaks that are similar between each pair of samples are shown in the same row, although the m/z values may be slightly different due to the mass accuracy of the instrument. The red values indicate peaks of similar masses between the skin and lymph node samples

In figure 4-5 profiles a and b show the spectra, in black, of cell lines derived from patients that survived less than 8 months, overlaid on spectra, in red, of cell lines derived from patients that survived greater than 8 months. Profiles c and d show the spectra of melanoma tissue taken from patients that survived less than 8 months (black), overlaid on spectra of melanoma tissue taken from patients that survived longer than 8 months (red). Profile a) shows cell line Ma Mel 57 derived from a patient surviving less than 8 months overlaid on cell line Ma Mel 39a, profile b) shows cell line Ma Mel 62 derived from a patient surviving less than 8 months overlaid on cell line Ma Mel 48a, profile c) shows tissue 476 overlaid on tissue 9 and profile d) shows tissue 374 overlaid on tissue 567 taken from a melanoma tumour in the lymph node. Table 4-3 shows all peaks that are similar between the cell lines and their paired tissues, highlighted in red.

It should also be noted that peak values at approximately 5650 Da, 10840 Da, 13780 Da and 14000 Da are present in all of the samples shown in figures 4-3, 4-4 and 4-5 regardless of the sample type or origin.

The similarities and differences highlighted in figures 4-3, 4-4 and 4-5 and tables 4-1, 4-2 and 4-3 represent only those cell lines shown, and may not be the same for other cell lines. The identification of these peaks was performed visually and therefore does not reflect the complexity of the cell line or tissue proteome. Further analysis was performed using ANNs in order to interrogate the entire spectra of all the cell lines and tissues available.

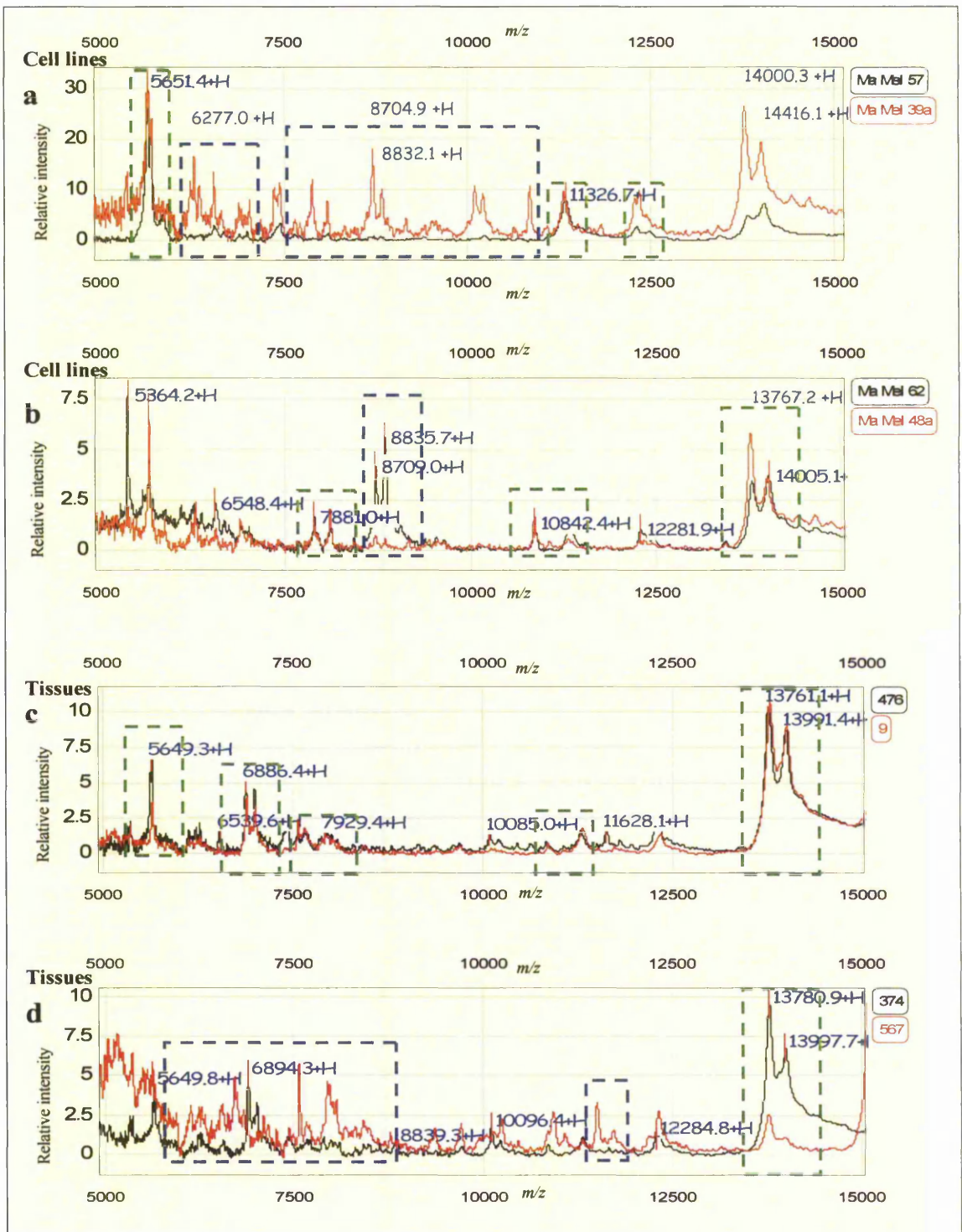


Figure 4-5 SELDI spectra showing the differences and similarities of cell lines and tissues taken from patients with a survival, after sample collection, of less than or more than 8 months. All four spectra show a profile of a sample taken from a patient with a survival time of less than 8 months (in black) and a profile of a sample taken from a patient with a survival of more than 8 months (in red). The green boxes show the similarities between the samples and the blue boxes show the differences.

Ma Mel 57 (<8m survival)	Cell lines			Tissues			
	Ma Mel 39a (>8m survival)	Ma Mel 62 (<8m survival)	Ma Mel 48a (>8m survival)	476 (<8m survival)	9 (>8m survival)	374 (<8m survival)	567 (>8m survival)
-	5362.3	5364.2	5359.9	5358.8	5359.9	5317.2	-
5651.4	5653.8	5653.6	5652.9	5649.3	5653.1	-	5357.3
5849.6	-	-	5862	6270.5	6281.3	-	5626.7
-	6277	6085	-	6539.6	-	5649.8	-
6545.3	6548.1	6255.8	6276.7	6886.4	6898.3	6248.6	-
-	6893.6	6548.4	6552	7001.2	7011.1	-	6715.4
7004.5	-	6893.3	6882.8	7437.3	-	6894.3	-
-	7034	7873.7	7877.3	-	7572.2	7006.6	-
7446.7	-	8096	8095.6	7660.6	-	-	7095.6
-	7447.8	8412.6	-	7929.4	7944.3	7448.9	-
-	7886.2	8708.9	8705.2	-	8418.7	-	7566.3
8101.9	8094.5	8835.7	-	-	9695.2	7668.8	-
-	8704.9	9041.7	-	10085	-	7887.7	-
-	8832.1	-	9153	10202.5	-	-	7935.2
-	9158.3	9525.6	-	10447	-	8407.7	8408.6
-	10093	10842.4	10839.5	10607.9	-	8621.8	-
-	10204.8	-	11043.6	10828.8	-	-	8768.3
-	10624.9	11316	11298.5	-	10849.3	8839.3	-
-	10842.5	12281.9	-	11289.7	11312.6	-	9348.9
11326.7	11302.8	13430.2	-	11628.1	-	9705.7	9694.8
-	11824.8	13783.8	13767.2	11804.8	-	10096.4	-
12302.5	-	14005.1	13986.7	12328.8	12344.8	10211.4	10235.3
-	12342.9	-	14633.8	13761.1	13783.1	10843.2	-
12471.8	-	-	-	13991.4	14006.8	-	10916.8
13821.1	13774.7	-	-	-	-	-	11066.5
14032.9	14000.3	-	-	-	-	11309.5	-
-	14416.1	-	-	-	-	-	11501
-	-	-	-	-	-	-	11720.3
-	-	-	-	-	-	12284.8	-
-	-	-	-	-	-	-	12319.2
-	-	-	-	-	-	13780.9	13771.3
-	-	-	-	-	-	13997.7	-

Table 4-3 Values of the main peaks observed in the spectra shown in figure 4-5

The table shows the main peaks (in Daltons) observed in the spectra obtained from cell lines and tissues originating from patients that either survived less than or more than 8 months, in the 5000-15000 Da mass range. Peaks that are similar between each pair of samples are shown in the same row, although the m/z values may be slightly different due to the mass accuracy of the instrument. The red values indicate peaks of similar masses between the samples taken from patients that survived less than or more than 8 months.

4.3.3 Bioinformatic analysis

Anonymised clinical data of patients from which the cell line and tissue samples were obtained was received from Dr Selma Ugurel, DKFZ, Germany, as part of the OISTER programme. In order to determine if cell lines are suitable tools to use to identify clinically relevant biomarkers, 2 parameters were chosen with which to mine the SELDI MS data.

The first parameter used was the site from where the tissue sample was taken; the second parameter was the survival of the patients after tissue extraction. A list of all the cell line and paired tissue samples received, along with clinical information, that were analysed by SELDI MS can be found in Appendix 2. Table 4-4 below lists the paired cell lines and tissues used in this aspect of the study; only patients with relevant clinical information were included in the bioinformatics analysis.

Tissue	Cell line	Tissue	Cell line
97	Ma-Mel-05	466	Ma-Mel-55
147	Ma-Mel-06	470	Ma-Mel-56
9	Ma-Mel-07	492	Ma-Mel-57
379	Ma-Mel-08b	476	Ma-Mel-59a
159	Ma-Mel-13	510	Ma-Mel-60
169	Ma-Mel-15	598	Ma-Mel-61c
208	Ma-Mel-19	525	Ma-Mel-62
171	Ma-Mel-20	548	Ma-Mel-65
285	Ma-Mel-26a	546	Ma-Mel-66a
293	Ma-Mel-26b	547	Ma-Mel-66b
252	Ma-Mel-27	561	Ma-Mel-71
284	Ma-Mel-28	568	Ma-Mel-74
288	Ma-Mel-33	620	Ma-Mel-81
304	Ma-Mel-36	657	Ma-Mel-82
313	Ma-Mel-37b	658	Ma-Mel-85
414	Ma-Mel-48a	656	Ma-Mel-90
449	Ma-Mel-51	628	Ma-Mel-91
395	Ma-Mel-52	652	Ma-Mel-93
453	Ma-Mel-54a	692	Ma-Mel-96

Table 4-4 Table showing the paired cell lines and tissues used for the bioinformatics part of this study.

The SELDI MS data was exported into EXCEL as csv files and converted into xls files before being merged into one file and smoothed to 3 Da, giving a total of 8000 ions between 1000-25000 Da. An ANN model using a multi-layer perceptron was trained (Statistica), on the smoothed mass/intensity data. Detail of the ANN method used can be found in section 2.2.6.1.

4.3.3.1 *Prediction of the site of growth of the tumour for both cell lines and tissues (Skin vs Lymph Node)*

The first analysis was to compare the predictive capability of the ANNs to discriminate between skin and lymph node as the site of tumour growth by analysing the SELDI MS profiles (H50 chips) of paired cell lines and tissues. A back propagation algorithm followed by a conjugate gradient descent algorithm was used and 22 samples were randomly selected for training, 8 for the test set and 8 for the blind dataset for each model, with 100 models run. The ANNs were first trained using the cell line data and then with the tissue data; the results were then compared to determine how well clinical parameters could be predicted using information from cell lines. Cell lines and tissues originating from the skin were assigned an output of 1 and cell lines and tissues originating from the lymph nodes were assigned an output of 2.

Cell lines

The ANNs correctly classified the cell lines originating from either the skin or the lymph nodes with an accuracy of $81.6\% \pm 0.4$, a sensitivity of 100% and a specificity of 58.8% using the top 1000 ions from the cell line data as identified by sensitivity analysis. Figure 4-6 shows the population distribution of the predicted outputs for all 38 samples and shows that 7 of the cell lines were misclassified; Ma Mel 26b, Ma Mel 6, Ma Mel 62, Ma Mel 74, Ma Mel 57, Ma Mel 37b and Ma Mel 7 were classified incorrectly as originating from the skin but were in fact cell lines originating from melanomas taken from the lymph node.

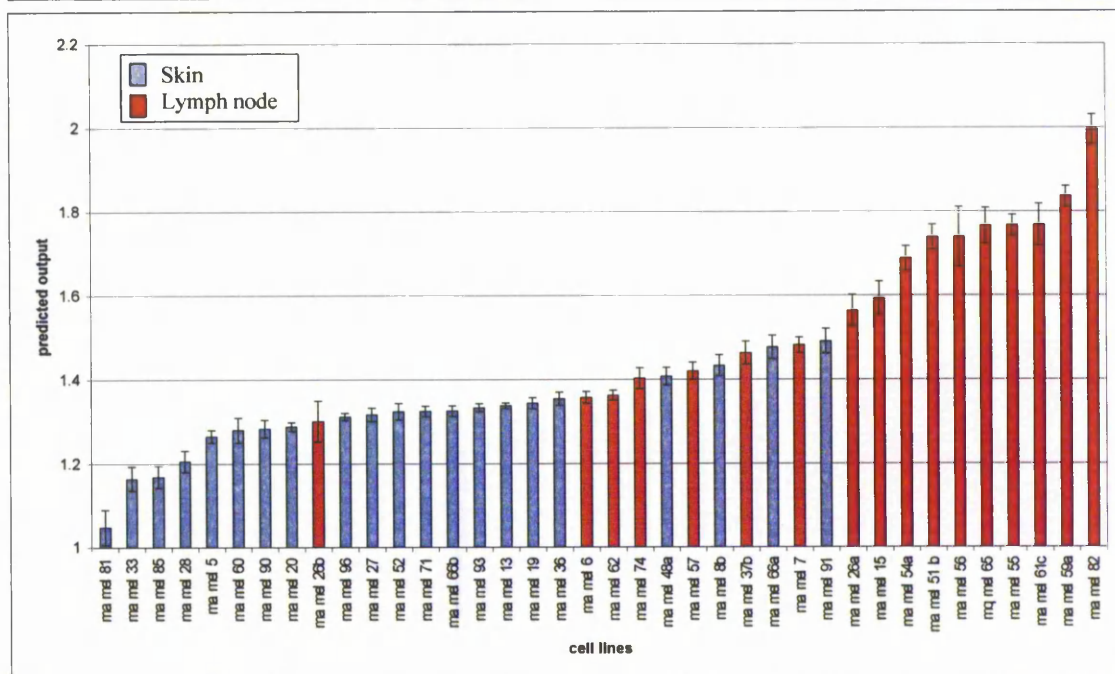


Figure 4-6 Classification by ANN analysis of 38 cell lines as either originating from skin or from lymph node.

The figure shows the classification of 38 cell lines as predicted by ANN analysis as either originating from skin or from lymph node. The predicted output shown is the average over 100 models; the error bars show the standard error of the mean with 95% confidence. The blue columns represent cell lines originating from skin, and therefore should have an output of 1, and the red columns represent cell lines originating from lymph nodes, and therefore should have an output of 2. Cell lines Ma Mel 26b, Ma Mel 6, Ma Mel 62, Ma Mel 74, Ma Mel 57, Ma Mel 37b and Ma Mel 7 were misclassified as originating from the skin when they had originated from melanomas taken from the lymph nodes.

Figure 4-7 shows the sensitivity analysis of the top 1000 ions from the training dataset, these 1000 ions were used to train the ANNs for all 100 models and enabled the ANNs to predict the site of tumour sample with 81.6% accuracy. The sensitivity analysis shows ratio values very close to 1 for all of the top 1000 ions, thus indicating that the removal of any of these ions does not greatly affect the classification error. There were 51 ions, however, with ratio values less than 1, suggesting that they are likely to have a negative influence on the analysis and that by removing these ions and retraining the ANNs, it may be possible to improve upon the predictive capability as well as possibly highlighting ions important in classifying the samples.

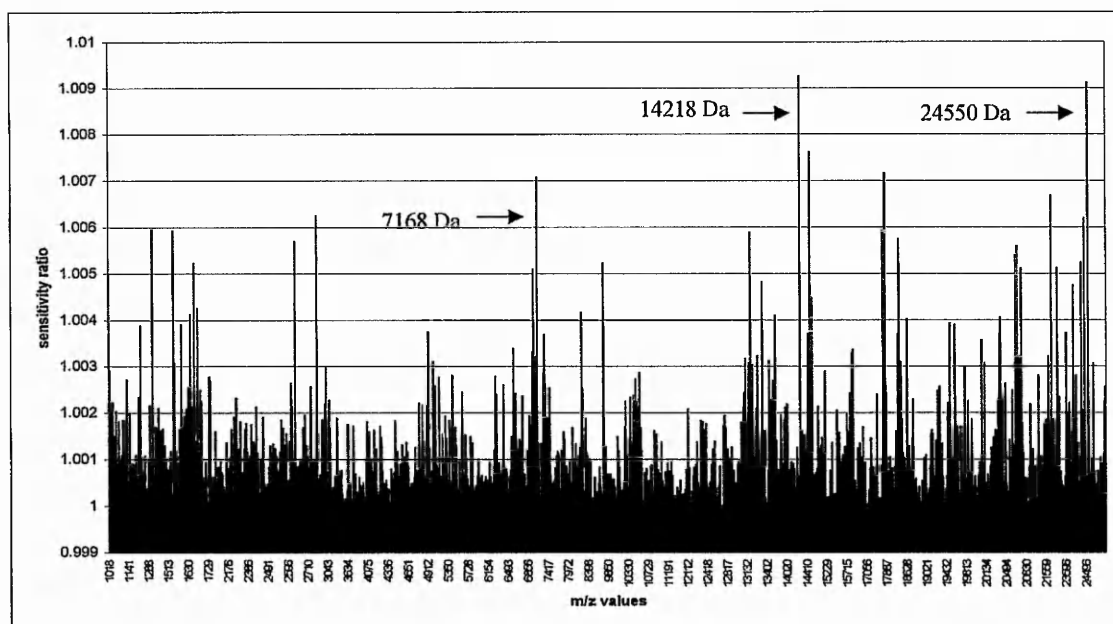


Figure 4-7 Sensitivity analysis of the top 1000 ions used to classify cell lines that originated from melanomas taken from either the skin or lymph nodes.

A list of the top 1000 ions used to classify melanoma cell lines according to their tissue of origin is obtained using a sensitivity analysis. The data presented shows the sensitivity ratios of the top 1000 ions averaged over the 100 models trained.

Tissues

Correct classification of tissues samples originating from either the skin or the lymph nodes was achieved with an accuracy of $86.8\% \pm 0.3$ and a sensitivity and specificity of 95.2% and 76.5%, respectively, using the top 1000 ions as identified by sensitivity analysis. This is illustrated in figure 4-8 and shows that five of the tissues were misclassified; tissues 285, 453, 293 and 598 derived from lymph nodes were classified as originating from the skin and tissue 620 was misclassified as originating from the lymph node.

Figure 4-9 shows the sensitivity analysis of the top 1000 ions and shows several ions that have a greater influence on error for classifying the samples and so are the most important ions for classification. There were, however, 200 ions with ratio values of less than 1 that would have a negative influence on the analysis and their removal should result in an improved classification.

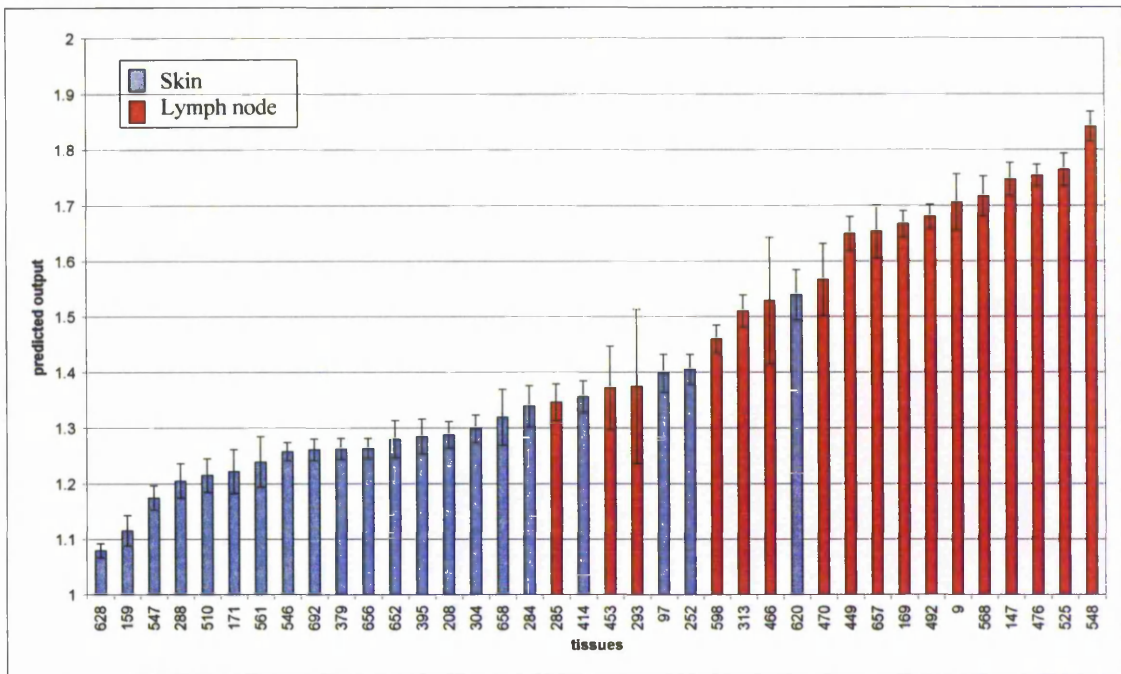


Figure 4-8 Classification by ANN analysis of 38 melanoma tissues as either originating from skin or lymph nodes.

The figure shows the classification of 38 tissues as predicted by ANN analysis as either originating from skin or from lymph node. The predicted output shown is the average over 100 models; the error bars show the standard error of the mean with 95% confidence. The blue columns represent cell lines originating from skin, and therefore should have an output of 1, and the red columns represent cell lines originating from lymph nodes, and therefore should have an output of 2. Tissues 285, 453, 293, 598 and 620 were misclassified.

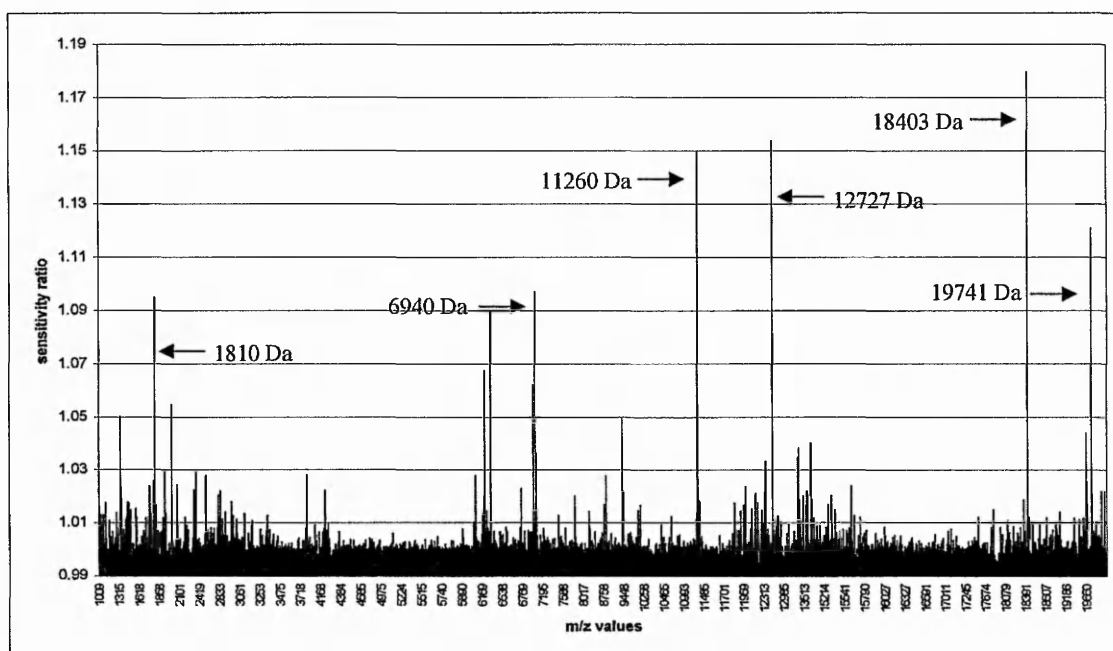


Figure 4-9 Sensitivity analysis of the top 1000 ions used to classify melanoma tissues that originated from either the skin or the lymph nodes.

The graph shows the sensitivity ratios of the top 1000 ions averaged over the 100 models trained.

Only 1 misclassified tissue (293) also had its paired cell line (Ma Mel 26b) misclassified. The performance of the cell line and tissue analyses revealed that although the accuracy of predicting the site of the tumour was similar using spectra obtained from cell lines and tissues, the specificities were different with the analysis using the cell line spectra giving a specificity of 58.8% and the analysis using the tissue spectra giving a specificity of 76.5%.

4.3.3.2 Prediction of patient survival for both cell lines and tissues

Cell lines and tissues were also analysed to predict patient survival, with 43 of the paired cell line and tissue profiles being included in the analysis. It was decided, at first, to use survival as a continuous variable; i.e. to predict the survival time of each patient in months. The survival of the 43 patients varied from 0.23 - 51 months. Unfortunately, using this approach, the ANNs could not correctly predict patient survival with either the cell line data or the tissue data, probably due to a low sample number but a broad range in survival time. It was therefore decided to use survival as a categorical variable and set 8 months as a cut off time for survival and determine if the ANNs could correctly assign patients as

surviving less or more than 8 months after the sample was collected. Patients that survived less than 8 months were assigned an output of 1 and patients that survived more than 8 months were assigned an output of 2.

Cell line analysis

The ANNs correctly classified cell lines originating from patients that survived less than or greater than 8 months with an accuracy of $79.1\% \pm 0.4$ a sensitivity of 96.4% and a specificity of 46.7% using the top 1000 ions. The predicted outputs of the cell lines are shown in figure 4-10 and shows that 9 of the cell lines from patients surviving longer than 8 months were incorrectly classified by the ANNs; Ma Mel 74, Ma Mel 93, Ma Mel 7, Ma Mel 24, Ma Mel 61, Ma Mel 91, Ma Mel 22 and Ma Mel 65; Ma Mel 26b was misclassified as originating from a patient surviving longer than 8 months.

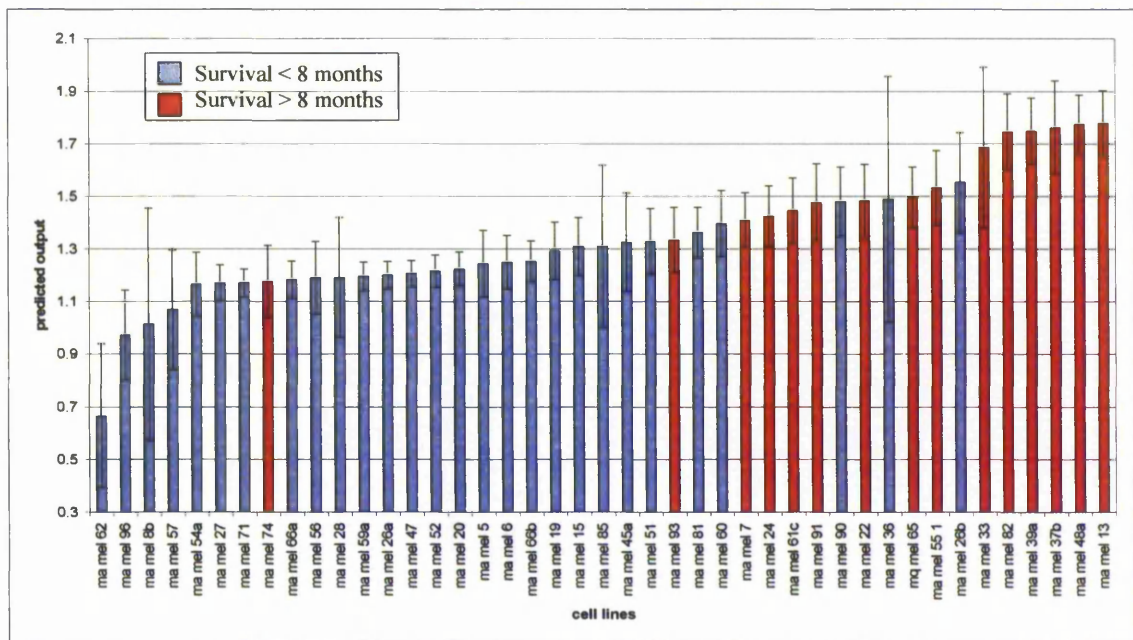


Figure 4-10 Classification by ANN analysis of 43 melanoma cell lines originating from patients with a survival, after sample collection, of less than or more than 8 months.

The figure shows the classification of 43 cell lines as predicted by ANN analysis originating from patients that survived less than or more than 8 months after tissue extraction. The predicted output shown is the average over 100 models; the error bars show the standard error of the mean with 95% confidence. The blue columns represent cell lines originating from patients with less than 8 month survival, and therefore should have an output of 1, and the red columns represent cell lines originating from patients that survived longer than 8 months, and therefore should have an output of 2. Cell lines Ma Mel 74, Ma Mel 93, Ma Mel 7, Ma Mel 24, Ma Mel 61, Ma Mel 91, Ma Mel 22 and Ma Mel 65 were misclassified as originating from a patient with a survival time of less than 8 months; Ma Mel 26b was misclassified as originating from a patient surviving longer than 8 months.

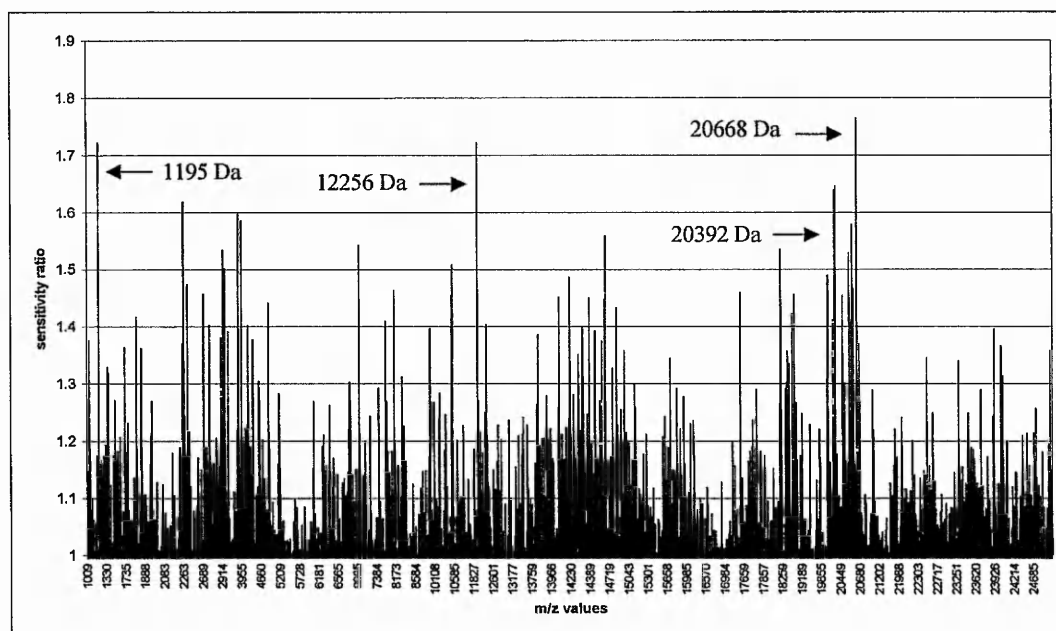


Figure 4-11 Sensitivity analysis of the top 1000 ions used to classify cell lines originating from patients with a survival, after sample collection, of less than or more than 8 months.

A list of the top 1000 ions used to classify the cell lines in each ANN model is generated using a sensitivity analysis. The data presented shows the sensitivity ratios of the top 1000 ions averaged over the 100 models trained.

The sensitivity analysis of the top 1000 ions, shown in figure 4-11, demonstrated large clusters of ions that were important for classifying the samples and only 4 of the top 1000 ions had a ratio value of less than 1.

Tissues

After training the ANNs with the top 1000 ions, correct classification of the tissue samples originating from patients that survived less than or greater than 8 months was achieved with an accuracy of $79.1\% \pm 0.384$ and a sensitivity and specificity of 100% and 40%, respectively. Figure 4-12 shows the population distribution of the predicted outputs for all 43 samples and reveals nine tissue samples that were incorrectly classified by the ANNs; samples 652, 313, 414, 288, 466, 598, 159, 192 and 568 were classified as originating from patients surviving less than 8 months when they originated from patients surviving longer than this time.

The sensitivity analysis of the top 1000 ions is shown in figure 4-13 and reveals that a few of the top 1000 ions are important for classifying the samples although the ratio values

overall are still very low; the most important ion still only had a ratio value of 1.017. Forty of the top 1000 ions were found to have sensitivity ratio values of less than 1 and so their removal from the training set may help towards improving the predictions.

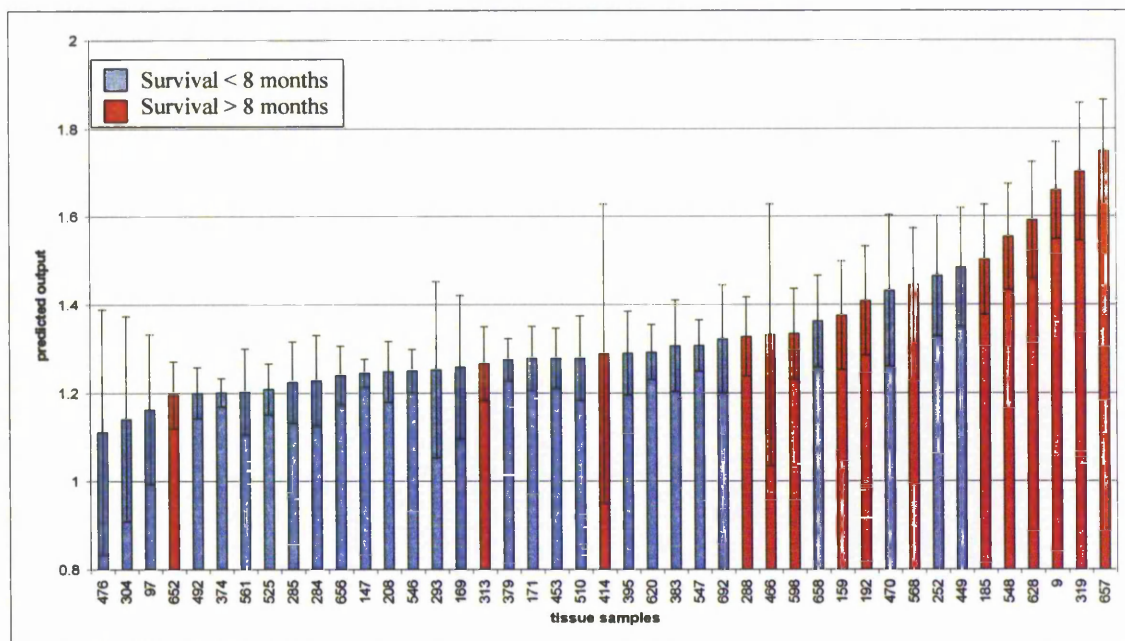


Figure 4-12 Classification by ANN analysis of 43 tissues originating from patients that survived less than or more than 8 months after tissue extraction.

The figure shows the classification of 43 tissues as predicted by ANN analysis originating from patients that survived less than or more than 8 months after tissue extraction. The predicted output shown is the average over 100 models; the error bars show the standard error of the mean with 95% confidence. The blue columns represent cell lines originating from patients with less than 8 month survival, and therefore should have an output of 1, and the red columns represent cell lines originating from patients that survived longer than 8 months, and therefore should have an output of 2. Tissues 652, 313, 414, 288, 466, 598, 159, 192 and 568 were misclassified as originating from patients surviving less than 8 months.

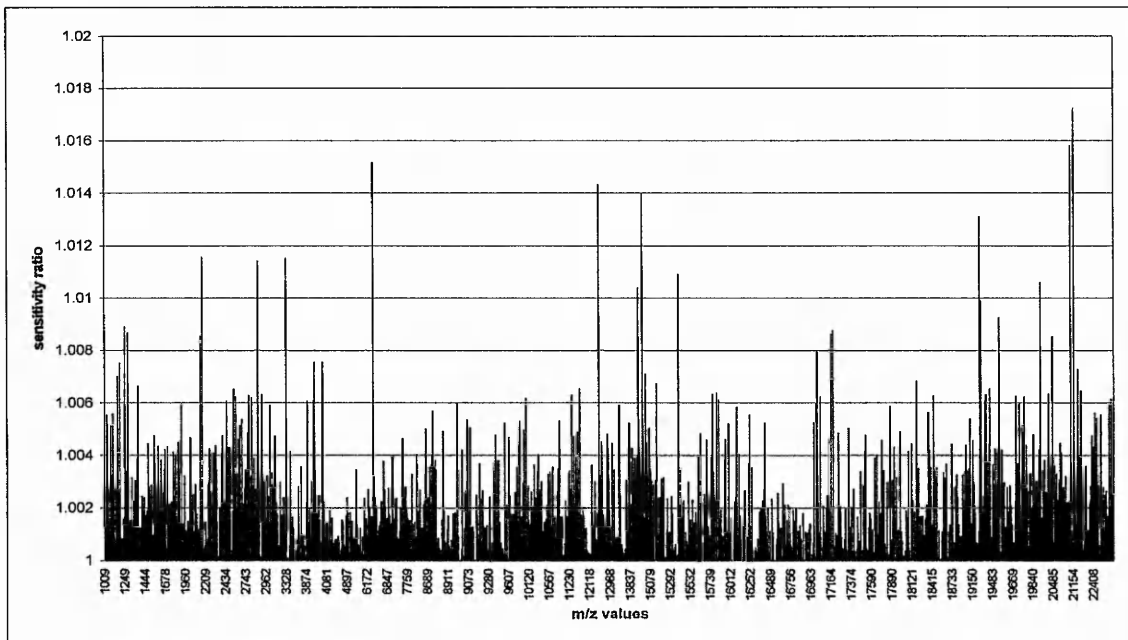


Figure 4-13 Sensitivity analysis of the top 1000 ions used to classify tissue samples that originated from patients that survived less than or more than 8 months after tissue extraction.

The figure shows the sensitivity ratios of the top 1000 ions averaged over the 100 models trained. The ratio values are all very low showing that there is not a small subset of ions that is of particular importance for the classification of these samples.

Four of the misclassified tissues and also had their paired cell lines misclassified; tissue 293/cell line Ma Mel 93, 598/Ma Mel 61c, 568/Ma Mel 74 and tissue 192/cell line Ma Mel 22. These four samples originated from patients that survived longer than 8 months after the sample was extracted. The following table (table 4-5) shows the survival time of the patients as obtained from the patient database, samples that were misclassified by ANN analysis are highlighted in red.

None of the top ions from the sensitivity analysis of the training set of the tissue data were the same as those from the cell line data, even when taking the mass accuracy of the SELDI instrument (found to be 0.05%) into consideration.

Histo-number	Cell Line Code	Survival since tissue extraction	Histo-number	Cell Line Code	Survival since tissue extraction
208	Ma-Mel-19	0.23	284	Ma-Mel-28	4.59
147	Ma-Mel-06	0.33	692	Ma-Mel-96	6.23
395	Ma-Mel-52	0.53	374	Ma-Mel-47	6.25
470	Ma-Mel-56	0.66	285	Ma-Mel-26a	6.30
546	Ma-Mel-66a	0.79	510	Ma-Mel-60	7.40
547	Ma-Mel-66b	0.79	252	Ma-Mel-27	7.56
169	Ma-Mel-15	1.02	466	Ma-Mel-55	8.09
171	Ma-Mel-20	1.25	652	Ma-Mel-93	9.28
379	Ma-Mel-08b	1.38	657	Ma-Mel-82	9.73
97	Ma-Mel-05	1.97	628	Ma-Mel-91	10.20
449	Ma-Mel-51	2.07	185	Ma-Mel-24	11.18
620	Ma-Mel-81	2.13	288	Ma-Mel-33	11.21
476	Ma-Mel-59a	2.66	9	Ma-Mel-07	11.84
658	Ma-Mel-85	3.31	319	Ma-Mel-39a	12.64
293	Ma-Mel-26b	3.61	598	Ma-Mel-61c	13.76
453	Ma-Mel-54a	3.78	414	Ma-Mel-48a	13.84
492	Ma-Mel-57	3.88	568	Ma-Mel-74	14.45
304	Ma-Mel-36	4.03	548	Ma-Mel-65	17.93
561	Ma-Mel-71	4.04	159	Ma-Mel-13	37.72
525	Ma-Mel-62	4.21	192	Ma-Mel-22	42.75
383	Ma-Mel-45a	4.37	313	Ma-Mel-37b	51.00
656	Ma-Mel-90	4.49			

Table 4-5 Survival data (in months) of the patients from which the cell lines and tissues analysed in this study were taken.

The table shows the survival time (in months) of the patients from which the cell lines and tissues analysed originated, as obtained from the clinical database made available by Dr Selma Ugurel as part of the OISTER project. Samples that were misclassified by ANN analysis are highlighted in red.

4.4 Discussion

The aim of this study was to determine whether cell lines could be used as a suitable model to investigate protein profiles that could indicate progression of the disease and identify clinically relevant biomarkers. Forty-four melanoma tissue samples were taken from patients with a known clinical history and with at least 5 years follow-up. Cell lines were also derived from the samples and a comparison between the cell lines and their original tumour of origin was performed via SELDI MS and ANN analysis.

The SELDI spectra (figures 4-4 and 4-5) show peaks at approximately 5650 Da, 10840 Da, 13780 Da and 14000 Da in all of the samples regardless of sample type and origin. It can be assumed that these peaks are likely to represent high abundant and/or house keeping proteins. If so then in future studies sample integrity could also be determined by the presence or absence of these peaks: samples not having these peak ions would not be included in the analysis.

The anonymised clinical information of patients was made available through OISTER. The SELDI MS data obtained from the cell lines was mined to determine if the site of tumour could be predicted from a protein profile and this was then repeated for the tissue samples, thus revealing whether the protein profiles obtained from cell lines hold clinically relevant information. The data showed, using this limited series of samples, that both the cell line profiles and the tissue profiles contained sufficient information to classify between a melanoma sample taken from the skin and sample taken from the lymph node; this was achieved with an 81.6% accuracy in the cell lines and 86.8% accuracy in the tissues. The fact that samples could be classified according to tumour site shows that the proteins (protein ions) present in cells originating from these locations must be significantly different. The influence of other factors, aside from the genome of differentiated cells, such as the skin and lymph nodes, allows for tissue specificity and this is evidence when cells are isolated in culture as they lose most functional differentiation (Bissell 1981). If cancers arose exclusively as a result of genetic mutations, then it could be expected that every organ could become cancerous. As this is not the case, then the presence of other molecules, including ECM, diffusible growth factors, cytokines, endothelial cells, lymphocytes, macrophages and fibroblasts, must also play key roles in cellular decision making and homeostasis (Bhowmick *et al.* 2004; Bissell *et al.* 2005). As a result of this,

predicting the site of a tumour within the tissue sample proteomes would be expected due to the molecules within the microenvironment (present within the samples) allowing for tissue specificity. This is also reflected in the specificity of prediction using tissues being higher (76.5%) than the specificity of prediction using the cell lines (58.8%). However, the fact that the site of tumour could also be predicted using the cell line samples must also mean that the cancer cells themselves must be affected by the tumour from which they originated. The predicted ions were also different between the cell lines and tissues, suggesting that the cell lines may have diversified in an 'artificial' *in vitro* microenvironment. However, given the fact that the cell lines and tissues originate from the same biopsy, it is not surprising that the predictability with regard to the site of tumour, is similar.

To further investigate whether cell lines could be used to identify clinically relevant biomarkers, SELDI profiles were mined according to patient survival. The survival of patients is the most important prognostic factor and accurate for the identification of patients at risk so would influence the selection of chemotherapeutic treatments. The survival of patients of less than or more than one year is an important parameter but due to a low number of patients surviving past 1 year in the sample group available, an 8 month cut off was used to provide approximately equal number of samples in the two groups. The data showed that both the cell line and tissue profiles contained relevant biomarkers that enabled the ANNs to classify a melanoma sample as originating from patients that survived less or more than 8 months. This was achieved with a 79.1% accuracy for both the cell line and tissue data. Misclassification of patients that survived more than 8 months was shown for 8 cell lines and 9 tissue samples. Only 43 samples were analysed in this study, 15 of which had a broad range of survival, from 8-51 months. The specificity of predicting survival using either cell lines (46.7%) or tissues (40%) was low and this is probably due to a number of factors; firstly the number of samples included in the analysis was low, especially in the second group which also had a broad range of survival; secondly, survival in a patient is affected by a number of factors including the symptoms of the cancer cachexia syndrome and quality of life, rather than just the size of the tumour, tumor grade and extent of the disease (Glare 2005), therefore the death of the patients included in this study could have been due to a number of different factors. To look at patient survival in this way, a larger cohort of samples would be required to allow for a more generalised

predictive pattern to be identified. One significant finding from this study has shown that the profiles from the cell lines allowed for classification of blind data with the same accuracy as the tissue data, again showing that it may be possible to use cell lines to study the biological differences between samples obtained from patients with different clinical outcomes. However, as the sensitivity analyses showed that the most important ions for classifying samples were different for the cell lines and tissues analysed, it is not possible in this instance to use cell lines to identify clinically relevant biomarkers. This is not unexpected as both the cell line lysates and tissue lysates would contain different concentrations and possibly different types of proteins due to the heterogenous nature of the tissue sample and genetic and phenotypic diversity that would have resulted from *in vitro* culture of the melanoma cells. Even though different ions were identified as important, these may represent fragments of the same protein or different proteins either with the same biological function or originating from the same family. It will, therefore, be important to identify the proteins associated with the key ions in order to understand their importance with regard to disease status and biology.

This study provides a “proof of principle” that clinically relevant parameters can be predicted through an integrated use of mass spectrometry and bioinformatics. In support of the study, Voss *et al* (2001) undertook expression profiling at the protein level using 2-D PAGE on 24 patients with B-cell chronic lymphocytic leukaemia (B-CLL). The analysis allowed the identification of proteins that clearly discriminated between patient survival (Voss *et al*. 2001). Another study also used 2D PAGE and MALDI-TOF combined with cluster analysis to analyse 85 glioma tissue samples in conjunction with survival. Clustering of proteins significantly correlated with patient survival and discriminant analysis extracted a set of proteins that were differentially expressed between histological grades and identified novel biomarkers for survival prediction (Iwadate *et al*. 2004). Schwartz *et al* (2005) obtained protein profiles from 108 human gliomas of various grades via direct tissue profiling using MALDI-TOF. By applying statistical algorithms to the profiles, protein patterns were identified that correlated with tumour histology and patient survival and also identified 2 patient populations, short-term and long term survivors (Schwartz *et al*. 2005). These studies demonstrate that it is possible to correlate SELDI MS profiles with patient survival.

As discussed, the tumour microenvironment may have an important role in tumour progression, caused by reciprocal signalling between the tumour cells and stromal cells such as endothelial, inflammatory, immune cells and CAFs. Thus, the tumour microenvironment may affect the progression of a tumour and the biomarker profile (Mueller & Fusenig 2004). Cell lines are homogeneous and express biomarkers that are tumour related. However, the selective pressure on melanoma cells *in situ* and *in vitro* is quite different and it would be expected that differences would occur at the protein levels. A comparison between microdissected tumour cells and cultured cells from the same tissue would go some way to highlighting differences and similarities, although there are obvious disadvantages.

Laser capture microdissection (LCM) is a microscope based dissection technique that relies on previous fixing and staining of tissue sections to identify areas of interest and as such there is a risk of artefacts arising from this procedure (Craven *et al.* 2002). LCM has been used as a tool for obtaining specific cell types from tissues for gene expression and proteomic based studies. A study by Kurose *et al.* (2001) used microdissected breast tissue to analyse LOH in both malignant breast epithelial cells of intra-ductal carcinoma and the surrounding stromal cells. The results showed frequent LOH in both the epithelial and stromal compartments, suggesting that genetic alterations in both the tumour and its microenvironment may be required for breast cancer progression (Kurose *et al.* 2001). Studies using microdissected tissue for SELDI MS analysis have also been reported. Potential biomarkers associating with prostate cancer were found that discriminated between benign prostatic hyperplasia (BPH), prostatic intraepithelial neoplasia (PIN) and prostate cancer (PCA). BPH, PIN and PCA cell populations were obtained from 9 prostatectomy specimens using LCM. The lysates from these cell populations were analysed by SELDI MS revealing differences in the protein profiles that discriminated between the different disease states. Also, differentially expressed proteins could be used as potential markers for screening for the early development of prostate cancer (Cazares *et al.* 2002). A study analysing tissue extracted from 10 hepatocellular carcinoma patients by SELDI MS revealed that differentiation between central tumour regions, peripheral tumour regions and normal tissue could only be detected in microdissected tissue (Melle *et al.* 2004). By using LCM to provide homogenous tissue samples, the potential problem of

'diluting out' of important biomarkers due to the heterogeneity of whole tissue lysates could be avoided.

Chapter 5 Breast cancer tissue proteome profiles associating with a basal phenotype and p53 expression

5.1 Introduction

Breast cancer is the most common malignant disease in Western women and metastases to distant sites are the main cause of death in patients. Abnormalities in *BRCA1* and *BRCA2* account for approximately 90-95% of familial breast cancers and the loss of heterozygosity (LOH) in the 17q21 region of *BRCA1* account for over 50% of sporadic breast and ovarian cancers (Peto 2002). The biology of breast cancer and the treatment options currently available have been discussed (see section 1.3). Although combinations of these treatments can induce tumour response in 80-90% of breast cancer patients, the statistical survival of patients with metastatic disease is estimated in months not years, despite the fact that prolonged disease remission is seen in a few patients. Adjuvant therapy, which can include chemotherapy and hormone therapy, can help eradicate tumour cells that may have already spread to distant sites when the patient is first diagnosed. Chemotherapy has a wide range of short and long-term side effects that can affect a patient's quality of life. Although only 40% of patients relapse and ultimately die of metastatic breast cancer, 80% of patients receive adjuvant therapy as it is currently not possible to accurately predict a patient's risk of developing metastases. The problem then arises in women who receive chemotherapy unnecessarily when they could be cured of disease by local treatment such as surgery and radiotherapy; and as a result needlessly suffers the side effects of chemotherapy. Traditional prognostic markers for breast cancer include tumour size, lymph node involvement, histological grade and *erb-b2* amplification and protein expression. These markers can identify approximately 30% of patients that are likely to have a favourable or poor clinical outcome but of the remaining 70% of patients, another 30% will still develop metastases. Therefore the discovery of new prognostic markers to identify these high risk patients is needed (Weigelt *et al.* 2005).

5.1.1 Gene expression profiling of Breast Cancer

DNA microarray studies revealed that primary breast tumours that develop metastases could be distinguished from those that remained localised by their gene expression profile (Weigelt *et al.* 2005). Recent gene expression studies have also shown that one type of breast cancer, invasive ductal breast carcinoma, can be divided into many subgroups which are not apparent with conventional histopathologic examination. These subgroups are known as luminal A, luminal B, normal breast-like, HER2-overexpressing and basal-like, each of which are associated with different clinical outcomes (Livasy *et al.* 2006). One of the most interesting phenotypes is the basal epithelial-like phenotype which can be identified as being ER negative, progesterone receptor negative and HER2 negative (triple negative) (Bryan *et al.* 2006). This phenotype can also be characterised by the positive staining for antibodies raised against cytokeratins such as cytokeratin 5/6, cytokeratin 8/18 and cytokeratin 14, which are known to be restricted to basal cells within the breast, as well as vimentin and EGFR (Livasy *et al.* 2006). The basal-like phenotype has been frequently found in *BRCA1* associated breast cancers and these tumours are shown to be aggressive and have a tendency to metastasise to the viscera. They have also been associated with poor prognosis in patients that did not receive adjuvant therapy (Rodriguez-Pinilla *et al.* 2006). Elevated levels of cyclin E and p53 and low levels of KIP1 are also associated with this subtype (Arnes *et al.* 2005).

These gene expression studies have shown that a global approach to biomarker discovery studies can reveal previously unknown facets of the biology of different cancers, including breast cancer. Without these studies the different subtypes of breast cancer may not have been discovered, for example basal-like breast cancer is one subtype that has a poor clinical outcome and a high risk of developing metastases. As proteins are the functional elements of the cell it can be assumed that protein expression studies could also reveal biomarkers of clinical relevance.

5.1.2 Proteomic profiling of Breast Cancer

The majority of proteomic studies to date, involving breast cancer, have utilized cell lines tumour tissue, serum, plasma and nipple aspirates as the source of material for analysis. These studies have focused on the identification of differentially expressed proteins that can

define the molecular and biochemical pathways that cause normal cells to become cancerous. They have also identified potential biomarkers and therapeutic targets which could be used for the treatment of this disease. These studies have analysed the breast cancer proteome using 2D-DIGE (2-D differential in-gel electrophoresis) and confirming their findings using immunohistochemistry (Somari *et al.* 2005). For example, Annexin V, L-plastin, and the 78 kDa glucose regulated protein (GRP78) were found to be differentially expressed between breast carcinoma (DCIS and IDCA) and normal breast tissue (Somari *et al.* 2003; Wulfkuhle *et al.* 2001).

Proteomic analysis of whole tissue lysates of normal and DCIS tissue and microdissected normal and DCIS epithelial cells revealed proteins that were differentially expressed in DCIS compared to the normal samples. The study was performed using 2-D PAGE and sequencing via mass spectrometry and identified proteins that were not previously connected with breast cancer (Wulfkuhle *et al.* 2002). Another 2-D PAGE based approach has revealed a number of proteins that are differentially expressed between breast infiltrating ductal carcinoma and matched normal tissue (Somari *et al.* 2003). The same group have also identified proteins that are differentially expressed between infiltrating ductal and infiltrating lobular carcinomas and their associated lymph nodes (Somari *et al.* 2005). Various proteomic studies analysing breast cancer cell lines have been undertaken including the effects of doxorubicin on chemosensitive and chemoresistant MCF-7 cell lines on the proteome (Chen *et al.* 2002; Mian *et al.* 2003). These studies revealed that the effects of therapeutic treatment could be identified by the protein expression profiles generated by 2-D PAGE and SELDI MS. Tissue microarrays (TMAs) have also been widely used to study the expression patterns in breast cancer. One such study revealed the differential expression of hormone receptors, p53, erb-b2, cell cycle regulators, apoptosis markers and basal markers between *BRCA1* and *BRCA2* tumours (Palacios *et al.* 2005) and have been used to confirm gene expression studies that identified different subtypes of breast cancer. Two studies have confirmed the known subtypes and a third study has identified luminal and basal tumour clusters (Callagy *et al.* 2003; Korsching *et al.* 2002; Li *et al.* 2002a; Zhang *et al.* 2003). SELDI MS has been used to investigate serum, plasma, nipple aspirate fluid and tumour tissue for the presence of diagnostic biomarkers and to predict the clinical outcome or response to treatment of patients. Many of these studies have shown that biomarkers are present within serum and plasma that can be used to

differentiate between breast cancer and benign/healthy patient samples. Li *et al* (2002a) screened 169 serum samples where 103 were from breast cancer patients, 41 were control samples and 25 were from patients with benign breast disease. By analysing the samples using SELDI MS and the software package ProPeak, 3 biomarkers were identified that could discriminate between cancer and non-cancer samples with a sensitivity and specificity of 93% and 91%, respectively. One study analysed serum samples taken from 45 breast cancer patients, 42 patients with benign breast disease and 47 normal volunteers using SELDI MS and classification trees. The classification of cancer versus normal samples was achieved with 80% sensitivity and 79% specificity and between cancer and benign disease, the sensitivity and specificity was 78% and 83% (Vlahou *et al.* 2003). Hu *et al* (2005) analysed serum obtained from 49 breast cancer patients, 51 with benign breast disease and 33 healthy women using SELDI MS, ANNs and discriminant analysis. The data showed that discrimination between cancer and benign breast disease could be achieved with 81.6% sensitivity and 78.4% specificity. Also, discrimination between cancer and non-cancer samples was achieved with a sensitivity and specificity of 79.6% and 77.4% (Hu *et al.* 2005).

The aim of this aspect of the study was to use SELDI MS and ANNs to identify proteomic patterns from 155 breast cancer tissues that correlate to clinical and TMA information associated with these samples, (details given in appendix 3). Due to time constraints two parameters were taken for ANN analysis, the identification of a basal phenotype within the samples and the presence of p53 expression.

5.2 Methods

All of the tissue samples processed by SELDI MS were processed as part of a larger study, Biopattern (a 6th framework EU funded programme), by tissue microarray. Tissue microarray and anonymised clinical data of patients from which the breast cancer samples were obtained was received from Prof Ian Ellis, City Hospital, Nottingham. The 165 samples analysed came from a study based on a consecutive series of 1944 cases of invasive breast carcinomas presenting between 1986 and 1998 that were entered into the Nottingham Tenovus Primary Breast Carcinoma Series. This is a well-characterised series of patients under the age of 70 presenting with primary operable breast carcinoma. The

patients were treated in a conventional uniform manner and the series has been used previously to study a wide range of proteins (Abd El-Rehim *et al.* 2004; Abd El-Rehim *et al.* 2005; Rakha *et al.* 2005). The patient's clinical history and tumour characteristics were obtained from the pathology database and the median follow-up period was 58 months (range 1–192 months). The tissue microarrays were prepared as described (Kononen *et al.* 1998) with each case sampled twice, from the centre and the periphery of the tumour, to form an array of 100 cases per block. Immunohistochemical staining of basal cytokeratins (CK5/6 and CK14) was performed (Abd El-Rehim *et al.* 2004; Abd El-Rehim *et al.* 2005) using the standard streptavidin–biotin complex method and TMA technology and where only staining of the malignant cells within the tissue cores was considered. Immunohistochemical scoring was performed using the modified Histo-score (H-score) that includes a semi-quantitative assessment of both the intensity of staining and the percentage of positive cells giving the range of possible scores as 0–300. Positivity was defined as the detection of 10% or more of invasive malignant cells positive for CK5/6, CK14, SMA, and p53 staining (Rakha *et al.* 2006).

A more detailed explanation of the methods used for is given in Chapter 2; a schematic flowchart showing how the tissue lysates were processed for mass spectrometry and ANN analysis is given below (figure 5-1). The bioinformatics analysis was performed as described in section 2.2.6.1 but with some parameters adjusted slightly in order to improve upon the predictions; only a BP algorithm was used for training 10,000 epochs, not CGD, the momentum was set to 0.5 instead of 0.3 and the 'window' set to 3000.

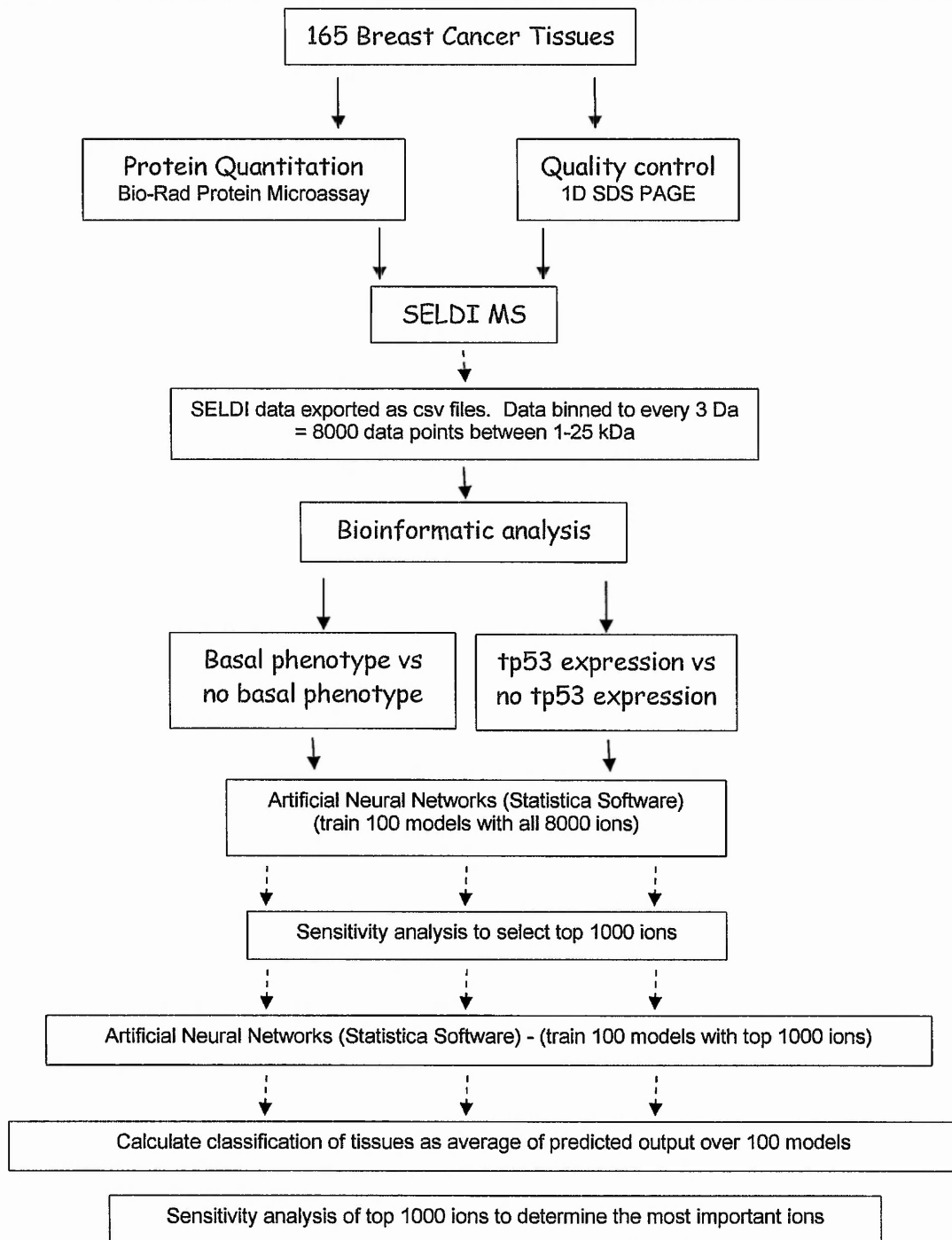


Figure 5-1 A schematic showing the methods for sample processing and analysis used in this study.

5.3 Results

5.3.1 Protein quantification and sample quality control.

Protein quantification was performed using the Bio-Rad protein microassay as described in section 2.2.2. A standard curve was determined by plotting the absorbance at 595nm versus BSA protein standard concentration (μg). The calculation of the protein concentrations is explained in detail in section 3.2.1. On average, concentrations of between 3 and 9 $\mu\text{g}/\mu\text{l}$ were obtained from the breast cancer tissue lysates. The samples were then run through a 12.5% resolving gel to determine sample integrity; an example is shown in figure 5-2. This figure shows consistent bands in all of the samples run through this 12.5% resolving gel. Prominent bands can be seen at approximately 70 kDa, 50 kDa, 45 kDa, 40 kDa, 35 kDa, 25 kDa and 11 kDa in all of the samples, concluding that no protein degradation has occurred during the sample preparation.

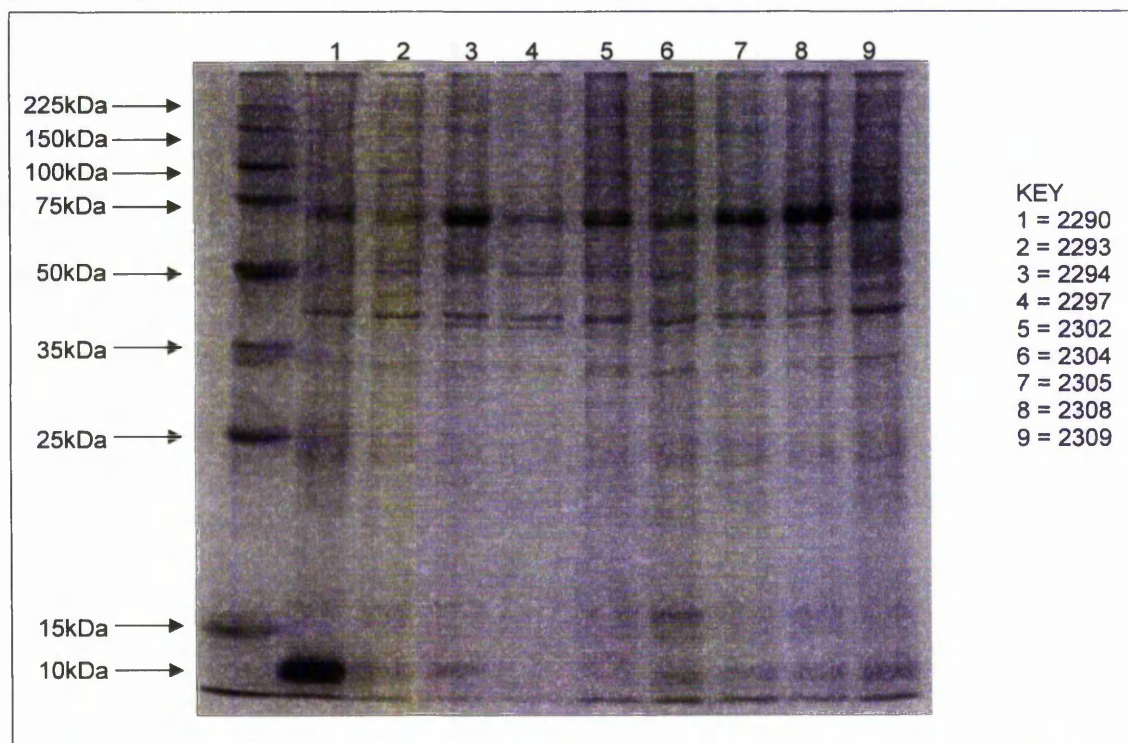


Figure 5-2 1-D PAGE of breast cancer tissue lysates representing sample quality of all samples analysed as part of the breast cancer study.

This figure represents an example of the data obtained from all of the breast cancer tissue lysates used in this study. 40 μg of each lysate sample was resolved in a 12.5% resolving gels and stained with coomassie blue to determine if degradation of the proteins occurred during sample preparation. The figure shows that band integrity is maintained between these samples and was observed in all of the lysates.

5.3.2 SELDI MS

The protocol used is described in section 2.2.4.1, but briefly 155 breast cancer tissue lysates were bound to the hydrophobic surface of a H50 SELDI chip and analysed using a PBS II mass analyzer. The following figures show some of the SELDI spectra obtained for this study. Figure 5-3 shows some differences, shown by the blue boxes, and some similarities, shown by the green boxes, that were noted by subjective visual assessment in the protein profiles of breast cancer samples with or without a basal-like phenotype. Sample 2123, a sample without a basal phenotype (shown in red) was overlaid with samples with a basal phenotype (shown in black); 2113 (a), 2080 (b), 2233 (c) and 2236 (d). Table 5-1 shows all of the peaks that can be seen in these profiles, showing peaks that are similar between the tissues with or without a basal-phenotype highlighted in red. Peaks at 7007.6 and 8463.4 Da can be found in the sample that does not have the basal phenotype but not in the samples that have a basal-like phenotype. Peak values at approximately 3100 and 3361 Da can be seen in all of the samples shown.

Figure 5-4 shows some of the differences, shown by the blue boxes, and some of the similarities, shown by the green boxes, of the protein profiles of breast cancer samples either expressing or not expressing p53. Profiles a-d are derived from samples that express p53, shown in black, overlaid on a profile of a sample not expressing p53, shown in red. Sample 2162 was overlaid with 2107 (a), 2202 (b), 2063 (c) and 2211 (d); the details are shown in table 5-2, with peaks that are similar between the tissues with or without p53 expression highlighted in red. Figure 5-3 and table 5-2 show that peak values at 5210.4, 6149.9 and 8498.6 Da can be found in the sample that does not express p53 but not in the samples that do express p53 and that peaks with approximate values at 5365 and 8640 Da can be seen in all of the samples shown here. These results do not however depict the complexity of the data, therefore a detailed analysis was performed using ANNs.

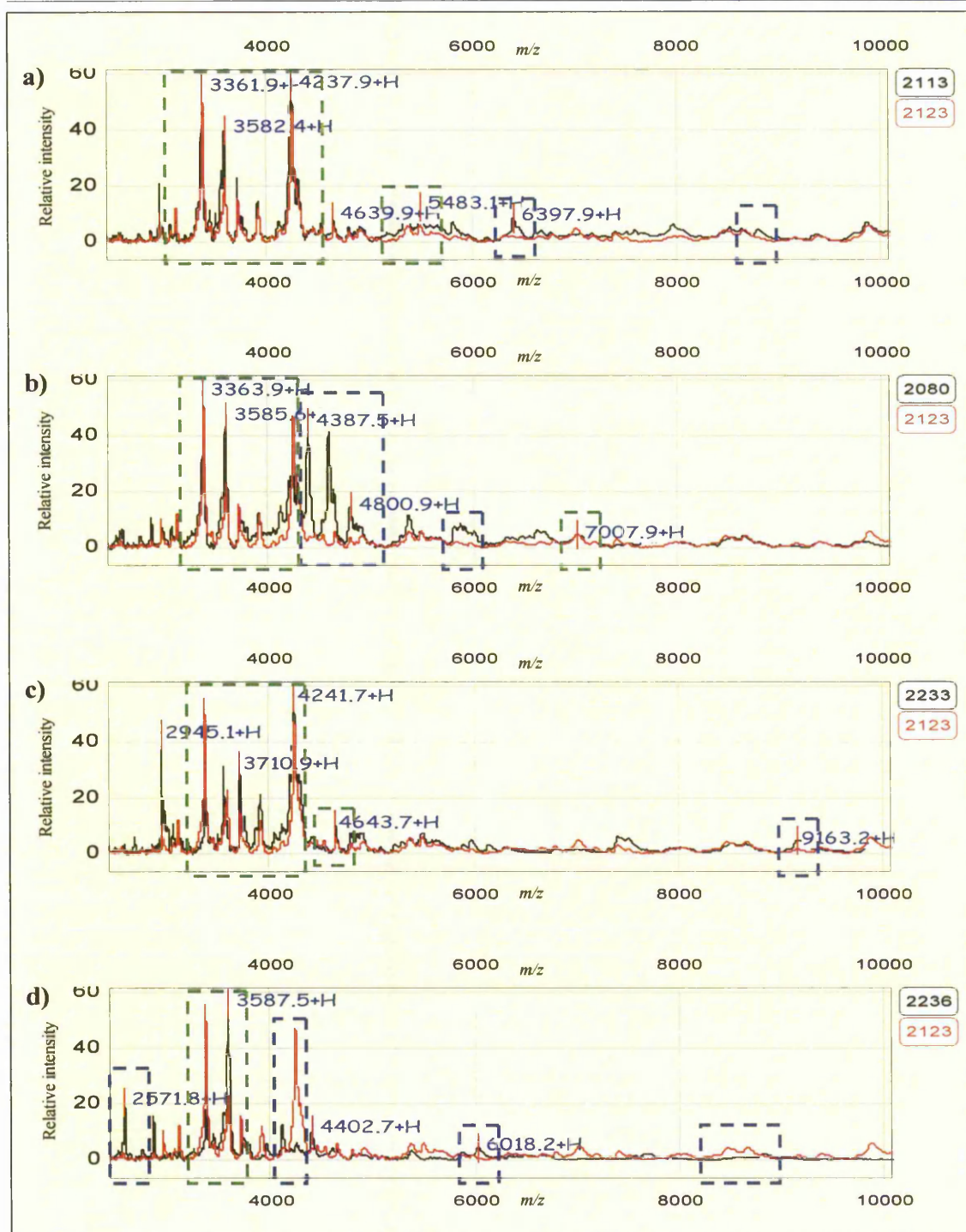


Figure 5-3 SELDI spectra obtained from breast cancer tissue lysates highlighting the differences and similarities between samples with different basal phenotypes, in the 2500 – 10000 Da range. The four spectra show a sample with a basal-like phenotype (in black) overlaid on sample 2123 which is an example of a sample without a basal-like phenotype (in red). The green boxes show the similarities between the samples and the blue boxes show the differences.

2123 (basal-like phenotype)	2113 (no basal-like phenotype)	2080 (no basal-like phenotype)	2233 (no basal-like phenotype)	2236 (no basal-like phenotype)
-	-	2571.8	-	2571.8
-	-	-	2679.1	-
-	-	2741.2	-	2747
-	-	2855.9	-	2856.2
2945.4	2941.5	2945.4	2945.2	-
3104.7	3098.8	3103.6	3103.1	3103
-	-	3248.5	-	-
3365.6	3361.9	3365.2	3366.8	3366.3
3586.5	3582.4	3587	3552.7	3587.5
3711.8	3706.6	-	3711	-
-	-	3748.6	-	3736.1
3914.1	3908.9	3912.6	3911.9	-
-	-	-	-	3962.2
-	-	4115.2	-	4114.5
4244.8	4237.9	4247.8	4241.8	4251.1
-	-	4404.2	-	4402.7
4448.8	4442.9	-	-	-
4513.3	4525.6	-	-	-
-	-	-	-	4607.2
4646.3	4639.8	4643.9	4643.7	-
4800.2	4797.2	4800.9	4798.5	-
-	-	4863.5	-	-
4913.3	4906	-	4911.6	-
5372	5366.8	5372.8	-	5374.1
5491	5483.1	-	5492.5	-
-	-	5590.5	-	-
-	5808	-	-	-
-	-	5871.4	-	5870.7
-	-	-	5963.6	-
-	-	-	-	6018.2
-	-	-	6137.1	-
-	6397.9	-	-	-
-	6670.3	-	-	-
7007.6	-	-	-	-
-	-	-	7406.9	-
-	-	-	7508	-
-	7962.6	-	-	-
8463.4	-	-	-	-
-	8517.8	-	-	-
8655.7	-	-	8654.7	-
-	8795	-	-	-
-	-	-	9103.3	-
-	9369.4	-	-	-
-	9858.3	-	-	-
9874	-	-	9882	-

Table 5-1 Values of the main peaks observed in the spectra presented in figure 5-3.

The table shows the main peaks (in Daltons) observed in the spectra obtained from breast cancer tissues with or without a basal-like phenotype, in the 2500 - 10000 Da mass range. Peaks that are similar between samples are shown in the same row, although the m/z values may be slightly different due to the mass accuracy of the instrument. Peaks that are observed in both the spectra of a lysate with a basal-like phenotype and sample 2123 (without a basal-like phenotype) are highlighted in red.

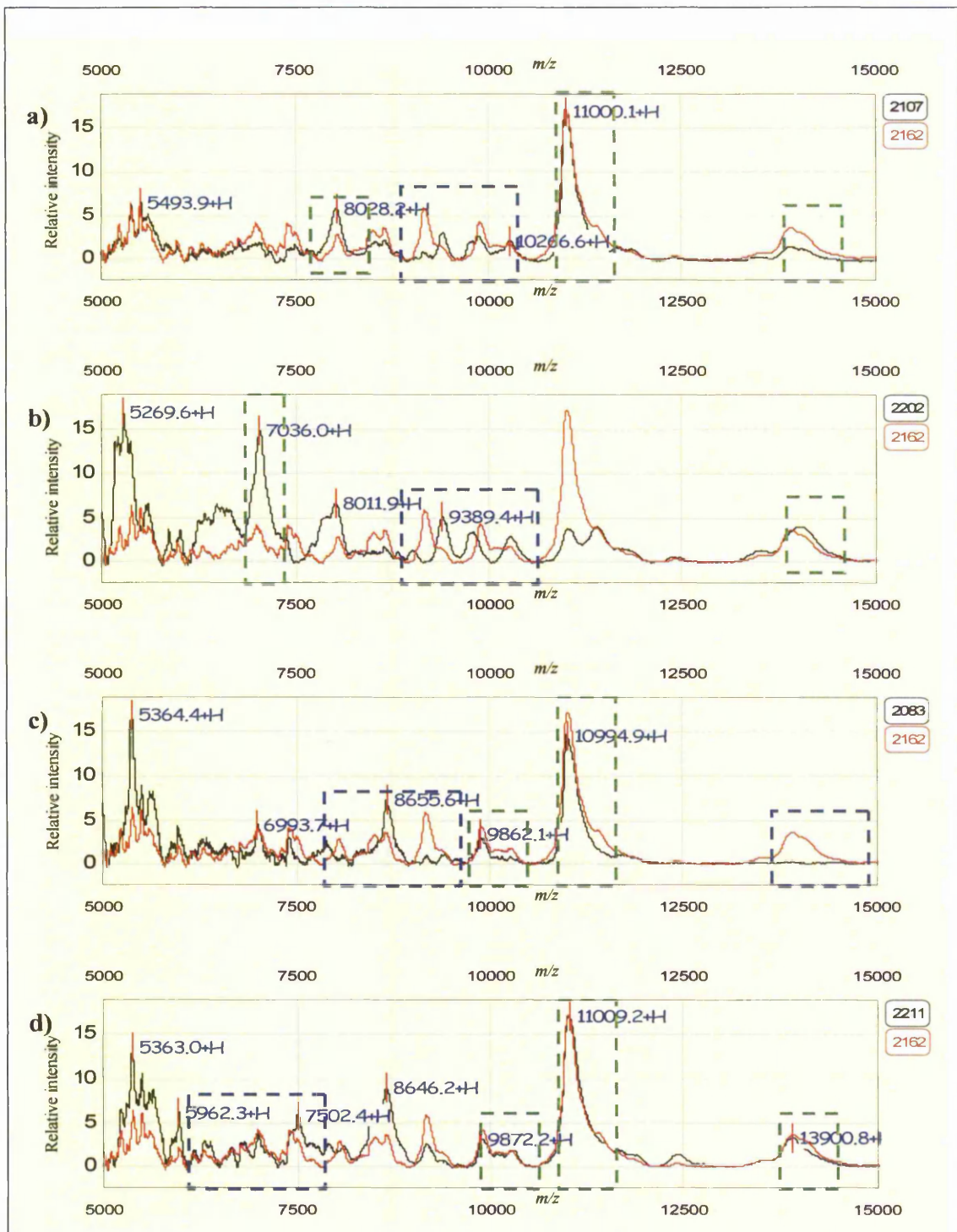


Figure 5-4 SELDI spectra obtained from breast cancer tissue lysates highlighting differences and similarities between samples that do or do not express p53, in the 5000-15000 Da mass range.

The four spectra show samples expressing p53 (in black) overlaid on sample 2162, a lysate that does not express p53 (in red) representing the differences, highlighted by blue boxes, and similarities, highlighted by green boxes between the two types of samples.

2162 (no p53 expression)	2107 (p53 expression)	2202 (p53 expression)	2063 (p53 expression)	2211 (p53 expression)
5210.4	-	-	-	-
-	-	5269.6	-	-
5367.8	5373.6	-	5364.4	5363
5486.5	5493.9	-	5491.2	5486.7
-	-	5587	-	-
-	-	-	5634.4	5633.2
-	-	5854.8	-	-
5960.1	5960.8	-	5963.9	5962.3
-	-	6001.6	-	-
6149.9	-	-	-	-
6289.9	-	-	6268.8	6296.8
-	-	6322.9	-	-
-	-	6480.6	-	-
-	-	-	6574.5	-
-	-	-	-	6656.2
-	6770.3	-	-	-
-	-	-	-	6815.2
6985.7	-	-	6993.7	6989.3
-	-	7036	-	-
7398.4	-	-	7389.7	-
-	7534.7	-	-	7502.4
-	-	-	7880.8	7899.9
8038.9	8028.2	8011.9	-	-
-	-	-	-	8099.4
8498.6	-	-	-	-
-	8534.5	-	-	-
8641.4	8663	8634.7	8655.6	8642.2
-	8809	-	-	-
-	8994.5	-	-	-
9158.9	-	-	9179	9166.7
-	9402.2	9389.4	9399.9	-
-	-	9774.6	-	-
9875.7	9859.7	-	9862.1	9872.2
-	10266.6	10262.1	-	10265.7
11001	11000.1	11016.4	10994.9	11009.2
-	-	11389.3	-	-
-	-	-	-	12418.5
13904	13910.4	-	-	13900.8
-	-	14015.5	-	-

Table 5-2 Values of the main peaks observed in the spectra presented in figure 5-4.

The table shows the main peaks (in Daltons) observed in the spectra obtained from breast cancer tissues with or without p53 expression, in the 5000-15000 Da mass range. Peaks that are similar between the samples are shown in the same row, although the m/z values may be slightly different due to the mass accuracy of the instrument. Peaks that are observed in both the spectra of samples with p53 expression samples without p53 expression are highlighted in red.

5.3.3 Bioinformatics Analysis

In order to determine whether SELDI MS and ANNs could be used to identify protein expression patterns in breast cancer samples relevant to diagnostic and/or prognostic outcomes, 2 parameters were chosen with which to mine the SELDI MS data. The first analysis was to determine if protein expression correlated to the basal-like phenotype. The second analysis was to examine whether there were patterns correlating with the presence of p53 expression. SELDI MS data in the 1-20 kDa mass range was exported into EXCEL as csv files and converted into xls files before being merged into one file and smoothed to 3Da for every sample analysed. A multi-layer perceptron was trained, using Statistica software, on the smoothed mass/intensity data, where only one hidden layer was used. A more detailed explanation of the ANN method used can be found in section 2.2.6.

5.3.3.1 Basal Phenotype

The first analysis was to compare the predictive capability of the ANNs to discriminate between breast cancer samples with a basal-like phenotype and those without by analysing the SELDI MS profiles obtained from H50 chips. Out of the 155 samples analysed, 140 SELDI spectra were included in this analysis due to the lack of TMA data available for the remaining samples.

A back propagation algorithm was used and 84 samples were randomly selected for training, 28 for the test set and 28 for the blind dataset for each model, with 100 models run. Samples that were not shown to have a basal-like phenotype by TMA were assigned an output of 1 and samples that did have this phenotype were assigned an output of 2.

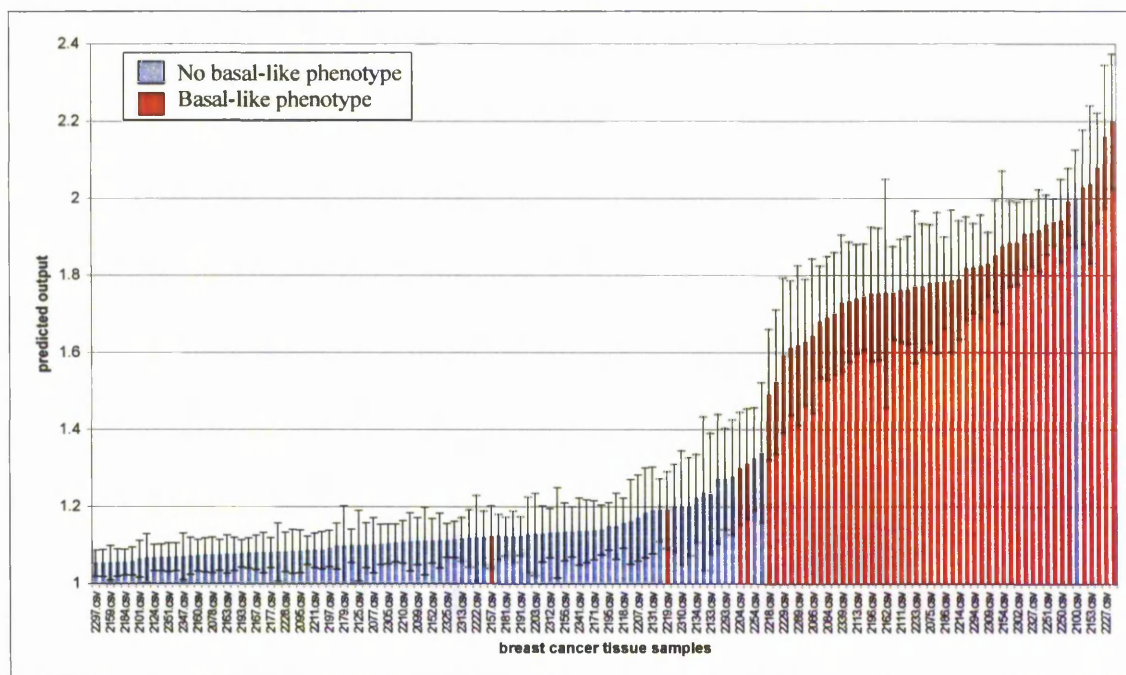


Figure 5-5 Classification by ANN analysis of 140 breast cancer tissues displaying a basal-like or non-basal-like phenotype.

The figure shows the average predicted outputs of 140 breast cancer samples as calculated by ANN analysis over 100 models. Samples that did not have a basal-like phenotype were assigned an output of 1, blue columns, and samples that were shown to have a basal-like phenotype were assigned an output of 2, red columns. The error bars show the 95% confidence intervals. Samples 2157, 2219, 2204, 2311 and 2218 were misclassified as not displaying a basal-like phenotype and tissue 2100 was misclassified as displaying a basal-like phenotype when TMA analysis revealed that it did not.

The ANNs correctly classified the tissues samples as having a basal-like phenotype or not with an accuracy of $95.7\% \pm 0.029$ and a sensitivity and specificity of 98.9% and 90.2%, respectively, using the top 1000 ions from the tissue data as identified by sensitivity analysis. Figure 5-5 shows the population distribution of the predicted outputs for all 140 samples with the samples that should have been classified as not having a basal-like phenotype, and therefore having a predicted output of less than 1.5, highlighted in blue. The samples that should have been classified as having a basal-like phenotype and therefore having a predicted output of greater than 1.5 are highlighted in red. The figure shows that six of the tissues were misclassified; tissues 2157, 2219, 2204, 2311 and 2218 were classified as not having a basal-like phenotype when TMA analysis showed that they

did. Tissue 2100 was classified as having a basal-like phenotype when it did not with a predicted output of 1.999.

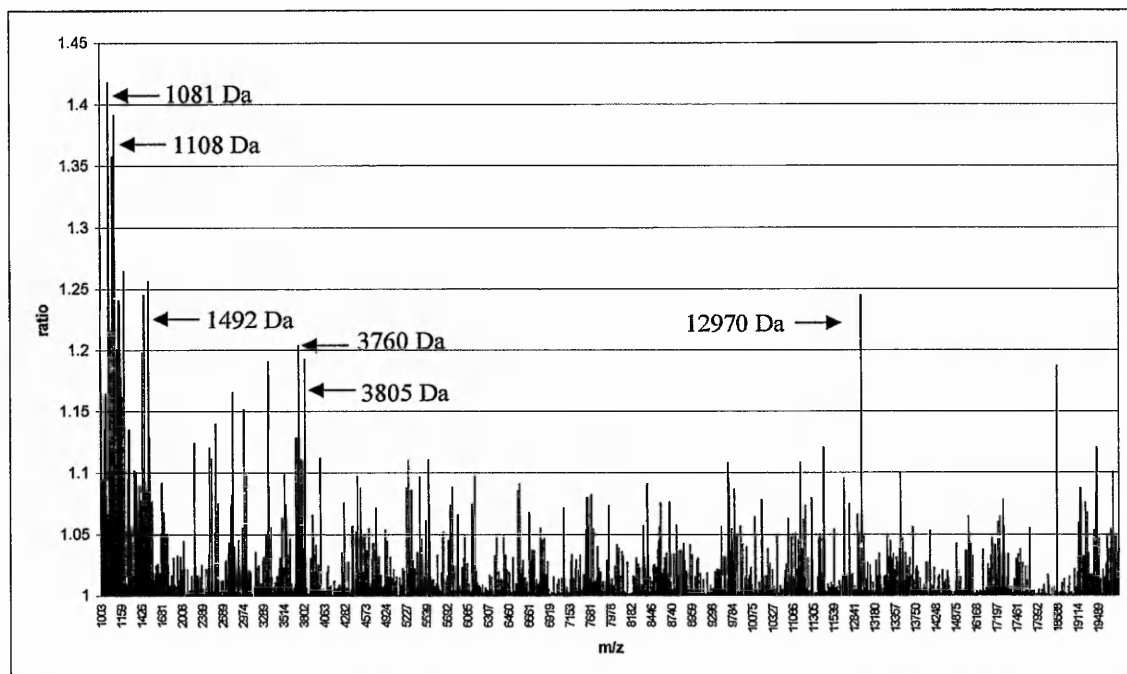


Figure 5-6 Graph showing the results of a sensitivity analysis of the top 1000 ions used to classify samples as those with or those without a basal-like phenotype.

A list of the top 1000 ions used to classify the breast cancer samples into either those with or those without a basal-like phenotype in each model is obtained using a sensitivity analysis. The above figure shows the sensitivity analysis of the top 1000 ions averaged over the 100 models trained. The analysis reveals two ions in particular that are very important for classifying the samples, these ions have masses of 1081 Da and 1108 Da.

Figure 5-6 shows the sensitivity analysis of the top 1000 ions. These 1000 ions were used to train the ANNs for all 100 models and enabled the ANNs to predict the phenotype of the tissue samples with 95.7% accuracy, using the data obtained from the tissue samples. The sensitivity analysis shows that the ions that have the greatest influence on error for classifying the samples are in the lower end of the mass range. There were 10 ions that had sensitivity ratios of less than 1 and their removal from analysis may improve the accuracy of classification even further.

5.3.3.2 P53 expression

The second analysis was to compare the predictive capability of the ANNs to discriminate between breast cancer samples that did or did not express p53 by analysing the SELDI MS profiles obtained from H50 chips. Out of the 155 samples analysed, 135 SELDI spectra were included in this analysis, the remaining 20 samples did not have relevant TMA data available.

A back propagation algorithm was used with 81 samples randomly selected for training, 27 for the test set and 27 for the blind dataset for each model, with 100 models run. During TMA analysis of these samples, they were assigned a value of between 0 and 300 according to the observed fluorescence which correlated with p53 expression. Samples that had a fluorescence value of 0 (did not express p53) were assigned an output of 1 for ANN analysis. Samples that had a fluorescence value of greater than 0 (expressed varying amounts of p53) were assigned an output value of 2 for ANN analysis.

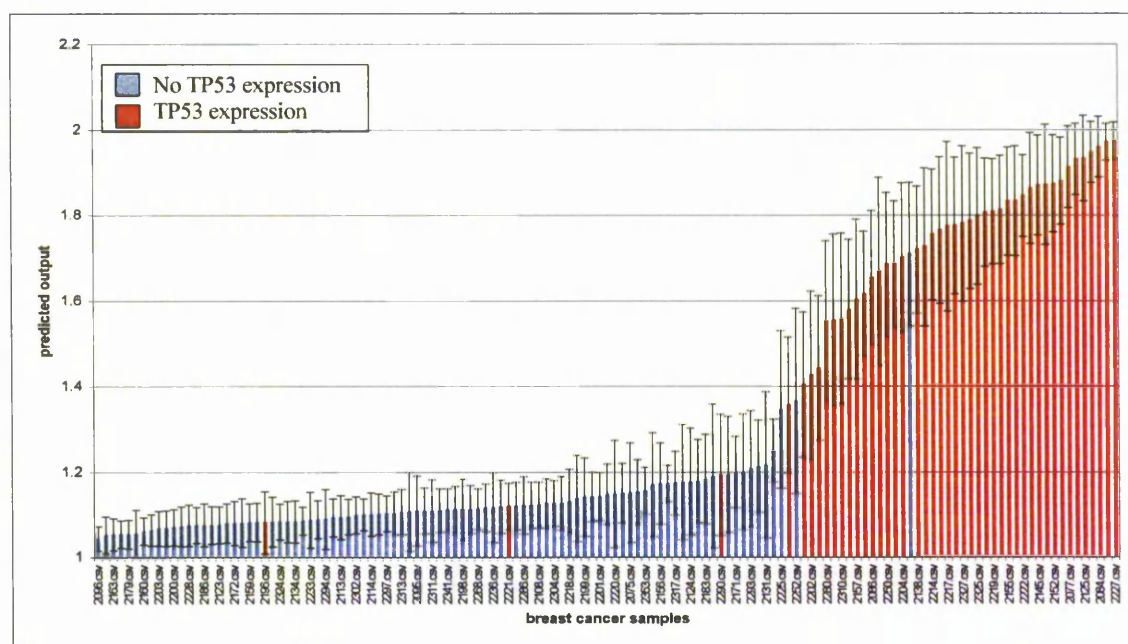


Figure 5-7 Classification by ANN analysis of 135 breast cancer tissue samples according to whether or not they express p53.

The figure shows the average predicted outputs of 135 breast cancer samples as calculated by ANN analysis over 100 models. Samples that did not express p53 were assigned an output of 1, blue columns, and samples that did express p53 were assigned an output of 2, red columns. The error bars show the 95% confidence intervals. Samples 2099, 2196, 2202, 2221, 2265, 2290 and 2339 were misclassified as not expressing p53 and sample 2153 was misclassified as expressing p53 when it did not.

Correct classification of tissues samples as expressing or not expressing p53 was achieved with an accuracy of $94.1\% \pm 0.041$ and a sensitivity and specificity of 98.9% and 84.4%, respectively, after training the ANNs using the top 1000 ions. Figure 5-7 shows the population distribution of the predicted outputs for all 135 samples and reveals that seven of the tissues expressing p53 were misclassified; tissues 2099, 2196, 2202, 2221, 2265, 2290 and 2339; tissue 2153 was classified as expressing p53 when TMA analysis showed that it did not.

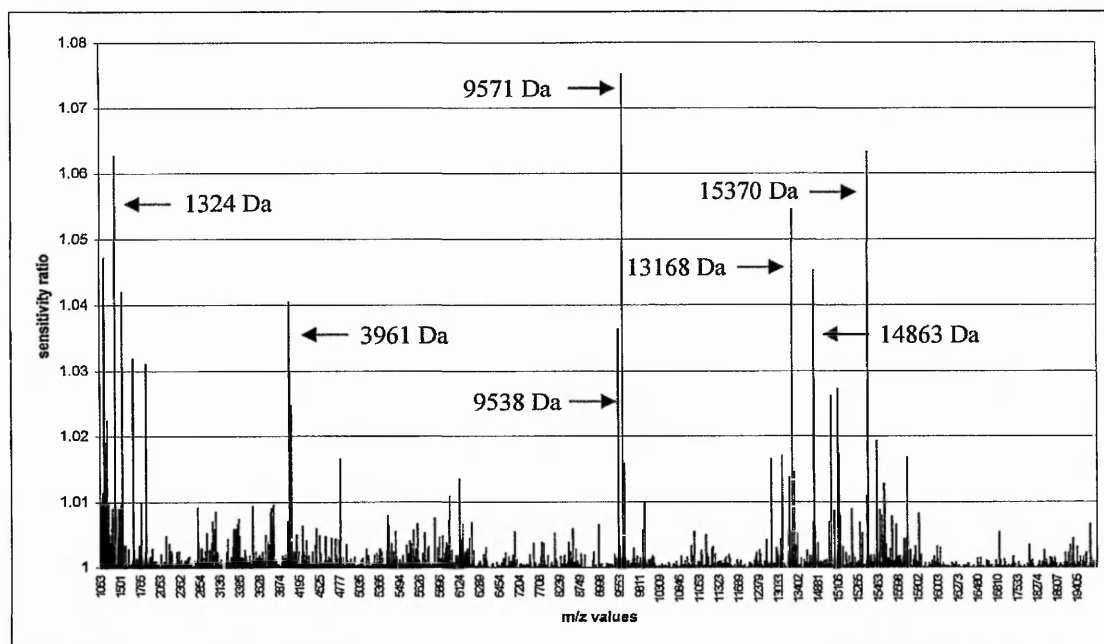


Figure 5-8 Sensitivity analysis of the top 1000 ions used by ANNs to classify breast cancer samples according to whether or not they express p53.

The above figure shows the sensitivity analysis of the top 1000 ions, from the tissue data, used to train the ANN models. The graph shows sensitivity ratio as an average over the 100 models trained. The graph reveals a few ions that stand out above the rest but the ratio values are all very close to 1 revealing that there are not any ions particularly important for classification of the samples, that all of the top 1000 ions are as important as each other.

Figure 5-8 shows the sensitivity analysis of the top 1000 ions and identifies that 10-15 of the ions have a greater influence on error for classification; 130 ions which had sensitivity ratios of less than 1, may influence the accuracy of classification and which could improve if they were removed.

5.4 Discussion

The study of the molecular changes that occur in breast cancer has revealed some of the pathways involved in breast carcinogenesis as well as identifying some of the key genes that are involved including *erb-b2*, *tp53*, *ccnd1*, *BRCA1* and *BRCA2*. High-throughput approaches that enable the profiling of large numbers of samples has allowed the complexity of this disease to be investigated. It is widely thought that a combination of markers is likely to be more sensitive and specific for patient diagnosis and predicting their prognosis and response to treatment than a single biomarker (Bertucci *et al.* 2006).

The aim of this study was to determine if predictive patterns in the SELDI MS data from 155 breast cancer tissues could be derived using ANN analysis. The samples were run in duplicate on H50 chips, although the duplicates were run randomly on different chips.

In this study 1-D PAGE was performed to determine sample quality and as long as consistent banding patterns were seen for the major bands present in every sample, the sample was deemed as being of good enough quality for SELDI MS analysis. Peaks at 3100 and 3361 Da were present in every sample and their presence could be considered indicative of sample quality and their suitability for MS and bioinformatics analysis. A number of differences and similarities are shown in figure 5-3 between samples with or without a basal-like phenotype and in figure 5-4 between samples that do or do not express p53. Certain peaks, seen visually in these two figures, are listed in table 5-1 and 5-2 together with peaks that are similar between the two sample types (highlighted in red). It is difficult to draw conclusions from this analysis and therefore ANN analysis was performed in order to interrogate in detail the complete protein profiles generated by SELDI MS.

SELDI MS data obtained from the tissue samples was analysed to determine if a basal-like phenotype could be predicted from protein profiles. This analysis utilised a cohort of 140 breast cancer tissue samples and could successfully predict samples that did or did not have this phenotype with a 95.7% accuracy. If time had allowed cluster analysis could have been performed to determine if the six misclassified samples belonged to a subset within the basal-like phenotype group to explain their misclassification. It would also be important to perform a TMA analysis on those six samples in case an error occurred when recording the phenotype data. The presence of a basal-like phenotype has been the subject of other studies in breast cancer. Lakhani *et al* (2005) analysed the histopathological

material obtained from breast cancer tumours in 182 *BRCA1* mutation carriers, 63 *BRCA2* carriers and 109 controls. The study was performed to investigate the proportion of *BRCA1* and *BRCA2* mutated tumours expressing basal markers to determine whether a predictive test could be developed to identify high-risk patients. The study found that positive staining for basal markers cytokeratin 14, cytokeratin 5/6, EGFR and cell morphology provided a more accurate predictor of *BRCA1* mutation than previously available and could be used for selecting patients for *BRCA1* mutation testing (Lakhani *et al.* 2005). The basal-like phenotype has been shown to be aggressive and is associated with poor prognosis in patients (see section 5.1.1). It has also been shown, in retrospective studies, that standard adjuvant therapy seems to be less effective in these tumours (Banerjee *et al.* 2006). By identifying those patients with this particular phenotype, the selection of patients for adjuvant treatment would be more successful and possibly lead to a better prognosis.

One concern of the study presented here, is that the sensitivity analysis revealed that the majority of the most important ions were found to be less than 1500 Da which is in the area where matrix peaks can be found. It is therefore possible that some of these 'important' ions may have been matrix peaks, although it is difficult to distinguish noise from real peaks of low intensity. The data from 1 kDa upwards was analysed to minimise the effects of noise in the ANN analysis but the fact that several ions of high importance were found in the 1-1.5 kDa range should be viewed cautiously. It is also important to note that although the mass accuracy was calculated to be approximately 0.6%, at the low mass end this could affect the actual m/z value of important peaks more than at higher mass values which could influence attempts to identify these particular ions.

The *tp53* oncogene encodes a 53 kDa protein, p53, which is involved in several critical pathways including cell cycle arrest, apoptosis, DNA repair, and cellular senescence, which are essential for normal cellular homeostasis and maintaining genome integrity. Mutations in the *tp53* gene have been found to occur in over 50% of human tumours and the alteration of this gene or the posttranslational modification in the p53 protein can alter its response to cellular stress (Hussain *et al.* 2006). P53 and MDM2 are linked through an autoregulatory negative feedback loop that maintains low p53 protein levels in the absence of stress (Moll *et al.* 2003). This feedback loop is disrupted, however, in many tumours that contain mutations of *tp53*. Loss of p53 function is indicated by *tp53* mutation or by the nuclear

accumulation of functionally impaired p53 protein and is observed in 20-40% of all breast cancers depending on the stage of the disease (Thor *et al.* 1992). Most *tp53* alterations found in breast carcinomas are point mutations that lead to the synthesis of a stable, malfunctioning and non-degradable protein that accumulates in tumour cells and so can be detected by immunohistochemistry (IHC) or tissue microarrays. Wild-type p53 protein does not normally accumulate within cells and has a half-life of 20 minutes, therefore it is not normally detected by IHC (Callahan 1992).

The study presented here shows that it is possible to identify patients that have a p53-positive breast tumour with a 94.1% accuracy. This study provides “proof-of-principle” that it is possible to identify p53-positive breast cancer patients using SELDI MS and ANNs. Additional studies could be undertaken to correlate the expression of more than 1 protein to SELDI MS profiles in order to develop a prognostic test for disease progression, patient survival or response to therapy. For example, a study assessing the TMA data of breast cancer tissues demonstrated that AR, c-erbB2, cytokeratin 18, MUC1, cytokeratin 5/6, p53, nuclear BRCA1, ER and E-cadherin were the key markers for the identification of different clusters of breast cancer samples (Abd El-Rehim *et al.* 2005). It has been shown that p53 positive breast tumours and tumours with a basal positive phenotype are both associated with aggressive behaviour of the cancer and poor patient outcome (Abd El-Rehim *et al.* 2004; Sorlie *et al.* 2001; Yamashita *et al.* 2004). By developing a test to identify patients with both of these indicators, it may be possible to provide a more accurate prognosis for breast cancer patients.

Tissue based proteomic studies are useful for identifying biomarkers directly related to disease and tissue type. One problem of using fresh or frozen tissue samples is that they can be difficult to obtain through routine clinical practice and can be expensive to store in a stable form. There are a number of alternative approaches that could be considered, including the use of cell lines, the use of formalin fixed tissue or the application of methods that allows the use of small amounts of protein for analysis; all of these have their pros and cons. Formalin fixation and paraffin embedding of tissue is the standard method used for processing tissue in pathology laboratories and results in a highly stable form of tissue that can be kept at room temperature. However, this type of processing is associated with a high degree of covalently cross-linked proteins as well as the effects of formalin fixation and storage on the samples and so IHC is currently the only published method that can

provide proteomic information from these tissues; although this method lacks sensitivity and quantitation. Hood *et al* (2005) have developed a method of protein extraction from formalin fixed prostate samples that can then be used for proteomic analysis. 10µm sections of a formalin fixed paraffin embedded (FFPE) prostate tissue were placed on slides, the paraffin removed by solvent treatment followed by tissue rehydration. Approximately 200,000 cells from different cellular regions including prostate cancer (PCa), benign prostatic hyperplasia (BPH) and stromal regions were microdissected using a novel tissue microdissection technique developed specifically for FFPE tissue, termed ExCellerator™. The proteins were extracted and trypsinised before protein microarray and NanoRPLC-MS/MS analysis. This analysis resulted in the identification of thousands of unique proteins in various histological regions of the tissue, although it is unclear whether low abundant proteins could be extracted from these samples (Hood *et al*. 2005).

The data presented here has shown that it is possible to extract proteins from snap-frozen breast cancer tissue and analyse these samples using SELDI MS and ANNs to identify differential protein expression patterns in breast tumours. Further studies could reveal the identity of proteins that may allow new biomarkers to be identified or the biological pathways involved in the disease to be further understood.

Chapter 6 SELDI MS and Bioinformatic approaches for the evaluation of serum biomarkers in stage I-IV melanoma

6.1 Introduction

Melanoma is the most aggressive form of skin cancer and the incidences of this disease is still continuing to rise. The five year survival in patients whose melanoma is localised at the site of the tumour (AJCC Stage I) is over 90% whereas it is only 5% in patients whose melanoma has spread and metastasised (AJCC Stage IV). In Stage III patients, 50% will survive longer than 5 years and as such stage III patients are the focus of adjuvant treatment trials. With no decline in death rates and no progression into its treatment, diagnosing patients with early disease, where the invasion of the tumour is much shallower, will result in much better prognosis and the possibility of complete cure with surgical excision of the tumour. There are currently no clinical assays that can accurately predict a patient's response to adjuvant therapy or if their disease will progress to stage IV and as such have a greatly decreased chance of survival. Therefore, patients that are identified as potentially being at risk of non-response to adjuvant treatment might then be considered for an alternative aggressive treatment (Wascher *et al.* 2003).

As has already been discussed, the construction of new models that will incorporate more than one validated biomarker is an important step towards obtaining accurate diagnosis and prognosis of patients in the clinic. Given the fact that metastatic melanoma is a clinically heterogenous disease would it be possible to develop a multiple biomarker model for the most aggressive subsets and how would we go about investigating this (Linette *et al.* 2005)? The previous chapters have highlighted the potential of SELDI MS and ANNs to identify patterns correlating to gene expression and certain clinical phenotypes obtained from cell lines and tissues, both of which provide suitable sample material for analysis. However, tumour tissue has to be obtained using an invasive procedure whereas bodyfluids such as serum, plasma and urine are more easily obtainable and have been shown to provide a rich source of biomarkers for the detection of cancer. A diseased organ can modify the blood proteome as a result of over-expression or abnormal shedding of proteins, the removal of proteins due to abnormal activation of proteolytic degradation or by the modification of proteins. Body fluids that circulate through the diseased tissues will

acquire proteins that are produced by the tumour and the microenvironment (Petricoin *et al.* 2002c).

6.1.1 Human blood proteome based studies

Both plasma and serum (a product of plasma coagulation) are widely available body fluids that contain active proteins that could provide information regarding disease status. Blood plasma is a complex body fluid, which contains a large diversity of proteins ranging in concentrations by at least 15 orders of magnitude. This pool contains intact as well as partially degraded proteins or protein fragments that circulate in the blood; as well as salts, lipids, amino acids, vitamins, and carbohydrates. The core plasma protein is albumin, which represents approximately 50% of the total plasma protein content, being present at high concentration (in the order of 30 g/L). The other main plasma proteins include immunoglobulins (Igs), fibrinogen, transferrin, haptoglobin, and lipoproteins. The remaining 1% of proteins are made up of approximately 10,000 lower abundance proteins, which are usually proteins of clinical interest (Adkins *et al.* 2002; Veenstra *et al.* 2005). Although the removal of these high abundant proteins may theoretically allow for lower abundant proteins to be detected, in practice this may result in the removal of information of clinical importance. Albumin is a carrier and transporter of proteins within the blood and has been shown to bind molecules such as hormones, cytokines and lipoproteins. One study isolated albumin, immunoglobulins, apolipoprotein and transferrins from the serum and showed that 209 unique proteins bound to these carriers; 12 of these proteins were clinical biomarkers currently in use such as PSA and meningioma-expressed antigen (Zhou *et al.* 2004). Serum is the product of blood clotting, where prothrombin is cleaved to thrombin, fibrinogen is removed to form the clot and a number of other protein changes occur including proteolytic cleavages (Anderson *et al.* 2002). The profiling of plasma, therefore, would provide a true snapshot of undegraded proteins present within the blood. A study by Villanueva *et al.* (2006) analysed low molecular weight (<20 kDa) serum biomarkers that could discriminate between breast, colon and prostate cancer. It was concluded that these biomarkers were not expressed directly by the diseased tissue but were in fact generated *ex vivo* by proteinase-mediated enzymatic cleavage as part of the coagulation process. It was suggested that fragments of endogenous blood proteins generated *ex vivo* served as a substrate pool for disease-specific proteinases that are

generated either from the tumour microenvironment or by the tumour itself. It was also suggested that the specific substrates cleaved by proteinases were themselves degradation products of the clotting cascade and so it was hypothesized that cancer specific protein profiles were an indirect snapshot of the enzymatic activity of tumor cells (Villanueva *et al.* 2006). If the detection and classification of certain types of cancer is a result of surrogate markers generated *ex vivo*, then these markers cannot be classified as tumour derived (Liotta *et al.* 2006).

6.1.2 Serum-based studies using SELDI MS

The best known cancer biomarker used in clinical practice to detect early stage disease is prostate-specific antigen (PSA). The screening of patient serum for PSA has been widely used for identifying prostate cancer for more than a decade and has increased the early detection rate for this disease, although presence of prostate cancer is still confirmed by biopsy (Barry 2001). Cancer antigen 125 (CA125) is another serum marker that is used to detect ovarian cancer in post-menopausal women and is the best available clinical marker for this group of patients, but a poor diagnostic marker for early stage ovarian cancer (Liede *et al.* 2002). It is hoped that the analysis of the serum proteome, biomarkers will be discovered that can be used to diagnose early stage disease, predict susceptibility and monitor disease progression (Xiao *et al.* 2005).

Attempts to analyse the serum proteome identified albumin and gamma-globulin peaks due to their relatively high concentration. The use of 2-D gel studies identified lower abundance proteins such as apolipoproteins, clotting factors and members of the complement system. Along with albumin and globulin, these proteins make up 98% of all serum proteins. 2-D gel analysis and western blotting have been performed in breast cancer, hepatocellular carcinoma and lung cancer patients in the search for serum biomarkers. These studies used autoantibodies against tumour cell proteins and revealed a novel circulating antigen, RS/DJ-1, in breast cancer, a repertoire of hepatocellular carcinoma specific autoantibodies and a protein that induces a humoral response in lung cancer, all of which have potential clinical use in the screening and diagnosis of patients (Brichory *et al.* 2001; Le Naour *et al.* 2001; Le Naour *et al.* 2002). Immunoassays have been used to quantify commonly used cancer serum markers at concentrations below 1ng/ml, (below the detection limit of 2-D gels). Advances in MS technology allows the display of hundreds of

low abundant proteins in a single spectra obtained from microlitre quantities of serum and provides clues as to the health status of a patient and whether the serum “peptidome” can reveal biomarkers that correspond to specific stages of diseases (Robbins *et al.* 2005). Petricoin *et al.* (2002b) were the first group to investigate the serum proteome using SELDI MS and bioinformatics and reported unique “fingerprints” that correlated with the presence or absence of ovarian carcinoma. This study claimed that a diagnostic pattern in serum could correctly classify 50/50 ovarian cancer samples, including 18 stage I patients, and 63/66 unaffected women or women with non-malignant disorders. Correct classification of control and ovarian cancer samples was achieved with 100% sensitivity and 95% specificity. Following publication, a plethora of studies were reported where SELDI MS and bioinformatics were used for the investigation of the serum proteome. In another ovarian cancer study, three panels of biomarkers were revealed that could distinguish between benign and ovarian cancer with sensitivities and specificities greater than 80% (Kozak *et al.* 2003). Ye *et al.* (2003) identified haptoglobin-alpha subunit as an independent ovarian cancer biomarker, via SELDI MS, showing a sensitivity of 64% and a specificity of 90%. When used in conjunction with an already known ovarian cancer biomarker, CA125, the sensitivity and specificity for identifying ovarian cancer patients increased to 91% and 95%, respectively (Ye *et al.* 2003). Another study also identified three candidate biomarkers for the detection of early stage ovarian cancer and when these were used in conjunction with CA125 early stage patients were distinguished from controls with a sensitivity of 83% and a specificity of 94% (Zhang *et al.* 2004b). The sensitivity of CA125 for detecting early stage ovarian cancer is 50-60%; thus combining CA125 with SELDI MS analysis has improved upon the current screening methods for ovarian cancer (Menon *et al.* 2000). Serum based studies using SELDI MS have also been reported in prostate cancer where protein patterns have been identified that distinguish between prostate cancer and benign disease. One study showed that 36/38 (95%) prostate cancer patients and 177/228 (78%) of patients with benign conditions were classified correctly (Petricoin *et al.* 2002d).

6.1.3 Current problems associated with serum proteome analysis

The use of SELDI MS for serum based studies is popular due to the ease of obtaining serum from patients and the high-throughput nature of the SELDI MS method. However,

there has been criticism of this method, regarding the reproducibility of the protein profiles obtained and whether proposed biomarkers are cancer-associated or are 'epiphenomena', produced by other organs in response to the presence of the cancer or due to the generalised condition of the patient, for example malnutrition, infection or acute-phase reaction. These critics have highlighted the need for validation and reproducibility in multiple sample sets and multiple centres as well as the need for the identification of these proposed biomarkers (Baggerly *et al.* 2005; Diamandis *et al.* 2005). The results in prostate cancer, published in 4 papers and generated by 3 research groups, have shown that the discriminatory biomarkers identified using SELDI MS are very different in each study. This suggests that the SELDI MS technique may not be reproducible and that the suggested discriminatory markers are not consistent for the same cancer in multiple centres analysed using the same technology: these studies reported different values for the sensitivity and specificity of the prostate cancer diagnosis (Diamandis 2003a). Given the complexity of the data generated by SELDI MS, differences in serum collection, storage and experimental handling could account for the different results. Thus, sample collection, pre-processing and storage should be standardised in order to eliminate experimental artifacts (Sorace *et al.* 2003). Other differences between samples could also occur as a result of patient differences, for example, gender, sex, ethnic background, menopausal status, nutritional status, drug use, and so on (Diamandis 2004b). In Petricoin *et al.* 2002a published in *The Lancet*, there was a 10 year difference in the mean age of the control and the ovarian cancer group so that differences observed in the proteome profiles could be due to the age or menopausal status of the patients rather than the presence or absence of the cancer.

6.1.4 Current serum biomarkers in melanoma

S-100 is an acidic calcium-binding protein identified in serum from melanoma patients and is used as an immunohistochemical marker for the diagnosis of melanoma. The serum concentration of this protein has been shown to be elevated in patients with stage III and IV melanoma compared with controls and can be used to detect advanced disease in melanoma patients with 82% sensitivity and 91% specificity. By combining the serum concentration of this protein with the Breslow thickness of the tumour, the sensitivity and specificity of identifying the metastatic spread was improved to 91% and 95%, respectively (Abraha *et*

al. 1997). This serum marker cannot be used for the primary diagnosis of melanoma as only 9% of patients with stage I and II disease have serum concentrations significantly higher than controls. The S-100 protein has also been found at high concentrations in normal tissues as well as in other benign and malignant tumours and cannot therefore be considered as melanoma specific (Kligman *et al.* 1988; Molina *et al.* 2002). Lactate dehydrogenase (LDH) is an independent prognostic indicator of stage IV melanoma. Elevated serum levels of this marker indicate poor survival in melanoma patients when used in combination with other indicators as levels of LDH can be elevated due to haemolysis and other factors unrelated to melanoma (Balch *et al.* 2004). Elevated serum levels of 5-S-Cysteinyldopa (5-SCD) have been reported as a marker for predicting response to systemic therapy. Elevated levels were found in all patients with metastatic disease with a 50-fold increase in levels in patients with advanced metastatic disease. During chemotherapy, levels of 5-SCD decreased in responders, where survival times were higher, but not in non-responders (Wimmer *et al.* 1997). All of these serum markers have shown promise for the diagnosis and prognosis of melanoma patients but the levels of these markers are affected by other factors aside from melanoma and there is a need to develop more accurate diagnostic and prognostic tests.

The aim of this study was to attempt to interrogate the serum proteome of melanoma patients and to identify patterns indicative of early and late stage disease and disease progression.

6.2 Methods

The following scheme (figure 6-1) was used for the analysis of serum samples collected from patients with melanoma, as described in section 2.1.2.3. The methods used are described in sections 2.2.4.2 and 2.2.6, however the parameters set for the bioinformatics analysis were adjusted slightly from those described in section 2.2.6, in order to improve upon the predictions: no CGD algorithm was used, the learning rate was set to 0.01 instead of 0.1 and the 'window' was set to 3000 instead of 2000. Two questions were addressed: firstly a comparison of biomarkers associating with stage I versus stage IV melanoma and secondly biomarkers associating with stage III patients with regard to disease progression.

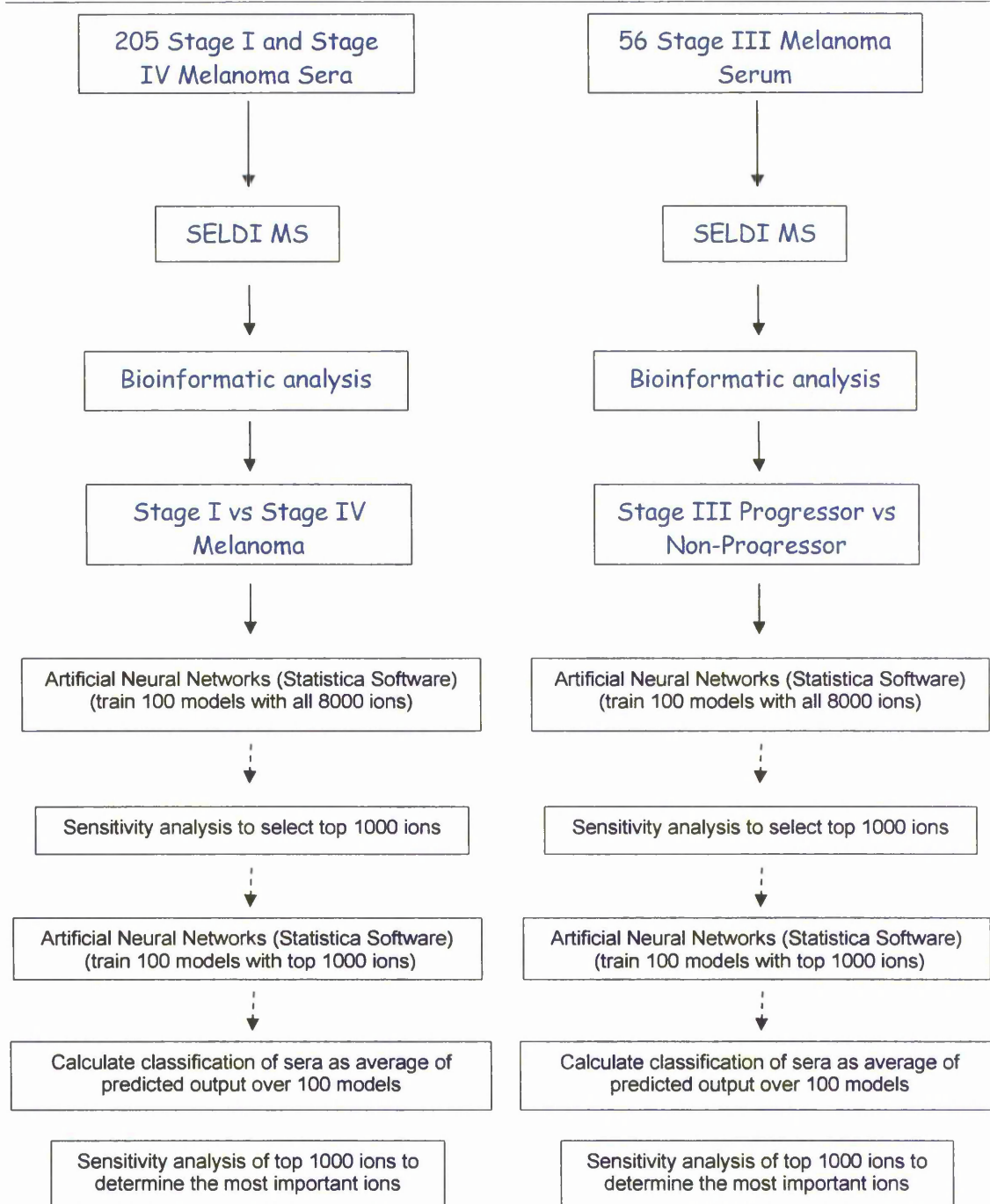


Figure 6-1 Scheme showing the methods used in this study for the analysis of melanoma serum samples.

6.3 Results

6.3.1 SELDI MS analysis

Figure 6-2 shows the protein profiles, obtained by SELDI MS, of sera from stage I melanoma patients overlaid on a profile of a serum sample taken from a stage IV melanoma patient; all three spectra show the 5000-10000 Da mass range. Table 6-1 gives the values of the peaks observed in the profiles shown in figure 6-2, with peaks that are similar between the stage I and stage IV sera highlighted in red. Peaks at 6449, 6646, 7775, 8152 and 8945 Da are present in all of the serum samples, however an ion of approximately 8630 Da is present only in stage IV samples, conversely a peak of approximately 9165 Da is present in all of the stage I but not the stage IV sera.

Figure 6-3 shows the protein profiles, obtained by SELDI MS, of sera taken from stage III melanoma patients. Spectra a) shows a sample taken from a patient whose disease progressed to stage IV overlaid on a profile of a serum sample taken from a patient whose melanoma did not progress. The spectra of the two further progressor samples overlaid with the same non-progressor sample are also given (figure 6-3 b and c). All three of the spectra show the 5000-10000 Da mass range. As in figure 6-2, differences between the profiles of progressing and non-progressing patient sera are observed, and although the profile patterns are similar, the intensities of the peaks are slightly different. Table 6-2 shows the major peaks that are similar between stage I and stage IV sera (highlighted in red). Whereas prominent protein peaks at masses of approximately 6470, 7810, 8180 and 8980 Da are observed in all of the spectra shown in figure 6-3, peaks at 5930 and 9330 Da are only present in stage III non-progressor samples.

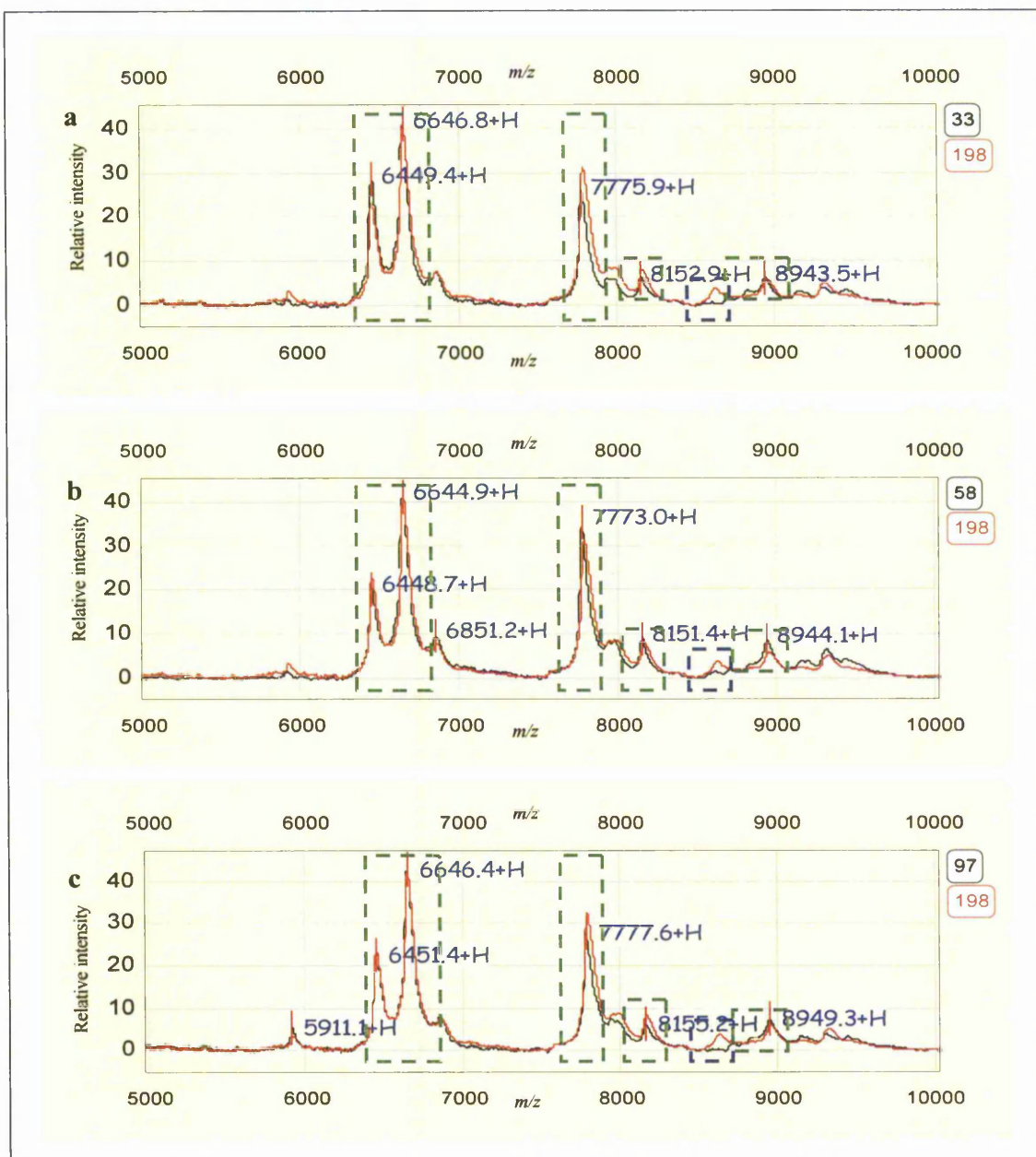


Figure 6-2 SELDI spectra of serum samples highlighting the differences and similarities between samples taken from patients with either stage I or stage IV melanoma.

The figure shows the SELDI spectra of three sera, 33, 58 and 97, taken from stage I melanoma patients (in black) overlaid on a spectra of a serum sample, 198, taken from a stage IV patient (in red), in the 5000–10000 Da mass range. Similarities in peaks between the spectra are highlighted by green boxes and peak differences are highlighted by blue boxes.

198 (stage IV sample)	33 (stage I sample)	58 (stage I sample)	97 (stage I sample)
5915.1	-	-	5911.1
6448.1	6449.4	6448.7	6451.4
6647.4	6647.2	6644.9	6646.4
6856.9	6853.2	6851.2	
7785	7775.9	7773	7777.6
-	7938.4	-	7941.4
7980.8	-	7976.6	-
8152.9	8161.3	8151.5	8155.2
-	8630.3	-	-
8943.5	8955.4	8944	8949.3
9166.1	-	9199.3	9160.9
9307.2	9308.1	9317.7	-
9482.8	-	-	-

Table 6-1 Values of the main peaks observed in the spectra shown in figure 6-2.

The table shows the main peaks (in Daltons) observed in the spectra obtained from stage I and stage IV melanoma serum, as shown in figure 6-2, in the 5000-10000 Da mass range. Peaks that are similar between the samples are shown in the same row, although the m/z values may be different due to the mass accuracy of the instrument. Peaks that are observed in both the spectra of stage I and stage IV melanoma serum samples are highlighted in red.

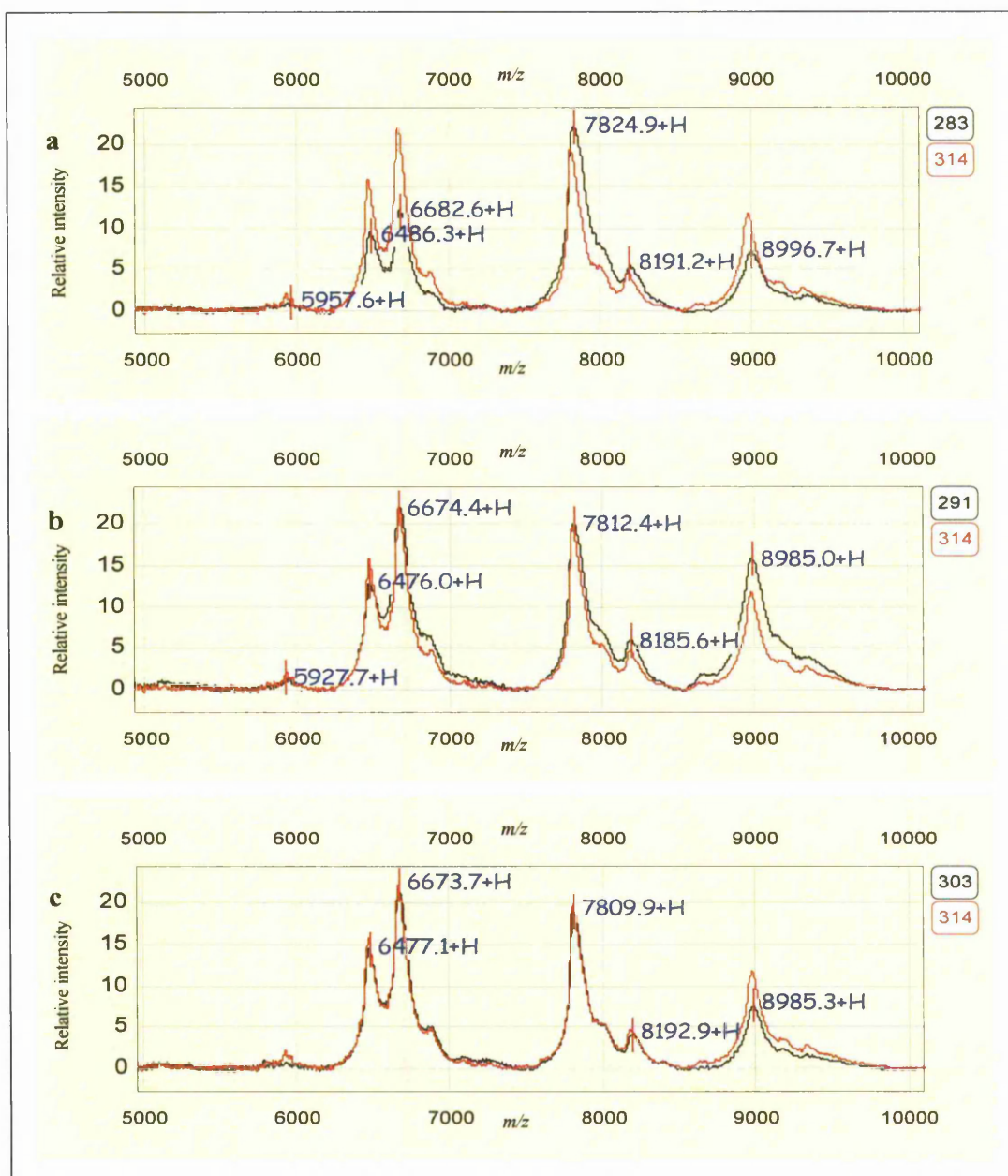


Figure 6-3 SELDI spectra of sera taken from stage III melanoma patients comparing the peak patterns of spectra generated from patients that progress to stage IV melanoma and patients that do not.

The figure shows the SELDI spectra of three sera 283, 291 and 303 taken from patients whose melanoma progressed to stage IV (in black) overlaid on a spectra of a serum sample, 314, taken from a patient whose melanoma did not progress to stage IV (in red).

314 non-progressor	283 Stage III progressor	291 Stage III progressor	303 Stage III progressor
5929.5	-	-	-
6471.7	6486.3	6476	6471.7
6668.2	-	6674.4	6665.8
-	6882.6	-	6884.7
7801.8	7824	7812.3	7805.2
8178.1	8191.2	8185.6	8187.2
8975.6	8996.7	8984.8	8984.2
9333.3	-	-	-

Table 6-2 Values of the main peaks observed in the spectra of serum samples taken from stage III progressing or non-progressing melanoma patients.

The table shows the main peaks (in Daltons) observed in the spectra obtained from stage III melanoma patients that did or did not progress to stage IV melanoma. Peaks that are similar are shown in the same row, although the m/z values may be different due to the mass accuracy of the instrument. Peaks that are observed in both the spectra of progressing and non-progressing patients are highlighted in red.

The data from all of the serum profiles was exported and analysed by ANNs to determine if ions could be identified that could be used to indicate either stage I or stage IV disease, or identify patients that progress their disease from stage III to stage IV melanoma.

6.3.2 Bioinformatics Analysis

The anonymised clinical data for patients was obtained from Dr Selma Ugurel and provided as part of the OISTER project. In order to determine if SELDI MS and ANNs could be used to identify protein expression patterns in melanoma serum samples 2 parameters were chosen for mining the SELDI MS data. The first analysis was to determine if the analysis could discriminate between early or late stage disease and secondly to attempt to discover patterns correlating with progression of disease in patients with stage III melanoma.

The data presented here was generated from 205 serum samples, 101 stage I and 104 stage IV, via SELDI MS using reverse phase protein chips for sample capture. This data was originally analysed in a previous study in order to determine protein patterns that could be discriminate between stage I and stage IV serum samples and between stage III patients that do or do not progress to stage IV melanoma (Mian *et al.* 2005); the methods used and the results obtained are described in detail in this publication. The published data was analysed by ANNs in mass ranges of 2000-5000 Da, 10000-15000 Da and 15000-20000 Da, and was restricted due to the limitations of the Neuroshell software package used. Exploratory analysis of the data in the 2000-30000 Da mass range revealed a peak at 11700 Da that was present at higher intensities in the stage IV serum samples compared to the stage I samples. For ANN analysis, the samples were divided so that 25 stage I and 25 stage IV samples were randomly selected as the independent validation set, and the remaining 155 samples were split into 94 samples for training, 15 for testing and 46 as additional blind samples. ANN analysis revealed that the best predictive capability was obtained from the 2000-5000 Da mass range. After training 50 models, the best performing model predicted the class assignment for 96 blind stage I and stage IV melanoma samples with an 88% accuracy and a sensitivity and specificity of 84% and 92%, respectively. With the stage III progressor and non-progressor samples, correct class assignment was obtained for 80% of the samples using the 2000-5000 Da mass range to train 51 models.

The same data set was analysed in this present study using Statistica software, enabling the full data range from 2000-30000 Da to be utilised for the training the ANNs, to discriminate between stage of disease and progressing and non-progressing patients. SELDI MS data in the 1000-30000 Da mass range was exported into EXCEL as csv files and converted into .xls files before being merged into one file and smoothed to 3 Da for

every sample analysed. A multi-layer perceptron was trained, using Statistica software, on the smoothed mass/intensity data, where only one hidden layer was used. A more detailed explanation of the ANN method is given in section 2.2.6.

6.3.2.1 *Stage I vs Stage IV serum analysis*

The first analysis was to compare the predictive capability of the ANNs to discriminate between 205 serum samples obtained from patients with either stage I or stage IV melanoma by analysing their SELDI MS profiles (H50 chips).

A back propagation algorithm was used to identify patterns in the data and 123 samples were randomly selected for training, 41 for the test set and 41 for the blind dataset for each model, with 100 models run. The serum samples originating from patients with stage I melanoma were assigned an output of 1 and the serum samples originating from patients with stage IV melanoma were assigned an output of 2.

The ANNs correctly classified the serum samples originating from either stage I or stage IV melanoma patients with an accuracy of $98.5\% \pm 0.007$ and a sensitivity and specificity of 98% and 99%, respectively, using the top 1000 ions from the serum data as identified by sensitivity analysis. Figure 6-4 shows the population distribution of the predicted outputs for all 205 samples. Samples originating from stage I patients are highlighted in blue and samples stage IV patients are highlighted in red. The figure shows that 3 of the serum samples were misclassified; sample 118 (stage IV patient) was classified as originating from a stage I patient with an output of 1.002 whilst samples 95 and 76 (stage I patients) which were given predicted outputs of 1.737 and 1.998, respectively.

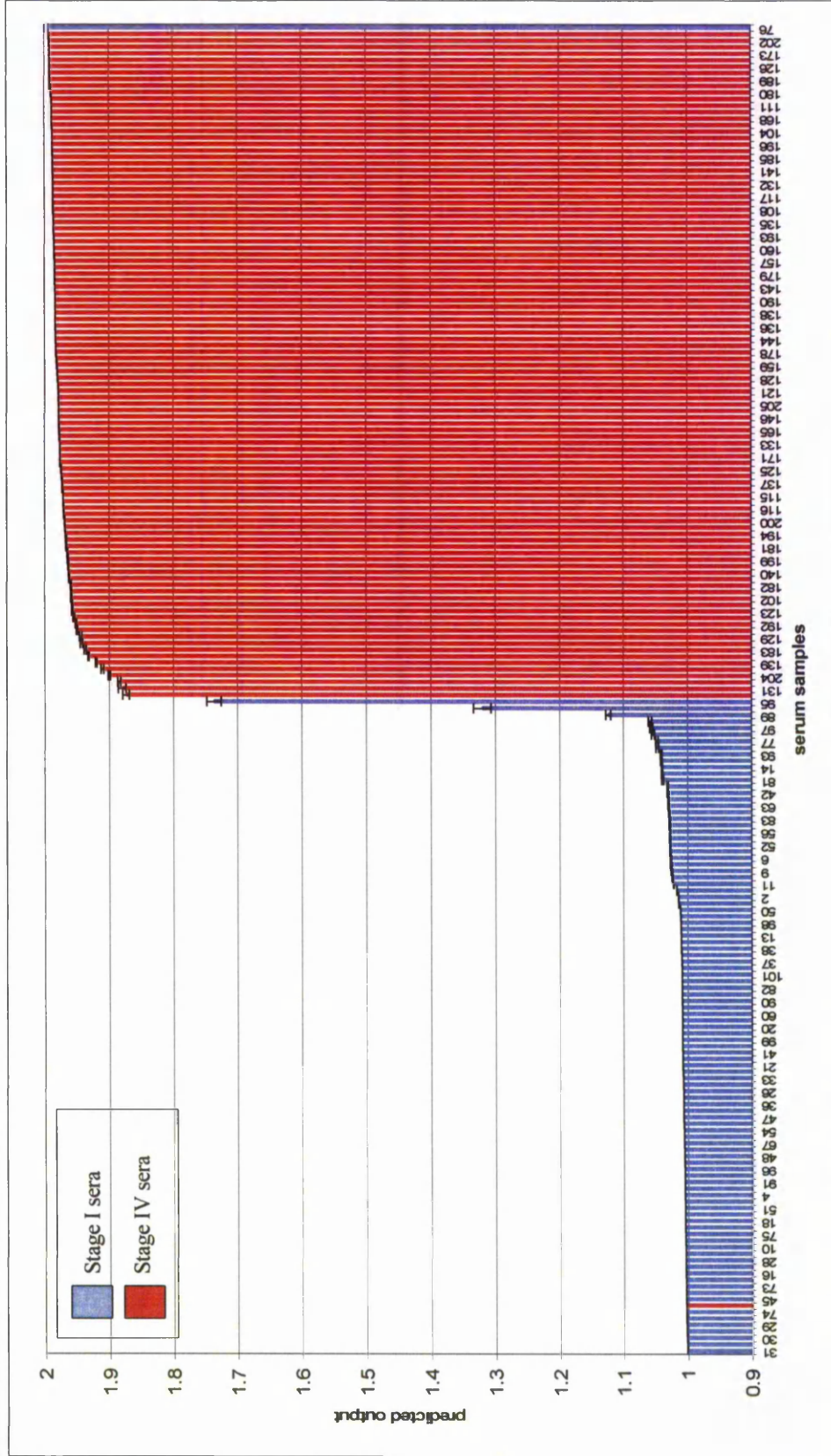


Figure 6-4 Classification by ANN analysis of 205 stage I and IV melanoma serum samples. The figure shows the average predicted outputs of 205 stage I and IV melanoma serum samples as calculated by ANN analysis over 100 models. Samples that were taken from stage I patients were assigned an output of 1, blue columns, and samples that were taken from stage IV patients were assigned an output of 2, red columns. The error bars show the 95% confidence intervals. Sample 118 was misclassified as a stage I sample and samples 95 and 76 were misclassified as stage IV samples.

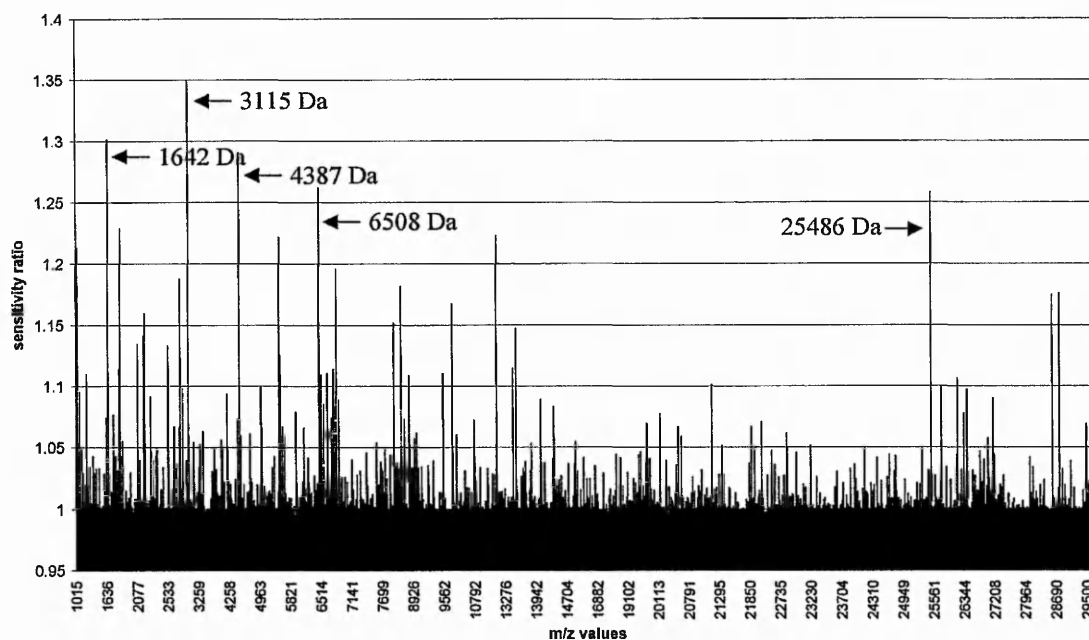


Figure 6-5 Sensitivity analysis of the top 1000 ions used to classify samples as stage I or stage IV melanoma serum samples.

The above figure shows the sensitivity analysis of the top 1000 ions, from the stage I vs stage IV serum data, used to train the ANN models. The graph shows the sensitivity ratio as an average over the 100 models trained. Approximately 20 ions are shown as being the most important of the top 1000 for classifying the serum samples according to their stage.

Figure 6-5 shows the sensitivity analysis of the top 1000 ions. These 1000 ions were used to train the ANNs for all 100 models and enabled the ANNs to predict stage of disease with 98.5% accuracy for stage I and IV serum samples. The sensitivity analysis shows that a few of the top 1000 ions stand out as being more important than the rest for classifying the samples, although the ratio values overall are still not very high; the most important ion has a ratio value of 1.35. It is also important to note that the most important ions, as identified by the sensitivity analysis, are all within the low molecular weight range. 185 of the top 1000 ions were found to have sensitivity ratio values of less than 1 and so their removal from the training set may help towards improving the predictions.

6.3.2.2 Stage III progressors vs non-progressors

The second analysis was to compare the predictive capability of the ANNs to discriminate between serum samples obtained from 56 stage III melanoma patients where 28 of the

patients progressed to stage IV within the first year of follow-up and the remaining 28 patients did not.

A back propagation algorithm was used to identify patterns in the data and 34 samples were randomly selected for training, 11 for the test set and 11 for the blind dataset for each model, with 100 models run. The serum samples originating from progressing patients were assigned an output of 1 and the serum samples originating from patients who did not progress to a further stage of melanoma were assigned an output of 2.

Serum samples were correctly classified as originating from either stage III progressing or non-progressing melanoma patients with an accuracy of $94.6\% \pm 0.091$ and a sensitivity and specificity of 96.4% and 92.9%, respectively. Figure 6-6 shows the population distribution of the predicted outputs for all 56 samples with samples from stage III patients who progressed to stage IV highlighted in blue and samples from stage III patients who did not progress to stage IV highlighted in red. The figure shows that 3 of the serum samples were misclassified; samples 306 and 319 (non-progressing patients) and sample 299 (a progressing patient).

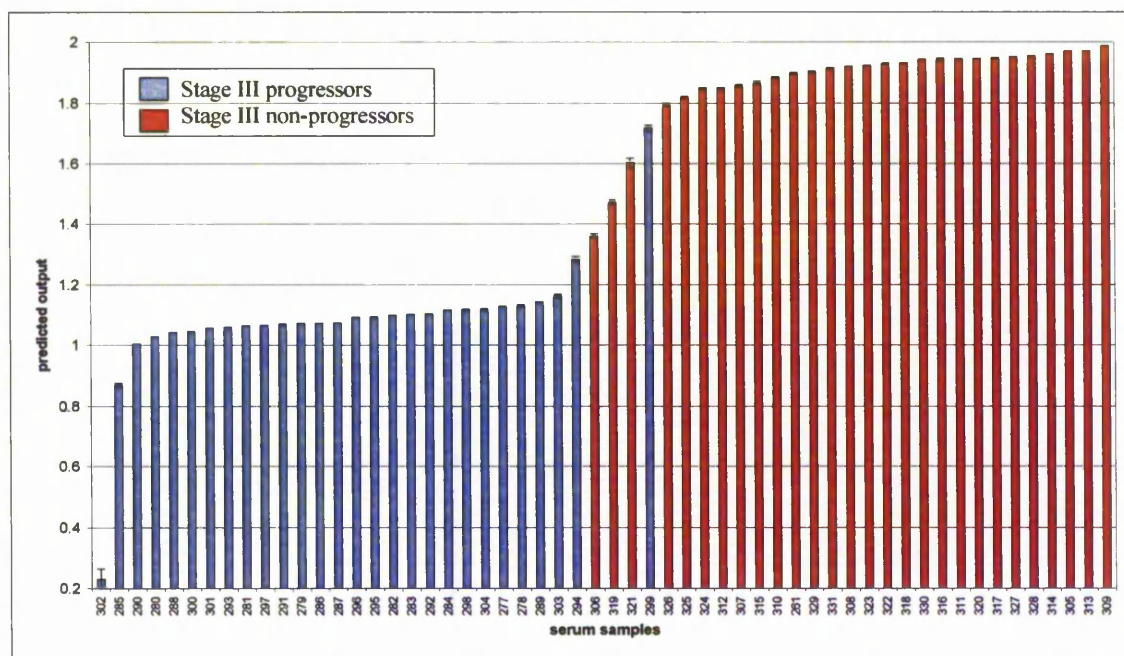


Figure 6-6 Classification by ANN analysis of 56 stage III patients according to whether the patients progressed to stage IV melanoma or not.

The figure shows the average predicted outputs of 56 stage III melanoma serum samples as calculated by ANN analysis over 100 models. Samples that were taken from patients who progressed to stage IV were assigned an output of 1, blue columns, and samples that were taken from patients that did not progress to stage IV were assigned an output of 2, red columns. The error bars show the 95% confidence intervals. Samples 306 and 319 (highlighted in red) and sample 299 (highlighted in blue) were misclassified.

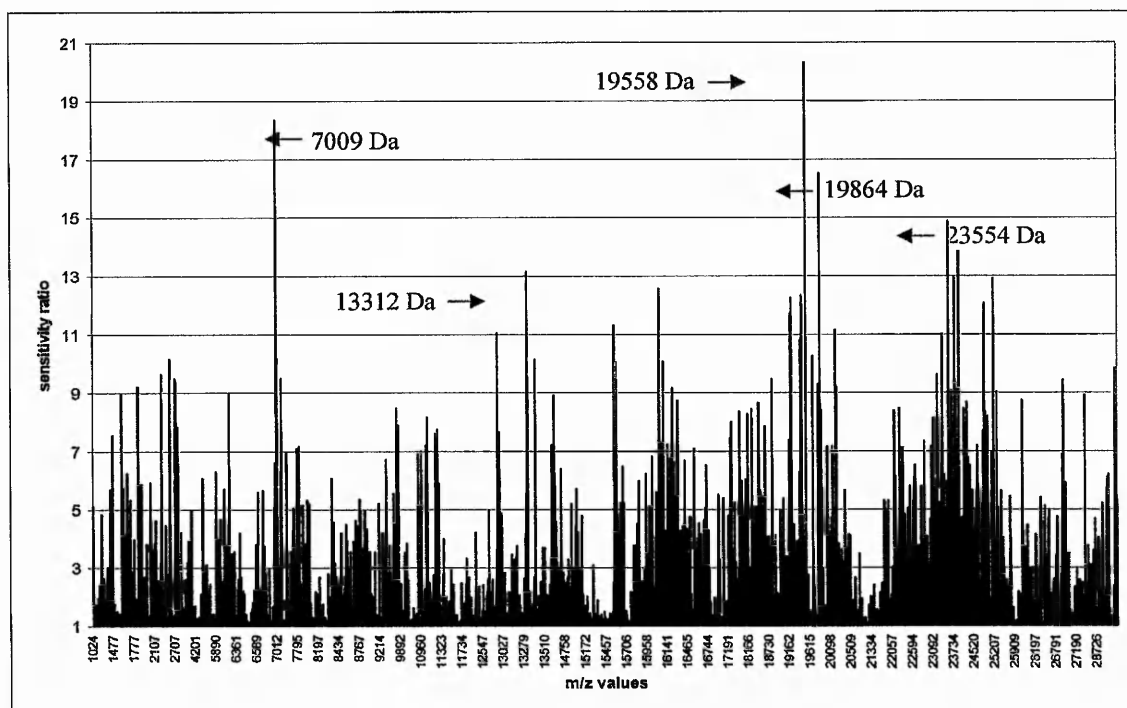


Figure 6-7 Sensitivity analysis of the top 1000 ions used to classify stage III sera taken from progressing or non-progressing melanoma patients.

The above figure shows the sensitivity analysis of the top 1000 ions, from the stage III serum data, used to train the ANN models. The graph shows sensitivity ratio as an average over the 100 models trained.

The sensitivity analysis, illustrated in figure 6-7, shows that many of the top 1000 ions have high a sensitivity ratio showing that those particular ions are important for the identification of stage III patients that will or will not progress to stage IV melanoma. All of the top 1000 ions were found to have sensitivity ratio values of greater than 1.

6.4 Discussion

The initial study (Mian *et al.* 2005) revealed that different potential biomarkers could be identified using ANN analysis, with the most correct predictions obtained from the 2000-5000 Da dataset and not the 10000-15000 Da dataset. When testing the equality of variances between stage I and stage IV samples, the F score revealed that a peak at 11700 Da was significant.

The same data was re-analysed using a different software package, Statistica, where the full data set from 1000-30000 Da could be included. Interestingly, this analysis revealed that diagnostic patterns between stage I and IV melanoma serum can be identified with a 98.5% accuracy, which is greater than what was previously reported. The best performing ions, as revealed by the sensitivity analysis were seen at the low molecular weight range and this correlates with the analysis performed by NeuroShell, where the best predictive capability came from the 2000-5000 Da mass range. Ferrari *et al.* (2000) obtained sera from 17 melanoma patients at different stages of the disease and, following ultrafiltration using spin-columns, analysed them by MALDI MS. They reported that protein ions in the mass range of 2.5-3.5 kDa were completely absent in control sera compared to sera taken from stage I melanoma patients and that the abundance of these low molecular weight proteins increased in sera taken from stage II melanoma patients (Ferrari *et al.* 2000). However, another study evaluated sera from 27 melanoma patients with differing stages of the disease compared with sera taken from 10 healthy subjects using ultrafiltration, MALDI MS and cluster analysis and revealed that proteins in the 10000-15000 Da mass range were particularly useful for identifying patients with malignant melanoma (Ragazzi *et al.* 2003). Ferrari *et al.* compared the relative abundances of peaks observed by eye while Ragazzi *et al.* used cluster analysis for their data in order to determine which proteins discriminated between different patient groups.

In the present study, the analysis of the stage III sera revealed that discriminatory patterns were present that could be used to identify patients that progressed to stage IV melanoma. This was achieved with a 94.6% accuracy using 56 serum samples; the sensitivity analysis revealed that some ions had very high ratios, indicative of their importance for determining progressing or non-progressing patients within 1 year of follow. What would have been interesting to see is if the two misclassified non-progressing samples were from patients

that actually progressed to stage IV melanoma soon after the 1 year follow-up and therefore should have originally been classed as progressors. Unfortunately data on these patients after 1 year follow up is unavailable and so this is not possible to determine.

As stated in section 6.1.1, some of the discriminating ions at low molecular weights may be degradation products or fragments of larger proteins and further interrogation of these would be required to confirm their true identity. Although the identity of a marker is not necessary for its clinical use in identifying high-risk patients, information on its identity would increase confidence in the biological basis of the discriminatory pattern (Jacobs *et al.* 2004). It is important to note that these biomarker ions may not all be cancer related. A study identified 3 biomarkers that could be used for detecting ovarian cancer with high sensitivity and specificity, one of which was identified as transthyretin (Zhang *et al.* 2004b). It was revealed that transthyretin is a protein that is synthesised in the liver in response to nutritional supply and the levels of this protein are affected by acute and chronic diseases inducing an acute phase response (Schweigert *et al.* 2005). Schweigert *et al.* highlighted the fact that effects caused by nutritional status, inflammatory processes and possible hepatic diseases should be taken into consideration when selecting control groups in such studies. It has been suggested that using serum protein profiling and bioinformatics analysis would identify only high-abundance proteins that are not released by the tumour which are instead representative of non-specific epiphenomena caused by the presence of the cancer (Diamandis 2004a). Many of the “serum profiling” studies published so far have failed to identify the discriminatory peaks and in studies that have identified the discriminatory peaks, these biomarkers are acute-phase reactants produced by the liver in high concentrations in response to inflammation (Diamandis & van der Merwe 2005). One study identified haptoglobin as a potential biomarker for pancreatic cancer but further investigation using a biochemical test to identify haptoglobin, failed to show differences between patients and controls (Koomen *et al.* 2005). It is important, therefore, to identify discriminating ions that are truly disease related.

The melanoma data presented here has been analysed by ANNs but with additional software packages and different processing parameters. The first analysis, performed by S Mian (2005), used Neuroshell, whilst the current analysis was performed using Statistica. These two different approaches used ANNs to analyse the same dataset but in different ways, with the approach presented here classifying samples with greater accuracy. Thus,

classification of blind data can be improved by different data mining procedures and optimisation and standardisation of the statistical approach is important and essential.

Following Petricoin's publication in the *Lancet* in 2002, additional studies concluded that sophisticated learning algorithms could be used to correctly identify complex serum patterns from patients with or without breast cancer, prostate cancer and ovarian cancer with sensitivities and specificities of greater than 85%, giving a higher diagnostic accuracy than could be achieved with conventional serum biomarkers. However, Baggerley *et al* (2004) highlighted their concern with regard to reproducibility when they analysed the Petricoin data (original datasets were freely available on the internet). They found inconsistencies in the way the SELDI MS experiments were performed, particularly in relation to the baseline correction of the spectra and mass calibration; they also found that multiple spectra obtained from the same sample could not be overlaid. The author's concluded that the high accuracy of identifying patients with or without ovarian cancer was probably due to artefacts from sample processing and shifts in the mass accuracy of the spectra. This paper re-inforced the need for standard procedures to be implemented for this approach to be widely accepted as a diagnostic technique for clinical use. Specifically, preanalytical sample preparation, parameters used for obtaining MS spectra, the use of baseline correction and peak extraction as well as the type of bioinformatic programme used are all important areas for consideration in order to standardise and validate this approach.

Chapter 7 Comparison of mass spectrometry platforms for protein biomarker identification in melanoma cell lines carrying specific genetic mutations

7.1 Introduction

As has already been discussed in previous chapters the field of proteomics has grown rapidly over the past few years, especially in the area of proteomic profiling and biomarker identification using mass spectrometry. One pioneering study was published by Petricoin *et al* (2002b) and claimed that patterns in the mass spectral data generated by SELDI MS from patient serum could be used to identify and diagnose patients with ovarian cancer. This 'breakthrough' paved the way for several of studies utilising SELDI MS spectra in order to identify discriminatory patterns in a number of diseases, including cancer, with the hope that clinically relevant diagnostic and prognostic biomarkers could be identified.

When analysing complex mixtures by MALDI MS, there is a competition between the different proteins and salts for ionisation and as a result a phenomenon known as ion suppression occurs, this can lower the number of ions detected as well as reduce the sensitivity of detection and as such most proteins that are detected in complex mixtures such as plasma have concentrations of $>1\mu\text{mol/L}$ (Annesley 2003). This limit of detection can be increased if the sample is pre-fractionated before MALDI MS analysis and this can be employed using a number of common approaches such as solid-phase extraction, dilution, chromatography, fractionation or capture on magnetic particles (Gilar *et al.* 2001; Hortin 2006; Jin *et al.* 2005). This is where SELDI MS has the advantage over MALDI as although the MS analysis is MALDI, the sample clean up is via the use of a target surface (chip) with different retentate chromatographies. This chromatography is modified to contain ion-exchange, hydrophobic, hydrophilic or metal-chelate groups allowing for the fractionation of a complex sample according to the properties of the proteins. Once unbound proteins and salts have been washed away, the chip can then be inserted directly into the MS instrument and analysed. This ease of sample clean up and the fact that only a small amount of sample is required has made this technology extremely popular, especially for serum based studies (Simpkins *et al.* 2005; Tang *et al.* 2004). SELDI MS is a variant of

MALDI-TOF and both techniques have provided the most sensitive detection of low molecular mass proteins whereas traditional methods such as 2-D PAGE and liquid chromatography have an increased or equivalent sensitivity for detecting large proteins. The advantages and disadvantages of using SELDI MS or MALDI MS are detailed in table 7-1.

Technology	Principle	Advantages	Disadvantages
MALDI MS	Application of a protein sample onto a target plate; desorption of proteins from the plate by a laser and measurement of the protein masses and their peak intensities by time-of-flight; comparison of peak intensities between multiple samples	High throughput (up to 1536 samples per plate); post translational modifications can be determined	Need for (offline) sample fractionation of complex samples; more starting material needed for offline sample fractionation; unsuitable for high molecular weight proteins (<100 kDa)
SELDI MS	Variant of MALDI-TOF MS in which certain proteins from a complex sample are bound to a specific chromatographic surface and the rest washed away	High throughput (up to 96 samples if a Bioprocessor is used); direct application of neat sample onto target (fast on-chip sample cleanup); post translational modifications can be determined; only small amount of starting material needed	Unsuitable for high molecular weight proteins (<100 kDa); only proteins bound to the chip surface are detected; lower resolution and mass accuracy than, for example, MALDI-TOF

Table 7-1 Table comparing MALDI MS and SELDI MS technologies, adapted from Engwegen *et al.* 2006.

In the last few years there has been much debate regarding SELDI MS reproducibility which has caused concern over some published studies. One of the main reasons it has been difficult to compare studies from different groups is that the datasets that are compared and are generated using different assay procedures (Baggerly *et al.* 2004). To date only a few studies have identified proteins belonging to ions which were detected using SELDI MS; one of the reasons being that this instrument does not have the capacity for PSD function

however, with the advent of hybrid systems that are able to house SELDI MS sample platforms identification has been possible. Reproducibility studies (shown in chapter 2) have shown that the chip surfaces were very reproducible with a CV of 14-21% after normalisation which is comparable with MALDI MS studies. The difference in the 2 technologies however, is the sensitivity and resolution in instrumentation; SELDI MS leads to more coalesced peaks and maybe to loss of resolution of closely related m/z peaks (Koomen *et al.* 2005) whereas recent improvements in MALDI machines has lead to greater resolution of closely related proteins/peptides.

The use of MALDI MS with other sample fractionation methods may prove to be more beneficial to biomarker discovery studies if these methods can be shown to be reproducible as well as high-throughput. A relatively new technique using C8-functionalised magnetic beads for enriching a specific subset of proteins within plasma and serum before MALDI MS is a promising development (Zhang *et al.* 2004a). Zhang *et al* have shown that by using these magnetic beads to enrich plasma samples taken from 12 patients with asthma and 12 controls, significant differences could be seen between the two groups and the coefficients of variation for intraday reproducibility was 18% and the interday reproducibility was 26%. Another study analysed 56 serum samples, 34 of which were from glioma patients and 22 from healthy volunteers (Villanueva *et al.* 2004). The samples were subjected to magnetic bead based sample processing using automated liquid handling robot followed by MALDI-TOF MS. The use of the liquid handling robot allowed for high throughput and reproducibility of sample processing and although only a small set of serum samples were used, patterns were identified within the MALDI spectra that could be used to distinguish between the control and glioma patients with 96.4% accuracy (53 out of 55 were correctly classified). The use of magnetic beads holds promise as a sample clean up technique but it has yet to be shown to be high throughput and reproducible for biomarker discovery studies that would require the processing of large numbers of samples. Another method that can be employed to reduce the complexity of samples for MALDI MS analysis is solid phase extraction using ZipTips. This approach has been in used for the clean up of tryptically digested spots extracted from 2-D gels before MS analysis (Zhang *et al.* 2006b). Bronchoalveolar lavage fluid samples were investigated, 75 of which were from patients with bronchiolitis obliterans syndrome (BOS) and 49 samples were from healthy volunteers. The samples were subjected to ZipTip clean up followed by MALDI MS

analysis and analysis revealed protein profile changes that were indicative of a high chance of developing BOS. This analysis provided a rapid and inexpensive method to identify many new proteins that are linked to chronic lung transplant rejection which is currently being validated in a large prospective study.

The aim for this study was to compare SELDI MS with MALDI MS instrumentation whilst using the same bioinformatics analysis to determine if the 2 technologies would provide comparative biomarker profiles ions. The sample clean-up methods used in this study were different due to the available chromatographies. However, we chose similar chromatographies using C6-C12 chips for SELDI MS and C18 ZipTips for MALDI analysis. The fact that the resolution of the MALDI instrument compared to the SELDI instrument is better would mean that important ions identified by ANN analysis would have mass values that are more reliable. Also with the increased sensitivity of the MALDI instrument it is thought that more peaks would be detected.

7.2 Methods

72 of the cell line lysates analysed in chapter 3 were used for this SELDI vs MALDI study as these were available in a large enough quantity for analysis in duplicate. One spectra from the duplicate analysis of each cell line was then randomly chosen for bioinformatic interrogation in order to identify patterns relating to the presence or absence of mutations in the *tp53*, *braf* or *p16^{INK4A}* genes.

The following scheme (figure 7-1) was used for the analysis of the cell lines with the methods described in section 2.2.4.3 for the SELDI MS analysis, section 2.2.5 for the MALDI analysis and for the bioinformatics method, section 2.2.6. The SELDI MS data was analysed from 1000-25000 Da but the MALDI data was only analysed from the 1000-15000 Da mass range as no peaks were seen above this range. The SELDI MS and MALDI MS data was analysed by ANNs using exactly the same parameters: back propagation algorithm for 10000 epochs followed by conjugate gradient descent algorithm for 5000 epochs, window set at 2000, with a learning rate of 0.01 and a momentum of 0.3 and weight decay regularisation. Random resampling was used with 100 models trained where samples were randomly selected so that 60% were used in the training, 20% were used in

the test and 20% were used in the blind datasets for each model. A stepwise approach was also performed to show how the SELDI MS and MALDI MS approaches compare when using a different type of analysis; this was achieved as described in section 2.2.6.2. For all of the studies, cell lines possessing a mutation were assigned an output of 1 and cell lines not containing a mutation were assigned an output of 2.

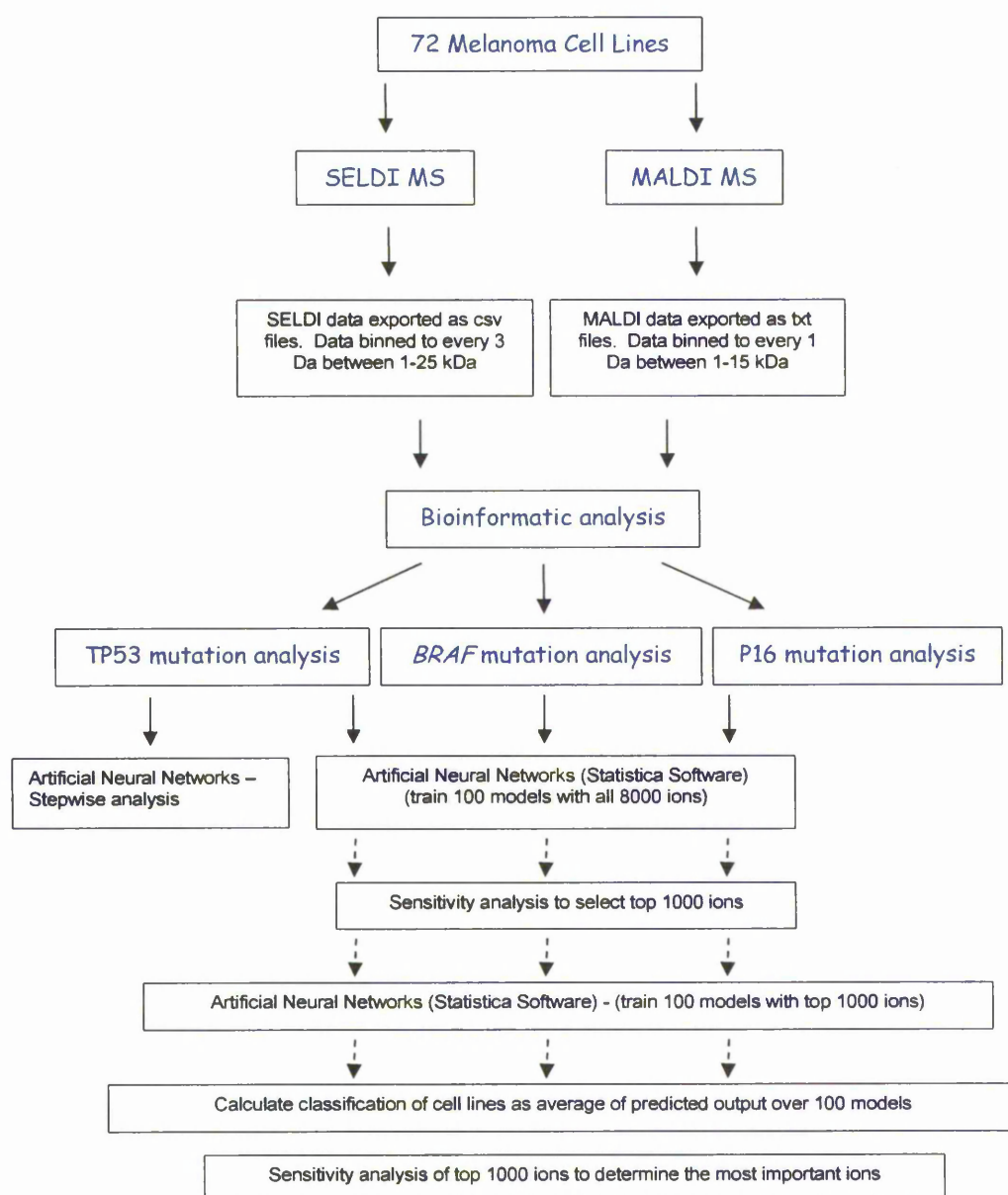


Figure 7-1 Figure showing a flow diagram of the method followed in order to compare SELDI MS and MALDI MS technologies.

7.3 Results

7.3.1 Mass spectrometry

Cell line lysates were analysed by both SELDI MS using H4 SELDI chips (C16 chromatography) and by MALDI MS using C18 ZipTips for sample clean up. Figures 7-2 shows the spectra obtained from cell line FM 93/2 using both methods and figure 7-3 shows the spectra obtained from cell line ESTDAB 027 using both methods.

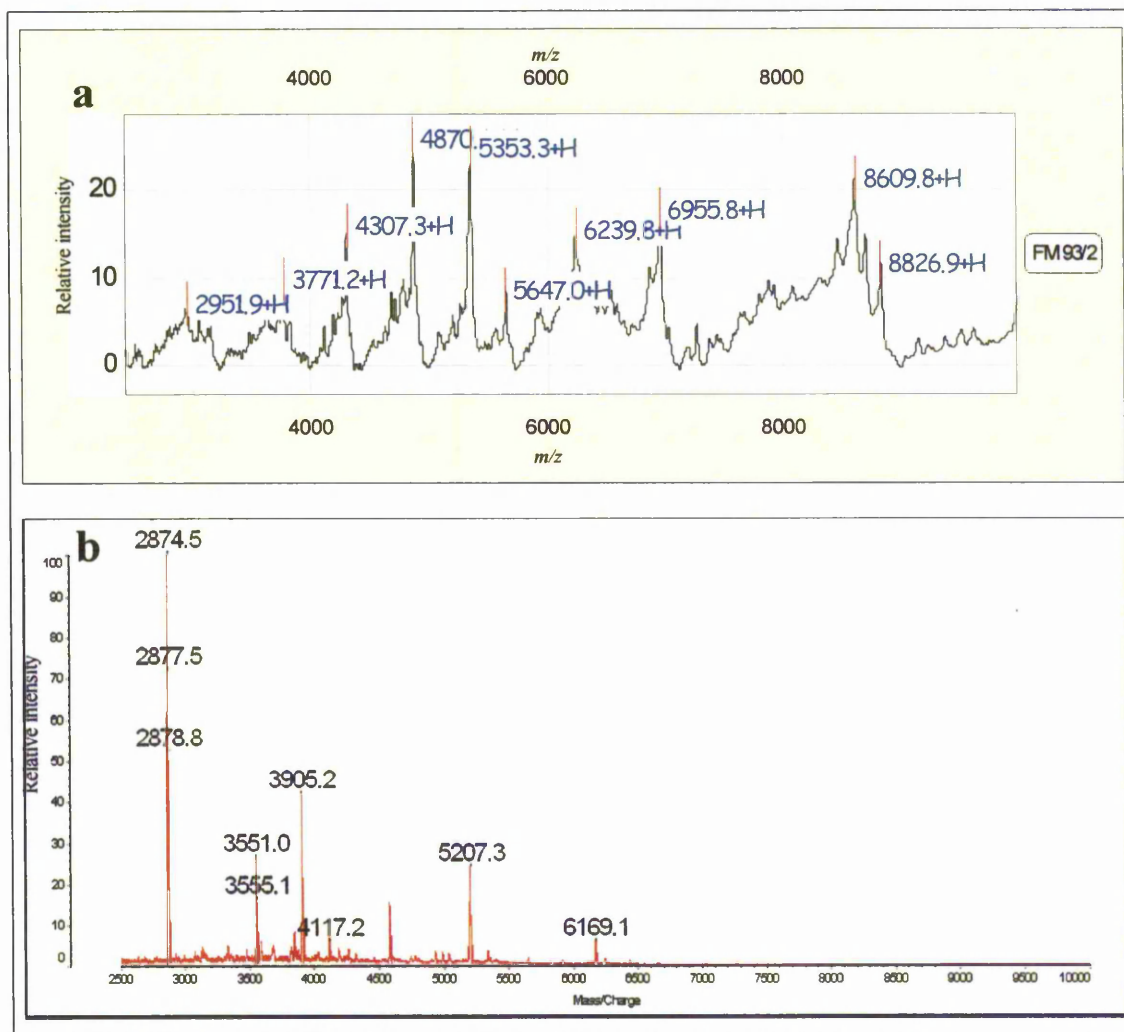


Figure 7-2 SELDI and MALDI spectra generated from melanoma cell line FM 93/2.

The figure shows cell line FM 93/2 analysed by SELDI MS in the top spectra (a), and analysed by MALDI MS in the bottom spectra (b), in the 2500 – 10000 Da mass range.

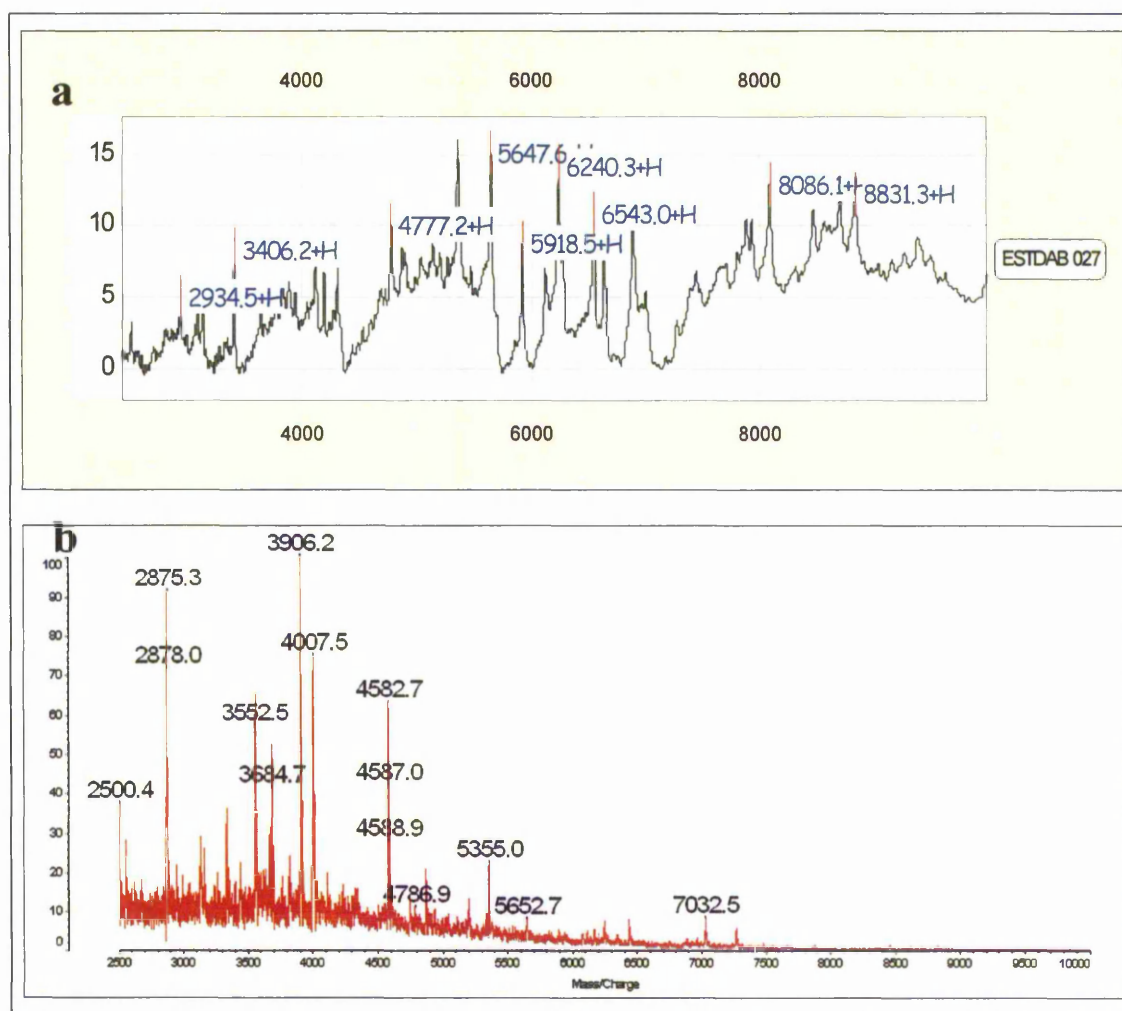


Figure 7-3 SELDI and MALDI spectra generated from analysis of melanoma cell line ESTDAB 027. The figure shows cell line ESTDAB 027 analysed by SELDI MS in the top spectra (a), and analysed by MALDI MS in the bottom spectra (b), in the 2500 – 10000 Da mass range.

The figures show a large difference in resolution between the two MS techniques with SELDI profiles being lower in resolution compared with the MALDI generated spectra. However, visual inspection of the spectra shows that there are more peaks generated at the higher mass range using SELDI MS compared to MALDI MS. It is not possible to determine visually if any peaks detected in the SELDI spectra are the same as those detected by MALDI analysis.

7.3.2 Bioinformatic analysis

Bioinformatic analysis was used to compare SELDI MS data (C16 coated H4 chips) with MALDI data (C18 ZipTips) in order to determine how different profiling methods affect the ability of ANNs to identify discriminatory patterns within the spectra. The cell line data with genetic mutation information was analysed for the presence or absence of the *Braf*, *p16^{INK4A}* and *tp53* genes as before.

7.3.2.1 *Braf* analysis

The m/z/intensity SELDI MS data from 1000-25000 Da of 72 cell lines were used to train 100 ANN models, where 44 cell lines were randomly selected for the training set, 14 were randomly selected for the test set and 14 were randomly selected for the blind data set for each model, i.e. every sample will appear in the training, test and blind dataset at least once over the 100 models. After the training had completed the predictions for the blind dataset of all 100 models were averaged and the standard error of the mean calculated. With the top 1000 ions the ANNs correctly classified cell lines with or without *Braf* mutations in the blind data set with 94.4% accuracy and a sensitivity of 98% and a specificity of 85.7%. The classification of the cell lines as predicted by the ANNs is illustrated in figure 7-4a. This figure shows that 4 cell lines were misclassified; Ma Mel 39a, FM 3.26, FM 79 and WM 451. The sensitivity ratio was calculated for all of the top 1000 ions used to train the 100 models and the average ratio for each ion across the 100 models is illustrated in figure 7-4b.

Five ions at 4192, 12703, 12931, 24925 and 24997 Da have ratio values significantly higher than the other ions.

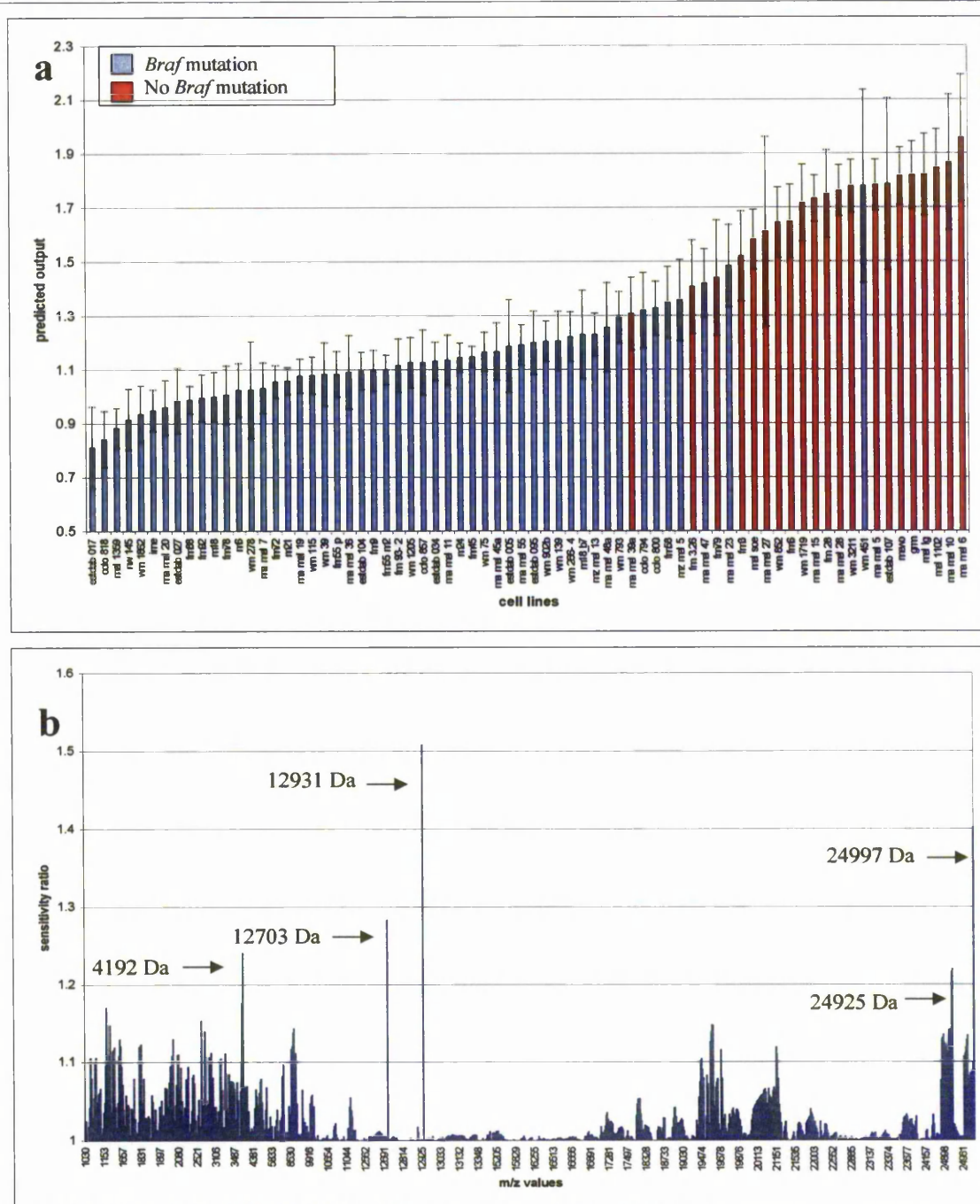


Figure 7-4 Classification by ANN analysis and sensitivity analysis of SELDI data generated from cell lines with or without a *braf* mutation.

The above figure shows results of ANN analysis of SELDI MS data generated from cell lines with or without a *Braf* mutation. Figure a) shows the classification of 72 cell lines as predicted by ANN analysis, averaged over 100 models. The blue bars show the cell lines which have a *Braf* mutation and the red bars show the cell lines which do not have a *Braf* mutation. The error bars were determined by calculating the standard error of the mean with 95% confidence intervals. Figure b) shows a graph of the average sensitivity ratio for the top 1000 ions over the 100 models trained.

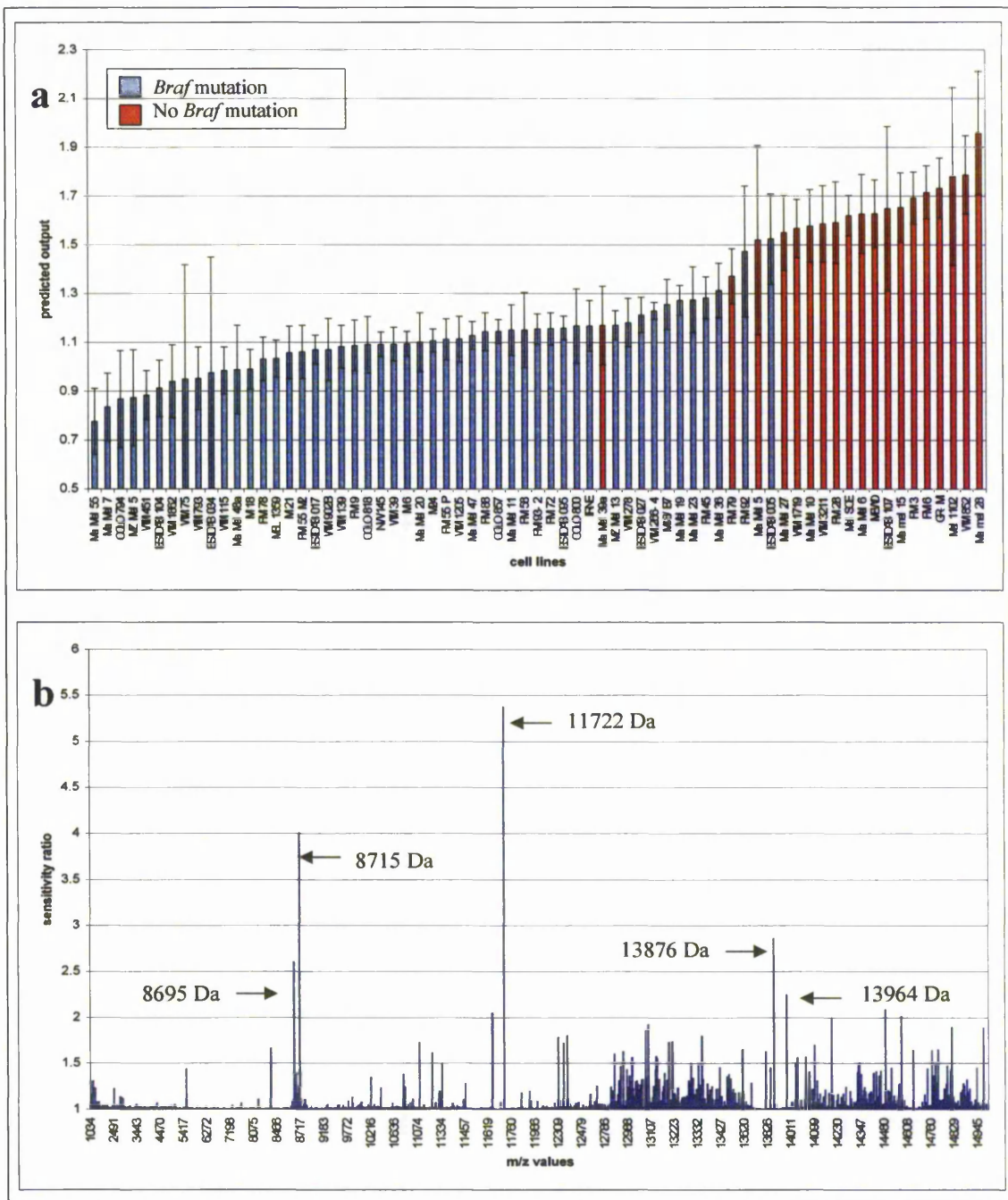


Figure 7-5 Classification by ANN analysis and sensitivity analysis of MALDI data generated from cell lines with or without a *braf* mutation.

The above figure shows results of ANN analysis of MALDI data generated from cell lines or without a *Braf* mutation. Figure a) shows the classification of 69 cell lines as predicted by ANN analysis, averaged over 100 models. The blue bars show the cell lines which have a *Braf* mutation and the red bars show the cell lines which do not have a *Braf* mutation. The error bars were determined by calculating the standard error of the mean with 95% confidence intervals. Figure b) shows a graph of the average sensitivity ratio for the top 1000 ions over the 100 models trained.

Data from three of the cell line samples had to be excluded from analysis due to poor sample spotting onto the MALDI target plate, therefore the m/z intensity MALDI data from 1-15 kDa of 69 cell lines were used to train 100 ANN models, where 41 cell lines were used for training, 14 for testing and 14 for validation (blind). After the analysis had completed the predictions for the blind dataset the models were averaged; this is illustrated in figure 7-5a where the error bars show the standard error of the mean. Using the top 1000 ions, the ANNs correctly classified cell lines with or without *Braf* mutations in the blind data set with 97.7% accuracy, with a sensitivity of 98% and a specificity of 89.5%. Figure 7-5a shows that 3 cell lines were misclassified; Ma Mel 39a, FM-79 and ESTDAB 005. The average sensitivity ratio for all of the top 1000 ions is shown in figure 7-5b, with ions at 8695, 8715, 11722, 13876 and 13964 Da with ratio values significantly higher than the rest.

7.3.2.2 *P16^{INK4A}* analysis

Mutational analysis data of the *p16^{INK4A}* gene was only available for 68 of the 72 cells lines analysed by SELDI MS therefore these 68 cell lines were used to train 100 models (40 cell lines for the training set, 14 for the test set and 14 for the blind data set for each model). After training, the predictions for the blind dataset of all 100 models were averaged and the standard error of the mean calculated, as illustrated in figure 7-6a. The ANNs correctly classified the cell lines with 91.2% accuracy, a sensitivity of 93.2% and a specificity of 87.5%. As can be seen in figure 7-6a, 6 cell lines were misclassified; Ma Mel 28, Mel SOE, Ma Mel 15, Ma Mel 48a, Ma Mel 27 and MZ Mel 5. The sensitivity ratio was calculated for all of the top 1000 ions for each model, the average was determined and illustrated in figure 7-6b. This graph shows that an ion at 22189 Da has a ratio much larger than the rest of the top 1000 showing that it must be extremely important for prediction.

Poor spotting of three of the cell line samples onto the MALDI plate left as well as incomplete mutational analysis data left MALDI data for only 65 of the 72 cell lines. Figure 7-7a shows the average predicted output for each cell line after training the 100 models with the error bars representing the standard error of the mean after analysis of the MALDI data from 65 cell lines. To train 100 ANN models, 39 cell lines were randomly

selected for the training set, 13 were randomly selected for the test set and 13 were randomly selected for the blind data set for each model. The ANNs correctly classified the cell lines with an accuracy of 90.5%, a sensitivity of 96% and a specificity of 74%. The figure shows that 5 cell lines were misclassified; WM 75, M21, ESTDAB 005, WM 793 and Ma Mel 55. Figure 7-7b shows the average sensitivity ratio of the top 1000 ions where a number of ions have ratios significantly higher than the rest of the population.

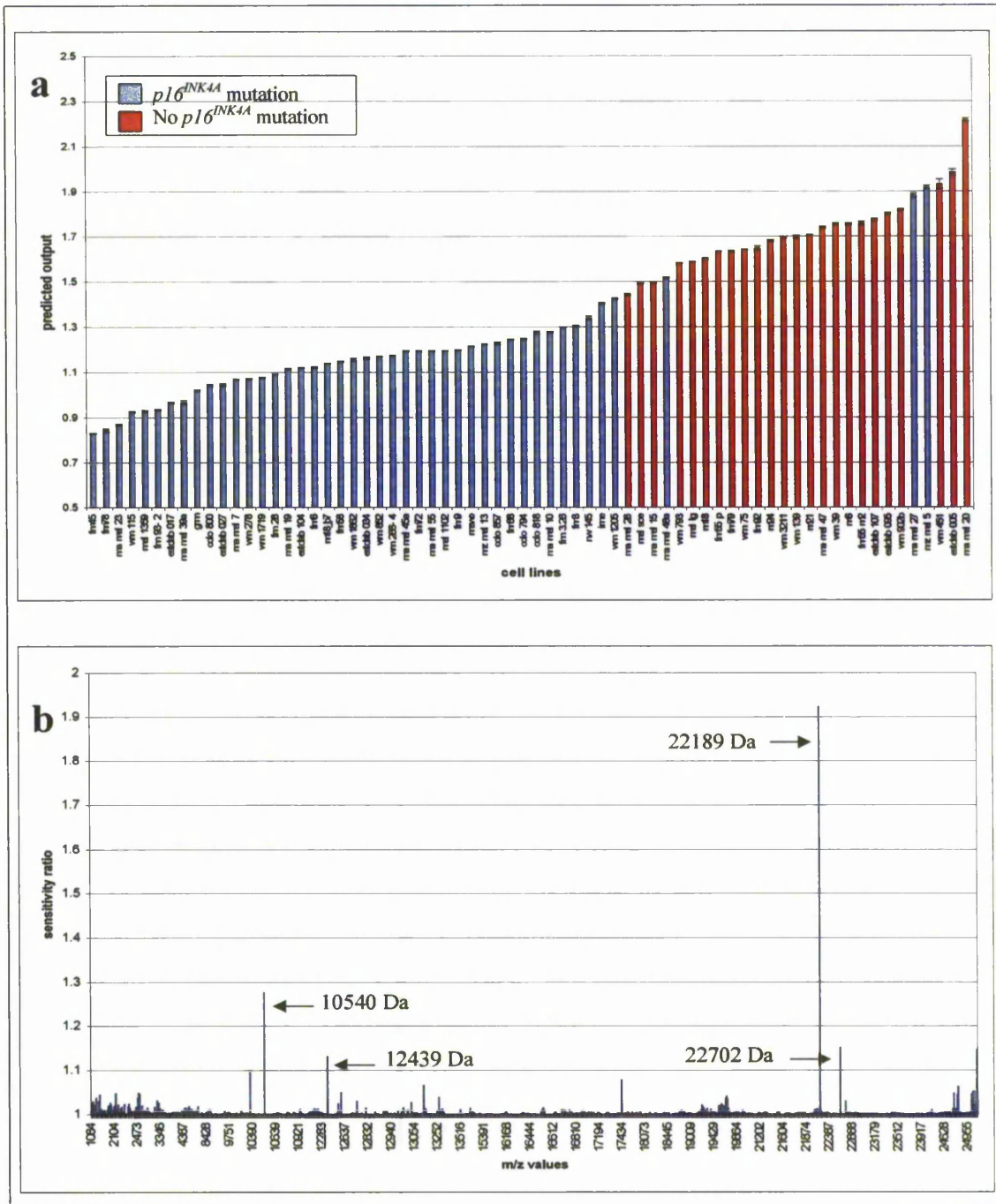


Figure 7-6 Classification by ANN analysis and sensitivity analysis of SELDI data generated from cell lines with or without a $p16^{INK4A}$ mutation.

The above figure shows results of ANN analysis of SELDI MS data generated from cell lines with or without a $p16^{INK4A}$ mutation. Figure a) shows the classification of 68 cell lines as predicted by ANN analysis, averaged over 100 models. The blue bars show the cell lines which have a $p16^{INK4A}$ mutation and the red bars show the cell lines which do not have a $p16^{INK4A}$ mutation. The error bars were determined by calculating the standard error of the mean with 95% confidence intervals. Figure b) shows a graph of the average sensitivity ratio for the top 1000 ions over the 100 models trained.

7.3.2.3 *Tp53* analysis

Seventy-two cell lines of which 19 expressed a *tp53* mutation were used to train the ANN model using the SELDI MS data. After training using the top 1000 ions, the ANNs correctly classified the cell lines with 98.1% accuracy and a sensitivity and specificity of 92.9% and 100%, respectively. The population classification is shown in figure 7-8a, and it shows that cell lines ESTDAB 107, FM 93-2 and WM 451 were misclassified. The sensitivity ratio for the top 1000 ions is shown in figure 7-8b and reveals 2 ions with ratio values much higher than the rest, 10510 and 21529 Da.

Due to poor spotting of three cell line samples onto the MALDI plate, the MALDI data constituted spectra from 69 cell lines, of which 19 contained a *tp53* mutation, to train 100 ANN models. Forty-one of the cell lines were randomly selected for the training set, 14 for the test set and 14 for the blind data set for each model. The predictions for the blind dataset of all 100 models were averaged and the standard error of the mean calculated. The ANNs correctly classified the cell lines in the blind dataset with an accuracy of 92.8%, a sensitivity of 78.9% and a specificity of 98%, this is illustrated in figure 7-9a. The sensitivity ratio was calculated for all of the top 1000 ions for each model, the average was determined and illustrated in figure 7-9b.

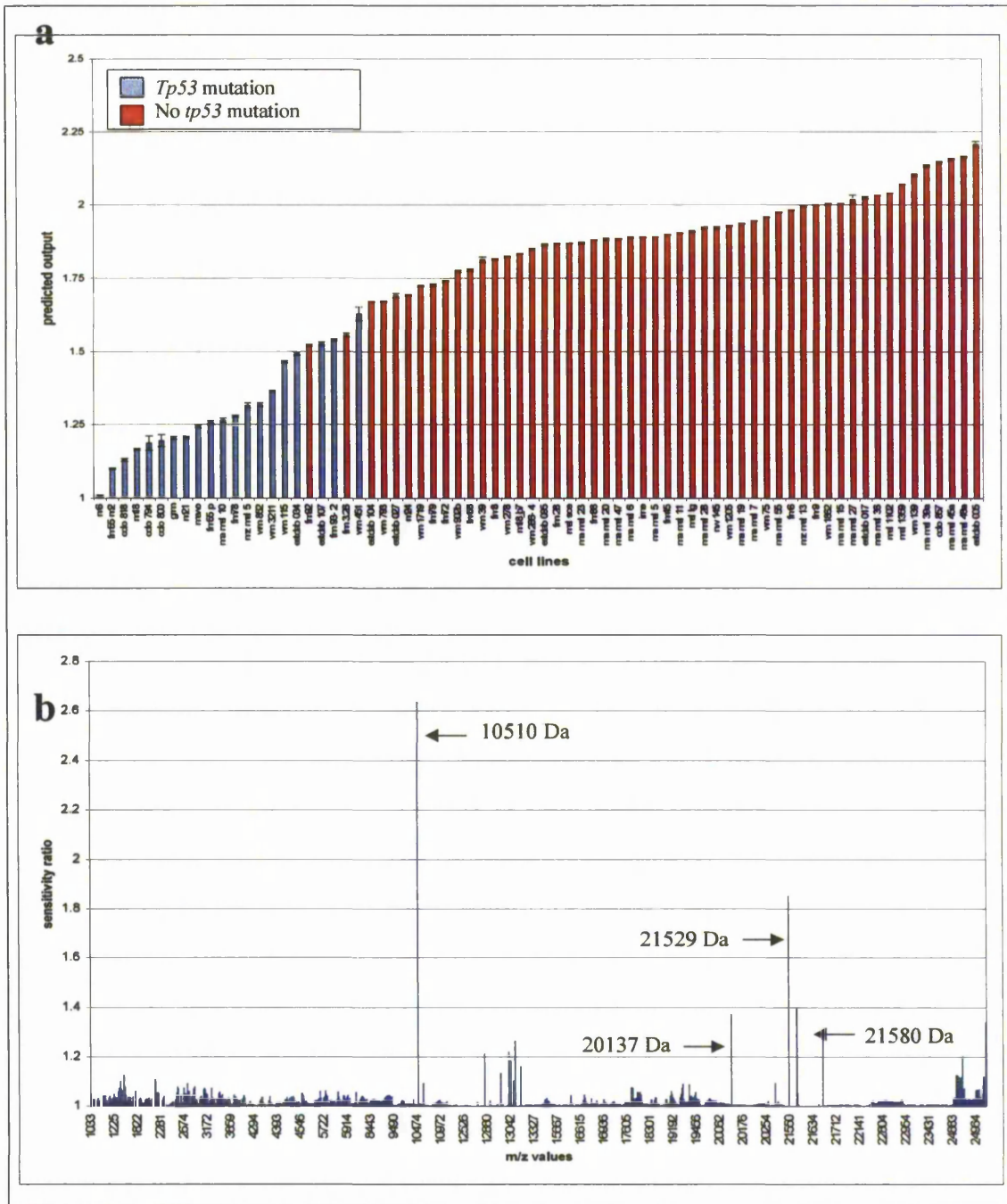


Figure 7-8 Classification by ANN analysis and sensitivity analysis of SELDI data generated from cell lines with or without a *tp53* mutation.

The above figure shows results of ANN analysis of SELDI MS data generated from cell lines with or without a *tp53* mutation. Figure a) shows the classification of 72 cell lines as predicted by ANN analysis, averaged over 100 models. The blue bars show the cell lines which have a *tp53* mutation and the red bars show the cell lines which do not have a *tp53* mutation. The error bars were determined by calculating the standard error of the mean with 95% confidence intervals. Figure b) shows a graph of the average sensitivity ratio for the top 1000 ions over the 100 models trained.

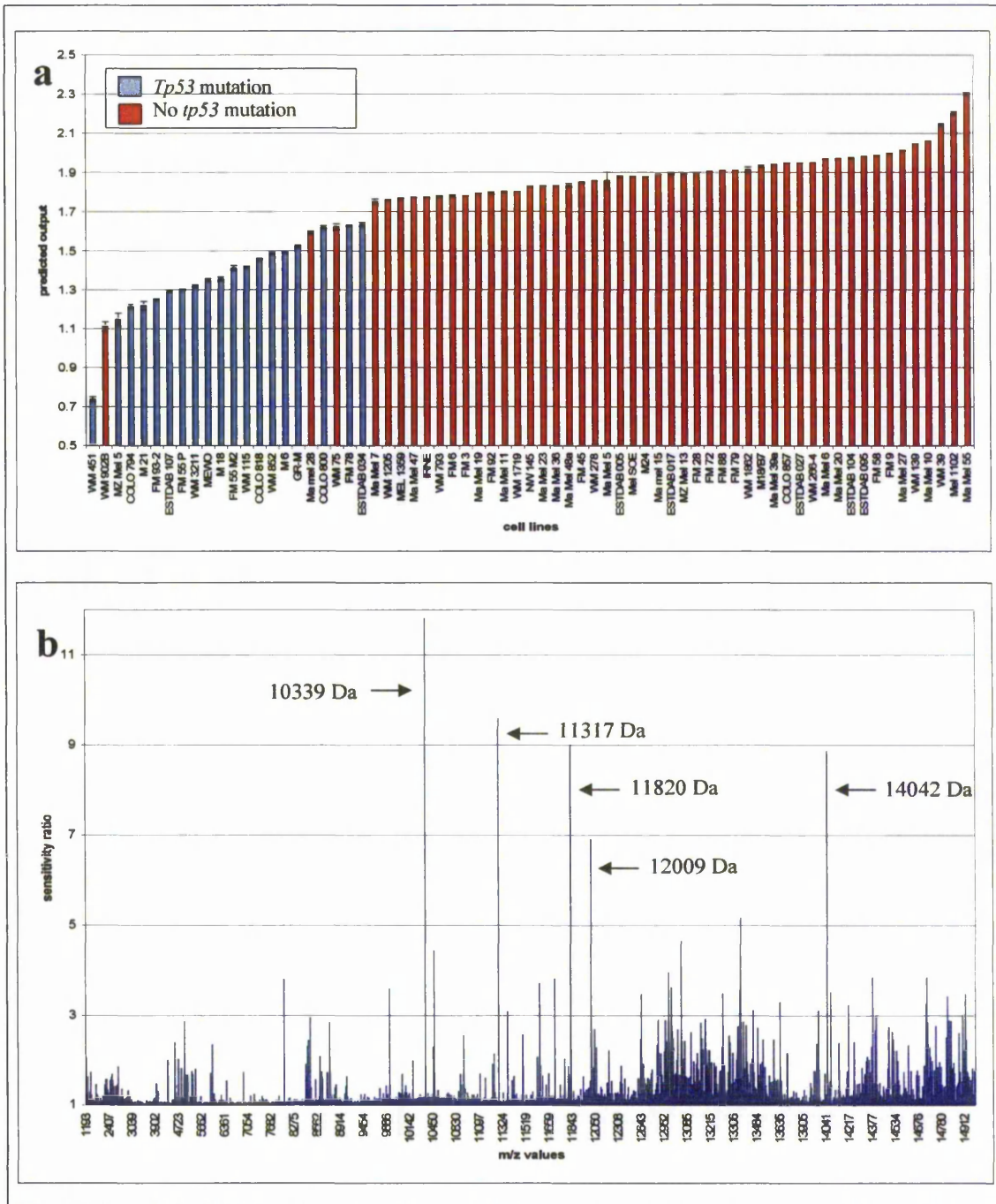


Figure 7-9 Classification by ANN analysis and sensitivity analysis of MALDI data generated from cell lines with or without a *tp53* mutation.

The above figure shows results of ANN analysis of MALDI data generated from cell lines or without a *tp53* mutation. Figure a) shows the classification of 69 cell lines as predicted by ANN analysis, averaged over 100 models. The blue bars show the cell lines which have a *tp53* mutation and the red bars show the cell lines which do not have a *tp53* mutation. The error bars were determined by calculating the standard error of the mean with 95% confidence intervals. Figure b) shows a graph of the average sensitivity ratio for the top 1000 ions over the 100 models trained.

A stepwise analysis was also performed on the SELDI MS and MALDI MS data comparing the *tp53* mutated or non-mutated cell lines in order to compare the two platforms using another bioinformatics approach. Ions were sequentially added to the training set until the error for the test data no longer improved. The stepwise analysis using the cell line spectra generated by SELDI MS showed that no improvement upon the test error occurred after using 21 ions for training. The accuracy of classifying cell lines with or without *tp53* mutations when only using these 21 ions for training was 93.2% and 89.7% for the blind dataset. Table 7-2 shows how the performance and error changed for the training, test and blind datasets after the addition of each ion into the ANN models. Figure 7-10 illustrates how the median accuracy for prediction increases (with upper and lower quartiles shown as error bars) and how the sum of squared errors decreases (with the 95% confidence intervals shown as error bars) as more ions are added to the model.

input num	m/z value	Train Perf.	Test Perf.	Blind Perf.	Train Error	Test Error	Blind Error
1	5536	0.717	0.764	0.682	0.432	0.383	0.453
2	13393	0.792	0.857	0.750	0.386	0.335	0.422
3	9775	0.833	0.866	0.760	0.339	0.312	0.427
4	3208	0.845	0.880	0.777	0.348	0.291	0.402
5	12490	0.855	0.883	0.772	0.331	0.286	0.410
6	3898	0.837	0.876	0.755	0.348	0.286	0.436
7	18046	0.850	0.893	0.747	0.344	0.292	0.449
8	2902	0.872	0.884	0.764	0.319	0.286	0.418
9	9709	0.856	0.890	0.765	0.326	0.286	0.413
10	15124	0.857	0.873	0.793	0.320	0.298	0.403
11	10987	0.884	0.857	0.772	0.308	0.302	0.414
12	17449	0.899	0.876	0.756	0.268	0.284	0.428
13	21940	0.901	0.913	0.768	0.264	0.245	0.420
14	18868	0.914	0.910	0.822	0.248	0.252	0.372
15	9778	0.912	0.926	0.833	0.252	0.228	0.366
16	15088	0.917	0.916	0.851	0.257	0.226	0.340
17	2101	0.934	0.926	0.841	0.221	0.212	0.352
18	22573	0.942	0.933	0.850	0.217	0.210	0.331
19	18103	0.937	0.934	0.867	0.218	0.181	0.303
20	13942	0.936	0.937	0.901	0.214	0.186	0.271
21	15085	0.932	0.947	0.897	0.227	0.174	0.280
22	17378	0.935	0.940	0.881	0.220	0.182	0.276

Table 7-2 Stepwise analysis of SELDI data generated from cell lines with or without *tp53* mutations.

The table represents the data obtained from the stepwise analysis of SELDI MS data generated from cell lines with or without *tp53* mutations. The table shows a summary of the median accuracies and the mean squared error for the training, test and blind data sets as each input is added to the model. The row highlighted in red shows at which point the accuracy and the error failed to improve with subsequent addition of inputs.

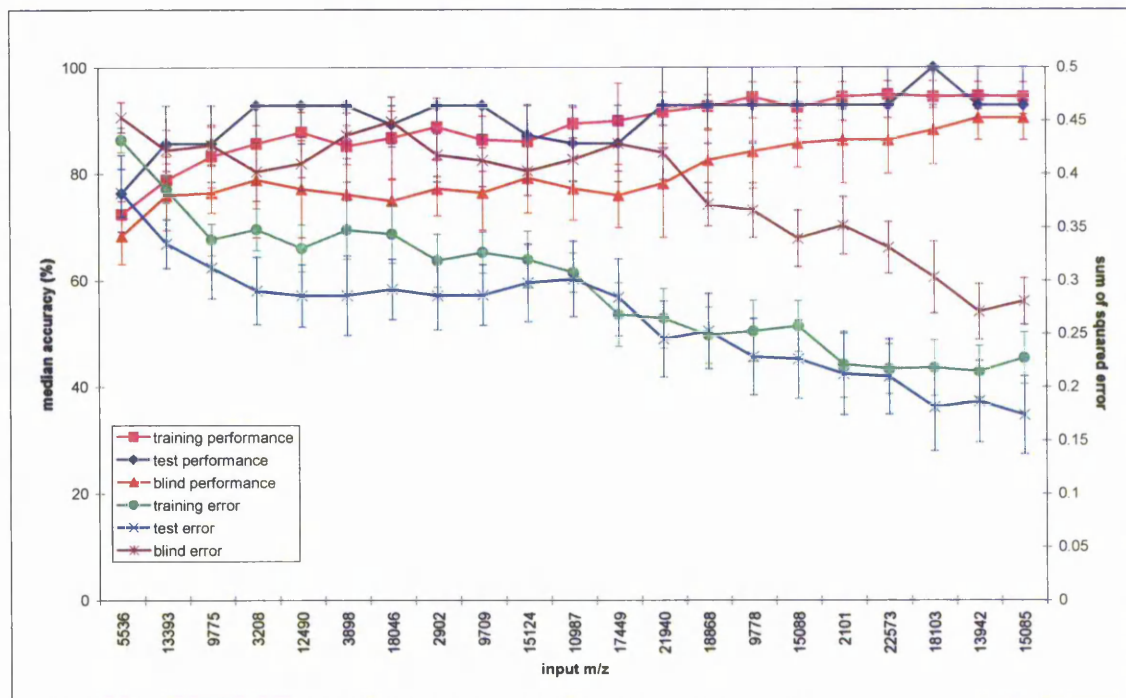


Figure 7-10 Stepwise model performance after addition of each input during the analysis of *tp53* mutation data generated by SELDI MS.

The figure shows the model performance with each input addition for the stepwise analysis of the *tp53* mutation data. The pink line shows the median accuracy for the training data set, with the lower and upper quartile ranges represented as error bars. The dark blue line shows the median accuracy for the test data and the red line shows the median accuracy for the blind data. The green line shows the sum of squared errors for the predictions at each step with the error bars representing the 95% confidence intervals. The light blue line shows the mean squared error for the test and the purple line shows the mean squared error for the blind data.

The stepwise analysis using the cell line spectra generated by MALDI MS showed that no improvement upon the test error occurred after using 23 ions for training. The accuracy of classifying cell lines as those with or without *tp53* mutations only using these 23 ions for training was 99.1% and 97.7% for the blind dataset. Table 7-3 shows how the performance and error changed for the training, test and blind datasets after the addition of each ion into the ANN models. Figure 7-11 illustrates how the median accuracy for prediction increases (with upper and lower quartiles shown as error bars) and how the sum of squared errors decreases (with the 95% confidence intervals shown as error bars) as more ions are added to the model.

input num	m/z value	Train Perf.	Test Perf.	Blind Perf.	Train Error	Test Error	Blind Error
1	12106	0.740	0.784	0.688	0.425	0.386	0.460
2	13285	0.789	0.811	0.728	0.391	0.358	0.432
3	14396	0.836	0.860	0.778	0.360	0.314	0.403
4	13349	0.837	0.889	0.799	0.349	0.277	0.399
5	3129	0.880	0.896	0.833	0.315	0.272	0.360
6	13305	0.903	0.934	0.838	0.292	0.215	0.364
7	12955	0.913	0.930	0.849	0.249	0.219	0.339
8	12393	0.919	0.941	0.870	0.233	0.178	0.294
9	6919	0.943	0.940	0.877	0.193	0.173	0.307
10	3678	0.942	0.951	0.871	0.211	0.168	0.306
11	10648	0.950	0.936	0.853	0.209	0.209	0.322
12	3996	0.942	0.939	0.874	0.198	0.189	0.297
13	11346	0.957	0.934	0.867	0.180	0.179	0.301
14	6253	0.942	0.944	0.837	0.209	0.173	0.358
15	5201	0.956	0.959	0.921	0.179	0.146	0.241
16	13608	0.968	0.969	0.897	0.144	0.115	0.241
17	11190	0.974	0.979	0.948	0.136	0.091	0.181
18	5598	0.976	0.974	0.935	0.127	0.101	0.190
19	11030	0.993	0.981	0.931	0.048	0.088	0.182
20	10759	0.984	0.989	0.969	0.065	0.059	0.116
21	12313	0.982	0.989	0.965	0.077	0.058	0.131
22	9288	0.989	0.990	0.967	0.052	0.048	0.130
23	14383	0.991	0.991	0.977	0.049	0.053	0.100
24	6966	0.988	0.987	0.969	0.053	0.051	0.105
25	13566	0.865	0.881	0.781	0.331	0.289	0.411

Table 7-3 Stepwise analysis of MALDI data generated from cell lines with or without *tp53* mutations. The table represents the data obtained from the stepwise analysis of MALDI data generated from cell lines with or without *tp53* mutations. The table shows a summary of the median accuracies and the mean squared error for the training, test and blind data sets as each input is added to the model. The row highlighted in red shows at which point the accuracy and the error failed to improve with subsequent addition of inputs.

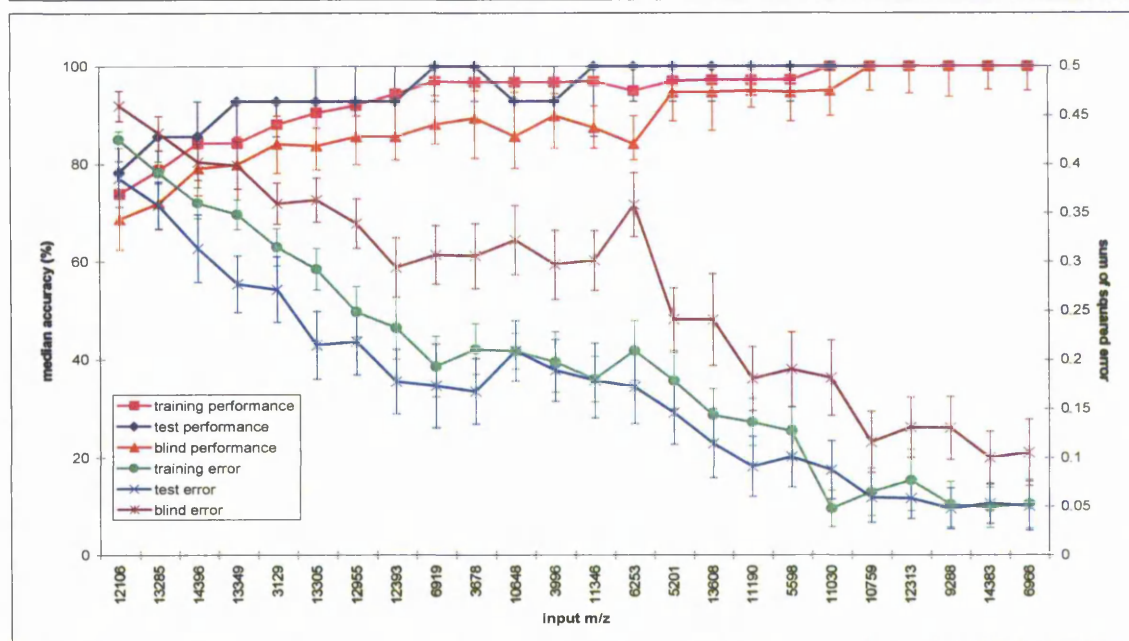


Figure 7-11 Stepwise model performance after addition of each input during the analysis of *tp53* mutation data generated using MALDI MS.

The figure shows the model performance with each input addition for the stepwise analysis of the *tp53* mutation data. The pink line shows the median accuracy for the training data set, with the lower and upper quartile ranges represented as error bars. The dark blue line shows the median accuracy for the test data and the red line shows the median accuracy for the blind data. The green line shows the mean squared error for the predictions at each step with the error bars representing the 95% confidence intervals. The light blue line shows the mean squared error for the test and the purple line shows the mean squared error for the blind data.

7.4 Discussion

The use of SELDI MS and ANN to distinguish between patients with a specific disease state compared with control patients has been proven in many different diseases in the literature. The main criticism that has come from using this approach is the lack of validation studies that look at the reproducibility and accuracies of the mass and intensity of peaks as well as sensitivity of the SELDI MS and MALDI MS instrument. This study begins to address these issues by comparing the SELDI and MALDI approach, looking at the mass accuracies and intensities of peaks and predictive capability of the ANNs to determine regions of the profile that change when the same samples are processed using both instruments. Petricoin *et al* (2002b) have shown that when analysing ovarian cancer serum using H4 SELDI chips, the coefficients of variation (CV's) for eight selected peaks within a run was 10% and <10% between runs, showing that there is little variation day to

day with the chips and instrument. De Noo *et al* (2005) have shown that by using C8 magnetic beads for clean up of 16 normal serum samples prior to analysis by MALDI-TOF the CV's were between 20-30% for both the peptides and proteins. The CV's for both the SELDI MS and MALDI MS techniques used in this study are shown in chapter 2, where aliquots of a normal serum sample were spotted on a H4 SELDI chip and analysed, and also subjected to C18 ZipTip clean up before MALDI MS analysis. The CV for peak location of the five most prominent peaks in the SELDI MS method was 0.09 – 0.13% and 0.07 – 0.09% for the MALDI method. The CV for peak intensity for SELDI MS was between 14.5 – 20.6% and between 25.7 – 47.7% for the MALDI method. One reason why the CV's between the SELDI and MALDI shown here and those of de Noo/Petricoin's study are different is due to the sample preparation methods used. In the SELDI MS method used here and in Petricoin's paper (Petricoin *et al.* 2002b), the samples were processed by binding to the protein chip and unbound samples were washed off. The MALDI method used here utilised ZipTips for sample cleanup, where the samples were bound to a C18 ZipTip, unbound proteins were washed through and the proteins of interest were then eluted and spotted onto the MALDI target, loss of sample could have occurred due to the sample not binding to or eluting from the ZipTip efficiently. In de Noo's paper (De Noo *et al.* 2005), the samples were subjected to C8 magnetic bead prefractionation which also could lead to reproducibility problems as samples are bound to beads, washed three times, and then eluted by incubating the beads and sample for 1 minute in 50% acetonitrile.

The aim of this study was to compare ZipTip sample clean up and MALDI analysis with a commonly used analytical platform, SELDI MS, in order to determine if the limitations of the SELDI MS technique can be overcome and therefore improve upon biomarker studies already being performed. Tables 7-4, 7-5 and 7-6 show the accuracy of the ANNs for classifying cell lines with or without specific genetic mutations by using data collected from the same sets of samples but analysed using different techniques. As the same sample set was also used in chapter 3 using a C₆ – C₁₂ hydrophobic H50 chip, this data has also been used for comparison. The data shows that there is no difference between which of the platforms is used to generate the data in order for the ANNs to classify the samples, although each platform produces slightly different spectra for the same sample. Also, only Ma Mel 39a was consistently misclassified in the comparison of cell lines with or without

braf mutations across all three techniques. There were no other similarities seen in the other misclassified cell lines. Due to the low resolution and sensitivity of the SELDI spectra compared to the high resolution of the MALDI spectra it is difficult to determine if some of the protein peaks seen by MALDI are similar to those seen by SELDI MS. Different spectra are seen using the different platforms because each sample preparation technique binds slightly different proteins due to the differing hydrophobic capacity of the sample clean-up procedures used. This could mean that a large number of proteins are affected as a result of a specific mutation and although not all of the same protein peaks are seen in each type of analysis there are sufficient numbers seen within the spectra for the ANNs to make accurate classifications. This is also reflected by the fact that the most important ions identified as a result of sensitivity analysis or stepwise analysis are different within each genetic mutation analysis. Another explanation for the different spectra could be that these different peaks are either fragments of the same protein, members of the same protein family or members of the same signalling pathway that only bind to one of the sample surfaces used in these studies. Also, because they are related and so are similarly affected by the genetic mutations they will allow the ANNs to correctly classify the samples with similar accuracies. Unfortunately, without identification of these protein peaks, this is only speculation.

	SELDI H50	SELDI H4	MALDI
Misclassified cell lines	Ma Mel 39a FM-3 FM 93/2	Ma Mel 39a FM 3.26 FM 79 WM 451	Ma Mel 39a FM 79 ESTDAB 005
Accuracy of classification (%)	96.5	97.7	97.7
Sensitivity (%)	98.3	98	98
Specificity (%)	92.6	85.7	89.5

Table 7-4 ANN data generated by the analysis of SELDI and MALDI spectra of cell lines with or without *braf* mutations.

This table compares ANN data generated by SELDI analysis, using H50 and H4 SELDI chips, and MALDI analysis of cell lines with or without *braf* mutations. The accuracy of classification and the sensitivities and specificities are similar across all the three sampling techniques used. Only 1 cell line was misclassified in all three analyses, Ma Mel 39a.

	SELDI H50	SELDI H4	MALDI
Misclassified cell lines	FM 92 WM 39	Ma Mel 28 Mel SOE Ma Mel 15 Ma Mel 48a Ma Mel 27 MZ Mel 5	WM 75 M21 ESTDAB 005 WM 793 Ma Mel 55
Accuracy of classification (%)	97.2	91.2	90.5
Sensitivity (%)	100	93.2	96
Specificity (%)	92.3	87.5	74

Table 7-5 ANN data generated by the analysis of SELDI and MALDI spectra of cell lines with or without *p16^{INK4a}* mutations.

This table compares ANN data generated by SELDI analysis, using H50 and H4 SELDI chips, and MALDI analysis of cell lines with or without *p16^{INK4a}* mutations. The data shows that ANN analysis of SELDI spectra generated using H50 chips classify the cell lines with greater accuracy than the other two methods. The misclassification of cell lines is different across the three techniques.

	SELDI H50	SELDI H4	MALDI
Misclassified cell lines	FM 92	ESTDAB 107 FM 93/2 WM 451	WM 902B GR-M COLO 800 FM 78 ESTDAB 034
Accuracy of classification (%)	98.8	98.1	92.8
Sensitivity (%)	100	92.9	78.9
Specificity (%)	98.4	100	98

Table 7-6 ANN data generated by the analysis of SELDI and MALDI spectra of cell lines with or without *tp53* mutations.

This table compares ANN data generated by SELDI analysis, using H50 and H4 SELDI chips, and MALDI analysis of cell lines with or without *tp53* mutations. The data shows that ANN analysis of SELDI spectra generated using H50 and H4 chips classify the cell lines with greater accuracy than data generated using MALDI MS. The misclassification of cell lines is different across the three techniques.

ZipTips are a reverse-phase capturing technique that exist as a manual, microtip format and are used in this study as an alternative to the SELDI chips. However, this technique has proved to be largely impractical due to viscosity problems and poor reproducibility of packing the tips. Other sample preparation methods could have been used in this study for comparison such as the C8-coated magnetic beads, as discussed earlier, which have shown promise for reducing the complexity of samples prior to MALDI MS analysis. The

advantage of using beads is that they have large accessible surface areas on a per. weight basis, allow elution in small volumes without centrifugation or a microcolumn style operation and have minimal losses during washes and sample transfers which are favorable attributes for automation (Villanueva *et al.* 2004).

The main reason that SELDI MS has and still is proving popular for biomarker discovery studies is that the on-chip sample clean up is quick and high throughput, but most MALDI instruments have much better resolution and sensitivity than SELDI MS making the identification of biomarkers more accurate. Qiagen have recently released mass spectrometry focus desalting chips that are compatible with the Shimadzu/Kratos MALDI. They bind hydrophobic proteins onto their surface allowing hydrophilic proteins and salts to be washed away, thereby providing a sample clean up method which can also be used to reduce the sample complexity. The use of these chips could then combine the easy and high-throughput sample preparation methods with a much improved MALDI MS instrument.

Chapter 8 Conclusions and Further Study

Proteomics has proven to be a useful tool in a number of medical areas including infectious diseases, cardiovascular diseases and neurological diseases. As cancer is a disease caused by mutational defects in the DNA and therefore causes changes in protein expression, cancer proteomics could lead to the discovery of biomarkers of potential significance for diagnosis, to define a patient's prognosis and determine a suitable treatment regimen, (Ludwig & Weinstein 2005).

The research presented in this thesis has introduced the use of MS and bioinformatics in the analysis of the cancer proteome, in particular melanoma and breast cancer. Cell lines, tissue and serum samples were used in these studies demonstrating the type of proteomic information that can be obtained from different types of samples and an ANN approach was used for the analysis of these complex datasets.

8.1 The potential use of cancer cell lines in proteomic analysis

Proteomic approaches in biomedical research are influenced by ethics, the availability of adequate samples and privacy concerns. The use of biological specimens is expected to increase knowledge of human diseases, however there is concern that the use of genetic or proteomics information found in these specimens, as well as the use of related clinical data, may infringe upon the privacy of the patient if misused. Since there is no amplification technique available in proteomics, and that some proteomic based approaches require large amounts of sample, the demand for human tissues and/or body fluids is expected to increase (Nestler *et al.* 2004). Cell lines may prove to be a suitable alternative for genetic and proteomic studies to minimise the usage of precious patient material and the impact that would have on patient privacy. Thus, large amounts of material can be acquired from cultured cell lines for the optimisation of methods and potentially offer a continuous and extensive source of material for biological studies. A number of proteomic studies have been published where the authors have used cell lines to identify proteins of possible clinical significance. For example, a study which profiled oral keratinocytes and oral squamous cell carcinoma cell lines by 2-D differential in-gel electrophoresis (2-D DIGE) revealed 22 proteins that were differentially expressed between the two sample types that

could have relevance as diagnostic/prognostic markers (Koike *et al.* 2005). Another study compared the protein profiles of human non-small cell lung cancer cell lines, a normal bronchial epithelial cell line and primary human bronchial epithelial cells and revealed the differential expression of 12 proteins between the cancer and the normal cells, already known to be upregulated in lung cancer (Liu *et al.* 2004). A study by Mian *et al.* (2003) identified proteomic patterns that could be used to correctly distinguish between a drug treated and a control breast cancer cell line and between chemosensitive and chemoresistant cells (Mian *et al.* 2003). A study on the effect of etoposide on neuroblastoma cell lines revealed the overexpression of a number of proteins that could prove useful in predicting the prognosis of patients with neuroblastoma after chemotherapy (Urbani *et al.* 2005). Furthermore, it is generally accepted that patients receiving chemotherapy may respond at first but subsequently develop resistance to the treatment, however it is not clear which pathways are involved in developing chemoresistance. Studies using cell lines may uncover the biological pathways involved, providing further insight into the mechanisms involved and allowing for the development of more successful treatments. One advantage of using cancer cell lines for biomarker discovery is that they are less heterogenous (free of stromal components) and so analysis of these samples could reveal biomarkers that would be tumour cell-specific. It is well known that tumour cells diversify in culture and so they may not be an exact representation of the original population, however, if they retain the essential genetic mutations then this is likely to be reflected in all of the cells as a 'protein fingerprint'.

The cell line studies presented in this thesis have revealed proteomic profiles that can be used to identify the presence of a certain genetic mutation. To the best of my knowledge this is the first study to combine genetic and proteomic information, demonstrating a high degree of association as it was possible to classify cell lines with or without a mutated *Braf* gene, *p16^{INK4A}* gene or *tp53* gene with over 90% accuracy. *Braf*, *p16^{INK4A}* and *tp53* are all important genes involved in the initiation and progression of melanoma and so a study such as this could provide a further insight into the biological mechanisms of the cancer as the presence of a number of protein peaks, not just one specific protein, allowed the discrimination of the cell lines. The proteome is likely to be the result of downstream protein changes/events and so the identities of the discriminatory protein peaks could possibly be indicative of pathways affected by the mutations.

8.2 Analysis of the Cancer Tissue Proteome

Cell lines do not provide an exact model of a tumour, as the tumour microenvironment invokes changes upon the tumour genome and proteome and the molecules it expresses and secretes. Tissue based studies are therefore still of prime importance. The studies presented here show that proteins extracted from tumour tissue provide information on the clinical nature of the disease, for example proteomic patterns indicating the site of the primary tumour and survival of the patient could be demonstrated within melanoma tissue. In breast cancer tissue, patterns indicating the basal phenotype and the expression of p53 were revealed. These studies provide a proof-of-principle for using SELDI MS to extract clinically relevant information from tumour tissue. Future studies would necessitate the use of larger cohorts of samples and multiple centres in order to elucidate additionally clinically relevant proteomic patterns, for example, an important finding would be to identify the proteome predicting patient response to therapy.

The method of tissue analysis used in this study involved the extraction and processing of proteins which did not allow for the spatial localisation of proteins within the tissue to be determined. The localization of unique proteins identified by MALDI mass spectrometry could provide additional information on the function of the protein as well as a better insight into the biology of the disease. Immunohistochemistry may be used to determine the spatial localization of proteins but requires prior knowledge of the protein as well as the production of labeled antibodies specifically reactive with the protein *in situ*. Tissue profiling by MALDI mass spectrometry can be performed on intact tissue sections that are either transferred directly onto a cold MALDI target plate after cryostat sectioning or transferred after laser capture microdissection (LCM). In the latter method, tissue sections are laid onto a heat sensitive thermoplastic film allowing areas of interest to be cut from the section. The area (bound to the film) can then be transferred to the MALDI target using double sided conductive tape (Caldwell *et al.* 2005). Direct tissue profiling by mass spectrometry was first introduced in 1999 where the analysis of murine tissue sections revealed unique signals from different parts of the same organ (Chaurand *et al.* 1999b). Since then there have been many direct tissue profiling studies; a study by Zheng *et al.* (2003) used LCM to isolate cancerous and non-cancerous human prostate cancer cells and

revealed the over-expression of a protein in the cancerous cells but not in non-cancer cells that could be used as a prostate cancer marker (Zheng *et al.* 2003). A new method involves imaging mass spectrometry (IMS) whereby molecular weight specific maps or images of tissue sections can be obtained rapidly at high resolution and sensitivity. This method is still in the early stages of development and improvements in sample preparation, instrumentation and data analysis are still ongoing (Caldwell & Caprioli 2005). So far this method has been applied to the profiling of glioblastoma, prostate and colon cancers, resulting in the discovery of numerous disease specific biomarkers and their spatial localisation within the tumours being determined (Stoeckli *et al.* 2001).

8.3 The Cancer Serum Proteome

Tissue material is limited and can be difficult to obtain from a patient without the use of invasive surgery. Alternatively, serum is a much easier sample type to obtain and can be acquired in fairly large quantities and as a result, serum-based proteomic studies are widespread. One main disadvantage with serum analyses is the amount of albumin and IgG present, which can affect MS studies; ion suppression can occur due to the presence of these high abundant molecules and so low abundant proteins associated with cancer may be missed. The removal of high-abundant proteins in serum sample preparation strategies is becoming increasingly widespread, providing higher sensitivity for achieving broader proteome coverage, particularly of low-abundance proteins that are normally present in the concentration range of ng/mL and lower (Anderson & Anderson 2002). The removal of these high abundant proteins, however, also poses a problem since important (low abundant) proteins may also be removed due to the fact that these high abundant proteins act as carrier proteins. Antibody-based immunoaffinity subtraction is highly efficient and is used for the removal of high abundant proteins (Liu *et al.* 2006).

There have been several reports where MS analysis of serum has identified cancer from non-cancer but have been unable to classify tumours further or predict the tumour behaviour (Yanagisawa *et al.* 2003). The present study, using MS and ANN analysis resulted in the correct classification of melanoma stage and correctly classified a high proportion of patients that would progress from stage III to stage IV melanoma, revealing important prognostic information. This study cohort was small and studies involving larger

numbers of patients from different centres would determine whether MALDI/SELDI methodologies and ANN analyses can be used routinely in clinical practice.

8.4 Bioinformatic studies

The evaluation of multidimensional datasets, such as mass spectra, has led to the development of several new statistical methods over the past decade (Lisboa *et al.* 2006). These include clustering, where clustering algorithms are used to find similarities between features in the dataset; support vector machines, where the classification is based upon decision planes and has proven very popular for the analysis of high dimensional problems; principal components analysis, where the aim is to derive a small number of features that can be used to explain the whole dataset; and ANNs that can be used to learn and to adapt to the dataset and apply the model to unseen data, allowing for generalization across a population (Lisboa & Taktak 2006). There are two ways in which mass spectra are used for analysis using computer algorithms in order to identify m/z values of importance; either a list of peaks found across all spectra are taken or the entire spectra is used as inputs, in the present study the whole spectra is used. The process used for the identification of key ions as potential biomarkers must be sensitive whilst limiting false-discovery and should guarantee that those key ions are biologically meaningful allowing for further analysis and validation (Carlson *et al.* 2005). An element of data pre-processing is also incorporated into the bioinformatics analysis in order to improve peak detection and can include baseline subtraction, scaling and normalisation. The quality of the data pre-processing will, therefore, directly impact the downstream learning algorithm and so improvement in the peak detection will improve the power of biomarker discovery (Carlson *et al.* 2005).

The application of ANNs to identify patterns correlating with clinical parameters allows us to gain further understanding of the biological diversity of different cancers. The two cancers studied here, melanoma and breast cancer, are recognised as heterogeneous diseases whose biology is still not fully understood. The emergence of subtypes of disease also makes it difficult to determine the prognosis of patients and decide upon the most appropriate treatment. Classifying tumours into distinct groups according to their protein profiles may provide additional information of diagnostic and prognostic benefit to the

patient. ANN analysis also identified sample outliers that could represent new subtypes of the cancers. The stepwise analytical approaches used for the studies presented here has resulted in the identification of multiple markers associated with, for example, the stage of the disease. This highlights the fact that using a simplistic, single marker approach to describe very complex and heterogenous diseases is unrealistic and that the use of multi-marker models will allow much more accurate conclusions to be drawn. Also, it has been revealed that predictive patterns can be identified to classify a high proportion of samples using different bioinformatics approaches. However, these approaches identify different groups of ions that could be used for prediction and so biomarker identification using these types of approaches is dependent upon which method is used for the analysis of the data. This present study has shown that ANNs can be used to determine predictive markers in melanoma and breast cancer, but it has also been used in studies involving prostate cancer. These studies have compared data analysis using ANNs with conventional multivariate logistic regression (LR) analysis and univariate analysis and showed that ANNs gave higher accuracies for predicting prostate cancer than LR and that both ANNs and LR gave higher accuracies than univariate analysis (Zhu *et al.* 2006). Three studies showed specificities of 33%, 67% and 68% at a sensitivity of 95% when using ANNs compared with 24%, 60% and 54% for LR and 19%, 40% and 33.5% for univariate analysis (Djavan *et al.* 2002; Finne *et al.* 2000; Remzi *et al.* 2003).

There have been a large number of papers published involving the use of ANNs in cancer studies but as yet few of these results have led to the improvement of healthcare for patients (Lisboa & Taktak 2006). There are a number of reasons for this including overoptimistic assessment of predictive performance, poor selection of the model when many variables are involved and poor evaluation, that is quoting the best performing model and so introducing bias (Schwarzer *et al.* 2000). Addressing these issues before embarking on future ANN-based studies would enable more confidence in the data generated and may then lead to their routine use in the clinic.

8.5 Identification of cancer biomarkers using proteomic profiling: A question of reproducibility and validity

Recently, there have been a number of published studies where specific cancer biomarkers revealed by SELDI MS analysis have been identified; these cover a wide range of cancer types and are summarised in table 8-1. Many of the biomarkers currently identified via proteome profiling experiments are inflammatory molecules and are not specific to a particular disease type. For example, Serum Amyloid A-1 has been identified as a biomarker in prostate cancer (Le *et al.* 2005), nasopharyngeal cancer (Cho *et al.* 2004), ovarian cancer (Moshkovskii *et al.* 2005), monitoring pneumonia (Yip *et al.* 2005) and lung cancer (Howard *et al.* 2003). It has been suggested by Diamandis (2004a) that the discriminatory peaks identified in serum are most likely representative of high abundance molecules. It is proposed that these discriminatory peaks may in fact represent acute phase reactants, as shown in table 8-1, or other proteins or protein fragments that are released by organs such as the liver in response to the presence of the tumour or cancer epiphenomena such as infection, inflammation or malnutrition (Diamandis 2002; Diamandis 2003b). It is also possible that some of these discriminatory peaks represent artifacts of sample collection, storage, pre-treatment or patient selection. It is possible, however, that high abundance proteins such as those shown in table 8-1 may still have a place as clinical biomarkers of disease. Cancer cells may produce proteases and other enzymes that are tumour specific, causing fragmentation and cleavage of common, abundant proteins, therefore the presence of these proteins could still be an amplified signal of a disease process that may have otherwise been undetected at an early stage. The concern is that the presence of a highly abundant protein could not be classed as a biomarker for a specific cancer, for example with Serum Amyloid A, but instead become part of a panel of multiple markers, each with their own clinical relevance to the disease, that could be used for diagnosis or prognosis (Robbins *et al.* 2005; Villanueva *et al.* 2006; Zhang *et al.* 2004).

Biomarker	Sample type	Biology of marker	Reference
Ovarian cancer			
Haptoglobin- α fragment	Serum	Acute phase protein, haemoglobin binding	Ye <i>et al.</i> 2003
Apolipoprotein A-I	Serum	Acute phase protein, lipid metabolism	Zhang <i>et al.</i> 2004b
N-terminal truncated transthyretin	Serum	Thyroid hormone binding protein	
ITIH4 fragment	Serum	Involved in acute phase reactions?	
Breast Cancer			
Haemoglobin- β chain	Nipple aspirate fluid	Oxygen transport	Sauter <i>et al.</i> 2005
High molecular weight kininogen	Serum	Blood coagulation, release of bradykinin	Heike <i>et al.</i> 2005
ApolipoA-II isoform	Serum	Lipid metabolism	
Prostate Cancer			
ApolipoA-II isoform	Serum	Lipid metabolism	Malik <i>et al.</i> 2005
Serum amyloid A-1	Serum	Acute phase protein	Le <i>et al.</i> 2005
Colorectal Cancer			
Prothymosin- α	Cell lines	Immune function mediator	Shiwa <i>et al.</i> 2003
α -Defensin	Tissue, Serum	Antibiotic, fungicidal and antiviral activities, concentration-dependent mediation of tumour lysis	Albrethsen <i>et al.</i> 2005; Melle <i>et al.</i> 2005
Nasopharyngeal Cancer			
Serum amyloid A-1	Serum	Acute phase protein	Cho <i>et al.</i> 2004
Head and Neck Cancer			
Annexin V	Tissue	Anticoagulant protein, tumour proliferation and metastases?	Melle <i>et al.</i> 2003

Table 8-1 Some of the identified biomarker proteins discovered using SELDI MS, adapted from Engwegen *et al.* 2006

The majority of biomarkers associating with diseases other than cancer have been identified in body fluids, using a number of different techniques. One interesting study by Novikova *et al.* (2006) describes how SELDI MS analysis of dorsolateral prefrontal cortex samples revealed biomarker peaks that were used for diagnosis and differentiation of schizophrenia, bipolar disorder and normal controls. The biomarker peaks from 24 proteins were identified after bands of the correct molecular weights were cut from 1-D gels, trypsinised and analysed by MALDI-TOF-PSD MS analysis. Thus, the combination of techniques was

able to identify the proteins of potential importance from the peak values obtained by MS (Novikova *et al.* 2006).

Using proteomics for the identification of new biomarkers has proven popular over recent years but standardisation of the sample handling, processing and data analysis has become a major concern and questions have been raised regarding reproducibility and reliability of data that has already been published (Check 2004; Diamandis 2004b; Garber 2004). Sample handling and processing, instrument noise and data analysis all contribute to the challenges of reproducibility in any proteomics experiment (White *et al.* 2004). Variability in sample handling and the mass spectrometer causes a baseline across the spectra as well as 'noise' and variability in the amount of protein bound onto the target *also* causes a fluctuation in the intensity scale of the spectra. Variability in the spectra has the largest effect on small peaks as the baseline and instrument noise can be as large as the peak itself, yet it is possible that these small peaks contain much of the biological information; hence the standardisation of sample and data processing procedures should be of utmost importance when embarking upon biomarker discovery studies (Rodland 2004).

The question of reproducibility can be addressed in a single study setting, where a separate set of samples can be analysed under exactly the same conditions in order to establish that the findings are true and not the result of overfitting or chance (Ransohoff 2005b). In the present study the question of chance has been addressed by the splitting of the data into training and blind datasets prior to ANN analysis, showing that the models generated using the training data can then be applied to blind or unseen data. Bias is another problem that can occur when the samples and/or data being compared are handled differently causing the introduction of an extra signal into only one of the groups. This bias can come from differences in the collection of the two sets of samples, including the types of tubes used and time to storage; the number of freeze/thaw cycles; and the analysis on the MS, for example if one set of samples is run on a different day to the other set. Therefore the methods used, from selecting patients, obtaining samples, storage and processing, must be rigorously standardised in order to avoid bias in the findings (Ransohoff 2005a). In this present study all the samples were collected in the same way, the cell culture and protein extraction methods were optimised and standardised before the studies were undertaken. The samples were randomised prior to SELDI/MALDI MS analysis to avoid bias caused by

sample processing and all of the mass spectral data was handled in the same way before analysis by ANNs.

Once reproducibility and reliability of the data has been addressed and biomarkers have been identified it is important that appropriate validation of the results are obtained (Anderson & Anderson 2002), for example, the analysis should be repeated using additional 'blind' samples from other sets of patients obtained at different centres, to determine that the biomarker is present across the general population. Determining the timing of the appearance of a biomarker has been shown to be important in assessing at which point a biomarker has prognostic value, and so this would have to be considered when validating new biomarkers. During the validation process, the specificity of the biomarker should also be determined. In the case of cancer tissues, this is further complicated by the heterogeneity of the tumour and its microenvironment, and for serum by the presence of other 'normal' proteins. The use of tumor tissues or needle biopsies is problematic due to the fact that multiple tissue sampling is not always an option from tumors *in situ* and protein profiling can vary between different areas of the tumor mass. Using serum or urine is an alternative, where re-analysis of different samples is possible, although reports to date have mainly used very small numbers of samples with limited validation (Alaoui-Jamali *et al.* 2006).

8.6 Cancer diagnosis and treatment – personalised medicine

The clinical and pathological biomarkers that are currently used poorly predict early disease development and response to treatment. The aim of this study and others is to attempt to identify biomarkers that can improve upon markers currently in use.

Only a small number of the biomarkers generated by SELDI MS have been identified and these studies have mainly concentrated on biomarkers present in serum and appear to be isoforms of ubiquitous proteins that occur as a result of secondary tumour effects, for example altered enzymatic activity, and so the specificity of these isoforms requires further investigation (Engwegen *et al.* 2006). Like normal cells, most cancer cells use multiple intracellular signaling pathways to ensure the maintenance of functions that are critical to their survival. Thus, cellular pathways that are integral to cell function, survival, proliferation, and receptor expression are potential targets for therapeutic intervention but

are not necessarily tumour cell specific, one example being the epidermal growth factor receptor signaling pathway. Molecules that mediate the production of angiogenic and invasion factors that allows tumor growth and metastasis, such as the vascular endothelial growth factor and downstream events that result in cellular apoptosis, represent additional potential targeting pathways (Ajani *et al.* 2005).

Standard diagnostic methods, including tissue histopathology are being replaced or complimented by the use of molecular diagnosis, which can identify proteins and their posttranslational modifications that occur in disease conditions, and hence greatly accelerate progress toward novel diagnostic and predictive tools to track early disease and tailor treatments to specific patients (personalised medicine) (Alaoui-Jamali & Xu 2006). In the clinic, new patients could be tested (either using serum or by isolating protein from a biopsy) by SELDI MS or MALDI MS analysis to generate protein or peptide fingerprints. The profiles of patients could then be compared to control or 'baseline' profiles from normal subjects in order to determine the diagnosis, or prognosis or to predict response to a particular treatment. Continued investigations may result in the development of proteomic profiling databases through which a patient could be matched with protein profiles relevant to the disease and potential benefits of treatment. In this way, clinicians would be able to recommend combinations of molecularly targeted agents and therapies on the basis of an individual patient's proteomic profile (Ajani & Allgood 2005). Samples compared to profiles in an existing database would have to be applicable to the general population, taking into consideration factors such as age, sex, ethnicity and nutritional status. The proteins identified to date in SELDI MS based proteomic studies have revealed that patient populations can be determined by the up or down regulation of inflammatory proteins and so the diagnosis of a patient could be determined by the presence of molecules that are not tumour cell specific.

The SELDI MS approach utilised in the present research has the potential to be used as a clinical tool in the future for the diagnosis and prognosis of patients. However, further developmental work is required before this can become a reality, including the type of samples that could be used analysed, the bioinformatics approach used to analyse the data as well as the method used to generate the spectra. The work presented has shown that

different types of samples can be used to derive clinically relevant information from a proteomic profile. Different types of bioinformatics approaches can be used to analyse the data and to classify samples with the same accuracy; this is likely to be achieved using a panel of ions/biomarkers. It has also been shown that different sample preparation methods can be utilised, each generating a different set of prominent ions; despite this, however, the patterns within these spectra allow the classification of samples with a similar degree of accuracy. Further work should determine the best method to use for sample preparation, proteome profiling and bioinformatics analysis, using the same sample set derived from a large cohort of samples initially from a single centre, but extending to multiple centres. To date proteome profiling and analysis of samples to establish reproducibility in several centres has not been carried out.

The study presented here, as well as those published by other groups, shows the potential of proteomic profiling and bioinformatics analysis for the diagnosis, prognosis and therapeutic outcome in cancer and other diseases. Through the further development of both mass spectrometry and bioinformatics approaches, the profiling of tissue and/or body fluids from individual patients could potentially have a huge impact on personalised medicine and thereby improve upon the quality of life as well as mortality rates of patients with cancer and other diseases.

References

Abd El-Rehim DM, Pinder SE, Paish CE *et al.* Expression of luminal and basal cytokeratins in human breast carcinoma. *J Pathol* 2004; **203** (2):661-671.

Abd El-Rehim DM, Ball G, Pinder SE *et al.* High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. *Int J Cancer* 2005; **116** (3):340-350.

Abraha HD, Fuller LC, Du Vivier AW *et al.* Serum S-100 protein: a potentially useful prognostic marker in cutaneous melanoma. *Br J Dermatol* 1997; **137** (3):381-385.

Adam BL, Qu Y, Davis JW *et al.* Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res* 2002; **62** (13):3609-3614.

Adkins JN, Varnum SM, Auberry KJ *et al.* Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. *Mol Cell Proteomics* 2002; **1** (12):947-955.

Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003; **422** (6928):198-207.

Agarwal R, Kaye SB. Expression profiling and individualisation of treatment for ovarian cancer. *Curr Opin Pharmacol* 2006; **6** (4):345-349.

Ahmed A, Tollefsbol T. Telomeres, telomerase, and telomerase inhibition: clinical implications for cancer. *J Am Geriatr Soc* 2003; **51** (1):116-122.

Ajani J, Allgood V. Molecular mechanisms in cancer: what should clinicians know? *Semin Oncol* 2005; **32** (6 Suppl 8):2-4.

Al-Hajj M, Becker MW, Wicha M *et al.* Therapeutic implications of cancer stem cells. *Curr Opin Genet Dev* 2004; **14** (1):43-47.

Alaoui-Jamali MA, Xu YJ. Proteomic technology for biomarker profiling in cancer: an update. *J Zhejiang Univ Sci B* 2006; **7** (6):411-420.

Albrethsen J, Bogebo R, Gammeltoft S *et al.* Upregulated expression of human neutrophil peptides 1, 2 and 3 (HNP 1-3) in colon cancer serum and tumours: a biomarker study. *BMC Cancer* 2005; **5**:8.

Ali S, Coombes RC. Endocrine-responsive breast cancer and strategies for combating resistance. *Nat Rev Cancer* 2002; **2** (2):101-112.

Alonso SR, Ortiz P, Pollan M *et al.* Progression in cutaneous malignant melanoma is associated with distinct expression profiles: a tissue microarray-based study. *Am J Pathol* 2004; **164** (1):193-203.

Amari M, Suzuki A, Moriya T *et al.* LOH analyses of premalignant and malignant lesions of human breast: frequent LOH in 8p, 16q, and 17q in atypical ductal hyperplasia. *Oncol Rep* 1999; **6** (6):1277-1280.

An HX, Beckmann MW, Reifemberger G *et al.* Gene amplification and overexpression of CDK4 in sporadic breast carcinomas is associated with high tumor cell proliferation. *Am J Pathol* 1999; **154** (1):113-118.

Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* 2002; **1** (11):845-867.

Annesley TM. Ion suppression in mass spectrometry. *Clin Chem* 2003; **49** (7):1041-1044.

Anon. UK Breast Cancer Mortality Statistics. In., Series UK Breast Cancer Mortality Statistics. 2006.

Arnes JB, Brunet JS, Stefansson I *et al.* Placental cadherin and the basal epithelial phenotype of BRCA1-related breast cancer. *Clin Cancer Res* 2005; **11** (11):4003-4011.

Awada A, Cardoso F, Atalay G *et al.* The pipeline of new anticancer agents for breast cancer treatment in 2003. *Crit Rev Oncol Hematol* 2003; **48** (1):45-63.

Baak JPA, Path FRC, Hermsen MAJA *et al.* Genomics and proteomics in cancer. *European Journal of Cancer* 2003; **39**:1199-1215.

Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* 2004; **20** (5):777-785.

Baggerly KA, Morris JS, Edmonson SR *et al.* Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *J Natl Cancer Inst* 2005; **97** (4):307-309.

Balch CM, Reintgen DS, Kirkwood JM. Cutaneous Melanoma. In: VT DeVita; S Hellmann; SA Rosenberg, editors *Cancer: Principles and Practice of Oncology*. 5th edn. Philadelphia, Pa: Lippincott; 1997; p. 1935-1993.

Balch CM, Soong SJ, Atkins MB *et al.* An evidence-based staging system for cutaneous melanoma. *CA Cancer J Clin* 2004; **54** (3):131-149.

Ball G, Mian S, Holding F *et al.* An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics* 2002; **18** (3):395-404.

Banerjee S, Reis-Filho JS, Ashley S *et al.* Basal-like breast carcinomas: clinical outcome and response to chemotherapy. *J Clin Pathol* 2006; **59** (7):729-735.

Barry MJ. Clinical practice. Prostate-specific-antigen testing for early diagnosis of prostate cancer. *N Engl J Med* 2001; **344** (18):1373-1377.

Bartek J, Lukas J. Pathways governing G1/S transition and their response to DNA damage. *FEBS Lett* 2001; **490** (3):117-122.

Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. *J Microbiol Methods* 2000; **43** (1):3-31.

Bastian BC, Kashani-Sabet M, Hamm H *et al.* Gene amplifications characterize acral melanoma and permit the detection of occult tumor cells in the surrounding skin. *Cancer Res* 2000; **60** (7):1968-1973.

Bastian BC, Olshen AB, LeBoit PE *et al.* Classifying melanocytic tumors based on DNA copy number changes. *Am J Pathol* 2003; **163** (5):1765-1770.

Beer I, Barnea E, Ziv T *et al.* Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* 2004; **4** (4):950-960.

Behrens J, von Kries JP, Kuhl M *et al.* Functional interaction of beta-catenin with the transcription factor LEF-1. *Nature* 1996; **382** (6592):638-642.

Bertram JS. The molecular biology of cancer. *Molecular Aspects of Medicine* 2001; **21**:167-223.

Bertucci F, Birnbaum D, Goncalves A. Proteomics of breast cancer: principles and potential clinical applications. *Mol Cell Proteomics* 2006; **5** (10):1772-1786.

Bhatia M, Wang JC, Kapp U *et al.* Purification of primitive human hematopoietic cells capable of repopulating immune-deficient mice. *Proc Natl Acad Sci U S A* 1997; **94** (10):5320-5325.

Bhowmick NA, Neilson EG, Moses HL. Stromal fibroblasts in cancer initiation and progression. *Nature* 2004; **432** (7015):332-337.

- Bianchi NO, Bianchi MS, Richard SM. Mitochondrial genome instability in human cancers. *Mutat Res* 2001; **488** (1):9-23.
- Bischoff R, Luider TM. Methodological advances in the discovery of protein and peptide disease markers. *J Chromatogr B Analyt Technol Biomed Life Sci* 2004; **803** (1):27-40.
- Bissell MJ. The differentiated state of normal and malignant cells or how to define a "normal" cell in culture. *Int Rev Cytol* 1981; **70**:27-100.
- Bissell MJ, Radisky D. Putting tumours in context. *Nat Rev Cancer* 2001; **1** (1):46-54.
- Bissell MJ, Labarge MA. Context, tissue plasticity, and cancer: are tumor stem cells also regulated by the microenvironment? *Cancer Cell* 2005; **7** (1):17-23.
- Bittner M, Meltzer P, Chen Y *et al*. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 2000; **406** (6795):536-540.
- Bogenrieder T, Herlyn M. Cell-surface proteolysis, growth factor activation and intercellular communication in the progression of melanoma. *Critical Reviews in Oncology/Hematology* 2002; **44**:1-15.
- Bonnet D, Dick JE. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med* 1997; **3** (7):730-737.
- Bos JL. Ras oncogenes in human cancer: a review. *Cancer Res* 1989; **49** (17):4682-4689.
- Bos JL. All in the family? New insights and questions regarding interconnectivity of Ras, Rap1 and Ral. *EMBO Journal* 1998; **17** (23):6776-6782.
- Brembeck FH, Rosario M, Birchmeier W. Balancing cell adhesion and Wnt signaling, the key role of beta-catenin. *Curr Opin Genet Dev* 2006; **16** (1):51-59.
- Brennan M, Wilcken N, French J *et al*. Management of early breast cancer--the current approach. *Aust Fam Physician* 2005; **34** (9):755-760.
- Brenner S, Tamir E. Early Detection of Melanoma: The Best Strategy for a Favourable Prognosis. *Clinics in Dermatology* 2002; **20**:203-211.
- Breslow A. Tumor thickness, level of invasion and node dissection in stage I cutaneous melanoma. *Ann Surg* 1975; **182** (5):572-575.

- Brichory F, Beer D, Le Naour F *et al.* Proteomics-based identification of protein gene product 9.5 as a tumor antigen that induces a humoral immune response in lung cancer. *Cancer Res* 2001; **61** (21):7908-7912.
- Brochez L, Naeyaert JM. Serological markers for melanoma. *Br J Dermatol* 2000; **143** (2):256-268.
- Bruce WR, Van Der Gaag H. A Quantitative Assay For The Number Of Murine Lymphoma Cells Capable Of Proliferation In Vivo. *Nature* 1963; **199**:79-80.
- Bryan BB, Schnitt SJ, Collins LC. Ductal carcinoma in situ with basal-like phenotype: a possible precursor to invasive basal-like breast cancer. *Mod Pathol* 2006; **19** (5):617-621.
- Cairns P, Polascik TJ, Eby Y *et al.* Frequency of homozygous deletion at p16/CDKN2 in primary human tumours. *Nat Genet* 1995; **11** (2):210-212.
- Caldwell RL, Caprioli RM. Tissue profiling by mass spectrometry: a review of methodology and applications. *Mol Cell Proteomics* 2005; **4** (4):394-401.
- Callagy G, Cattaneo E, Daigo Y *et al.* Molecular classification of breast carcinomas using tissue microarrays. *Diagn Mol Pathol* 2003; **12** (1):27-34.
- Callahan R. p53 mutations, another breast cancer prognostic factor. *J Natl Cancer Inst* 1992; **84** (11):826-827.
- Campbell SL, Khosravi-Far R, Rossman KL *et al.* Increasing complexity of Ras signaling. *Oncogene* 1998; **17** (11 Reviews):1395-1413.
- Carlson SM, Najmi A, Whitin JC *et al.* Improving feature detection and analysis of surface-enhanced laser desorption/ionization-time of flight mass spectra. *Proteomics* 2005; **5** (11):2778-2788.
- Castellano M, Pollock PM, Walters MK *et al.* CDKN2A/p16 is inactivated in most melanoma cell lines. *Cancer Res* 1997; **57** (21):4868-4875.
- Castellano M, Parmiani G. Genes involved in melanoma: an overview of INK4a and other loci. *Melanoma Res* 1999; **9** (5):421-432.
- Cazares LH, Adam B-L, Ward MD *et al.* Normal, Benign, Preneoplastic, and Malignant Prostate Cells have Distinct Protein Expression Profiles Resolved by Surface Enhanced Laser Desorption/Ionization Mass Spectrometry. *Clinical Cancer Research* 2002; **8**:2541-2552.

Celis JE, Wolf H, Ostergaard M. Bladder squamous cell carcinoma biomarkers derived from proteomics. *Electrophoresis* 2000; **21** (11):2115-2121.

Chauhan H, Abraham A, Phillips JR *et al.* There is more than one kind of myofibroblast: analysis of CD34 expression in benign, in situ, and invasive breast lesions. *J Clin Pathol* 2003; **56** (4):271-276.

Chaurand P, Luetzenkirchen F, Spengler B. Peptide and protein identification by matrix-assisted laser desorption ionization (MALDI) and MALDI-post-source decay time-of-flight mass spectrometry. *J Am Soc Mass Spectrom* 1999a; **10** (2):91-103.

Chaurand P, Stoeckli M, Caprioli RM. Direct profiling of proteins in biological tissue sections by MALDI mass spectrometry. *Anal Chem* 1999b; **71** (23):5263-5270.

Chaurand P, Schwartz SA, Caprioli RM. Profiling and imaging proteins in tissue sections by MS. *Anal Chem* 2004; **76** (5):87A-93A.

Chaurand P, Cornett DS, Caprioli RM. Molecular imaging of thin mammalian tissue sections by mass spectrometry. *Curr Opin Biotechnol* 2006; **17** (4):431-436.

Check E. Proteomics and cancer: running before we can walk? *Nature* 2004; **429** (6991):496-497.

Chen ST, Pan TL, Tsai YC *et al.* Proteomics reveals protein profile changes in doxorubicin--treated MCF-7 human breast cancer cells. *Cancer Lett* 2002; **181** (1):95-107.

Chevallet M, Santoni V, Poinas A *et al.* New zwitterionic detergents improve the analysis of membrane proteins by two-dimensional electrophoresis. *Electrophoresis* 1998; **19** (11):1901-1909.

Chin L. The genetics of malignant melanoma: lessons from mouse and man. *Nat Rev Cancer* 2003; **3** (8):559-570.

Cho WC, Yip TT, Yip C *et al.* Identification of serum amyloid a protein as a potentially useful biomarker to monitor relapse of nasopharyngeal cancer by serum proteomic profiling. *Clin Cancer Res* 2004; **10** (1 Pt 1):43-52.

Cho Y, Gorina S, Jeffrey PD *et al.* Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science* 1994; **265** (5170):346-355.

Christensen C, Guldberg P. Growth factors rescue cutaneous melanoma cells from apoptosis induced by knockdown of mutated (V 600 E) B-RAF. *Oncogene* 2005; **24** (41):6292-6302.

Christofori G, Semb H. The role of the cell-adhesion molecule E-cadherin as a tumour-suppressor gene. *Trends Biochem Sci* 1999; **24** (2):73-76.

Chudnovsky Y, Khavari PA, Adams AE. Melanoma genetics and the development of rational therapeutics. *J Clin Invest* 2005; **115** (4):813-824.

Clark RE, Dodi IA, Hill SC *et al*. Direct evidence that leukemic cells present HLA-associated immunogenic peptides derived from the BCR-ABL b3a2 fusion protein. *Blood* 2001; **98** (10):2887-2893.

Clark WH, Jr., From L, Bernardino EA *et al*. The histogenesis and biologic behavior of primary human malignant melanomas of the skin. *Cancer Res* 1969; **29** (3):705-727.

Clark WH, Jr., Elder DE, Guerry Dt *et al*. A study of tumor progression: the precursor lesions of superficial spreading and nodular melanoma. *Hum Pathol* 1984; **15** (12):1147-1165.

Clarke MF, Fuller M. Stem cells and cancer: two faces of eve. *Cell* 2006; **124** (6):1111-1115.

Conrads TP, Hood BL, Petricoin EF, 3rd *et al*. Cancer proteomics: many technologies, one goal. *Expert Rev Proteomics* 2005; **2** (5):693-703.

Consortium BCL. Pathology of familial breast cancer: differences between breast cancers in carriers of BRCA1 or BRCA2 mutations and sporadic cases. Breast Cancer Linkage Consortium. *Lancet* 1997; **349** (9064):1505-1510.

Coombes KR, Fritsche HA, Jr., Clarke C *et al*. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin Chem* 2003; **49** (10):1615-1623.

Craven RA, Totty N, Harnden P *et al*. Laser capture microdissection and two-dimensional polyacrylamide gel electrophoresis: evaluation of tissue preparation and sample limitations. *Am J Pathol* 2002; **160** (3):815-822.

Craven RA, Stanley AJ, Hanrahan S *et al*. Identification of proteins regulated by interferon-alpha in resistant and sensitive malignant melanoma cell lines. *Proteomics* 2004; **4** (12):3998-4009.

Crespo P, Leon J. Ras proteins in the control of the cell cycle and cell differentiation. *Cell Mol Life Sci* 2000; **57** (11):1613-1636.

- Davies AMC, Fearn T. Back to basics: the principles of principal component analysis. In: *Spectroscopy Europe*. Vol. 16, Series Back to basics: the principles of principal component analysis. 2005; p. 20-23.
- Davies H, Bignell GR, Cox C *et al*. Mutations of the BRAF gene in human cancer. *Nature* 2002; **417** (6892):949-954.
- De Noo ME, Tollenaar RA, Ozalp A *et al*. Reliability of human serum protein profiles generated with C8 magnetic beads assisted MALDI-TOF mass spectrometry. *Anal Chem* 2005; **77** (22):7232-7241.
- de Snoo FA, Hayward NK. Cutaneous melanoma susceptibility and progression genes. *Cancer Lett* 2005; **230** (2):153-186.
- Deichmann M, Thome M, Benner A *et al*. B-raf exon 15 mutations are common in primary melanoma resection specimens but not associated with clinical outcome. *Oncology* 2004; **66** (5):411-419.
- Diamandis EP. Proteomic patterns in serum and identification of ovarian cancer. *Lancet* 2002; **360** (9327):170; author reply 170-171.
- Diamandis EP. Point: Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? *Clin Chem* 2003a; **49** (8):1272-1275.
- Diamandis EP. Re: Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst* 2003b; **95** (6):489-490; author reply 490-481.
- Diamandis EP. Identification of serum amyloid a protein as a potentially useful biomarker for nasopharyngeal carcinoma. *Clin Cancer Res* 2004a; **10** (15):5293; author reply 5293-5294.
- Diamandis EP. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. *J Natl Cancer Inst* 2004b; **96** (5):353-356.
- Diamandis EP, van der Merwe DE. Plasma protein profiling by mass spectrometry for cancer diagnosis: opportunities and limitations. *Clin Cancer Res* 2005; **11** (3):963-965.
- DiCiommo D, Gallie BL, Bremner R. Retinoblastoma: the disease, gene and protein provide critical leads to understand cancer. *Seminars in Cancer Biology* 2000; **10**:255-269.
- Dittmer D, Pati S, Zambetti G *et al*. Gain of function mutations in p53. *Nat Genet* 1993; **4** (1):42-46.

- Djavan B, Remzi M, Zlotta A *et al.* Novel artificial neural network for early detection of prostate cancer. *J Clin Oncol* 2002; **20** (4):921-929.
- Downward J. Targeting RAS signalling pathways in cancer therapy. *Nat Rev Cancer* 2003; **3** (1):11-22.
- Drake R, Cazares LH, Corica A *et al.* Quality control, preparation, and protein stability issues for blood serum and plasma used in biomarker discovery and proteomic profiling assays. *Bioprocessing Journal* 2004; (3):45-50.
- Drees F, Pokutta S, Yamada S *et al.* Alpha-catenin is a molecular switch that binds E-cadherin-beta-catenin and regulates actin-filament assembly. *Cell* 2005; **123** (5):903-915.
- Druker BJ, Sawyers CL, Capdeville R *et al.* Chronic myelogenous leukemia. *Hematology (Am Soc Hematol Educ Program)* 2001:87-112.
- Eccles SA. Targeting key steps in metastatic tumour progression. *Current opinion in genetics & development* 2005; **15**:77-86.
- el-Deiry WS, Tokino T, Velculescu VE *et al.* WAF1, a potential mediator of p53 tumor suppression. *Cell* 1993; **75** (4):817-825.
- Engwegen JY, Gast MC, Schellens JH *et al.* Clinical proteomics: searching for better tumour markers with SELDI-TOF mass spectrometry. *Trends Pharmacol Sci* 2006; **27** (5):251-259.
- Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990; **61** (5):759-767.
- Fearon ER, Dang CV. Cancer genetics: tumor suppressor meets oncogene. *Curr Biol* 1999; **9** (2):R62-65.
- Fedi P, Tronick SR, Aaronson SA. Growth Factors. In: JF Holland; RC Bast; DL Morton *et al.*, editors *Cancer Medicine*. Baltimore, MD: Williams and Wilkins; 1997; p. 41-64.
- Ferrari L, Seraglia R, Rossi CR *et al.* Protein profiles in sera of patients with malignant cutaneous melanoma. *Rapid Commun Mass Spectrom* 2000; **14** (13):1149-1154.
- Finne P, Finne R, Auvinen A *et al.* Predicting the outcome of prostate biopsy in screen-positive men by a multilayer perceptron network. *Urology* 2000; **56** (3):418-422.
- Fisher B, Dignam J, Bryant J *et al.* Five versus more than five years of tamoxifen for lymph node-negative breast cancer: updated findings from the National Surgical Adjuvant Breast and Bowel Project B-14 randomized trial. *J Natl Cancer Inst* 2001; **93** (9):684-690.

- Ford D, Easton DF, Peto J. Estimates of the gene frequency of BRCA1 and its contribution to breast and ovarian cancer incidence. *Am J Hum Genet* 1995; **57** (6):1457-1462.
- Frebourg T, Friend SH. Cancer risks from germline p53 mutations. *J Clin Invest* 1992; **90** (5):1637-1641.
- Fung ET, Weinberger SR, Gavin E *et al.* Bioinformatics approaches in clinical proteomics. *Expert Rev Proteomics* 2005; **2** (6):847-862.
- Garber K. Debate rages over proteomic patterns. *J Natl Cancer Inst* 2004; **96** (11):816-818.
- Garnett MJ, Marais R. Guilty as charged: B-RAF is a human oncogene. *Cancer Cell* 2004; **6** (4):313-319.
- Gilar M, Bouvier ES, Compton BJ. Advances in sample preparation in electromigration, chromatographic and mass spectrometric separation methods. *J Chromatogr A* 2001; **909** (2):111-135.
- Glare P. Clinical predictors of survival in advanced cancer. *J Support Oncol* 2005; **3** (5):331-339.
- Godbout R, Dryja TP, Squire J *et al.* Somatic inactivation of genes on chromosome 13 is a common event in retinoblastoma. *Nature* 1983; **304** (5925):451-453.
- Golub TR, Slonim DK, Tamayo P *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; **286** (5439):531-537.
- Gonzalez-Zulueta M, Bender CM, Yang AS *et al.* Methylation of the 5' CpG island of the p16/CDKN2 tumor suppressor gene in normal and transformed human tissues correlates with gene silencing. *Cancer Res* 1995; **55** (20):4531-4535.
- Gray-Schopfer VC, da Rocha Dias S, Marais R. The role of B-RAF in melanoma. *Cancer Metastasis Rev* 2005; **24** (1):165-183.
- Green DR, Reed JC. Mitochondria and apoptosis. *Science* 1998; **281** (5381):1309-1312.
- Gronbaek K, de Nully Brown P, Moller MB *et al.* Concurrent disruption of p16INK4a and the ARF-p53 pathway predicts poor prognosis in aggressive non-Hodgkin's lymphoma. *Leukemia* 2000; **14** (10):1727-1735.
- Gross JH. Introduction. In: *Mass Spectrometry: A Textbook*. Berlin: Springer; 2004a; p. 1-12.

- Gross JH. Matrix-Assisted Laser Desorption/Ionization. In: Mass Spectrometry: A Textbook. Berlin: Springer; 2004b; p. 411-440.
- Guldborg P, thor Straten P, Birck A *et al.* Disruption of the MMAC1/PTEN gene by deletion or mutation is a frequent event in malignant melanoma. *Cancer Res* 1997; **57** (17):3660-3663.
- Gygi SP, Corthals GL, Zhang Y *et al.* Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A* 2000; **97** (17):9390-9395.
- Haber DA, Fearon ER. The promise of cancer genetics. *Lancet* 1998; **351 Suppl 2**:SII1-8.
- Hanahan D, Folkman J. Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. *Cell* 1996; **86** (3):353-364.
- Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 2000; **100** (1):57-70.
- Hansson LO, von Schoultz E, Djureen E *et al.* Prognostic value of serum analyses of S-100 protein beta in malignant melanoma. *Anticancer Res* 1997; **17** (4B):3071-3073.
- Harbour JW, Luo RX, Dei Santi A *et al.* Cdk phosphorylation triggers sequential intramolecular interactions that progressively block Rb functions as cells move through G1. *Cell* 1999; **98** (6):859-869.
- Harlozinska A. Progress in molecular mechanisms of tumor metastasis and angiogenesis. *Anticancer Res* 2005; **25** (5):3327-3333.
- Harper JW, Elledge SJ, Keyomarsi K *et al.* Inhibition of cyclin-dependent kinases by p21. *Mol Biol Cell* 1995; **6** (4):387-400.
- Harris CC. p53 tumor suppressor gene: from the basic research laboratory to the clinic--an abridged historical perspective. *Carcinogenesis* 1996; **17** (6):1187-1198.
- Harris RA, Yang A, Stein RC *et al.* Cluster analysis of an extensive human breast cancer cell line protein expression map database. *Proteomics* 2002; **2** (2):212-223.
- Healy E, Rehman I, Angus B *et al.* Loss of heterozygosity in sporadic primary cutaneous melanoma. *Genes Chromosomes Cancer* 1995; **12** (2):152-156.
- Heike Y, Hosokawa M, Osumi S *et al.* Identification of serum proteins related to adverse effects induced by docetaxel infusion from protein expression profiles of serum using SELDI ProteinChip system. *Anticancer Res* 2005; **25** (2B):1197-1203.

- Herbert BR, Molloy MP, Gooley AA *et al.* Improved protein solubility in two-dimensional electrophoresis using tributyl phosphine as reducing agent. *Electrophoresis* 1998; **19** (5):845-851.
- Herbst RA, Weiss J, Ehnis A *et al.* Loss of heterozygosity for 10q22-10qter in malignant melanoma progression. *Cancer Res* 1994; **54** (12):3111-3114.
- Herlyn M. Emerging concepts and technologies in melanoma research. *Melanoma Res* 2002; **12** (1):3-8.
- Herman JG, Merlo A, Mao L *et al.* Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers. *Cancer Res* 1995; **55** (20):4525-4530.
- Herman JG, Graff JR, Myohanen S *et al.* Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc Natl Acad Sci U S A* 1996; **93** (18):9821-9826.
- Heubeck B, Wendler O, Bumm K *et al.* Tumor-associated antigenic pattern in squamous cell carcinomas of the head and neck - Analysed by SEREX. *Eur J Cancer* 2006.
- Hoek K, Rimm DL, Williams KR *et al.* Expression profiling reveals novel pathways in the transformation of melanocytes to melanomas. *Cancer Res* 2004; **64** (15):5270-5282.
- Hollstein M, Shomer B, Greenblatt M *et al.* Somatic point mutations in the p53 gene of human tumors and cell lines: updated compilation. *Nucleic Acids Res* 1996; **24** (1):141-146.
- Holyoake T, Jiang X, Eaves C *et al.* Isolation of a highly quiescent subpopulation of primitive leukemic cells in chronic myeloid leukemia. *Blood* 1999; **94** (6):2056-2064.
- Hong H, Dragan Y, Epstein J *et al.* Quality control and quality assessment of data from surface-enhanced laser desorption/ionization (SELDI) time-of flight (TOF) mass spectrometry (MS). *BMC Bioinformatics* 2005; **6 Suppl 2**:S5.
- Hood BL, Darfler MM, Guiel TG *et al.* Proteomic analysis of formalin-fixed prostate cancer tissue. *Mol Cell Proteomics* 2005; **4** (11):1741-1753.
- Hope KJ, Jin L, Dick JE. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nat Immunol* 2004; **5** (7):738-743.
- Hortin GL. The MALDI-TOF mass spectrometric view of the plasma proteome and peptidome. *Clin Chem* 2006; **52** (7):1223-1237.

Hoshino R, Chatani Y, Yamori T *et al.* Constitutive activation of the 41-/43-kDa mitogen-activated protein kinase signaling pathway in human tumors. *Oncogene* 1999; **18** (3):813-822.

Houssami N, Cuzick J, Dixon JM. The prevention, detection, and management of breast cancer. *Med J Aust* 2006; **184** (5):230-234.

Howard BA, Wang MZ, Campa MJ *et al.* Identification and validation of a potential lung cancer serum biomarker detected by matrix-assisted laser desorption/ionization-time of flight spectra analysis. *Proteomics* 2003; **3** (9):1720-1724.

Howell A, Wardley AM. Overview of the impact of conventional systemic therapies on breast cancer. *Endocr Relat Cancer* 2005; **12 Suppl 1**:S9-S16.

Hu Y, Zhang S, Yu J *et al.* SELDI-TOF-MS: the proteomics and bioinformatics approaches in the diagnosis of breast cancer. *Breast* 2005; **14** (4):250-255.

Hussain SP, Harris CC. p53 biological network: at the crossroads of the cellular-stress response pathway and molecular carcinogenesis. *J Nippon Med Sch* 2006; **73** (2):54-64.

Hussein MR, Wood GS. Molecular aspects of melanocytic dysplastic nevi. *J Mol Diagn* 2002; **4** (2):71-80.

Hussein MR, Roggero E, Tuthill RJ *et al.* Identification of novel deletion Loci at 1p36 and 9p22-21 in melanocytic dysplastic nevi and cutaneous malignant melanomas. *Arch Dermatol* 2003; **139** (6):816-817.

Hussein MR. Genetic pathways to melanoma tumorigenesis. *J Clin Pathol* 2004; **57** (8):797-801.

Hutchens TW, Yip TT. New desorption strategies for the mass spectrometric analysis of macromolecules. *Rapid Communications in Mass Spectrometry* 1993; **7**:576-580.

Ingvarsson S. Molecular genetics of breast cancer progression. *Semin Cancer Biol* 1999; **9** (4):277-288.

Issaq HJ, Conrads TP, Janini GM *et al.* Methods for fractionation, separation and profiling of proteins and peptides. *Electrophoresis* 2002a; **23** (17):3048-3061.

Issaq HJ, Veenstra TD, Conrads TP *et al.* The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification. *Biochem Biophys Res Commun* 2002b; **292** (3):587-592.

- Iwadata Y, Sakaida T, Hiwasa T *et al.* Molecular classification and survival prediction in human gliomas based on proteome analysis. *Cancer Res* 2004; **64** (7):2496-2501.
- Jacobs IJ, Menon U. Progress and challenges in screening for early detection of ovarian cancer. *Mol Cell Proteomics* 2004; **3** (4):355-366.
- Jain KK. Personalized medicine. *Curr Opin Mol Ther* 2002; **4** (6):548-558.
- Jain KK. Role of oncoproteomics in the personalized management of cancer. *Expert Rev Proteomics* 2004; **1** (1):49-55.
- Jenkins RE, Pennington SR. Arrays for protein expression profiling: towards a viable alternative to two-dimensional gel electrophoresis? *Proteomics* 2001; **1** (1):13-29.
- Jin Y, Manabe T. Direct targeting of human plasma for matrix-assisted laser desorption/ionization and analysis of plasma proteins by time of flight-mass spectrometry. *Electrophoresis* 2005; **26** (14):2823-2834.
- Johnson RE, Sundquist BUR. Laser-pulse Ejection of Organic Molecules from a Matrix - Lessons from Fast-ion-induced Ejection. *Rapid Communications in Mass Spectrometry* 1991; **5**:574-578.
- Jones MB, Krutzsch H, Shu H *et al.* Proteomic analysis and identification of new biomarkers and therapeutic targets for invasive ovarian cancer. *Proteomics* 2002; **2** (1):76-84.
- Jordan VC. The development of tamoxifen for breast cancer therapy: a tribute to the late Arthur L. Walpole. *Breast Cancer Res Treat* 1988; **11** (3):197-209.
- Kalkman E, Baxter G. Melanoma. *Clin Radiol* 2004; **59** (4):313-326.
- Kamb A, Gruis NA, Weaver-Feldhaus J *et al.* A cell cycle regulator potentially involved in genesis of many tumor types. *Science* 1994a; **264** (5157):436-440.
- Kamb A, Shattuck-Eidens D, Eeles R *et al.* Analysis of the p16 gene (CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus. *Nat Genet* 1994b; **8** (1):23-26.
- Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* 1988; **60** (20):2299-2301.
- Kawakami Y, Fujita T, Matsuzaki Y *et al.* Identification of human tumor antigens and its implications for diagnosis and treatment of cancer. *Cancer Sci* 2004; **95** (10):784-791.

- Khan SA, Rogers MA, Khurana KK *et al.* Estrogen receptor expression in benign breast epithelium and breast cancer risk. *J Natl Cancer Inst* 1998; **90** (1):37-42.
- Kim CJ, Dessureault S, Gabrilovich D *et al.* Immunotherapy for Melanoma. *Cancer Control* 2002a; **9** (1):22-30.
- Kim CJ, Reintgen DS, Balch CM. The New Melanoma Staging System. *Cancer Control* 2002b; **9** (1):9-15.
- Kligman D, Hilt DC. The S100 protein family. *Trends Biochem Sci* 1988; **13** (11):437-443.
- Koike H, Uzawa K, Nakashima D *et al.* Identification of differentially expressed proteins in oral squamous cell carcinoma using a global proteomic approach. *Int J Oncol* 2005; **27** (1):59-67.
- Kononen J, Bubendorf L, Kallioniemi A *et al.* Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 1998; **4** (7):844-847.
- Koomen JM, Shih LN, Coombes KR *et al.* Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins. *Clin Cancer Res* 2005; **11** (3):1110-1118.
- Korsching E, Packeisen J, Agelopoulos K *et al.* Cytogenetic alterations and cytokeratin expression patterns in breast cancer: integrating a new model of breast differentiation into cytogenetic pathways of breast carcinogenesis. *Lab Invest* 2002; **82** (11):1525-1533.
- Kouzarides T. Transcriptional control by the retinoblastoma protein. *Semin Cancer Biol* 1995; **6** (2):91-98.
- Kozak KR, Amneus MW, Pusey SM *et al.* Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: potential use in diagnosis and prognosis. *Proc Natl Acad Sci U S A* 2003; **100** (21):12343-12348.
- Kumar R, Angelini S, Czene K *et al.* BRAF mutations in metastatic melanoma: a possible association with clinical outcome. *Clin Cancer Res* 2003; **9** (9):3362-3368.
- Kurose K, Hoshaw-Woodard S, Adeyinka A *et al.* Genetic model of multi-step breast carcinogenesis involving the epithelium and stroma: clues to tumour-microenvironment interactions. *Hum Mol Genet* 2001; **10** (18):1907-1913.
- Lakhani SR. The transition from hyperplasia to invasive carcinoma of the breast. *J Pathol* 1999; **187** (3):272-278.

- Lakhani SR, Reis-Filho JS, Fulford L *et al.* Prediction of BRCA1 status in patients with breast cancer using estrogen receptor and basal phenotype. *Clin Cancer Res* 2005; **11** (14):5175-5180.
- Lakin N, Jackson S. Regulation of p53 in response to DNA damage. *Oncogene* 1999; **18**:7644 - 7655.
- Lancashire L, Schmid O, Shah H *et al.* Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis and principal components analysis. *Bioinformatics* 2005; **21** (10):2191-2199.
- Le L, Chi K, Tyldesley S *et al.* Identification of serum amyloid A as a biomarker to distinguish prostate cancer patients with bone lesions. *Clin Chem* 2005; **51** (4):695-707.
- Le Naour F, Misek DE, Krause MC *et al.* Proteomics-based identification of RS/DJ-1 as a novel circulating tumor antigen in breast cancer. *Clin Cancer Res* 2001; **7** (11):3328-3335.
- Le Naour F, Brichory F, Misek DE *et al.* A distinct repertoire of autoantibodies in hepatocellular carcinoma identified by proteomic analysis. *Mol Cell Proteomics* 2002; **1** (3):197-203.
- Legoix P, Bluteau O, Bayer J *et al.* Beta-catenin mutations in hepatocellular carcinoma correlate with a low rate of loss of heterozygosity. *Oncogene* 1999; **18** (27):4044-4046.
- Lehrer S, Roboz J, Ding H *et al.* Putative protein markers in the sera of men with prostatic neoplasms. *BJU Int* 2003; **92** (3):223-225.
- Lerebours F, Lidereau R. Molecular alterations in sporadic breast cancer. *Crit Rev Oncol Hematol* 2002; **44** (2):121-141.
- Lewis JK, Wei JS, Siuzdak G. Matrix-assisted Laser Desorption/Ionization Mass Spectrometry in Peptide and Protein Analysis. In: C Schoneich, editor *Encyclopedia of Analytical Chemistry*; John Wiley & Sons; 2000.
- Li G, Herlyn M. Dynamics of intercellular communication during melanoma development. *Mol Med Today* 2000; **6** (4):163-169.
- Li G, Miles A, Line A *et al.* Identification of tumour antigens by serological analysis of cDNA expression cloning. *Cancer Immunol Immunother* 2004; **53** (3):139-143.
- Li J, Zhang Z, Rosenzweig J *et al.* Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem* 2002a; **48** (8):1296-1304.

- Li N, Mangini J, Bhawan J. New prognostic factors of cutaneous melanoma: a review of the literature. *J Cutan Pathol* 2002b; **29** (6):324-340.
- Liede A, Karlan BY, Baldwin RL *et al.* Cancer incidence in a population of Jewish women at risk of ovarian cancer. *J Clin Oncol* 2002; **20** (6):1570-1577.
- Lilley KS, Razzaq A, Dupree P. Two-dimensional gel electrophoresis: recent advances in sample preparation, detection and quantitation. *Curr Opin Chem Biol* 2002; **6** (1):46-50.
- Linette GP, Carlson JA, Slominski A *et al.* Biomarkers in melanoma: stage III and IV disease. *Expert Rev Mol Diagn* 2005; **5** (1):65-74.
- Liotta LA, Kohn EC. The microenvironment of the tumour-host interface. *Nature* 2001a; **411**:375-379.
- Liotta LA, Kohn EC, Petricoin EF. Clinical Proteomics: Personalised Molecular Medicine. *Journal of the American Medical Association* 2001b; **286** (18):2211-2214.
- Liotta LA, Petricoin EF. Serum peptidome for cancer detection: spinning biologic trash into diagnostic gold. *J Clin Invest* 2006; **116** (1):26-30.
- Lipinski M, Jacks T. The retinoblastoma gene family in differentiation and development. *Oncogene* 1999; **18**:7873 - 7882.
- Lisboa PJ, Taktak AF. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural Netw* 2006; **19** (4):408-415.
- Liu T, Qian WJ, Mottaz HM *et al.* Evaluation of multi-protein immunoaffinity subtraction for plasma proteomics and candidate biomarker discovery using mass spectrometry. *Mol Cell Proteomics* 2006.
- Liu Y, West SC. Distinct functions of BRCA1 and BRCA2 in double-strand break repair. *Breast Cancer Res* 2002; **4** (1):9-13.
- Liu Y, Chen Q, Zhang JT. Tumor suppressor gene 14-3-3sigma is down-regulated whereas the proto-oncogene translation elongation factor 1delta is up-regulated in non-small cell lung cancers as identified by proteomic profiling. *J Proteome Res* 2004; **3** (4):728-735.
- Livasy CA, Karaca G, Nanda R *et al.* Phenotypic evaluation of the basal-like subtype of invasive breast carcinoma. *Mod Pathol* 2006; **19** (2):264-271.
- Loeb LA, Loeb KR, Anderson JP. Multiple mutations and cancer. *Proc Natl Acad Sci U S A* 2003; **100** (3):776-781.

- Lohrum MA, Ludwig RL, Kubbutat MH *et al.* Regulation of HDM2 activity by the ribosomal protein L11. *Cancer Cell* 2003; **3** (6):577-587.
- Lomuto M, Calabrese P, Giuliani A. Prognostic signs in melanoma: state of the art. *J Eur Acad Dermatol Venereol* 2004; **18** (3):291-300.
- Lopez MF. Proteome analysis. I. Gene products are where the biological action is. *J Chromatogr B Biomed Sci Appl* 1999; **722** (1-2):191-202.
- Ludwig JA, Weinstein JN. Biomarkers in cancer staging, prognosis and treatment selection. *Nat Rev Cancer* 2005; **5** (11):845-856.
- Lugovic L, Situm M, Kos L. Malignant Melanoma - Future Prospects. *Acta Dermatovenerologica Croatica* 2005; **13** (1):36-43.
- Lutzker SG, Levine AJ. A functionally inactive p53 protein in teratocarcinoma cells is activated by either DNA damage or cellular differentiation. *Nat Med* 1996; **2** (7):804-810.
- Malik G, Ward MD, Gupta SK *et al.* Serum levels of an isoform of apolipoprotein A-II as a potential marker for prostate cancer. *Clin Cancer Res* 2005; **11** (3):1073-1085.
- May P, May E. Twenty years of p53 research: structural and functional aspects of the p53 protein. *Oncogene* 1999; **18**:7621-7636.
- Melanoma Statistics and Prognosis.
[<http://www.cancerhelp.org.uk/help/default.asp?page=5436>] 2006 [cited 2006 August].
- Melle C, Ernst G, Schimmel B *et al.* Biomarker discovery and identification in laser microdissected head and neck squamous cell carcinoma with ProteinChip technology, two-dimensional gel electrophoresis, tandem mass spectrometry, and immunohistochemistry. *Mol Cell Proteomics* 2003; **2** (7):443-452.
- Melle C, Kaufmann R, Hommann M *et al.* Proteomic profiling in microdissected hepatocellular carcinoma tissue using ProteinChip technology. *Int J Oncol* 2004; **24** (4):885-891.
- Melle C, Ernst G, Schimmel B *et al.* Discovery and identification of alpha-defensins as low abundant, tumor-derived serum markers in colorectal cancer. *Gastroenterology* 2005; **129** (1):66-73.
- Menon U, Jacobs IJ. Recent developments in ovarian cancer screening. *Curr Opin Obstet Gynecol* 2000; **12** (1):39-42.

Mercer KE, Pritchard CA. Raf proteins and cancer: B-Raf is identified as a mutational target. *Biochim Biophys Acta* 2003; **1653** (1):25-40.

Mian S, Ball G, Hornbuckle J *et al.* A prototype methodology combining surface-enhanced laser desorption/ionization protein chip technology and artificial neural network algorithms to predict the chemoresponsiveness of breast cancer cell lines exposed to Paclitaxel and Doxorubicin under in vitro conditions. *Proteomics* 2003; **3** (9):1725-1737.

Mian S, Ugurel S, Parkinson E *et al.* Serum proteomic fingerprinting discriminates between clinical stages and predicts disease progression in melanoma patients. *J Clin Oncol* 2005; **23** (22):5088-5093.

Michener CM, Ardekani AM, Petricoin EF, 3rd *et al.* Genomics and proteomics: application of novel technology to early detection and prevention of cancer. *Cancer Detect Prev* 2002; **26** (4):249-255.

Michor F, Iwasa Y, Nowak MA. Dynamics of cancer progression. *Nat Rev Cancer* 2004; **4** (3):197-205.

Miles AK, Matharoo-Ball B, Li G *et al.* The identification of human tumour antigens: current status and future developments. *Cancer Immunol Immunother* 2006:1-8.

Misek DE, Imafuku Y, Hanash SM. Application of proteomic technologies to tumor analysis. *Pharmacogenomics* 2004; **5** (8):1129-1137.

Mittnacht S. Control of pRB phosphorylation. *Curr Opin Genet Dev* 1998; **8** (1):21-27.

Molina MA, Codony-Servat J, Albanell J *et al.* Trastuzumab (herceptin), a humanized anti-Her2 receptor monoclonal antibody, inhibits basal and activated Her2 ectodomain cleavage in breast cancer cells. *Cancer Res* 2001; **61** (12):4744-4749.

Molina R, Navarro J, Filella X *et al.* S-100 protein serum levels in patients with benign and malignant diseases: false-positive results related to liver and renal function. *Tumour Biol* 2002; **23** (1):39-44.

Moll UM, Petrenko O. The MDM2-p53 interaction. *Mol Cancer Res* 2003; **1** (14):1001-1008.

Momand J, Zambetti GP, Olson DC *et al.* The mdm-2 oncogene product forms a complex with the p53 protein and inhibits p53-mediated transactivation. *Cell* 1992; **69** (7):1237-1245.

Morin PJ, Sparks AB, Korinek V *et al.* Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. *Science* 1997; **275** (5307):1787-1790.

Mueller MM, Fusenig NE. Friends or foes - bipolar effects of the tumour stroma in cancer. *Nat Rev Cancer* 2004; **4** (11):839-849.

Myers MP, Pass I, Batty IH *et al.* The lipid phosphatase activity of PTEN is critical for its tumor suppressor function. *Proc Natl Acad Sci U S A* 1998; **95** (23):13513-13518.

Myers TG, Anderson NL, Waltham M *et al.* A protein expression database for the molecular pharmacology of cancer. *Electrophoresis* 1997; **18** (3-4):647-653.

Nambiar S, Mirmohammadsadegh A, Bar A *et al.* Applications of array technology: melanoma research and diagnosis. *Expert Rev Mol Diagn* 2004; **4** (4):549-557.

Negm RS, Verma M, Srivastava S. The promise of biomarkers in cancer screening and detection. *Trends in Molecular Medicine* 2002; **8** (6):288-293.

Nestler G, Steinert R, Lippert H *et al.* Using human samples in proteomics-based drug development: bioethical aspects. *Expert Rev Proteomics* 2004; **1** (1):77-86.

Nishizuka S, Chen ST, Gwadry FG *et al.* Diagnostic markers that distinguish colon and ovarian adenocarcinomas: identification by genomic, proteomic, and tissue array profiling. *Cancer Res* 2003; **63** (17):5243-5250.

Nobori T, Miura K, Wu DJ *et al.* Deletions of the cyclin-dependent kinase-4 inhibitor gene in multiple human cancers. *Nature* 1994; **368** (6473):753-756.

Novikova SI, He F, Cutrufello NJ *et al.* Identification of protein biomarkers for schizophrenia and bipolar disorder in the postmortem prefrontal cortex using SELDI-TOF-MS ProteinChip profiling combined with MALDI-TOF-PSD-MS analysis. *Neurobiol Dis* 2006; **23** (1):61-76.

Nyman TA. The role of mass spectrometry in proteome studies. *Biomol Eng* 2001; **18** (5):221-227.

Oehr P. Proteomics as a Tool for Detection of Nuclear Matrix Proteins and New Biomarkers for Screening of Early Tumour Stage. *Anticancer Research* 2003; **23**:805-812.

Ohta M, Nagai H, Shimizu M *et al.* Rarity of somatic and germline mutations of the cyclin-dependent kinase 4 inhibitor gene, CDK4I, in melanoma. *Cancer Res* 1994; **54** (20):5269-5272.

Okada T, Akada M, Fujita T *et al.* A novel cancer testis antigen that is frequently expressed in pancreatic, lung, and endometrial cancers. *Clin Cancer Res* 2006; **12** (1):191-197.

- Oliner JD, Kinzler KW, Meltzer PS *et al.* Amplification of a gene encoding a p53-associated protein in human sarcomas. *Nature* 1992; **358** (6381):80-83.
- Olumi AF, Grossfeld GD, Hayward SW *et al.* Carcinoma-associated fibroblasts direct tumor progression of initiated human prostatic epithelium. *Cancer Res* 1999; **59** (19):5002-5011.
- Omholt K, Platz A, Kanter L *et al.* NRAS and BRAF mutations arise early during melanoma pathogenesis and are preserved throughout tumor progression. *Clin Cancer Res* 2003; **9** (17):6483-6488.
- Ortega S, Malumbres M, Barbacid M. Cyclin D-dependent kinases, INK4 inhibitors and cancer. *Biochim Biophys Acta* 2002; **1602** (1):73-87.
- Osborne C, Wilson P, Tripathy D. Oncogenes and tumor suppressor genes in breast cancer: potential diagnostic and therapeutic applications. *Oncologist* 2004; **9** (4):361-377.
- Palacios J, Honrado E, Osorio A *et al.* Phenotypic characterization of BRCA1 and BRCA2 tumors based in a tissue microarray study with 37 immunohistochemical markers. *Breast Cancer Res Treat* 2005; **90** (1):5-14.
- Park WS, Vortmeyer AO, Pack S *et al.* Allelic deletion at chromosome 9p21(p16) and 17p13(p53) in microdissected sporadic dysplastic nevus. *Hum Pathol* 1998; **29** (2):127-130.
- Parsonage G, Filer AD, Haworth O *et al.* A stromal address code defined by fibroblasts. *Trends Immunol* 2005; **26** (3):150-156.
- Pearson A, Greenblatt J. Modular organization of the E2F1 activation domain and its interaction with general transcription factors TBP and TFIID. *Oncogene* 1997; **15** (22):2643-2658.
- Perez-Caro M, Sanchez-Garcia I. Killing time for cancer stem cells (CSC): discovery and development of selective CSC inhibitors. *Curr Med Chem* 2006; **13** (15):1719-1725.
- Perl AK, Wilgenbus P, Dahl U *et al.* A causal role for E-cadherin in the transition from adenoma to carcinoma. *Nature* 1998; **392** (6672):190-193.
- Perry A, Anderl K, Borell TJ *et al.* Detection of p16, RB, CDK4, and p53 gene deletion and amplification by fluorescence in situ hybridization in 96 gliomas. *Am J Clin Pathol* 1999; **112** (6):801-809.
- Peto J. Breast cancer susceptibility-A new look at an old model. *Cancer Cell* 2002; **1** (5):411-412.

- Petricoin EF, Ardekani AM, Hitt BA *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet* 2002a; **359**:572-577.
- Petricoin EF, Liotta LA. Proteomic analysis at the bedside: early detection of cancer. *Trends Biotechnol* 2002b; **20 (12 Suppl)**:S30-34.
- Petricoin EF, Zoon KC, Kohn EC *et al.* Clinical Proteomics: translating benchside promise into bedside reality. *Nature Reviews Drug Discovery* 2002c; **1**:683-695.
- Petricoin EF, 3rd, Ornstein DK, Paweletz CP *et al.* Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst* 2002d; **94 (20)**:1576-1578.
- Picksley SM, Lane DP. The p53-mdm2 autoregulatory feedback loop: a paradigm for the regulation of growth control by p53? *Bioessays* 1993; **15 (10)**:689-690.
- Pittet MJ, Zippelius A, Valmori D *et al.* Melan-A/MART-1-specific CD8 T cells: from thymus to tumour. *Trends in Immunology* 2002; **23 (7)**:325-328.
- Polakis P. The oncogenic activation of beta-catenin. *Curr Opin Genet Dev* 1999; **9 (1)**:15-21.
- Polsky D, Cordon-Cardo C. Oncogenes in melanoma. *Oncogene* 2003; **22 (20)**:3087-3091.
- Pomerantz J, Schreiber-Agus N, Liegeois NJ *et al.* The Ink4a tumor suppressor gene product, p19Arf, interacts with MDM2 and neutralizes MDM2's inhibition of p53. *Cell* 1998; **92 (6)**:713-723.
- Posadas EM, Simpkins F, Liotta LA *et al.* Proteomic analysis for the early detection and rational treatment of cancer--realistic hope? *Ann Oncol* 2005; **16 (1)**:16-22.
- Pritzker KP. Cancer biomarkers: easier said than done. *Clin Chem* 2002; **48 (8)**:1147-1150.
- Qu Y, Adam BL, Yasui Y *et al.* Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin Chem* 2002; **48 (10)**:1835-1843.
- Ragazzi E, Vogliardi S, Allegri G *et al.* Cluster analysis of serum proteins in malignant cutaneous melanoma: search for disease markers. *Rapid Commun Mass Spectrom* 2003; **17 (13)**:1511-1515.
- Rai AJ, Zhang Z, Rosenzweig J *et al.* Proteomic approaches to tumor marker discovery. *Arch Pathol Lab Med* 2002; **126 (12)**:1518-1526.

- Rakha EA, Abd El Rehim D, Pinder SE *et al.* E-cadherin expression in invasive non-lobular carcinoma of the breast and its prognostic significance. *Histopathology* 2005; **46** (6):685-693.
- Rakha EA, Putti TC, Abd El-Rehim DM *et al.* Morphological and immunophenotypic analysis of breast carcinomas with basal and myoepithelial differentiation. *J Pathol* 2006; **208** (4):495-506.
- Ranade K, Hussussian CJ, Sikorski RS *et al.* Mutations associated with familial melanoma impair p16INK4 function. *Nat Genet* 1995; **10** (1):114-116.
- Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer* 2005a; **5** (2):142-149.
- Ransohoff DF. Lessons from controversy: ovarian cancer screening and serum proteomics. *J Natl Cancer Inst* 2005b; **97** (4):315-319.
- Reed JA, Loganzo F, Jr., Shea CR *et al.* Loss of expression of the p16/cyclin-dependent kinase inhibitor 2 tumor suppressor gene in melanocytic lesions correlates with invasive stage of tumor progression. *Cancer Res* 1995; **55** (13):2713-2718.
- Remzi M, Anagnostou T, Ravery V *et al.* An artificial neural network to predict the outcome of repeat prostate biopsies. *Urology* 2003; **62** (3):456-460.
- Rieger PT. The biology of cancer genetics. *Semin Oncol Nurs* 2004; **20** (3):145-154.
- Righetti PG. Recent developments in electrophoretic methods. *J Chromatogr* 1990; **516** (1):3-22.
- Robbins RJ, Villanueva J, Tempst P. Distilling cancer biomarkers from the serum peptidome: high technology reading of tea leaves or an insight to clinical systems biology? *J Clin Oncol* 2005; **23** (22):4835-4837.
- Rocco JW, Sidransky D. p16(MTS-1/CDKN2/INK4a) in cancer progression. *Exp Cell Res* 2001; **264** (1):42-55.
- Rodland KD. Proteomics and cancer diagnosis: the potential of mass spectrometry. *Clin Biochem* 2004; **37** (7):579-583.
- Rodolfo M, Daniotti M, Vallacchi V. Genetic progression of metastatic melanoma. *Cancer Lett* 2004; **214** (2):133-147.

- Rodriguez-Pinilla SM, Sarrío D, Honrado E *et al.* Prognostic significance of basal-like phenotype and fascin expression in node-negative invasive breast carcinomas. *Clin Cancer Res* 2006; **12** (5):1533-1539.
- Ross DT, Scherf U, Eisen MB *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 2000; **24** (3):227-235.
- Ross JS, Fletcher JA. HER-2/neu (c-erb-B2) gene and protein in breast cancer. *Am J Clin Pathol* 1999; **112** (1 Suppl 1):S53-67.
- Ross RA, Biedler JL, Spengler BA. A role for distinct cell types in determining malignancy in human neuroblastoma cell lines and tumors. *Cancer Lett* 2003; **197** (1-2):35-39.
- Rotter V, Aloni-Grinstein R, Schwartz D *et al.* Does wild-type p53 play a role in normal cell differentiation? *Semin Cancer Biol* 1994; **5** (3):229-236.
- Roura S, Miravet S, Piedra J *et al.* Regulation of E-cadherin/Catenin association by tyrosine phosphorylation. *J Biol Chem* 1999; **274** (51):36734-36740.
- Ruiter D, Bogenrieder T, Elder D *et al.* Melanoma-stroma interactions: structural and functional aspects. *Lancet Oncol* 2002; **3** (1):35-43.
- Samowitz WS, Powers MD, Spirio LN *et al.* Beta-catenin mutations are more frequent in small colorectal adenomas than in larger adenomas and invasive carcinomas. *Cancer Res* 1999; **59** (7):1442-1444.
- Sasaki Y, Niu C, Makino R *et al.* BRAF point mutations in primary melanoma show different prevalences by subtype. *J Invest Dermatol* 2004; **123** (1):177-183.
- Sauter ER, Shan S, Hewett JE *et al.* Proteomic analysis of nipple aspirate fluid using SELDI-TOF-MS. *Int J Cancer* 2005; **114** (5):791-796.
- Scherf U, Ross DT, Waltham M *et al.* A gene expression database for the molecular pharmacology of cancer. *Nat Genet* 2000; **24** (3):236-244.
- Schmidt F, Schmid M, Jungblut PR *et al.* Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis. *J Am Soc Mass Spectrom* 2003; **14** (9):943-956.
- Schwartz SA, Weil RJ, Thompson RC *et al.* Proteomic-based prognosis of brain tumor patients using direct-tissue matrix-assisted laser desorption ionization mass spectrometry. *Cancer Res* 2005; **65** (17):7674-7681.

- Schwarzer G, Vach W, Schumacher M. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat Med* 2000; **19** (4):541-561.
- Schweigert FJ, Sehoul J. Transthyretin, a biomarker for nutritional status and ovarian cancer. *Cancer Res* 2005; **65** (3):1114.
- Scully R, Chen J, Ochs RL *et al*. Dynamic changes of BRCA1 subnuclear location and phosphorylation state are initiated by DNA damage. *Cell* 1997; **90** (3):425-435.
- Seibert V, Wiesner A, Buschmann T *et al*. Surface-enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI TOF-MS) and ProteinChip technology in proteomics research. *Pathol Res Pract* 2004; **200** (2):83-94.
- Seow TK, Ong SE, Liang RC *et al*. Two-dimensional electrophoresis map of the human hepatocellular carcinoma cell line, HCC-M, and identification of the separated proteins by mass spectrometry. *Electrophoresis* 2000; **21** (9):1787-1813.
- Shay JW, Bacchetti S. A survey of telomerase activity in human cancer. *Eur J Cancer* 1997; **33** (5):787-791.
- Shay JW, Wright WE. Senescence and immortalization: role of telomeres and telomerase. *Carcinogenesis* 2005; **26** (5):867-874.
- Sherr CJ, Roberts JM. CDK inhibitors: positive and negative regulators of G1-phase progression. *Genes Dev* 1999; **13** (12):1501-1512.
- Sherr CJ. The Pezcoller lecture: cancer cell cycles revisited. *Cancer Res* 2000; **60** (14):3689-3695.
- Shields JM, Pruitt K, McFall A *et al*. Understanding Ras: 'it ain't over 'til it's over'. *Trends Cell Biol* 2000; **10** (4):147-154.
- Shin BK, Wang H, Hanash S. Proteomics approaches to uncover the repertoire of circulating biomarkers for breast cancer. *J Mammary Gland Biol Neoplasia* 2002; **7** (4):407-413.
- Shinozaki M, Fujimoto A, Morton DL *et al*. Incidence of BRAF oncogene mutation and clinical relevance for primary cutaneous melanomas. *Clin Cancer Res* 2004; **10** (5):1753-1757.
- Shiwa M, Nishimura Y, Wakatabe R *et al*. Rapid discovery and identification of a tissue-specific tumor biomarker from 39 human cancer cell lines using the SELDI ProteinChip platform. *Biochem Biophys Res Commun* 2003; **309** (1):18-25.

Sidransky D, Von Eschenbach A, Tsai YC *et al.* Identification of p53 gene mutations in bladder cancers and urine samples. *Science* 1991; **252** (5006):706-709.

Sidransky D, Tokino T, Hamilton SR *et al.* Identification of ras oncogene mutations in the stool of patients with curable colorectal tumors. *Science* 1992; **256** (5053):102-105.

Sigal A, Rotter V. Oncogenic mutations of the p53 tumour suppressor: the demons of the guardian of the genome. *Cancer Research* 2000; **60**:6788-6793.

Simpkins F, Czechowicz JA, Liotta L *et al.* SELDI-TOF mass spectrometry for cancer biomarker discovery and serum proteomic diagnostics. *Pharmacogenomics* 2005; **6** (6):647-653.

Simpson RJ, Connolly LM, Eddes JS *et al.* Proteomic analysis of the human colon carcinoma cell line (LIM 1215): development of a membrane protein database. *Electrophoresis* 2000; **21** (9):1707-1732.

Singletary SE. Rating the risk factors for breast cancer. *Ann Surg* 2003; **237** (4):474-482.

Skobe M, Fusenig NE. Tumorigenic conversion of immortal human keratinocytes through stromal cell activation. *Proc Natl Acad Sci U S A* 1998; **95** (3):1050-1055.

Sobol H, Stoppa-Lyonnet D, Bressac-de-Paillerets B *et al.* Truncation at conserved terminal regions of BRCA1 protein is associated with highly proliferating hereditary breast cancers. *Cancer Res* 1996; **56** (14):3216-3219.

Somiari RI, Sullivan A, Russell S *et al.* High-throughput proteomic analysis of human infiltrating ductal carcinoma of the breast. *Proteomics* 2003; **3** (10):1863-1873.

Somiari RI, Somiari S, Russell S *et al.* Proteomics of breast carcinoma. *J Chromatogr B Analyt Technol Biomed Life Sci* 2005; **815** (1-2):215-225.

Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* 2003; **4**:24-37.

Sorlie T, Perou CM, Tibshirani R *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 2001; **98** (19):10869-10874.

Soufir N, Avril MF, Chompret A *et al.* Prevalence of p16 and CDK4 germline mutations in 48 melanoma-prone families in France. The French Familial Melanoma Study Group. *Hum Mol Genet* 1998; **7** (2):209-216.

- Souhami R, Tobias J. Skin Cancer. In: Cancer and its Management. 3rd edn: Blackwell Science; 1998; p. 373-380.
- StatSoft I. Electronic Statistics Textbook. In., Series Electronic Statistics Textbook. StatSoft, Tulsa, OK; 2006.
- Stoeckli M, Chaurand P, Hallahan DE *et al.* Imaging mass spectrometry: A new technology for the analysis of protein expression in mammalian tissues. *Nature Medicine* 2001; **7** (4):493-496.
- Tafeit E, Reibnegger G. Artificial Neural Networks in Laboratory Medicine and Medical Outcome Prediction. *Clinical Chemistry and Laboratory Medicine* 1999; **37** (9):845-853.
- Tanaka K, Waki H, Ido Y *et al.* Protein and Polymer Analyses up to M/z 100,000 by Laser Ionization-TOF-MS. *Rapid Communications in Mass Spectrometry* 1988; **2**:151-153.
- Tang N, Tornatore P, Weinberger SR. Current developments in SELDI affinity technology. *Mass Spectrom Rev* 2004; **23** (1):34-44.
- Thomas NE, Alexander A, Edmiston SN *et al.* Tandem BRAF mutations in primary invasive melanomas. *J Invest Dermatol* 2004; **122** (5):1245-1250.
- Thor AD, Moore DH, II, Edgerton SM *et al.* Accumulation of p53 tumor suppressor gene protein: an independent marker of prognosis in breast cancers. *J Natl Cancer Inst* 1992; **84** (11):845-855.
- Tibshirani R, Hastie T, Narasimhan B *et al.* Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics* 2004; **20** (17):3034-3044.
- Timperman AT, Aebersold R. Peptide electroextraction for direct coupling of in-gel digests with capillary LC-MS/MS for protein identification and sequencing. *Anal Chem* 2000; **72** (17):4115-4121.
- Tlsty TD. Stromal cells can contribute oncogenic signals. *Seminars in Cancer Biology* 2001; **11**:97-104.
- Tockman MS, Gupta PK, Pressman NJ *et al.* Considerations in bringing a cancer biomarker to clinical application. *Cancer Res* 1992; **52** (9 Suppl):2711s-2718s.
- Tsao H, Zhang X, Benoit E *et al.* Identification of PTEN/MMAC1 alterations in uncultured melanomas and melanoma cell lines. *Oncogene* 1998; **16** (26):3397-3402.

- Ueno I, Sakai T, Yamaoka M *et al.* Analysis of blood plasma proteins in patients with Alzheimer's disease by two-dimensional electrophoresis, sequence homology and immunodetection. *Electrophoresis* 2000; **21** (9):1832-1845.
- Unger T, Juven-Gershon T, Moallem E *et al.* Critical role for Ser20 of human p53 in the negative regulation of p53 by Mdm2. *Embo J* 1999; **18** (7):1805-1814.
- Urbani A, Poland J, Bernardini S *et al.* A proteomic investigation into etoposide chemoresistance of neuroblastoma cell lines. *Proteomics* 2005; **5** (3):796-804.
- van 't Veer LJ, Dai H, van de Vijver MJ *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415** (6871):530-536.
- van de Vijver MJ, Peterse JL, Mooi WJ *et al.* Neu-protein overexpression in breast cancer. Association with comedo-type ductal carcinoma in situ and limited prognostic value in stage II breast cancer. *N Engl J Med* 1988; **319** (19):1239-1245.
- van de Vijver MJ, He YD, van't Veer LJ *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002; **347** (25):1999-2009.
- van Dijk M, Sprenger S, Rombout P *et al.* Distinct chromosomal aberrations in sinonasal mucosal melanoma as detected by comparative genomic hybridization. *Genes Chromosomes Cancer* 2003; **36** (2):151-158.
- Vapnik V. The nature of statistical learning theory. Berlin: Springer; 1995.
- Veenstra TD, Conrads TP, Hood BL *et al.* Biomarkers: mining the biofluid proteome. *Mol Cell Proteomics* 2005; **4** (4):409-418.
- Verma M, Wright GL, Jr., Hanash SM *et al.* Proteomic approaches within the NCI early detection research network for the discovery and identification of cancer biomarkers. *Ann N Y Acad Sci* 2001; **945**:103-115.
- Vertes A, Levine RD. Sublimation versus Fragmentation in Matrix-assisted Laser Desorption. *Chemical Physics Letters* 1990; **171**:284-290.
- Villanueva J, Philip J, Entenberg D *et al.* Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry. *Anal Chem* 2004; **76** (6):1560-1570.
- Villanueva J, Shaffer DR, Philip J *et al.* Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J Clin Invest* 2006; **116** (1):271-284.

- Vlahou A, Schellhammer PF, Mendrinos S *et al.* Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. *Am J Pathol* 2001; **158** (4):1491-1502.
- Vlahou A, Laronga C, Wilson L *et al.* A novel approach toward development of a rapid blood test for breast cancer. *Clin Breast Cancer* 2003; **4** (3):203-209.
- Vlahou A, Fountoulakis M. Proteomic approaches in the search for disease biomarkers. *J Chromatogr B Analyt Technol Biomed Life Sci* 2005; **814** (1):11-19.
- Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. *Nat Med* 2004; **10** (8):789-799.
- Volpert OV, Dameron KM, Bouck N. Sequential development of an angiogenic phenotype by human fibroblasts progressing to tumorigenicity. *Oncogene* 1997; **14** (12):1495-1502.
- Voss T, Ahorn H, Haberl P *et al.* Correlation of clinical data with proteomics profiles in 24 patients with B-cell chronic lymphocytic leukemia. *Int J Cancer* 2001; **91** (2):180-186.
- Vuong GL, Weiss SM, Kammer W *et al.* Improved sensitivity proteomics by postharvest alkylation and radioactive labelling of proteins. *Electrophoresis* 2000; **21** (13):2594-2605.
- Wagner AJ, Fisher DE. Melanocyte signaling pathways and the etiology of melanoma. *Drug Discovery Today* 2005; **2** (2):179-183.
- Wang H, Hanash S. Multi-dimensional liquid phase based separations in proteomics. *J Chromatogr B Analyt Technol Biomed Life Sci* 2003; **787** (1):11-18.
- Wang JC, Dick JE. Cancer stem cells: lessons from leukemia. *Trends Cell Biol* 2005; **15** (9):494-501.
- Wascher RA, Morton DL, Kuo C *et al.* Molecular tumor markers in the blood: early prediction of disease outcome in melanoma patients treated with a melanoma vaccine. *J Clin Oncol* 2003; **21** (13):2558-2563.
- Webb A. Statistical Pattern Recognition. 2nd ed. Chichester, England: John Wiley & Sons Ltd; 2002.
- Weeraratna AT. A Wnt-er wonderland--the complexity of Wnt signaling in melanoma. *Cancer Metastasis Rev* 2005; **24** (2):237-250.
- Wei G, Lonardo F, Ueda T *et al.* CDK4 gene amplification in osteosarcoma: reciprocal relationship with INK4A gene alterations and mapping of 12q13 amplicons. *Int J Cancer* 1999; **80** (2):199-204.

- Weigelt B, Peterse JL, van 't Veer LJ. Breast cancer metastasis: markers and models. *Nat Rev Cancer* 2005; **5** (8):591-602.
- Weinberg RA. Tumor suppressor genes. *Science* 1991; **254** (5035):1138-1146.
- Weinberg RA. Oncogenes and tumor suppressor genes. *CA Cancer J Clin* 1994; **44** (3):160-170.
- Weinberg RA. The retinoblastoma protein and cell cycle control. *Cell* 1995; **81** (3):323-330.
- Werb Z. ECM and cell surface proteolysis: regulating cellular ecology. *Cell* 1997; **91** (4):439-442.
- White CN, Chan DW, Zhang Z. Bioinformatics strategies for proteomic profiling. *Clin Biochem* 2004; **37** (7):636-641.
- Widschwendter M, Jones PA. DNA methylation and breast carcinogenesis. *Oncogene* 2002; **21** (35):5462-5482.
- Wimmer I, Meyer JC, Seifert B *et al*. Prognostic value of serum 5-S-cysteinyl-dopa for monitoring human metastatic melanoma during immunochemotherapy. *Cancer Res* 1997; **57** (22):5073-5076.
- Winer EP, Hudis C, Burstein HJ *et al*. American Society of Clinical Oncology technology assessment on the use of aromatase inhibitors as adjuvant therapy for postmenopausal women with hormone receptor-positive breast cancer: status report 2004. *J Clin Oncol* 2005; **23** (3):619-629.
- Wistuba, II, Behrens C, Milchgrub S *et al*. Comparison of features of human breast cancer cell lines and their corresponding tumors. *Clin Cancer Res* 1998; **4** (12):2931-2938.
- Wright Jr GW, Cazares LH, Leung SM *et al*. Proteinchip(R) surface enhanced laser desorption/ionization (SELDI) mass spectrometry: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures. *Prostate Cancer Prostatic Dis* 1999; **2** (5/6):264-276.
- Wu B, Abbott T, Fishman D *et al*. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 2003a; **19** (13):1636-1643.
- Wu H, Goel V, Haluska FG. PTEN signaling pathways in melanoma. *Oncogene* 2003b; **22** (20):3113-3122.

- Wu X, Bayle JH, Olson D *et al.* The p53-mdm-2 autoregulatory feedback loop. *Genes Dev* 1993; **7** (7A):1126-1132.
- Wu X, Senechal K, Neshat MS *et al.* The PTEN/MMAC1 tumor suppressor phosphatase functions as a negative regulator of the phosphoinositide 3-kinase/Akt pathway. *Proc Natl Acad Sci U S A* 1998; **95** (26):15587-15591.
- Wulfkuhle JD, McLean KC, Paweletz CP *et al.* New approaches to proteomic analysis of breast cancer. *Proteomics* 2001; **1** (10):1205-1215.
- Wulfkuhle JD, Sgroi DC, Krutzsch H *et al.* Proteomics of human breast ductal carcinoma in situ. *Cancer Res* 2002; **62** (22):6740-6749.
- Wyllie AH, Kerr JF, Currie AR. Cell death: the significance of apoptosis. *Int Rev Cytol* 1980; **68**:251-306.
- Wysocki VH, Resing KA, Zhang Q *et al.* Mass spectrometry of peptides and proteins. *Methods* 2005; **35** (3):211-222.
- Xiao Z, Prieto D, Conrads TP *et al.* Proteomic patterns: their potential for disease diagnosis. *Mol Cell Endocrinol* 2005; **230** (1-2):95-106.
- Yamada S, Pokutta S, Drees F *et al.* Deconstructing the cadherin-catenin-actin complex. *Cell* 2005; **123** (5):889-901.
- Yamashita H, Nishio M, Toyama T *et al.* Coexistence of HER2 over-expression and p53 protein accumulation is a strong prognostic molecular marker in breast cancer. *Breast Cancer Res* 2004; **6** (1):R24-30.
- Yanagisawa K, Shyr Y, Xu BJ *et al.* Proteomic patterns of tumour subsets in non-small-cell lung cancer. *Lancet* 2003; **362** (9382):433-439.
- Yano S, Herbst RS, Shinohara H *et al.* Treatment for malignant pleural effusion of human lung adenocarcinoma by inhibition of vascular endothelial growth factor receptor tyrosine kinase phosphorylation. *Clin Cancer Res* 2000; **6** (3):957-965.
- Yasui Y, Pepe M, Thompson ML *et al.* A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* 2003; **4** (3):449-463.
- Yates JR, 3rd. Mass spectrometry. From genomics to proteomics. *Trends Genet* 2000; **16** (1):5-8.

- Yazdi AS, Palmedo G, Flaig MJ *et al.* Mutations of the BRAF gene in benign and malignant melanocytic lesions. *J Invest Dermatol* 2003; **121** (5):1160-1162.
- Ye B, Cramer DW, Skates SJ *et al.* Haptoglobin-alpha subunit as potential serum biomarker in ovarian cancer: identification and characterization using proteomic profiling and mass spectrometry. *Clin Cancer Res* 2003; **9** (8):2904-2911.
- Yoshida K, Miki Y. Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. *Cancer Sci* 2004; **95** (11):866-871.
- Zhang DH, Salto-Tellez M, Chiu LL *et al.* Tissue microarray study for classification of breast tumors. *Life Sci* 2003; **73** (25):3189-3199.
- Zhang HS, Postigo AA, Dean DC. Active transcriptional repression by the Rb-E2F complex mediates G1 arrest triggered by p16INK4a, TGFbeta, and contact inhibition. *Cell* 1999; **97** (1):53-61.
- Zhang M, Rosen JM. Stem cells in the etiology and treatment of cancer. *Curr Opin Genet Dev* 2006a; **16** (1):60-64.
- Zhang X, Leung SM, Morris CR *et al.* Evaluation of a novel, integrated approach using functionalized magnetic beads, bench-top MALDI-TOF-MS with prestructured sample supports, and pattern recognition software for profiling potential biomarkers in human plasma. *J Biomol Tech* 2004a; **15** (3):167-175.
- Zhang Y, Wroblewski M, Hertz MI *et al.* Analysis of chronic lung transplant rejection by MALDI-TOF profiles of bronchoalveolar lavage fluid. *Proteomics* 2006b; **6** (3):1001-1010.
- Zhang Z, Bast RC, Jr., Yu Y *et al.* Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. *Cancer Res* 2004b; **64** (16):5882-5890.
- Zheng Y, Xu Y, Ye B *et al.* Prostate carcinoma tissue proteomics for biomarker discovery. *Cancer* 2003; **98** (12):2576-2582.
- Zhou G, Li H, DeCamp D *et al.* 2D differential in-gel electrophoresis for the identification of esophageal scans cell cancer-specific protein markers. *Mol Cell Proteomics* 2002; **1** (2):117-124.
- Zhou M, Lucas DA, Chan KC *et al.* An investigation into the human serum "interactome". *Electrophoresis* 2004; **25** (9):1289-1298.

Zhu Y, Williams S, Zwiggelaar R. Computer technology in detection and staging of prostate carcinoma: a review. *Med Image Anal* 2006; **10** (2):178-199.

Zhukov TA, Johanson RA, Cantor AB *et al.* Discovery of distinct protein profiles specific for lung tumors and pre-malignant lung lesions by SELDI mass spectrometry. *Lung Cancer* 2003; **40** (3):267-279.

Ziegler A, Jonason AS, Leffell DJ *et al.* Sunburn and p53 in the onset of skin cancer. *Nature* 1994; **372** (6508):773-776.

Communications resulting from study

Publications:

Serum proteomic fingerprinting discriminates between clinical stages and predicts disease progression in melanoma patients. Shahid Mian, Selma Ugurel, **Erika Parkinson**, Iris Schlenzka, Ian Dryden, Lee Lancashire, Graham Ball, Colin Creaser, Robert Rees and Dirk Schadendorf. *J Clin Oncol.* 2005 Aug 1;23(22):5088-93

Abstracts:

The interplay of proteomics, genomics and bioinformatics approaches and their potential for cancer diagnosis and prognosis. **Erika Parkinson**, Balwir Matharoo-Ball, Graham Ball, Colin S. Creaser and Robert Rees. American Association of Cancer Research (AACR) conference 2006

Application of proteomics, genomics and bioinformatics to identify predictive markers in melanoma. East Midlands Proteomics Workshop 2005, Nottingham, UK. Delivered 20 minute short talk.

Application of proteomics and bioinformatics to identify discriminatory patterns in melanoma. British Mass Spectrometry Society (BMSS) Conference 2005, York, UK. Delivered a 20 minute short talk

Application of proteomics and bioinformatics to identify discriminatory patterns between preparation of tissue samples using modified protocols. House of Commons SET for Europe 2005 as part of National Science Week. Poster

Application of proteomics and bioinformatics to identify discriminatory patterns between preparation of tissue samples using modified protocols. East Midlands Proteomics Workshop 2004, Nottingham, UK. Delivered 20 minute short talk.

Appendix I

Melanoma cell line sample names and genetic information showing the type of mutation found in the *Braf*, *tp53* and *p16^{INK4A}* gene in each cell line used in the study.

<i>ESTDAB-Code</i>	<i>Cell line</i>	<i>Braf</i>	<i>tp53</i>	<i>p16^{INK4A}</i>
ESTDAB-001	COLO 794	V599E (GTG->GAG)	C135R (TGC->CGC)	HD (ex 1a, 2 & 3)
ESTDAB-002	COLO 800	V599E (GTG->GAG)	C135R (TGC->CGC)	HD (ex 1a, 2 & 3)
ESTDAB-003	COLO 818	V599E (GTG->GAG)	C135R (TGC->CGC)	HD (ex 1a, 2 & 3)
ESTDAB-004	COLO 857	V599E (GTG->GAG)		Del Nuc. 203&204 -> FS codon 68
ESTDAB-005	FM 2	V599E (GTG->GAG)		HD (ex 2 & 3)
ESTDAB-006	FM 28			HD (ex 1a, 1b, 2 & 3)
ESTDAB-007	FM 3			HD (ex 1a, 1b, 2 & 3)
ESTDAB-011	FM 45	V599E (GTG->GAG)		HD (ex 1a, 1b, 2 & 3)
ESTDAB-013	FM 55M2	V599E (GTG->GAG)	Q317X (CAG->TAG)	
ESTDAB-014	FM 55P	V599E (GTG->GAG)	Q317X (CAG->TAG)	
ESTDAB-017	FM 57	V599K (GTG->AAG)		
ESTDAB-018	FM 58	V599E (GTG->GAG)		HD (ex 1a, 1b, 2 & 3)
ESTDAB-019	FM 6			Methylation
ESTDAB-021	FM 72	V599E (GTG->GAG)		HD (ex 1a, 1b, 2 & 3)
ESTDAB-023	FM 78	V599E (GTG->GAG)		HD (ex 1a, 1b, 2 & 3)
ESTDAB-024	FM 79		V272G (GTG->GGG)	
ESTDAB-026	FM 81	V599E (GTG->GAA)		HD (ex 1a, 2 & 3)
ESTDAB-027	FM 82	V599E (GTG->GAG)		HD (ex 2 & 3)
ESTDAB-029	FM 88	V599K (GTG->AAG)		HD (ex 1a, 2 & 3)
ESTDAB-030	FM 9	V599E (GTG->GAG)		HD (ex 1a, 1b, 2 & 3)
ESTDAB-032	FM 92	V599E (GTG->GAG)		
ESTDAB-033	FM 93/2	V599E (GTG->GAG)	L265L (CTG->CTA)& G266R (GGA->AGA)	HD (ex 1a, 1b, 2 & 3)
ESTDAB-034		L596S (CTA->TCA)	R181C (CGC->TGC)	Methylation
ESTDAB-036	GR-M		K152Q (AAG->CAG)	HD (ex 1a, 2 & 3)
ESTDAB-038	IRNE	V599E (GTG->GAG)		HD (ex 1a, 1b, 2 & 3)

ESTDAB-Code	Cell line	Braf	tp53	p16^{INK4A}
ESTDAB-040	M18	V599E (GTG->GAG)	Q317X (CAG-->TAG)	R80X (CGA->TGA)
ESTDAB-041	M18/B7	V599E (GTG->GAG)		
ESTDAB-042	M21	V599E (GTG->GAG)	G262V (GGT->GTT)	
ESTDAB-043	M24	V599E (GTG->GAG)		
ESTDAB-044	M6	V599E (GTG->GAG)	Q317X (CAG-->TAG)	
ESTDAB-046	Mel 1102			HD (ex 1a, 1b, 2 & 3)
ESTDAB-047	Mel 1359			HD (ex 1a, 2 & 3)
ESTDAB-050	MEWO	V599E (GTG->GAG)	Q317X (CAG-->TAG)	R80X (CGA->TGA)&T79T (ACC->ACT)
ESTDAB-052	MZ Mel 13	V599E (GTG->GAG)		HD (ex 1a, 1b, 2 & 3)
ESTDAB-056	MZ Mel 5	V599E (GTG->GAG)	IVS7ntlg->c	HD (ex 1a, 2 & 3)
ESTDAB-057	NW 145 HLA-A2	V599E (GTG->GAG)		HD (ex 1a, 1b, 2 & 3)
ESTDAB-058	NW 145 ClonE7	V599E (GTG->GAG)		HD (ex 1a, 1b, 2 & 3)
ESTDAB-065	UKRV Mel 7			
ESTDAB-066	WM 115	V599D (GTG->GAT)	R181C (CGC --> TGC) Exon 5.II	HD (ex 1a & 1b)
ESTDAB-067	WM 1205 LU	V599E (GTG->GAG)		HD (ex 1a, 1b, 2 & 3)
ESTDAB-071	WM 139	V599E (GTG->GAG)		
ESTDAB-074	WM 1791A			
ESTDAB-075	WM 1862	V599E (GTG->GAG)		HD (ex 1a, 1b, 2 & 3)
ESTDAB-076	WM 266-4	V599K (GTG->AAG)		HD (ex 1a, 1b, 2 & 3)
ESTDAB-077	WM 278	V599E (GTG->GAG)		HD (1a & 1b)
ESTDAB-079	WM 3211		C242G (TGC->GGC)	HD (ex 1a, 2 & 3)
ESTDAB-080	WM 39	V599E (GTG->GAG)		
ESTDAB-081	WM 451 LU	V599E (GTG->GAG)	Y220C (TAT->TGT)	
ESTDAB-082	WM 75	V599E (GTG->GAG)		
ESTDAB-083	WM 793	V599E (GTG->GAG)		
ESTDAB-084	WM 852	V599E (GTG->GAG)	S241F (TCC->TTC)	HD (ex 1a, 1b, 2 & 3)
ESTDAB-085	WM 902B	V599E (GTG->GAG)		
ESTDAB-095	Mel-WL	V599E (GTG->GAG)		
ESTDAB-097	Mel SOE			
ESTDAB-102	Ma-mel-28			
ESTDAB-104	Ma-mel-30	D593N (GAT->AAT)		HD (ex 1a, 1b, 2 & 3)
ESTDAB-107	UKRV Mel 15a		R196X(CGA->TGA)	
ESTDAB-137	Ma-mel-7	V599E (GTG->GAG)		HD (ex 1a, 1b, 2 & 3)

<i>ESTDAB-Code</i>	<i>Cell line</i>	<i>Braf</i>	<i>tp53</i>	<i>p16^{INK4a}</i>	
ESTDAB-138	Ma-Mel-15				
ESTDAB-143	Ma-mel-22		C238F (TGT->TTT)		
ESTDAB-151	Ma-mel-10				
ESTDAB-152	Ma-mel-19	V599E (GTG->GAG)			Methylation
ESTDAB-153	Ma-Mel-20	V599E (GTG->GAG)			HD (ex 1a, 1b, 2 & 3)
ESTDAB-155	Ma-mel-23	V599E (GTG->GAG)			HD (ex 1a, 1b, 2 & 3)
ESTDAB-168	Ma-mel-27				Methylation
ESTDAB-172	Ma-mel-39a				HD (ex 2 & 3)
ESTDAB-173	Ma-mel-45a	V599E (GTG->GAG)			HD (ex 1a, 1b, 2 & 3)
ESTDAB-190	Ma-Mel-8b		R283C (CGC->TGC)		HD (ex 1a, 1b, 2 & 3)
ESTDAB-194	Ma-Mel-47	V599E (GTG->GAG)			
ESTDAB-195	Ma-Mel-48a	V599E (GTG->GAG)			Methylation
ESTDAB-196	Ma-Mel-51	G469R (GGA->AGA)			
ESTDAB-199	Ma-Mel-55	V599E (GTG->GAG)			Methylation
ESTDAB-200	Ma-Mel-56	V599E (GTG->GAG)			HD (ex 1a, 1b, 2 & 3)
ESTDAB-205	Ma-Mel-71				Methylation
ESTDAB-206	Ma-Mel-73a				HD (ex 1a, 1b, 2 & 3)
ESTDAB-207	Ma-Mel-73b				Methylation
	FM74				
156	Ma-Mel-26a		A278V (GCT->GTT)		ND
	Ma-Mel-26b		A278V (GCT->GTT)		ND
149	Ma-Mel-13	V599E (GTG->GAG)	R213X (CGA->TGA)		ND
	Ma-Mel-05				ND
	Ma-Mel-24				ND
150	Ma-Mel-06		N.D.		ND
	Ma-Mel-60				ND
	Ma-Mel-59a				ND
174	Ma-Mel-54a	V599E (GTG->GAG)			ND
	Ma-Mel-57	V599E (GTG->GAG)			ND
ESTDAB 108	Ma-Mel-11	V599E (GTG->GAG)			ND
	Ma-Mel-52	V599K (GTG->AAG)			ND
ESTDAB 144	Ma-Mel-36	V599E (GTG->GAG)			ND
ESTDAB-139	Ma-Mel-33	V599E (GTG->GAG)			ND
			N247K(AAC->AAA) & R248W(CGG->TGG)		ND

<i>ESTDAB-Code</i>	<i>Cell line</i>	<i>Braf</i>	<i>tp53</i>	<i>p16^{INK4A}</i>
Ma-Mel-66a		V599E (GTG->GAG)		ND
Ma-Mel-66b		V599E (GTG->GAG)		ND
Ma-Mel-68		N.D.	N.D.	ND
Ma-Mel-74			R273H (CGT->CAT)	ND
Ma-Mel-75				ND

Key

Ex - exon

HD - homozygous deletion

ND - no data

Example: V599E - substitution of valine for glutamine at amino acid position 599.

Appendix II

Clinical information used for the cell line vs tissue analysis, discussed in chapter 4.

Tissue number	Patient Number	Cell Line Code	Tissue origin	Survival since tissue extraction
97	MA 000332	Ma-Mel-05	C/SQ	1.97
147	MA 000334	Ma-Mel-06	LK	0.33
9	MA 000341	Ma-Mel-07	LK	11.84
379	-	Ma-Mel-08b	C/SQ	1.38
143	MA 000340	Ma-Mel-10		
168	MA 000346	Ma-Mel-11		
95	MA 000339	Ma-Mel-12	C/SQ	5.62
159	MA 000142	Ma-Mel-13	C/SQ	37.72
169	-	Ma-Mel-15	LK	1.02
208	MA 000345	Ma-Mel-19	C/SQ	0.23
171	MA 000330	Ma-Mel-20	C/SQ	1.25
180	MA 000335	Ma-Mel-21		
192	MA 000338	Ma-Mel-22	kidney	42.75
149	MA 000331	Ma-Mel-23		
185	MA 000333	Ma-Mel-24		11.18
285	MA 000263	Ma-Mel-26a	LK	6.30
293	MA 000263	Ma-Mel-26b	LK	3.61
252	MA 000347	Ma-Mel-27	C/SQ	7.56
284	MA 000336	Ma-Mel-28	C/SQ	4.59
287	-	Ma-Mel-30	C/SQ	13.66
288	-	Ma-Mel-33	C/SQ	11.21
304	MA 000190	Ma-Mel-36	C/SQ	4.03
313	-	Ma-Mel-37b	LK	51.00
319	MA 000350	Ma-Mel-39a		12.64
383	MA 000327	Ma-Mel-45a	brain	4.37
384	-	Ma-Mel-45b	brain	4.37
374	MA 000271	Ma-Mel-47		6.25
414	MA 000256	Ma-Mel-48a	C/SQ	13.84
449	MA 000337	Ma-Mel-51	LK	2.07
395	MA 000348	Ma-Mel-52	C/SQ	0.53
380	MA 000298	Ma-Mel-53	C/SQ	33.74
453	MA 000343	Ma-Mel-54a	LK	3.78
466	MA 000351	Ma-Mel-55	LK	8.09
470	MA 000344	Ma-Mel-56	LK	0.66
492	MA 000305	Ma-Mel-57	LK	3.88
476	MA 000258	Ma-Mel-59a	LK	2.66
510	MA 000342	Ma-Mel-60	C/SQ	7.40
598	-	Ma-Mel-61c	LK	13.76
525	-	Ma-Mel-62	LK	4.21
548	-	Ma-Mel-65	LK	17.93
546	-	Ma-Mel-66a	C/SQ	0.79
547	-	Ma-Mel-66b	C/SQ	0.79
550	-	Ma-Mel-67	LK	17.74

Tissue number	Patient Number	Cell Line Code	Tissue origin	Survival since tissue extraction
542	-	Ma-Mel-68	LK	6.05
561	-	Ma-Mel-71	C/SQ	4.04
566	-	Ma-Mel-73a	LK	7.57
567	-	Ma-Mel-73b	LK	7.57
568	-	Ma-Mel-74	LK	14.45
587	-	Ma-Mel-75	LK	12.00
601	-	Ma-Mel-76	LK	5.67
637	-	Ma-Mel-79	C/SQ	13.57
619	-	Ma-Mel-80b	LK	9.31
620	-	Ma-Mel-81	C/SQ	2.13
657	-	Ma-Mel-82	LK	9.73
617	-	Ma-Mel-83	LK	7.08
658	-	Ma-Mel-85	C/SQ	3.31
656	-	Ma-Mel-90	C/SQ	4.49
628	-	Ma-Mel-91	C/SQ	10.20
652	-	Ma-Mel-93	C/SQ	9.28
696	-	Ma-Mel-94	LK	3.97
692	-	Ma-Mel-96	C/SQ	6.23
715	-	Ma-Mel-97	LK	4.93

Key:

LK – lymph node

C/SQ – squamous cell (skin)

Survival is in months

Appendix III

Clinical information of breast cancer patients with at least a five year follow-up and TMA data relevant to the data presented in chapter 5.

sample num	p53 exp	basalphenotype	surviv5	dead5	dfi5	recurr5
2072			124	0	124	0
2074		0	132	0	132	0
2075	0	1	68	1	10	1
2076			135	0	135	0
2077	10	0	137	0	137	0
2078	20	0	124	0	124	0
2079			85	1	22	1
2080	0	1	122	0	122	0
2081			132	0	72	1
2082	0	0	27	1	24	1
2083	255	0	133	0	133	0
2084	155	1	130	0	130	0
2086	300	1	128	0	3	1
2087	0	0	140	0	11	1
2091			16	1	14	1
2093			45	1	45	0
2094	0	1	157	0	157	0
2095	0	0	154	0	30	1
2096	0	0	129	0	129	0
2097			76	1	53	1
2098			132	0	132	0
2099	160	0	26	1	26	0
2100	0	0	28	2	28	0
2101	210	0	78	1	78	0
2104	0	1	136	0	70	1
2105	0	0	63	1	60	1
2107	100	0	132	0	132	0
2108	0	0	98	1	98	0
2109		0	133	0	133	0
2111	40	1	135	0	67	1
2113	0	1	97	2	97	0
2114	0	0	128	0	128	0
2118	10	0	130	0	8	1
2120			133	0	133	0
2123	0	0	130	0	130	0
2124	0	0	50	2	36	1
2125	215	0	69	1	48	1
2127		0	129	0	129	0
2131	0	0	119	0	15	1
2133	0	0	127	0	127	0
2134	0	0	125	0	125	0
2136			133	0	133	0
2138	105	0	35	1	18	1
2145	275	1	17	1	11	1

sample num	p53 exp	basalphenotype	surviv5	dead5	dfi5	recurr5
2146			133	0	57	1
2149	0	0	85	1	18	1
2152	260	0	28	2	28	0
2153	0	1	130	0	130	0
2154	0	1	32	1	32	0
2155	215	0	134	0	134	0
2156	0	0	131	0	131	0
2157	15	1	95	1	95	0
2158	0	0	124	0	124	0
2159	0	0	123	0	123	0
2160	0	0	41	1	13	1
2162	0	1	132	0	132	0
2163	0	0	132	0	132	0
2166			132	0	132	0
2167	0	0	87	1	70	1
2171	0	0	25	1	21	1
2172	0	0	132	0	130	1
2174	0	0	125	0	125	0
2176	155	0	123	0	123	0
2177	0	0	120	0	120	0
2179	0	0	46	1	22	1
2180	40	0	131	0	131	0
2181	0	0	64	1	64	0
2183	0	1	41	1	11	1
2184	0	0	121	0	121	0
2185	0	1	128	0	25	1
2186	0	1	129	0	129	0
2187		0	131	0	33	1
2191	0	0	121	0	121	0
2193	0	0	128	0	128	0
2195	0	0	131	0	121	1
2196	300	1	120	0	120	0
2197	0	0	131	0	131	0
2198	0	0	123	0	123	0
2199	0	0	127	0	127	0
2200	0	1	77	1	36	1
2201	0	1	127	0	127	0
2202	225	1	120	0	120	0
2203	0	0	119	0	119	0
2204	30	1	126	0	126	0
2206	100	0	106	1	96	1
2207	0	0	113	0	113	0
2209	150	0	131	0	131	0
2210	40	0	86	1	74	1
2211	70	0	28	1	6	1
2214	15	1	130	0	74	1
2217	10	0	83	2	83	0
2218	0	1	28	1	25	1
2219	140	1	21	1	21	0

sample num	p53 exp	basalphenotype	surviv5	dead5	dfi5	recurr5
2220	0	0	126	0	126	0
2221	205	1	92	2	92	0
2222	95	0	130	0	130	0
2223	0	1	125	0	125	0
2225	0	0	127	0	127	0
2227	250	1	17	1	7	1
2228	0	0	124	0	124	0
2229	0	1	122	0	122	0
2230	0	0	72	1	72	0
2231	200	1	130	0	130	0
2233	0	1	110	2	110	0
2236	0	1	128	0	128	0
2237	0	1	122	0	122	0
2239	285	1	121	0	121	0
2241	0	1	60	1	52	1
2242	0	0	121	0	121	0
2246	155	1	117	0	117	0
2250	90	1	115	0	115	0
2251	150	1	127	0	127	0
2252	0	1	128	0	79	1
2254	0	0	126	0	126	0
2264	0	1	73	2	73	0
2265	70	1	83	1	14	1
2268	0	1	126	0	126	0
2270		0	123	0	123	0
2280	185	1	42	1	23	1
2281	155	0	116	0	116	0
2282	0	0	124	0	124	0
2286	0	0	116	0	116	0
2287			116	0	116	0
2289	0	1	123	0	123	0
2290	170	0	117	0	117	0
2293	0	0	17	1	9	1
2294	0	1	37	2	23	1
2297	0	0	112	0	112	0
2302	0	1	124	0	124	0
2304	0	0	122	0	122	0
2305	0	0	90	2	63	1
2308	0	0	118	0	118	0
2309	250	1	114	0	114	0
2310	100	0	7	1	7	1
2311	0	1	120	0	120	0
2312	0	0	116	0	116	0
2313	0	0	122	0	122	0
2317	0	0	110	0	110	0
2323	0	0	119	0	119	0
2324	0	0	126	0	126	0
2325	95	0	111	0	111	0
2327	10	1	107	0	107	0

sample num	p53 exp	basalphenotype	surviv5	dead5	dfi5	recurr5
2333			119	0	119	0
2334	0	0	94	1	94	0
2337	0	1	96	1	73	1
2339	250	1	121	0	121	0
2341	0	0	117	0	117	0
2342	0	0	107	0	107	0
2345	0	0	111	0	111	0
2347	0	0	118	0	118	0
2348	0	1	112	2	112	0
2349			89	0	89	0
2350	0	0	108	0	108	0
2351	0	0	123	0	67	1
2353	0	0	117	0	117	0

Key:

P53 expression is measured semi-quantitatively where both the assessment of the intensity of p53 staining as well as the number of positive cells gave scores in the range of 0-300.

Basal phenotype – 1 = positive score as determined by detection of 10% or more of invasive malignant cells positive for CK5/6, CK14, SMA and p53 staining. 0 = negative score.

Surviv5 – The time, in months, from the date of the primary surgery to the time of breast cancer related death or last follow-up.

Dead5 – 0 = patient alive at last follow-up, 1 = breast cancer related death, 2 = non-breast cancer related death

Dfi5 – The interval (in months) from the date of the primary surgery to the first loco-regional recurrence or distant metastasis

Recurr5 – Follow-up of patients where those whose disease recurred were given a value of 1. No recurrence of disease = 0.

Appendix IV

Melanoma serum samples analysed and discussed in chapter 6.

Sample Number	Melanoma Stage	Patient Initials	Gender
1	AJCC I	RD	m
2	AJCC I	KD	f
3	AJCC I	WB	m
4	AJCC I	AB	m
5	AJCC I	CD	f
6	AJCC I	BB	f
7	AJCC I	EB	f
8	AJCC I	UB	f
9	AJCC I	HB	f
10	AJCC I	RB	m
11	AJCC I	MB	f
12	AJCC I	KB	f
13	AJCC I	HA	m
14	AJCC I	MA	m
15	AJCC I	-A	m
16	AJCC I	EA	f
17	AJCC I	NA	f
18	AJCC I	RA	f
19	AJCC I	PB	m
20	AJCC I	SB	f
21	AJCC I	IB	f
22	AJCC I	WB	m
23	AJCC I	SD	f
24	AJCC I	HD	m
25	AJCC I	ED	f
26	AJCC I	BB	f
27	AJCC I	HE	m
28	AJCC I	ME	m
29	AJCC I	HF	f
30	AJCC I	RF	f
31	AJCC I	HF	f
32	AJCC I	WB	m
33	AJCC I	GB	m
34	AJCC I	GB	f
35	AJCC I	MB	f
36	AJCC I	MF	f
37	AJCC I	NF	m
38	AJCC I	HF	m
39	AJCC I	MF	f
40	AJCC I	VF	f
41	AJCC I	HF	m
42	AJCC I	H-JF	m
43	AJCC I	IG	f
44	AJCC I	BG	f
45	AJCC I	KG	f
46	AJCC I	RG	m
47	AJCC I	AG	f
48	AJCC I	TG	m
49	AJCC I	LG	f
50	AJCC I	-G	m
51	AJCC I	FG	m
52	AJCC I	GH	m

Sample Number	Melanoma Stage	Patient Initials	Gender
53	AJCC I	GH	f
54	AJCC I	DH	m
55	AJCC I	RH	m
56	AJCC I	HH	f
57	AJCC I	BH	m
58	AJCC I	-H	f
59	AJCC I	HH	m
60	AJCC I	AH	m
61	AJCC I	RJ	m
62	AJCC I	RJ	f
63	AJCC I	HJ	m
64	AJCC I	IJ	f
65	AJCC I	NJ	m
66	AJCC I	FJ	m
67	AJCC I	HK	m
68	AJCC I	UK	f
69	AJCC I	HK	f
70	AJCC I	RK	f
71	AJCC I	HK	m
72	AJCC I	KK	m
73	AJCC I	HK	m
74	AJCC I	-N-K	m
75	AJCC I	WK	m
76	AJCC I	HK	f
77	AJCC I	CK	f
78	AJCC I	AK	f
79	AJCC I	RM	f
80	AJCC I	KM	m
81	AJCC I	KM	f
82	AJCC I	EM	m
83	AJCC I	HL	m
84	AJCC I	HL	f
85	AJCC I	HL	f
86	AJCC I	ML	f
87	AJCC I	EM	f
88	AJCC I	WM	m
89	AJCC I	EM	f
90	AJCC I	CM	f
91	AJCC I	FM	m
92	AJCC I	NM	f
93	AJCC I	AN	f
94	AJCC I	LN	m
95	AJCC I	AN	f
96	AJCC I	-O	f
97	AJCC I	IP	f
98	AJCC I	RP	m
99	AJCC I	GR	m
100	AJCC I	KR	f
101	AJCC I	K-HR	m
102	AJCC IV	-A	m
103	AJCC IV	LA	m
104	AJCC IV	HA	m
105	AJCC IV	HA	m
106	AJCC IV	EB	m
107	AJCC IV	KvB	f
108	AJCC IV	-R-B	f
109	AJCC IV	HB	m

Sample Number	Melanoma Stage	Patient Initials	Gender
110	AJCC IV	KB	f
111	AJCC IV	DB	f
112	AJCC IV	HB	f
113	AJCC IV	HB	f
114	AJCC IV	-U-B	f
115	AJCC IV	HB	m
116	AJCC IV	AC	m
117	AJCC IV	GD	f
118	AJCC IV	FD	m
119	AJCC IV	BD	m
120	AJCC IV	EE	f
121	AJCC IV	SE	f
122	AJCC IV	DE	f
123	AJCC IV	KE	m
124	AJCC IV	KE	f
125	AJCC IV	WF	m
126	AJCC IV	AF	m
127	AJCC IV	JF	m
128	AJCC IV	GF	m
129	AJCC IV	GF	f
130	AJCC IV	IG	f
131	AJCC IV	NG	m
132	AJCC IV	IG	f
133	AJCC IV	JG	m
134	AJCC IV	RG	f
135	AJCC IV	PG	m
136	AJCC IV	-G	f
137	AJCC IV	HH	m
138	AJCC IV	KH	m
139	AJCC IV	KH	m
140	AJCC IV	EH	f
141	AJCC IV	GH	m
142	AJCC IV	HH	m
143	AJCC IV	AGK	f
144	AJCC IV	EK	f
145	AJCC IV	DK	f
146	AJCC IV	RK	f
147	AJCC IV	FK	m
148	AJCC IV	KK	m
149	AJCC IV	HL	m
150	AJCC IV	KL	f
151	AJCC IV	AM	f
152	AJCC IV	RM	m
153	AJCC IV	SM	f
154	AJCC IV	SM	f
155	AJCC IV	BM	m
156	AJCC IV	IM	f
157	AJCC IV	AM	f
158	AJCC IV	SM	m
159	AJCC IV	GN	f
160	AJCC IV	WO	m
161	AJCC IV	HP	m
162	AJCC IV	TR	m
163	AJCC IV	SR	m
164	AJCC IV	KR	f
165	AJCC IV	GR	f
166	AJCC IV	NR	f

Sample Number	Melanoma Stage	Patient Initials	Gender
167	AJCC IV	DS	m
168	AJCC IV	SS	m
169	AJCC IV	RP	m
170	AJCC IV	WS	m
171	AJCC IV	BS	m
172	AJCC IV	MS	f
173	AJCC IV	ES	f
174	AJCC IV	IS	f
175	AJCC IV	IS	f
176	AJCC IV	GS-F	f
177	AJCC IV	HS	m
178	AJCC IV	AS	f
179	AJCC IV	ES	f
180	AJCC IV	AS	f
181	AJCC IV	VS	m
182	AJCC IV	WS	m
183	AJCC IV	HS	m
184	AJCC IV	KS	m
185	AJCC IV	KS	m
186	AJCC IV	ES	m
187	AJCC IV	FT	m
188	AJCC IV	HV	m
189	AJCC IV	MW	f
190	AJCC IV	KB	f
191	AJCC IV	RW	m
192	AJCC IV	HW	m
193	AJCC IV	WW	m
194	AJCC IV	WC	m
195	AJCC IV	HH	f
196	AJCC IV	MB	m
197	AJCC IV	GF	f
198	AJCC IV	JU	m
199	AJCC IV	EG	m
200	AJCC IV	-HT	f
201	AJCC IV	GR	f
202	AJCC IV	AW	f
203	AJCC IV	WF	f
204	AJCC IV	HB	f
205	AJCC IV	BM	f

Appendix V

Chip No	Insulin B chain (3494.65 Da)		Apomyoglobin (+2H) (8476.15 Da)		Cytochrome C (12362 Da)		Apomyoglobin (16952.3 Da)		Trypsinogen (23981.1 Da)	
	mass (Da)	intensity	mass (Da)	intensity	mass (Da)	intensity	mass (Da)	intensity	mass (Da)	intensity
1	3494.4	3.8	8479.6	1.5	12363.5	8.6	16960.5	6.9	23987	17.2
2	3495.8	4	8476.4	2.1	12360.2	7.4	16957.6	8.6	23989.2	15.2
3	3499.3	3.9	8475.3	1.8	12358.2	6.2	16955.4	6.7	23985.7	15.8
4	3493.3	8.3	8473.9	5	12352.7	15.6	16943.2	16.7	23960.4	34.7
5	3494.8	5.9	8479.9	6.2	12360.9	12.5	16949	19.7	23961.6	40
6	3496.1	7.7	8485.1	21.4	12372	16.8	16976.4	23.5	24004.4	44.1
7	3496.2	7.6	8484	19.7	12366	16.4	16966.9	26.7	23987.5	46.3
8			8484.1	26	12371.7	14.8	16977.7	24.2	24006.5	39.2
9			8477.3	21.3	12368.6	16.7	16987.6	25	24034.2	41.5
10	3497.5	9.4	8473.2	20.7	12360.4	19.2	16974.2	27.4	24025.3	46.3
11	3497.9	7.4	8477.2	20.9	12369.4	17.9	16988.9	27.2	24038.4	43.2
12	3499.4	7.5	8476.9	25.7	12369.2	18.6	16987.7	28.5	24045.2	39.3
mean	3496.47	6.55	8478.58	14.36	12364.40	14.23	16968.76	20.09	24002.12	35.23
st dev	2.04	2.02	4.02	10.00	6.04	4.50	15.68	8.36	28.64	12.01
C of V (%)	0.06	30.85	0.05	69.68	0.05	31.63	0.09	41.59	0.12	34.08
	0.05% mass accuracy		0.03% mass accuracy		0.02% mass accuracy		0.1% mass accuracy		0.09% mass accuracy	

Table showing the coefficients of variation for the peak location and peak intensity for 5 calibrant peaks contained within a calibrant mix which was spotted onto one spot of every SELDI H50 chip used in the experiment. The mass accuracy of the calibrants is also shown, as determined by the difference between the mean peak location and the known mass of the calibrant