UNIVERSITY OF SOUTHAMPTON

Medicine

Clinical and Experimental Sciences

Factors within skin cancer that contribute to metastasis

By

Andrew George Shapanis

ORCID ID: 0000-0003-4147-6956

Thesis for the degree of Doctor of Philosophy

July 2019

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MEDICINE

Clinical and Experimental Sciences

Thesis for the degree of doctor of philosophy

FACTORS WITHIN SKIN CANCER THAT CONTRIBUTE TO METASTASIS

Andrew George Shapanis

Skin cancer is the most frequent cancer worldwide and accounts for 1 in every 3 cancers diagnosed. Skin cancer comprises of melanoma, arising from the melanocytes of the skin and keratinocyte carcinomas which arise in the keratinocytes and include cutaneous squamous cell carcinoma (cSCC) and basal cell carcinoma. cSCC predominantly affects the older generation and is one of the most common types of cancer capable of metastasising, with 5 year survival rates reported as <30%. Although less common, melanoma can affect all ages and has one of the highest rates of metastasis of any cancer, with 5 year survival rates as low as 23%, depending on whether or not distant metastasis has occurred. There are currently very few prognostic markers capable of predicting metastasis in these skin cancers. Presently in the UK, melanoma is graded according to the American Joint Committee on Cancer Guidelines (AJCC) whereas cSCCs are categorised as high or low risk according to the British Association of Dermatologists' guidelines (BAD). Other staging systems have been proposed but most of them also rely on histological features such as differentiation, diameter, depth, site and/or Clark's level, amongst others.

This study aimed to identify factors within cSCC and melanoma which contribute to metastasis using a mass spectrometry based proteomics approach. A method to extract protein from formalin fixed paraffin embedded (FFPE) samples was developed and optimised. Proteins were extracted from 24 FFPE surgically excised primary cSCC (P-NM) and melanoma (Pmel-NM) tumours which had not metastasised at 5 years post-operatively and from 24 FFPE surgically excised primary metastatic cSCC (P-M) and melanoma (Pmel-M) tumours which had metastasised.

A total of 144 and 31 significantly differentially expressed proteins between metastatic and non-metastatic samples were identified in the cSCC and melanoma groups respectively. KEGG, gene ontology, weighted gene co-expression network analysis (WGCNA) and ingenuity pathway analysis (IPA) highlighted several key pathways likely to be involved in development of metastasis in cSCC and melanoma. Multiple reaction monitoring (MRM) of two proteins, ANXA5 and DDOST, verified the original differences in levels of these proteins in cSCC and also validated these findings in an independent sample cohort. Additionally, MRM analysis and machine learning revealed that the combination of ANXA5 and DDOST levels could correctly predict metastasis better than any guideline in current clinical use, with an AUC of 0.929, sensitivity and specificity of 88.24% and 94.12% respectively. However, MRM was technically challenging in the melanoma group and was not able to verify the original melanoma mass spectroscopy results.

Machine learning and modelling of histological characteristics from cSCC samples was subsequently undertaken to see whether it was possible to improve on current prediction of prognosis with these readily available parameters. Surprisingly, this produced a prediction model with an Area Under the Curve (AUC) of 0.997 and a sensitivity and specificity of 94.1% and 100% respectively. Despite this model not requiring any additional work over and above that which is already currently reported histologically when cSCCs are routinely excised in the UK, it was better than the aforementioned ANXA5 and DDOST model and moreover, than any guideline in clinical use at the present time. Moreover, this model has the potential to be integrated into a clinical setting with relative ease and speed.

This study has identified a number of factors, including key pathways that likely contribute to metastasis in cSCC and melanoma. In addition, the combination of proteomics, machine learning and mathematical modelling has identified key prognostic indicators in cSCC and has demonstrated that this approach may have potential to do likewise in many other cancer types.

Table of Contents

List of Tables

List of Figures

DECLARATION OF AUTHORSHIP

I,..... [Andrew Shapanis]

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Title of thesis - **Factors within skin cancer that contribute to metastasis**.....

.....

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. [Delete as appropriate] None of this work has been published before submission [or] Parts of this work have been published as: [please list references below]:

Signed: .....

Date: .....

Acknowledgements

I would like to begin by thanking my supervisor Professor Eugene Healy. His help and guidance throughout my PhD has been essential to my progress and development. He has been extremely patient in his teachings and has taught me a great deal I will undeniably carry forward in life. His supervision is unmatchable and his dedication to helping his students achieve their full potential is something I am grateful to receive. I would also like to extend my gratitude to my second supervisor, Dr Paul Skipp. He has shown unwavering support in my research and helped me develop the skills necessary to becoming a competent scientist. His support throughout has boosted my confidence and motivated me to succeed.

Throughout my research, friend and colleague, Dr Chester Lai, has supported me in every aspect imaginable. From personal matters to professional queries, he has helped me progress and succeed. In addition to Chester, Dr Erika Parkinson has also played a pivotal role in helping me achieve my goals throughout my PhD. I would like to further thank the people of the Centre for Proteomic Research and Dermatopharmacology unit and the help and friendships they have provided.

Last but by no means least, I would like to thank my family. I would not be where I am today if it wasn't for their love and support. My sister, Natalie, is the best sister a sibling could hope for. She has been there for me when I needed someone most and (like any sibling) is capable of annoying me beyond belief. My step father, Ian, has helped me in many aspects of life and more ways than I am sure I'm aware of. My father (and friend), George, has been a role model any son would be proud to call their dad. Finally, my mother, Elaine. My mum has never stopped believing in me. Without my mum I would not be the person I am today. She has supported me through every aspect of life and there is no other person I could thank more for everything I have achieved.

Abbreviations

ABC – ammonium bicarbonate

AC – alternating current

ACTB - actin beta

ACTBL2 - actin, beta like 2

ACTN4 - actinin alpha 4

ACTR3 - ARP3 actin related protein 3 homolog

ACTG1 – Actin gamma 1

AJCC – American Joint Committee on Cancer

Akt – serine/threonine kinase

ANXA4 - annexin A4

ANXA5 - annexin A5

ANXA6 - annexin A6

APC - adenomatous polyposis coli

APES - 3-aminopropyltriethoxysilane

APP – antigen processing and presentation

APRT - adenine phosphoribosyltransferase

ARHGAP1 - Rho GTPase activating protein 1

ARPC4 - actin related protein 2/3 complex subunit 4

ATP2A2 - ATPase sarcoplasmic/endoplasmic reticulum Ca2+ transporting 2

AUC – area under the curve

BAD – British Association of Dermatologists

BCC – basal cell carcinoma

BGN - biglycan

BPH – benign prostate hyperplasia

BRAF - v-Raf murine sarcoma viral oncogene homolog B

BSA – bovine serum albumin

BWH – Brigham and Women's hospital

CALM1 - calmodulin 1

CALML3 - calmodulin like 3

CALML5 - calmodulin like 5

CANX – calnexin

CART – classification and regression trees

CAPG - capping actin protein, gelsolin like

CCR5 - C-C chemokine receptor type 5

CCR7 - C-C chemokine receptor type 7

CCR10 - C-C chemokine receptor type 10

CCT4 - chaperonin containing TCP1 subunit 4

CCT8 - chaperonin containing TCP1 subunit 8

CD – Cluster of differentiation

CDKN2A - cyclin-dependent kinase Inhibitor 2A

CHAPS - 3-[(3-Cholamidopropyl) dimethylammonio]-1-propanesulfonate hydrate

CID – collisionaly induced dissociation

CIEF – capillary isoelectric focusing

CKAP4 - cytoskeleton associated protein 4

CLIC1 - chloride intracellular channel 1

CSL - CBF1, Suppressor of Hairless, Lag-1

CM – carbon metabolism

COL12A1 - collagen type XII alpha 1 chain

COL14A1 - collagen type XIV alpha 1 chain

COL1A2 - collagen type I alpha 2 chain

COL6A1 - collagen type VI alpha 1 chain

COL6A2 - collagen type VI alpha 2 chain

COL6A3 - collagen type VI alpha 3 chain

COL7A1 - collagen type VII alpha 1 chain

CORO1A - coronin 1A

CPA3 - carboxypeptidase A3

cSCC- cutaneous squamous cell carcinoma

CTLA - cytotoxic T-lymphocyte-associated protein

CTNNB1 – beta-catenin

CNN – convolutional neural network

CXCR1 - C-X-C chemokine receptor type 1

CXCR2 - C-X-C chemokine receptor type 2

CXCR3 - C-X-C chemokine receptor type 3

CXCR4 - C-X-C chemokine receptor type 4

DAB - 3,3'-Diaminobenzidine

DC – direct current

DDA – data dependant acquisition

DDOST - dolichyl - diphosphooligosaccharide - protein glycosyltransferase non - catalytic subunit

DDX39A - DExD - box helicase 39A

DDX3X - DEAD - box helicase 3, X - linked

DIA – data independent acquisition

DLST - dihydrolipoamide S - succinyltransferase

DMBA - 9, 10-dimethyl-1,2-benzanthracene

DMEM – Dulbecco's modified eagle medium

DPX - distyrene, plasticizer, xylene

DTE – dithioerythritol

DTT - dithiothreitol

ECM – extracellular matrix

ECMRI – extracellular matrix receptor interaction

EDF – European dermatology forum

EDTA - Ethylenediaminetetraacetic acid

EEF1A1 - eukaryotic translation elongation factor 1 alpha 1

EEF1D - eukaryotic translation elongation factor 1 delta

EEF2 - eukaryotic translation elongation factor 2

EGFR – epidermal growth factor receptor

EIF – eukaryotic translation initiation factor

EMT – epithelial mesenchymal transition

ER – endoplasmic reticulum

ERK – extracellular signal-regulated kinases

ESI – electrospray ionisation

EVPL - envoplakin

FA – focal adhesion

FASP – filter aided separation protocol

FBN1 - fibrillin 1

FBS - foetal bovine serum

FDA –food and drug administration

FDR – false discovery rate

FFPE – formalin fixed paraffin embedded

FGA - fibrinogen alpha chain

FGB - fibrinogen beta chain

FLNA - filamin A

FLNB - filamin B

FN1 - fibronectin 1

FTICR - fourier transform ion cyclotron resonance

GANAB - glucosidase II alpha subunit

GBM – gradient boosted machine

GDI2 - GDP dissociation inhibitor 2

GLM – generalised linear model

GO – gene ontology

GSN - gelsolin

H2AFV - H2A histone family member V

HCL – hydrochloric acid

HDMS – high definition mass spectrometry

HIST1H1C - histone cluster 1 H1 family member c

HIST1H2AJ - histone cluster 1 H2A family member J

HIST4H4 - histone cluster 4 H4

HNRNPA1 - heterogeneous nuclear ribonucleoprotein A1

HNRNPA2B1 - heterogeneous nuclear ribonucleoprotein A2/B1

HNRNPF - heterogeneous nuclear ribonucleoprotein F

HNRNPH2 - heterogeneous nuclear ribonucleoprotein H2

HNRNPK - heterogeneous nuclear ribonucleoprotein K

HNRNPR - heterogeneous nuclear ribonucleoprotein R

HPF – high power field

Hsp70 - heat shock protein 70

Hsp70 - heat shock protein family 70

HSP90AA1 - heat shock protein 90 alpha family class A member 1

HSP90AB1 - heat shock protein 90 alpha family class B member 1

HSP90B1 - heat shock protein 90 beta family member 1

HSPB1 - heat shock protein family B

IAA – iodacetamide

ID – identification

IgH – immunoglobulin H

IGHM - immunoglobulin heavy constant mu

IHC – immunohistochemistry

IL – interleukin

ILK – integrin-linked kinase

IPA – ingenuity pathway analysis

IQGAP1 - IQ motif containing GTPase activating protein 1

IQR – interquartile range

ITGB4 - integrin subunit beta 4

JAK – janus kinase

KC – Keratinocyte carcinoma

KEGG – Kyoto encyclopaedia of genes and genomes

KNN – K's nearest neighbour

KRT1 - keratin 1

KRT16 - keratin 16

KRT2 - keratin 2

KRT20 - keratin 20

KRT6A - keratin 6A

KRT6B - keratin 6B

KRT6C - keratin 6C

KRT80 - keratin 80

KRT84 - keratin 84

KRT9 – Keratin 9

KS – Kolmogorov-Smirnov

LC – liquid chromatography

LCP1 - lymphocyte cytosolic protein 1

LDA – linear discriminant analysis

LDHA - lactate dehydrogenase A

LDHB - lactate dehydrogenase B

LGALS1 - galectin 1

LMD – laser microdissection

LMNA - lamin A/C

LMNB1 - lamin B1

LMNB2 - lamin B2

LTM – leukocyte transendothelial migration

LTQ – linear ion trap

LUM - lumican

m/z – mass to charge ratio

MALDI – matrix-assisted laser desorption ionisation

MAP4 - microtubule associated protein 4

MAPK – mitogen-activated protein kinase

MC1R – melanocortin 1 receptor

MCAM – melanoma cell adhesion molecule

MDA - melanoma differentiation associated gene-9/Syntenin

MDH2 – malate dehydrogenase 2

ME – module eigengene

MMP - matrix metalloproteinase

MRM – multiple reaction monitoring

MS – mass spectrometry

MSN – moesin

MTHFD1 - C-1-tetrahydrofolate synthase, cytoplasmic

MTIF - microphthalmia-associated transcription factor

MYC - myelocytomatosis

MYH10 - myosin heavy chain 10

MYL6 - myosin light chain 6

Nb – Naïve Bayes

NCL - nucleolin

NER – nucleotide excision repair

NF-κB – nuclear factor kappa-light-chain-enhancer of activated B cells

NMSC – non-melanoma skin cancer

OTUB1 - OTU deubiquitinase, ubiquitin aldehyde binding 1

P4HB - prolyl 4 - hydroxylase subunit beta

PARK7 - Parkinsonism associated deglycase

PDA – protein digestion and absorption

PDIA3 - protein disulphide isomerase family A member 3

PEG – polyethylene glycol

PGK1 - phosphoglycerate kinase 1

PHB2 - prohibitin 2

PI3K - Phosphatidylinositol-4,5-bisphosphate 3-kinase

PIC – proteoglycans in cancer

PLGS – protein lynx global server

P-M – primary metastasising cutaneous squamous cell carcinoma

PMEL – premelanosome protein

Pmel-M – primary metastasising melanoma

Pmel-NM – primary non-metastasising melanoma

PML - promyelocytic leukaemia

P-NM – primary non-metastasising cutaneous squamous cell carcinoma

POSTN - periostin

PPER – Protein processing in endoplasmic reticulum

PPIA - peptidylprolyl isomerase A

PPP1CB - protein phosphatase 1 catalytic subunit beta

PRDX5 - peroxiredoxin 5

PRDX6 - peroxiredoxin 6

PRIDE – proteomics identification database

PSMA – prostate specific membrane antigen

PTEN - phosphatase and tensin homolog

RAC1 - ras - related C3 botulinum toxin substrate 1

RAF – rapidly accelerated fibrosarcoma

RAS – rat sarcoma

RegAC – regulation of actin cytoskeleton

RF – random forest

RGN – regucalcin

ROC – receiver operating characteristic

ROS – reactive oxygen species

RP-HPLC – reverse phase high performance liquid chromatography

RPL31 - ribosomal protein L31

RPL4 - ribosomal protein L4

RPL6 - ribosomal protein L6

RPN2 - ribophorin II

RPS10 - ribosomal protein S10

RPS12 - ribosomal protein S12

RPS13 - ribosomal protein S13

RPS16 - ribosomal protein S16

RPS18 - ribosomal protein S18

RPS2 - ribosomal protein S2

RPS20 - ribosomal protein S20

RPS28 - ribosomal protein S28

RPS5 - ribosomal protein S5

RPS7 - ribosomal protein S7

S100A11 - S100 calcium binding protein A11

S100A14 - S100 calcium binding protein A14

SAX – strong anion exchange

SDS - sodium dodecyl sulfate

SEPT2 - septin 2

SERPINA1 - serpin family A member 1

SERPINA3 - serpin family A member 3

SERPINB3 - serpin family B member 3

SERPINH1 - serpin family H member 1

SFPQ - splicing factor proline and glutamine rich

SH3BGRL3 - SH3 domain binding glutamate rich protein like 3

SLC25A5 - solute carrier family 25 member 5

SND1 - Staphylococcal nuclease domain-containing protein 1

SNRPD1 - small nuclear ribonucleoprotein D1 polypeptide

SNRPD3 - small nuclear ribonucleoprotein D3 polypeptide

SOD – superoxide dismutase

SOD2 - superoxide dismutase 2, mitochondrial

SPDEF - SAM pointed domain-containing Ets transcription factor

STAT1 - signal transducer and activator of transcription 1

SVM – support vector machine

STX7 – syntaxin 7

TAGLN2 - transgelin 2

TBS – tris-buffered saline

TCR – T cell receptor

TDA – topological data analysis

TEAB - triethylamonium bicarbonate

TFA – triflouracetic acid

TGFBI - transforming growth factor beta induced

TKT – transketolase

Treg – T regulatory cell

TNC - tenascin C

TOF – time of flight

TPA - tetradecanoyl-phorbol acetate

TPI1 - triosephosphate isomerase 1

TUBB - tubulin beta class I

TXN - thioredoxin

TYW1 - tRNA - yW synthesizing protein 1 homolog

UICC – union for international cancer control

UPLC – ultra performance liquid chromatography

UV – ultraviolet

UVA – ultraviolet A

UVB – ultraviolet B

UVC – ultraviolet C

UVR – ultraviolet radiation

VCL - vinculin

VIM – vimentin

WGCNA – weighted gene co-expression network analysis

WHO – World Health Organisation

Xgb – extreme gradient boosting

XP – Xeroderma pigmentosum

XRCC5 - X - ray repair cross complementing 5

YWHAG - tyrosine 3 - monooxygenase/tryptophan 5 - monooxygenase activation protein gamma

YWHAZ - tyrosine 3 - monooxygenase/tryptophan 5 - monooxygenase activation protein zeta

Chapter 1: GENERAL INTRODUCTION

1.1 Structure and function of skin

The skin is the largest human organ and plays a vital role in many homeostatic functions including temperature regulation and water loss as well as acting as a physical barrier for protection against the environment (Proksch et al., 2008). The skin is comprised of three main layers, the subcutaneous tissue, the dermis (inner layer) and the epidermis (outer layer) (**Figure 1.1**). The dermis consists mainly of connective tissue such as collagen and elastin with some cells (fibroblasts, endothelial cells within blood vessels) and structures such as sweat glands and hair follicles embedded within. The epidermis is a cell rich structure composed mainly of keratinocytes, but also contains melanocytes and Langerhans cells. The basal layer comprises undifferentiated keratinocytes which proliferate and whose daughter cells become more terminally differentiated as they move to the upper layers of the skin. Above the basal layer is the suprabasal layer, granular cell layer and the outermost layer which is called the stratum corneum. The stratum corneum consists of dead keratinocytes which gradually desquamate from the skin surface.

The skin has a relatively high turnover rate of cells, with cells transiting from the basal layer to the surface of the stratum corneum within 48 days (Iizuka, 1995). This high turnover and rate of proliferation amongst cells of the skin increase their susceptibility to carcinogenesis development (Ratushny et al., 2012).

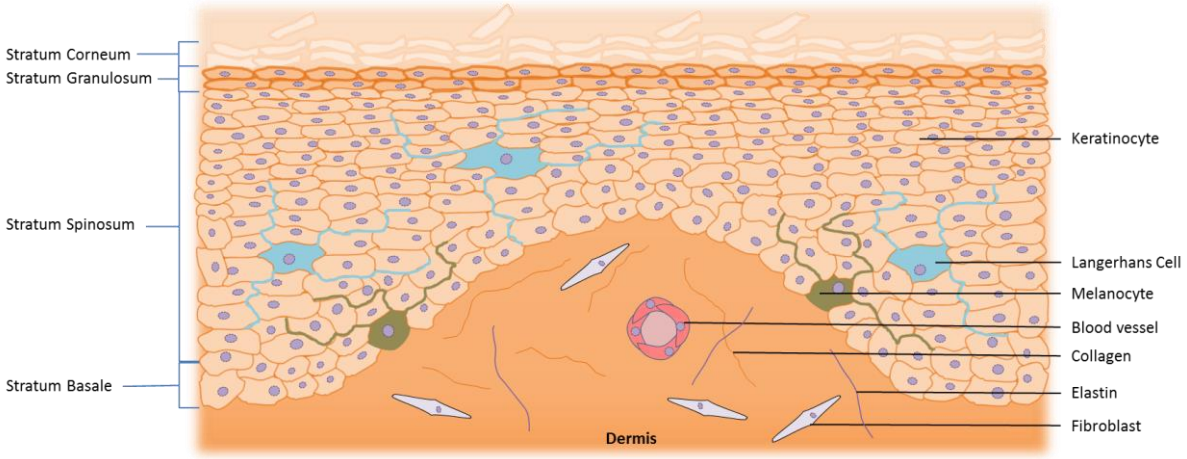


Figure 1:1 Schematic cross-sectional diagram of human skin

Chapter 1

1.2 Skin cancer

Skin cancer is the most common form of cancer worldwide and more frequently affects lighter pigmented individuals (World Health Organisation, 2016, Gloster and Neal, 2006). Skin cancer comprises of two main subgroups; melanoma and keratinocyte cancers (KCs, also known as non-melanoma skin cancer (NMSC)). Cutaneous melanoma arises when somatic gene mutations occur within the melanocytes in the basal layer of the epidermis. Within melanoma, there are 4 main histological sub-types, these include superficial spreading melanoma, nodular melanoma, lentigo maligna melanoma and acral melanoma (Bataille et al., 1996). Superficial spreading melanoma is the most common type of melanoma and as the name implies, usually refers to a melanoma which spread out across the skin. Conversely however, nodular melanomas often grow vertically up and down, being the second most common type of melanoma. Lentigo maligna melanoma is often found in older individuals at high sun exposed body sites. These themselves grow from a benign precancerous legion known as a lentigo maligna. Finally, the rarest form of melanoma is acral melanoma which is typically found in the palm of hands, soles of feet and under fingernails of patients. This type of melanoma is more prevalent than other forms of melanoma in individuals with black or brown skin (Farage et al., 2009).

KCs arise when keratinocytes of the epidermis develop somatic gene mutations and become cancerous (Armstrong and Kricker, 2001); (Albert and Weinstock, 2003). The two types of KCs are basal cell carcinoma (BCCs) and cutaneous squamous cell carcinoma (cSCCs). Although BCCs are very common, they rarely metastasise, unlike melanomas and cSCCs which can metastasise to other organs (Dinehart and Pollack, 1989, Thompson et al., 2005).

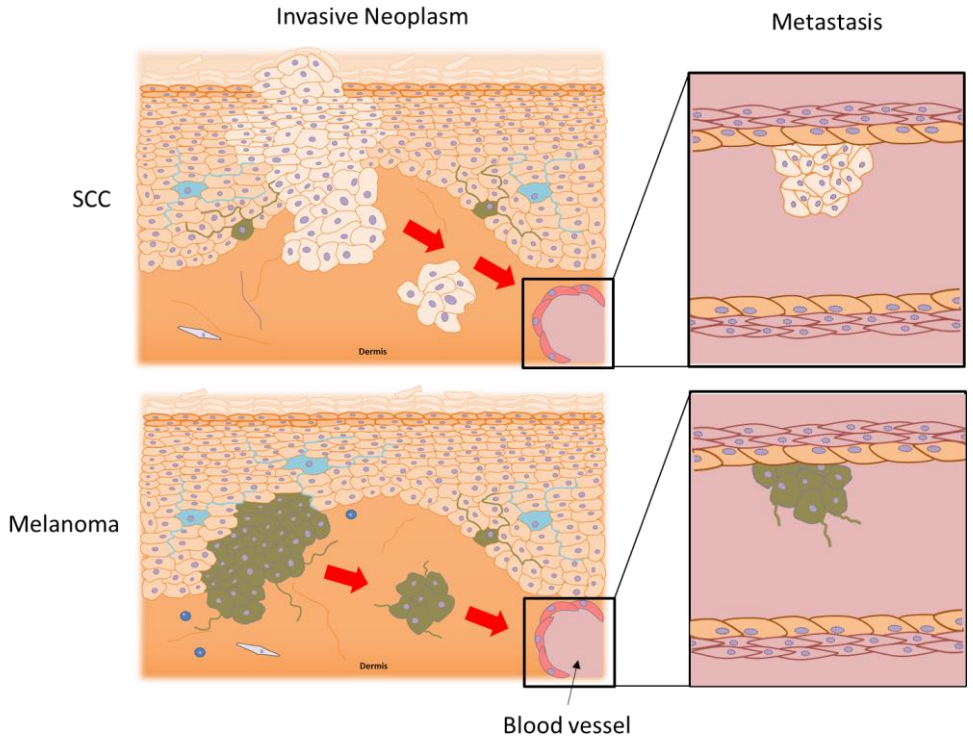


Figure 1:2 Schematic cross sectional diagram of squamous cell carcinoma (SCC) and melanoma development.

1.2.1 Incidence

The World Health Organisation (WHO) estimates that 1 in every 3 cancers diagnosed are skin cancers and that there are 130,000 melanoma and 2-3 million NMSC cases worldwide with a continual increase each year (World Health Organisation (2016). Skin cancer is one of the cancers that has increased by more than 50% from 1990 to 2010 (Murray et al., 2012). However, skin cancer numbers are believed to be greatly underestimated as it has been suggested that during 2006 in the US alone there were 3.5 million NMSCs in 2.1 million individuals (Rogers et al., 2010) and 5.4 million total NMSCs in 3.3 million individuals 6 years later in 2012 (Rogers et al., 2015). In addition, there was over 750,000 NMSC cases in Australia (Fransen et al., 2012) and an estimated 51,555 cSCC cases in the UK, which is more than double that of what it was less than a decade ago (Goon et al., 2016). Furthermore, figures presented at a UK TREND NMSC workshop suggested there was >200,000 keratinocyte carcinoma (KC) cases in the UK during 2015 (Rashbash, 2016).

The GLOBOCAN project, which is part of the International Agency for Research on Cancer, reports on incidence, mortality and prevalence of major types of cancer for 184 countries

of the world (GLOBOCAN, 2017). However, like many other cancer epidemiology registers, GLOBOCAN has not reported on the number of KC cases, mainly because recording of KCs is poor in many countries (Lomas et al., 2012). Related to this, it is thought that KC incidence is greatly underestimated because records are either incomplete or simply lacking; this is particularly the case in Europe, Australia and America, because registries often only include a patient's first KC, with subsequent tumours as well as multiple tumours not being counted individually (Lomas et al., 2012).

GLOBOCAN calculated the number of worldwide melanoma cases in 2008 to be 197,000 (Ferlay et al., 2010) and, given that this was over a decade ago coupled with the rising incidence rates (C.R.UK, 2015), it is likely that the actual figure today is much higher. Melanoma is most common in Caucasian populations and those who live in areas of higher sun exposure are at greater risk. It is for this reason that Australia, New Zealand and the USA have amongst the highest rates of melanoma in the world (Giblin and Thomas, 2007). Rates as high as 40-60 per 100,000 a year have been reported in Australia and New Zealand and 10-15 per 100,000 annually in central Europe (Garbe and Leiter, 2009). Although the UK does not have as high sun exposure as Australia, wealth and disposable income enables UK residents to use sun beds and/or to travel abroad to higher sun-exposed areas, potentially increasing the likelihood of developing melanoma (Giblin and Thomas, 2007, Godden et al., 2010). UK melanoma rates have been increasing for decades and continue to do so, despite intervention strategies such as public health announcements on avoiding excess exposure to sunshine (Diffey, 2004).

1.2.2 Economic burden

Due to the inconsistencies in skin cancer records, producing accurate estimations of the economic burden of skin cancer is inherently difficult (O'Dea, 2000). It has been reported that the annual economic costs of skin cancer to New Zealand in 2006 was NZ$123.1 million (£69 million) (O'Dea, 2000). Moreover, it has been estimated that total costs, including diagnosis, treatment and pathology for KCs in Australia during 2010 were AU$511 million (£299.5 million), predicted to increase to AU$703 million (£412 million) by 2015 (Fransen et al., 2012). In the USA, the total annual direct and indirect costs associated with skin cancer (including precancerous lesions) have been estimated at $6.6 billion (£5.1 billion)

(Bickers et al., 2006). More recently, the economic burden of skin cancer in the USA has been estimated by the U.S. Department of Health and Human Services to be $8.1 billion (£6.26 billion) each year, of which $4.8 billion (£3.7 billion) is for KC and $3.3 billion (£2.55 billion) for melanoma (Watson et al., 2014).

A study looking at the economic burden of skin cancer in England estimated the total cost of melanoma in 2002 at over £138 million (around £75 million direct cost to NHS) and KC at over £100 million (around £50 million direct cost to NHS) (Morris et al., 2009) with direct NHS costs expected to rise to £180 million in 2020 (Vallejo-Torres et al., 2014). The study by Vallejo-Torres et al (2014) used cancer registry data of 8,658 melanomas and 69,840 KCs. This figure of KCs is a lot lower than the >200,000 suggested by Prof Jem Rashbass, Director of the National Cancer Registration (Rashbash, 2016), therefore it is likely that KC costs in England and the UK are, and will be, much greater than these estimates.

1.2.3 Risk factors

There are a number of genetic and environmental risk factors which contribute to the development of skin cancer (**Figure 1.3**). The main environmental risk factor is exposure to ultraviolet radiation (UVR) from the sun and/or sun beds (Narayanan et al., 2010). Other environmental risk factors for skin cancer are smoking (De Hertog et al., 2001), exposure to arsenic (Yu et al., 2006), radiotherapy (Karagas et al., 1996), immunosuppression (for example via immunosuppressive drugs) (Alter et al., 2014, Euvrard et al., 2003) and use of certain medications for various diseases (for example oral steroids (Karagas et al., 2001) and the use of BRAF inhibitors in melanoma (Su et al., 2012)). In addition to these risk factors, some studies have suggested that that the use of sunscreen actually increases the risk of melanoma in latitudes greater than 40° (Gorham et al., 2007). This latter association may be due to the fact that older sunscreens protected mainly against UVB and thus also protected against sunburn, which may have resulted in people remaining in the sunshine for longer and obtaining more UVA exposure. However, fair-skinned people are at greater risk of skin cancer (see below), therefore the positive association between sunscreen use and melanoma may simply be due to the fact that fair-skinned people are more likely to use sunscreens.

Chapter 1

Phenotypic factors resulting from genetic inheritance such as red hair, freckles/fair complexion and a tendency to sunburn have a 2.4, 2.4 and 1.7 increase risk of developing melanoma (Thompson et al., 2005). The hair, skin and eye colour in mammals is determined by the content and composition of melanin pigment in these tissues (Ito and Wakamatsu, 2003). Melanin exists in two forms, the yellow/red pheomelanin and the dark brown/black eumelanin (Wakamatsu and Ito, 2002) which in different ratios produce the variation in human hair and skin colour. Although people with darker skin contain similar numbers of melanocytes to fair skinned individuals, they have higher amounts of melanin, specifically eumelanin, in their skin (Brenner and Hearing, 2008). Skin pigmentation, due to melanin content, plays a crucial role in defence against UVR induced DNA damage and many studies have shown that darker skinned people have more resistance to UVR-induced DNA damage compared to Caucasians (Tadokoro et al., 2003, Jablonski and Chaplin, 2010, Gallagher et al., 1995). Although there is little research surrounding the effect different melanin composition has on metastasis, it has been reported that high amounts of melanin can reduce the efficacy of radiotherapy (Brożyna et al., 2016) and furthermore that it could be a potential route for treatment using a 188-rhenium-labeled antibody, targeted to melanin (Klein et al., 2013, Schweitzer et al., 2007).

An important germline genetic factor which affects the ratio of eumelanin and pheomelanin in an individual is the melanocortin 1 receptor (*MC1R*) genotype, which also affects tanning response and susceptibility to skin cancer (Haddadeen et al., 2015, Robinson and Healy, 2002, Valverde et al., 1995). *MC1R* gene variants are frequent in the UK and Ireland (Gerstenblith et al., 2007) and the presence of two variant alleles in an individual causes red hair and fair skin, whereas a single variant allele results in fair skin (Raimondi et al., 2008, Flanagan et al., 2000, Healy et al., 2000). Certain diseases/syndromes can also lead to skin cancer development, for example oculocutaneous albinism type 2 (where skin melanin is lacking) and xeroderma pigmentosum (XP, where UV-induced DNA repair is compromised) (Setlow et al., 1969, Bradford et al., 2011).

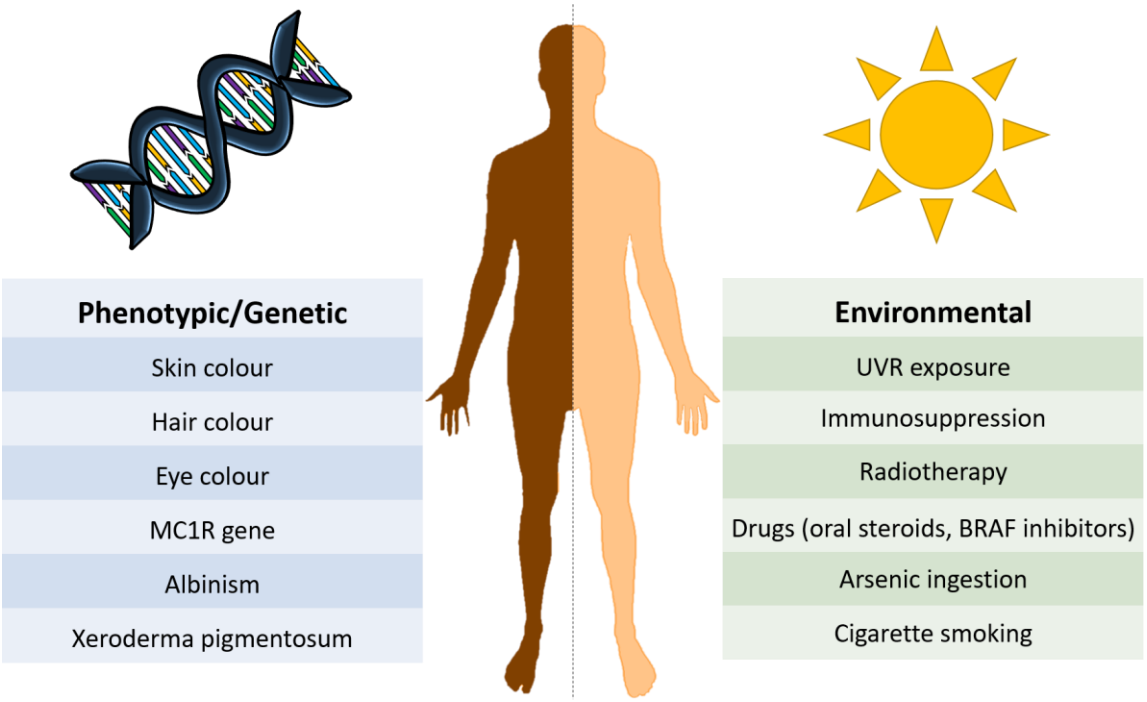


Figure 1:3. Phenotypic/genetic and environmental risk factors associated with skin cancer development.

MC1R, Melanocortin 1 receptor

1.2.4 Treatment

The main treatment for skin cancer is surgical excision (Madan et al., 2010). For small tumours, including BCCs and some cSCCs, it is not uncommon for curettage to be performed, a method also reported to achieve good results (Madan et al., 2010). Other treatment methods are ablation via CO_2 laser and cold induced destruction by liquid nitrogen cryosurgery, however, these are usually done on small tumours.

For cSCCs at low risk of metastasis a 4-5mm excision margin is advocated, whereas for high risk cSCCs a margin of at least 6mm is preferred (Madan et al., 2010). For SCCs of high risk, the draining lymph nodes can also be excised to assess for metastasis (Motley et al., 2003). An alternative mode of excision is via Mohs micrographic surgery whereby the tumour is removed and the surface of the residual open wound is then removed and stained before being viewed under a microscope. If tumour cells are seen in the wound skin under microscopy, further wound tissue is excised and again stained and viewed under microscopy. This process is repeated until the wound tissue is completely clear of tumour (at which stage all microscopic tumour has been removed).

Chapter 1

Following surgery, radiotherapy is occasionally used on high risk tumours which, under histological analysis, have been shown to have small groups of cancer cells away from the main tumour (i.e. microsatellite tumours) in order to reduce the risk of developing further metastasis. Treatment for metastatic cSCC is limited and radiotherapy is sometimes used on metastases which are difficult or impractical to remove (Rong et al., 2015). Chemotherapy has been used in some cases of metastatic cSCC but has proven to be limited in its efficacy (Weinberg et al., 2007). A study investigating the ability of retinol and isotretinoin to reduce new occurrences of skin cancer in 525 participants with a history of BCC or SCC revealed that using either of these chemotherapeutics had no benefit compared to a placebo group (Levine et al., 1997).

Similarly, the mainstay of treatment for melanoma is to excise the tumour. It is recommended that in situ melanomas are excised with a margin of 5mm, or those ≥2mm deep with a 2cm margin (Haigh et al., 2003). The British Association for Dermatologists (BAD) suggest a 1cm margin for those <1mm depth and between 1-2cm for those between 1 and 2mm deep (Marsden et al., 2010). Similar to cSCC, radiotherapy can be used after excision of high risk tumours with microsatellites present in histology in an attempt to prevent subsequent metastasis (Garbe et al., 2008). Radiotherapy is also sometimes used for metastases although mainly for palliative purposes, of which a response rate of 67% for the irradiated metastasis has been reported (Kirova et al., 1999). Chemotherapy has also been used for metastatic melanoma, but in recent European consensus-based interdisciplinary guidelines for diagnosis and treatment of melanoma, it was reported that adjuvant cytotoxic chemotherapy had no clear therapeutic advantage and suggested that this type of therapy should no longer be used (Garbe et al., 2016). Related to this, a study on 1256 patients published in the Lancet found that adjuvant PEGylated interferon α-2b significantly increases recurrence-free survival rates of patients with melanoma, but did not increase overall survival rates (Eggermont et al., 2008). By contrast, a systematic review and meta-analysis found that the use of adjuvant PEGylated interferon α-2b significantly increases disease free survival as well as the overall survival in melanoma, but noted that the adverse effects of the treatment could negatively affect quality of life of the patient (Mocellin et al., 2010). However, over recent years, newer therapies have included immunotherapy (e.g. anti cytotoxic T-lymphocyte-associated protein 4 (CTLA-4), anti-Programmed cell death protein 1 (PD1) antibodies) and BRAF inhibitors. Whereas

Ipilimumab, an anti CTLA-4 antibody has been shown to increase disease free survival of patients, it is unclear about the effect it has on overall survival (Eggermont et al., 2015, Garbe et al., 2016). Conversely, vemurafenib, a v-Raf murine sarcoma viral oncogene homolog B (BRAF) enzyme inhibitor has been found to significantly increase overall survival rates compared with dacarbazine, a systemic chemotherapy reagent (McArthur et al., 2014). More recently however, combination therapy using dabrafenib and trametinib have proven effective at treating BRAF mutant melanoma, increasing progression-free survival from 12% in the dabrafenib monotherapy group to 22% and furthermore increasing median overall survival from 18.7 to 25.1 months (Long et al., 2017, Long et al., 2015).

1.2.5 Ultraviolet radiation (UVR)

UVR is the most important exogenous factor that contributes to skin cancer development (Narayanan et al., 2010). There are three main types of UVR; the long wave UVA (315nm-400nm), the medium wave UVB (280nm-315nm) and the short wave UVC (100nm-280nm) (El Ghissassi et al., 2009). The amount of UVR that reaches the earth's surface consists of approximately 95% UVA and 5% UVB, whereas UVC is blocked by the stratospheric ozone layer (El Ghissassi et al., 2009, Narayanan et al., 2010). Tanning beds / sunbeds utilising artificial sunlight to induce suntans emit UVA and UVB to simulate sun exposure (Ting et al., 2007), and are used by the public to induce a suntan through production of melanin in the skin (Brenner and Hearing, 2008). It has been reported that individuals who use sun beds have a three times higher risk of developing melanoma (Chen et al., 1998, Ting et al., 2007) a 1.5 increase risk of developing BCC and a 2.5 increased risk of developing cSCC (Karagas et al., 2002).

For a long period of time it was believed that UVA was relatively harmless and that UVB was the main causative for skin cancer, particularly melanoma (Runger and Kappes, 2008, El Ghissassi et al., 2009, Narayanan et al., 2010). This was accredited to UVB's ability to cause more DNA damage, however, it is now recognised that UVA can also cause skin cancer through DNA mutations and that skin cancers (including cSCC and melanoma) can be induced by UVA in mice (Ikehata et al., 2008, Strickland, 1986, Kelfkens et al., 1991, Noonan et al., 2012).

Chapter 1

UVR causes direct DNA damage through the formation of cyclobutane pyrimidine dimers (CPDs), and 6-4 photoproducts (Brash et al., 1991, Balajee et al., 1999). Under normal circumstances, many of these photoproducts are repaired by nucleotide excision repair (NER), however, if not repaired they lead to development of DNA mutations, including C to T or CC to TT substitutions, which can result in skin carcinogenesis (Yokoyama and Mizutani, 2014). In patients who suffer from xeroderma pigmentosum (XP), who have a 10,000x increased risk of skin cancer (Kraemer et al., 1994), NER is deficient (Setlow et al., 1969) and therefore multiple DNA mutations arise, leading to skin cancer (Bradford et al., 2011). In normal skin and in the skin of XP patients, the development of mutations in tumour suppressor genes or oncogenes can affect the behaviour of the protein encoded by the gene, resulting in altered cell behaviour and uncontrolled proliferation, thus leading to tumorigenesis (Kramer et al., 1990, Setlow and Setlow, 1962, Kraemer et al., 1994).

1.2.6 Models of skin cancer

There are currently several different models used to study the development and progression of skin cancer. These differ according to the type of skin cancer being investigated. For example, development of melanoma has been observed and investigated using *Xiphophorus* fish (also known as swordtail fish) (Setlow et al., 1993), Sinclair swine (Millikan et al., 1974), horses (Rosengren Pielberg et al., 2008), dogs (Khanna et al., 2006), the *Monodelphis domestica* marsupial (also known as the South American Opossum) (Ley, 1984) and various transgenic mice (Kato et al., 1998, Becker et al., 2010). Some of these models, e.g. *Xiphophorus* fish, have been useful in trying to identify which wavelengths of UVR (i.e. UVA as well as UVB) are important in melanomagenesis (Setlow et al., 1993, Nairn et al., 1996). Furthermore, some of the genetic alterations leading to melanoma development in humans have also been seen in certain animal models such as cyclin-dependent kinase Inhibitor 2A (CDKN2A) mutations in *Xiphophorus* fish (Kazianis et al., 1999).

Another model, *M.domestica*, has allowed investigations into the role of pyrimidine dimers in melanoma. This is because *M.domestica* has a light activated DNA repair system, capable of repairing UVR-induced pyrimidine dimers (Ley, 1984) and because it is one of the few models where melanoma can be induced by UVR. However, all models of melanoma have

their limitations, for example in *M.domestica,* melanomas develop in the dermis rather than in the epidermis and only metastasise rarely (Ley, 2002). Sinclair swine have certain genetic similarities in their melanocytic tumours to those seen in humans, but the main difference is that a significant proportion of tumours spontaneously regress in this pig model (Millikan et al., 1974). Mouse models have become popular in studying melanoma development due to the ease of housing and ability for transgenic modification; examples include mice overexpressing Hepatocyte Growth Factor and mice with melanocortin 1 receptor alterations (Wolnicka-Glubisz et al., 2015). Another common animal model used to study melanoma, due to the ability to genetically manipulate them, is the zebrafish (van der Weyden et al., 2016). The Angora goat has been presented as a possible model for both melanoma and cSCC because, in one study, 2.2% and 3.8% of 1731 goats sampled had developed melanoma and cSCC respectively (Green et al., 1996).

Most animal models of cSCC are mice, including hairless mice which can develop cSCCs in response to UVR (de Gruijl and Forbes, 1995). Another mouse model is the chemical carcinogenesis model which involves the use of 9,10-dimethyl-1,2-benzanthracene (DMBA) and tetradecanoyl-phorbol acetate (TPA) which induces the formation of benign papillomas that then progress to cSCC (Abel et al., 2009). In this model, HRas, Kras and Tp53 genes are known to be mutated and more recent work suggests that altered expression of many genes in human cSCC are similarly modified in their expression in cSCCs in this model (Nassar et al., 2015). Furthermore, one study found that genomic drivers of SCC development could be identified in human and solar ultraviolet radiation-driven hairless mice (Chitsazzadeh et al., 2016). Although BCC is not a focus of this current thesis, there are also transgenic mouse models of BCC, particularly those affecting the hedgehog signalling pathway (Mao et al., 2006).

In addition to animal *in vivo* models, there are cell line *in vitro* models for skin cancer. Cell culture models have the advantage of being relatively cheap and easy to maintain in comparison to animal models (Beaumont et al., 2013). Furthermore, 2D cell models of melanoma and cSCC can be used effectively for many assays, including adhesion, migration and cell communication (Haass et al., 2005). More recently, 3D culture cell models have been employed; one such model by Commandeur *et al* (2009) was generated using a combination of cSCC cell lines, cSCC biopsies and fibroblast cultures to recreate an invasive

cSCC environment. Despite their limitations, many of the animal and cell models can help answer questions in specific research areas, and have improved our current understanding of melanoma and cSCC.

1.2.7 Oxidative stress

Exposure to UVR can cause oxidative stress, producing free radicals, collectively referred to as reactive oxygen species (ROS) in cells and tissues (Bayr, 2005). Free radicals are any molecules which possess one or more unpaired free electrons. These electrons can be passed onto, and consequently excite, nearby molecules potentially breaking bonds and/or making new bonds between atoms/molecules. ROS are common by-products produced by normal metabolic processes, typically neutralised by anti-oxidants such as superoxide dismutase (SOD) (Bickers and Athar, 2006, Bayr, 2005). In the event of a build-up or sudden increase in ROS, anti-oxidant defence mechanisms can become overwhelmed, resulting in an excess of free radicals. These free radicals can attack and damage nearby DNA and cause single and double strand breakages, base modifications such as 8-hydroxyguanine (Cheng et al., 1992) as well as by inducing cross linking between DNA and proteins (Athar, 2002). These free radical-induced alterations in DNA can affect tumour suppressor genes or genes regulating many aspects of cell function, including cell cycle, proliferation, and cell survival, and can ultimately lead to carcinogenesis (Sander et al., 2004). As well as effects on DNA, ROS can damage proteins resulting in loss of or gain of function (Bayr, 2005).

Enzymes involved in the detoxification of ROS produced by UVR include SOD and catalase (Rezvani et al., 2006). However, while catalase is known to protect against UVR-induced ROS (Rezvani et al., 2006, Rezvani et al., 2007), one study found that keratinocytes expressing higher levels of catalase had a notable increase in ROS after UVB exposure compared to those expressing lower levels of catalase, suggesting that catalase may not be protective under certain cellular and/or environmental conditions (Heck et al., 2003). ROS are thought to mediate a number of effects of UVR, and it has been reported that UVB induces cell cycle changes in keratinocytes similar to ROS and that both induce apoptosis by altering mitochondrial membrane permeability (Bickers and Athar, 2000). Other studies have found that UVR and ROS can induce a number of similar proteins and transcription

factors including NF-κB (Reelfs et al., 2004) and mitogen-activated protein kinase (MAPK) (Kim et al., 2005).

1.2.8 Genetic mutations in skin cancer

Melanoma and cSCC are two of the most highly mutated human malignancies, with melanoma harbouring around 20 mutations per mega base pair and cSCC containing up to 50 mutations per mega base pair (Durinck et al., 2011, Martincorena and Campbell, 2015, South et al., 2014, Nikolaev et al., 2011). It is thought that only a proportion of these mutations are needed for the development of cancer. This has led to the driver/passenger model of mutations in cancer, whereby key "driver" mutations are required for malignancy and many of the other mutations are "passenger" mutations that do not necessarily contribute to the development or growth of neoplasia (Greenman et al., 2007). This can be seen as an example of Darwinian evolution because those cells containing mutations with the "desirable" characteristics for neoplasia generally proliferate, survive and invade into surrounding tissue (Martincorena and Campbell, 2015). Examples of genes with driver mutations in cSCC and melanoma are provided in **Table 1.1**.

One key driver mutation for many cancers, including melanoma and cSCC is the *TP53* tumour suppressor gene, encoding the p53 protein which is a major cell cycle regulator (Benjamin and Ananthaswamy, 2007). p53 is often referred to as the guardian of the genome because it halts the cell cycle allowing the cell time to repair damaged DNA. A mutation in the *TP53* gene can cause a faulty p53 protein, reducing its ability to halt the cell cycle and thus allowing proliferation in the presence of DNA damage. It is reported that between 50-90% of cSCCs and around 35% of melanomas contain mutations in the *TP53* gene (Brash et al., 1991, Durinck et al., 2011, Leffell, 2000, Sparrow et al., 1995).

Another important signalling pathway known to potentially harbour driver mutations in cSCC is the NOTCH pathway. It has been reported that 82% of cSCCs have a mutation in either *NOTCH1* or *NOTCH2* genes (South et al., 2014). NOTCH is a signalling pathway that is activated in humans by Delta-like and Jagged ligands binding to NOTCH receptors 1, 2, 3 or 4, which causes the release of the NOTCH intracellular domain (D'Souza et al., 2008). The intracellular domain then travels to the nucleus where it regulates CBF1, Suppressor of Hairless, Lag-1 (CSL), a transcription factor for NOTCH target genes which are notably

involved in cell survival and growth (Mozuraitiene et al., 2015). Hyperactivity of NOTCH signalling can cause upregulation of β-catenin, another transcription factor, heavily involved in cell survival and proliferation (Moon et al., 2004). Although NOTCH mutations seem uncommon in melanoma, there is evidence that a C-to-G somatic mutation in the pre-microRNA, pre-miR-146a/C (leading to pre-miR-146a/G) in melanoma activates NOTCH signalling and promotes oncogenesis (Forloni et al., 2014) .

Table 1.1 Examples of genes with driver mutations in cSCC and melanoma

	Gene/Pathway	Protein involved	Effect of mutation	References
cSCC	NOTCH	NOTCH-1, NOTCH-2	Disruption between balance of growth and differentiation	(Wang et al., 2011, Saridaki et al., 2003, Zhang et al., 2016, South et al., 2014, Li et al., 2015, Pickering et al., 2014)
	RAS/RAF/MAPK	KRAS,HRAS, NRAS	Continuous activation of MAPK and PI3/AKT signalling pathways, increased survivability, growth and proliferation	(Wang et al., 2011, South et al., 2014, Durinck et al., 2011, Su et al., 2012, Li et al., 2015)
	P53	P53	TP53 mutations disable ability to halt cell cycle to allow DNA repair, enabling uncontrolled proliferation with development of mutations	(South et al., 2014, Durinck et al., 2011, Li et al., 2015, Pickering et al., 2014)
	CDKN2A	p16INK4a (p16)	P16 mutations disable ability to inhibit CDK4 and CDK6, resulting in less activation of retinoblastoma proteins, thus allowing more progression of cell cycle from G1 to S-phase.	(South et al., 2014, Durinck et al., 2011, Li et al., 2015, Pickering et al., 2014)
		p14arf (p14)	P14 mutations result in an inability to help activate p53. Promotes unregulated proliferation	(South et al., 2014, Durinck et al., 2011, Li et al., 2015, Pickering et al., 2014)
Melanoma				
	PI3k-AKT	PTEN	Loss of function causes activation of PI3/AKT pathway. (Often occurs with BRAF). Results in increased survivability, growth and proliferation	(Mozuraitiene et al., 2015, Goel et al., 2006, Haluska et al., 2006, Shull et al., 2012)
	RAS/RAF/MAPK	BRAF,MAPK, NRAS	Continuous activation of MAPK signalling, increased cell survival and growth	(Mozuraitiene et al., 2015, Goel et al., 2006, Hodis et al., 2012)
	CDKN2a	CDK4	Mutation causes inability for inhibition by P16, leading to cell cycle progression	(Hodis et al., 2012, Mozuraitiene et al., 2015)
		p16INK4a (p16)	P16 mutations disable ability to inhibit CDK4 and CDK6, resulting in less activation of retinoblastoma proteins, thus allowing more progression from G1 to S-phase.	(Mozuraitiene et al., 2015, Hodis et al., 2012)
		p14arf (p14)	P14 mutations results in an inability to help activate p53. Promotes unregulated proliferation	(Mozuraitiene et al., 2015, Hodis et al., 2012)
	Wnt/β-catenin	CTNNB1, APC, ICAT	Causes aberrant activation of wnt signalling, leading to increase in cell proliferation	(Mozuraitiene et al., 2015, Reifenberger et al., 2002)

The RAS signalling pathway is strongly associated with cell growth, differentiation and survival and mutations in this pathway are found in 20%-25% of human cancers (Downward, 2003). *RAS* genes are mutated in ~3-30% of cSCCs (South et al., 2014, Su et al., 2012), which is a much lower frequency than mutations in *TP53*, *NOTCH1* and *NOTCH2* in these tumours (South et al., 2014). Conversely, the *BRAF* gene which encodes for another member of the RAS signalling pathway is found to be mutated in ~60% - 70% of superficial spreading melanomas (Haluska et al., 2006). A single mutation at codon 600, where a valine residue is substituted for a glutamate (V600E), accounts for roughly 50% of total BRAF mutations in melanoma (Su et al., 2012). Whereas BRAF inhibitors, which target the oncogenic BRAF protein resulting from this mutation, are beneficial in melanoma, many patients can develop cSCCs as an adverse effect of this therapy in a process referred to as paradoxical MAPK activation (Gibney et al., 2013). Furthermore, in contrast to the lower level of RAS mutations in sporadic cSCC, approximately 60% of cSCCs arising secondary to vemurafenib (a BRAF inhibitor) have RAS mutations (Su et al., 2012).

Amongst the many other mutations that have been identified in cSCC and melanoma, those in *CDNK2A* and phosphatase and tensin homolog (*PTEN)* are also considered as driver mutations. For instance, somatic *CDKN2A* mutations can be found in 28% of cSCCs (South et al., 2014) whereas germline mutations in this gene are seen in melanoma patients with a strong family history of this cancer (Harland et al., 2014). In addition, between 10 and 30% of all melanomas have a loss of function in the *PTEN* tumour suppresser gene, resulting in phosphatidylinositol-4,5-bisphosphate 3-kinase/protein kinase B (PI3K/AKT) pathway activation; it is worth noting that mutation in PTEN are often found in conjunction with BRAF mutations (Davies et al., 2008, Haluska et al., 2006).

1.2.9 Proteins in skin cancer

Genes are segments of DNA which encode for proteins. The first step of protein synthesis is transcription and involves RNA polymerase reading a gene and creating a complimentary mRNA strand from the gene exons. Once the mRNA has been created, it exits the nucleus and travels to the ribosome where a small ribosomal subunit binds and moves along it until it reaches a sequence of three bases, known as a codon, which encodes a “start” signal. After the start signal, tRNA molecules which consist of an anti-codon (a complementary

codon to that found on the mRNA) and a specific amino acid, begin to bind to the mRNA. As more and more tRNA molecules are added, the amino acid chain gets longer, until a stop codon is reached, which signals for the ribosome to release the finished polypeptide chain. The amino acid sequence then undergoes folding to become a protein. Protein folding is dependent on a number of factors but usually results in a structure which is the most thermodynamically stable in its current environment (Dobson et al., 1998). Due to the immense number of possible structures one polypeptide can form it is also reasonable that folding favours those structures that are most efficient, that is those which require the least amount of energy to create (Dobson, 2003).

Synonymous mutations are genetic mutations which do not alter translated amino acid sequence. Although sometimes referred to as silent mutations, synonymous mutations can have effect on post translational modifications and splicing effecting downstream function or even cellular location and abundance. Non-synonymous mutations, however, are mutations which alter the amino acid sequence as the codon encodes a different amino acid when transcribed and translated. Changes in the amino acid sequence can result in issues with stability and the way the protein is folded (Lorch et al., 1999) which can subsequently lead to altered function (Yamada et al., 2006) as well as unintended protein-protein interactions (Jones et al., 2007).

1.2.10 The immune system involvement in skin cancer

It has been known that UVR has significant immunosuppressive effects since the 1970's when Margaret Kripke showed that skin tumours transplanted to un-irradiated mice resulted in rejection of the transplanted cancer whereas tumour rejection failed to occur in mice irradiated with UVR (Kripke, 1974, Kripke, 1977). Similar results have been reported by other groups (Sluyter and Halliday, 2001) and immune suppression resulting from UVR is evident in humans in other types of studies (O'Dell et al., 1980). The exact reasons for this UVR-induced immunosuppression are not fully understood. There is some evidence that UVR damages Langerhans cells, which are specialised antigen presenting cells in the skin, promoting apoptosis of these cells (Aberer et al., 1981). In addition, UVR causes Langerhans cells to migrate to the local lymph nodes where they induce production of T regulatory cells (Tregs), thus dampening immune responses (Schwarz et al., 2010).

Furthermore, keratinocytes release the immunosuppressive cytokine interleukin-10 (IL-10) after UVR exposure (Nishigori et al., 1996) and inhibit tumour antigen presentation by epidermal antigen presenting cells (Beissert et al., 1995) resulting in a reduced immune response against the tumour.

A role for a weakened immune system in cSCC development can be seen clearly in immunosuppressed individuals following organ transplantation, with some studies suggesting that the incidence of cSCC is 50-250 fold higher in transplant recipients compared to the general public (Alter et al., 2014, Euvrard et al., 2003). The reduction in immunosurveilance is believed to result in approximately 5%, 10-27% and 40-60% of renal transplant recipients developing NMSC within 2, 10 and 20 years following transplantation respectively (Ulrich et al., 2008). Although other malignancies have been reported to following organ transplantation, skin cancers account for the majority of malignancies in this group, with cSCC and BCC accounting for 90% of the tumours (Euvrard et al., 2003). An increased risk of melanoma in transplant recipients, by a factor of 1.6-3.4 in Europe and 2-4 in Australia, has also been noted (Euvrard et al., 2003). Furthermore, immunocompromised individuals are at significant risk of metastases from skin cancer (Martinez et al., 2003).

1.2.11 Metastatic skin cancer

The act of metastasis has been summarised into 8 major steps; 1-detachment from primary tumour, 2-invasion into surrounding tissue, 3-invasion into a vessel, 4-circulation in vessels (lymphatic or haematogenous), 5-stasis within the vessel, 6-extravasation, 7-invasion into new tissue and 8-finally proliferation (Brodland and Zitelli, 1992). Of the common skin malignancies, cSCC and melanoma are the two cancers most capable of metastasis.

1.2.11.1 Metastasis in cutaneous Squamous Cell Carcinoma (cSCC)

The mortality rate of cSCCs and melanoma vary depending on a number of factors but both have poor clinical outcomes after metastases have developed. The occurrence of metastasis for cSCC is approximately 4% (Brantsch et al., 2008) to 9.9% (Weinberg et al., 2007). In one study, it was found that 81.5% of cSCC metastases involved regional lymph nodes, 3.7% involved distal nodes and 14.8% involved distant metastases (Dinehart and

Pollack, 1989). It has been reported that common sites of distant metastases are the lungs, brain, liver, skin and bone (**Figure 1.4**) (Weinberg et al., 2007).

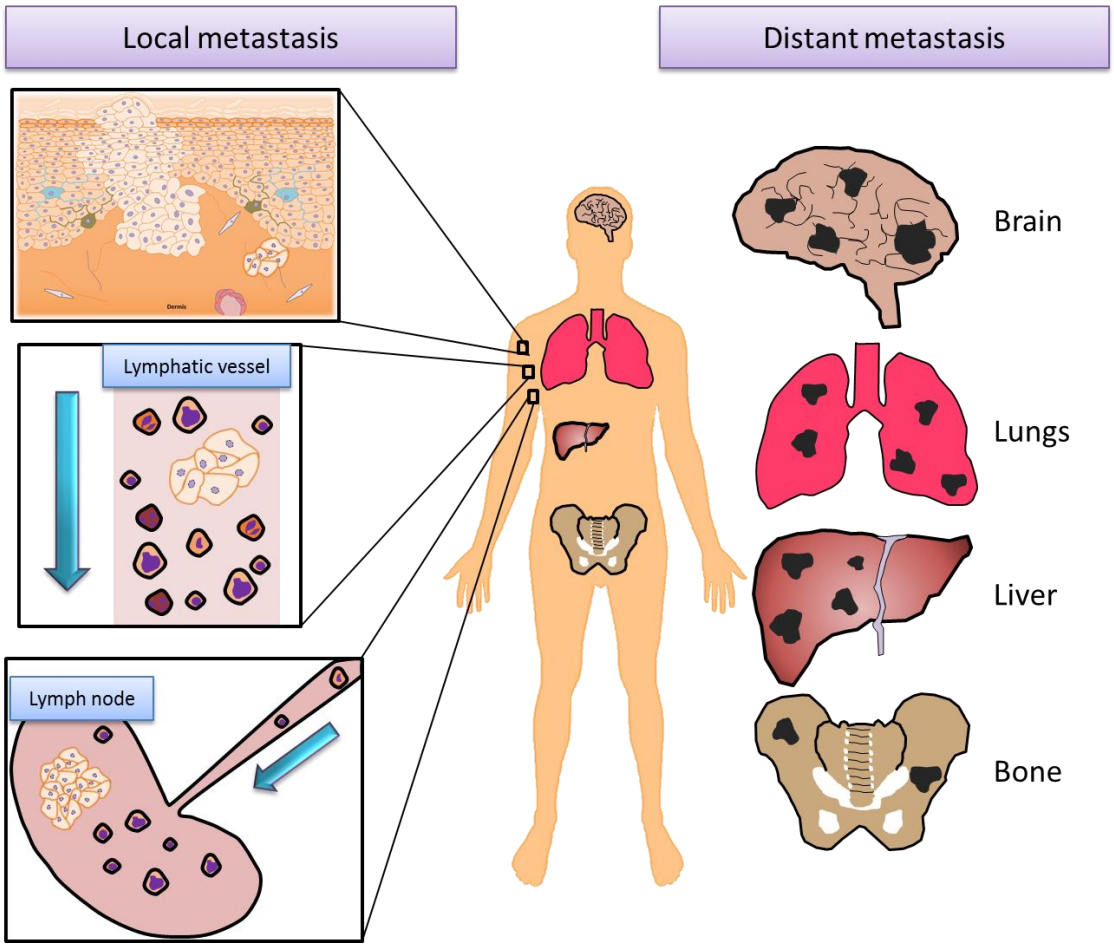


Figure 1:4 Local and distant cutaneous Squamous Cell Carcinoma (cSCC) metastases
Metastasis occurs by invasion into the dermis and subsequently into the lymphatic system, followed by metastatic deposition in local lymph node. Distant cSCC metastasis in brain, lungs, liver and bone may occur via haematogenous spread from the lymph node or from the primary cancer. cSCC cells are represented by the pale cells in the boxes on the left of the figure whereas the black areas represent metastatic deposits in distant organs on the right side of the figure.

Specific characteristics have been known to increase the risk of SCC metastasis in affected patients (Madan et al., 2010, Motley et al., 2003, Weinberg et al., 2007, Thompson et al., 2016, Veness, 2006). The site of the primary SCC has a major effect on the risk of developing metastasis. It has been reported that between 11 and 16% of SCC of the lip progress to metastasise (Rowe et al., 1992, Frierson and Cooper, 1986), which has been more recently confirmed in a systematic review and meta-analysis on risk factors for SCC metastasis in

which both lip and ear had a risk ratio of 2.28 and 2.33 in developing metastasis, respectively (Thompson et al., 2016). In addition to the site of the tumour, the size of the cSCC also plays a critical role in potential to metastasise; this effect of size is seen both in relation to diameter and depth of the cSCC. For example, cSCCs larger than 2cm metastasise in 30% - 42% of cases (Alam and Ratner, 2001, Rowe et al., 1992), whereas cSCCs exceeding 6mm in depth have been reported to metastasise in 16% of patients (Brantsch et al., 2008) with another study reporting that all cSCCs >6mm in depth metastasised (Stein and Tahan, 1994). In the systematic review by Thompson *et al* (2016), depth of tumour invasion (i.e. Breslow thickness) and invasion into subcutaneous fat were found to have the highest associated risk of both recurrence and metastasis of cSCC. Studies have found that patients with perineural invasion by the primary cSCC are more likely to suffer from nodal metastasis than those that don't have perineural invasion (Cherpelis et al., 2002). One large study found that patients diagnosed with perineural invasion were 20% more likely to develop regional metastasis and 11.7% more likely to develop distant metastases (Goepfert et al., 1984). This positive association between perineural invasion and metastases was also confirmed in the recent systematic review/meta-analysis (Thompson et al., 2016).

The differentiation status of the primary cSCC is also a factor influencing development of metastases (Motley et al., 2002). In the early part of the twentieth century, Broders designed a staging system for SCC, using stages 1 – 4, to categorise how differentiated tumours were (Broders, 1921). In Broders' system, stages 1-3 consist of a ratio of differentiated cells to undifferentiated cells of 3:1, 1:1 and 1:3 respectively, whereas stage 4 consists of no differentiated cells. It has been found that cSCCs with stage 2 or higher have a greatly increased risk of metastasis (Breuninger et al., 1990) with one study reporting that 92% of lip SCCs that metastasised were grade 4 (Frierson and Cooper, 1986). The rate of metastasis from poorly differentiated cSCCs is reported to be as high as triple the rates of well differentiated tumours (Weinberg et al., 2007). Poor differentiation is also associated with a higher disease-specific death rate (Thompson et al., 2016).

In addition to site, size, perineural invasion and differentiation, other factors which influence the metastasis rate of cSCC include immunosuppression (e.g. in organ transplant recipients as highlighted earlier) (Euvrard et al., 2003, Martinez et al., 2003, Ulrich et al.,

2008) and differences in treatment (e.g. type of surgery, radiotherapy etc.)(Karagas et al., 1996, Brantsch et al., 2008).

1.2.11.2 Metastasis in melanoma

While the incidence of melanoma is lower than that of cSCC, the rate of metastasis is higher and has a poor prognosis associated with it (Balch, 1992, Manola et al., 2000). Metastases in melanoma is dependent on a multitude of factors, one of which is depth of invasion, which is recorded as Breslow thickness (Breslow, 1970) and Clark's level (Thompson et al., 2005). Breslow thickness is simply the depth of the melanoma in mm but is well known to have a positive correlation with metastasis and indeed a worse prognosis (Breslow, 1979, Cornish et al., 2009). Clark's level is defined as the layer of skin that the tumour invades into (Clark et al., 1969). Although similar to Breslow depth in terms of reporting the thickness of the tumour, it takes into consideration the thickness of skin at different body sites (some being thicker than others) as all skin has an epidermis, dermis and subcutaneous tissue. A figure representing Clark's level and Breslow depth can be seen in **Figure 1:5**.

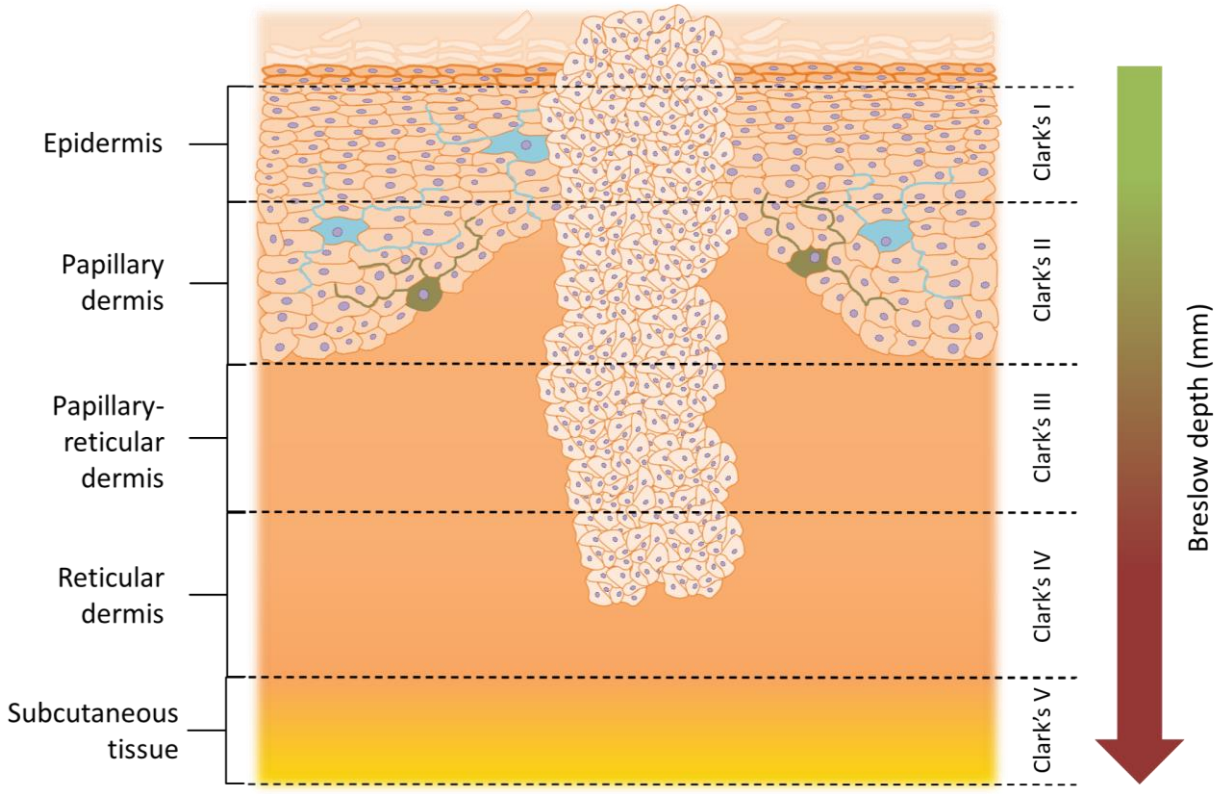


Figure 1:5: A Diagram of Clark's level and Breslow thickness

Chapter 1

A large study containing 3,001 patients with primary cutaneous melanoma reported that 466 (15.5%) progressed to metastatic melanoma (Meier et al., 2002). Of the 466 patients with metastatic melanoma, 21.7% developed satellite/in-transit metastases, 50.2% developed regional lymph node metastases, 28.1% developed distant metastases, with 51.5% of the satellite/in-transit metastases and 59% regional lymph node metastases subsequently developing into distant metastases (Meier et al., 2002). In the same study, 57.3% of the overall patients who developed distant metastases died (Meier et al., 2002). In a separate study of 1,521 patients with American Joint Committee on Cancer (AJCC) stage IV melanoma, a median survival time of 7.5 months was reported, which equated to an estimated 5 year survival rate of 6% (Barth et al., 1995). In that study, patients could be split into 3 prognostic groups based on site of metastasis, namely (i) nodal, cutaneous or gastrointestinal metastasis with a median survival of 12.5 months, (5 year survival rate of 14%), (ii) pulmonary metastasis with a median survival of 8.3 months (5 year survival rate of 4%) or (iii) liver, brain or bone metastasis with a mean survival time of 4.4 months (estimated 5 year survival rate of 3%) (Barth et al., 1995). In addition, one of the earlier melanoma meta-analysis of 15,000 patients with local melanoma and 2,116 with nodal metastasis highlighted ulceration of the primary tumour as an important prognostic marker, also noting that there was a positive correlation between ulceration and thickness (Balch, 1992). Recently, the American Society for Cancer Immunology (ASCO) has reported that melanoma involving nodal metastasis has a 5-year survival rate of 64% (depending on the number of nodes affected) and that melanoma involving distant metastasis has a 5-year survival rate of about 23% (ASCO). It is also known that the location of the melanoma plays an important role in prognosis, for instance an early study found that melanoma on the scalp had a worse prognosis than melanoma on the face or neck and that melanoma on the hand had a poorer prognoses than melanoma on the arms and legs (Balch, 1992).

Clinical parameters are vital in giving accurate prognostic information but with the advancement in technology / instruments and methodology, laboratory prognostic markers, more commonly referred to as biomarkers, are becoming more and more important in determining risk and prognosis (Manola et al., 2000). There are many studies that have looked for melanoma specific protein biomarkers (Ugurel et al., 2009, Griewank, 2016, Gogas et al., 2009).

There are two main types of biomarkers, these are diagnostic biomarkers and prognostic markers. Presently, it is common practise to carry out immunohistochemical staining on sections of tumours to first diagnose patients and to additionally aid in prognosis. Markers in use today are premelanosome protein (PMEL), Melanogenesis Associated Transcription Factor (MITF), S100 proteins family members (Weinstein et al., 2014), melanoma cell adhesion molecule (MCAM), PI16 and matrix metalloproteinase-2 (MMP2) which may also offer some limited predictive information in relation to clinical outcome (Gould Rothberg et al., 2009). This number of known biomarkers is not as high when looking at differences between primary tumours which will metastasise and those which will not. Nonetheless, a large systematic review and meta-analysis on tissue biomarkers for prognosis of melanoma revealed a number of markers associated with development of metastasis in melanoma (Gould Rothberg et al., 2009). These include Bcl-2 expression (Vlaykova et al., 2002), MCAM, MMP and tissue plasminogen activator which were all significantly different in melanomas which subsequently metastasised. Furthermore, the meta-analysis reported that although chemokine receptors C-X-C chemokine receptor type 1 (CXCR1), C-X-C chemokine receptor type 2 (CXCR2), C-X-C chemokine receptor type 3(CXCR3), C-X-C chemokine receptor type 4 (CXCR4), C-C chemokine receptor type 5 (CCR5), C-C chemokine receptor type 7 (CCR7) and C-C chemokine receptor type 10 (CCR10) have been associated with metastasis in melanoma, only CXCR4 is significantly associated with poorer prognosis (Scala et al., 2005, Gould Rothberg et al., 2009).

There have been several studies which have looked at markers in serum which suggest that cutaneous melanoma has metastasised. For example, increased expression of YKL-40 protein in patients with metastatic melanoma is associated with poorer prognosis (Schmidt et al., 2006). Another marker which has been investigated as a potential indicator of melanoma metastases is tyrosinase mRNA, but a meta-analysis has suggested that this offers limited potential as a biomarker (Tsao et al., 2001). Similarly, serum lactate dehydrogenase (LDH) and S100B levels are not sufficiently robust for use as biomarkers, despite being associated with poor prognosis in AJCC stage III/IV melanoma patients (Bougnoux and Solassol, 2013). A proteomic study utilising mass spectrometry was able to discriminate between clinical stages of melanoma in >80% of cases and suggested that proteomic profiling may become a valuable tool in identifying high risk melanomas (Mian et al., 2005). Several studies have utilised mass spectrometry (MS) based proteomics to

identify serum biomarkers for metastasis in melanoma. It has been reported that vitronectin and demcidin are potent serum biomarkers of survival in metastatic melanoma (Ortega-Martinez et al., 2016). Another MS proteomic study found that co-expression of MDA-9 and GRP78 are also good serum biomarkers indicative of lymph node metastasis (Guan et al., 2015). Furthermore, Regucalcin (RGN), Syntaxin-7 (STX7), methylenetetrahydrofolate dehydrogenase 1-like (MTHFD1L) have also been associated with different progression states of melanoma with high levels identified in recurring tumours and high Breslow's thickness (Bystrom et al., 2017).

There have been improvements in the systemic treatment of metastatic cutaneous melanoma, however, survival rates of patients with metastatic melanoma still remain poor (Garbe et al., 2011). In addition, there has been no major advance in the treatment of metastatic cSCC over recent years, however, a recent phase I/II trial using anti-PD1 therapy suggests that this may offer some hope for metastatic cSCC (Migden et al., 2018). In most cases of cutaneous melanoma and cSCC, the presence of metastasis generally results in disease related mortality. Thus, there is a need to discover new biomarkers which could be used for prognostic prediction of the future development of metastases from melanoma and cSCC at the time of excision of the primary tumour. Advances in this area of research could lead to better clinical management of patients, e.g. identifying those who require long term follow-up and potentially permitting earlier treatment with systemic anti-cancer therapies, as well as providing insight into key pathways which could be targets for development of novel treatments. Due to its incredible sensitivity, one of the methods at the forefront of biomarker discovery (and the focus of this thesis) is mass spectrometry based proteomics.

1.3 Proteomics

Proteomics is a diverse field encompassing the analysis of proteins within biological samples. It entails the process of identifying and quantifying proteins in a given sample and has a variety of methodologies dedicated to it. One method with exceptional sensitivity and accuracy at the forefront of proteomic research is mass spectrometry (MS) based proteomics (Aebersold and Mann, 2003).

Although many mass spectrometry systems differ in their setup, they largely consist of an ion source, a mass analyser to determine the mass to charge ratio (m/z) of the ionised analyte and a detector to establish the number of ions at each m/z value (Aebersold and Mann, 2003).

1.3.1 Ion source

Mass spectrometry measures ions in the gaseous phase, therefore the role of the ion source is to generate the ions that can then be introduced into the mass spectrometer in the gaseous phase. Different techniques can be used to ionise samples; two of the main methods used in proteomics are desorption ionisation and spray ionisation, of which the two most commonly employed methods are matrix assisted laser desorption ionisation (MALDI) and electrospray ionisation (ESI), respectively (Cole, 2011). MALDI requires the analyte being measured to be embedded into a low molecular weight matrix, which is sensitive to UV. A UV laser is then used to excite the matrix, which absorbs energy and passes it to nearby analytes, ionising them and liberating them from the matrix into the gaseous phase. MALDI is generally good for simple peptide mixtures and can allow specific ionisation for targeted and directed proteomics (Karas and Hillenkamp, 1988). Conversely, ESI is performed on samples in solution by applying a high voltage to create an aerosol spray. As the charged spray of droplets travel to the cone entrance of the mass spectrometer, they evaporate until the maximum amount of charge each droplet can hold is reached, this is known as their Rayleigh limit. At this point, the electrostatic repulsion of positive charges in a “size-decreasing” droplet becomes stronger than the surface tension of the water droplet itself, causing Coulomb fission whereby the droplet explodes into many smaller droplets, with less ions in each. This process continues until one analyte ion is present (Cole, 2011, Ho et al., 2003), (**Figure 1.5**). As ESI is carried out on samples in solution, it is often coupled with a liquid chromatography system (LC), enabling multidimensional fractionation prior to ionisation (Fenn et al., 1989).

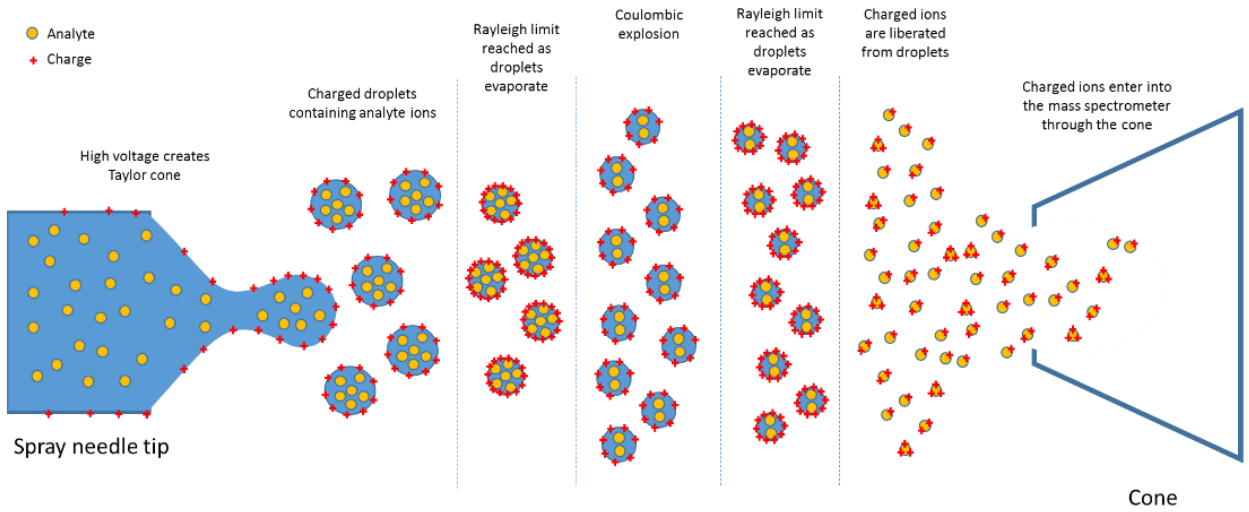


Figure 1:6 Diagram of electrospray ionisation (ESI).

Analytes in solution are forced through a capillary with a high voltage, creating a Taylor cone where charged droplets containing analytes are released. Charged droplets evaporate until their Rayleigh limit is reached and Coulombic fission takes place, causing a release of smaller droplets, each with less analytes within. These smaller droplets keep undergoing evaporation and Coulombic explosions until eventually charged ion analytes are liberated, where they enter the mass spectrometer through the cone

1.3.2 Mass analyser

There are four basic types of mass analysers used in mass spectrometry based proteomics; these are: - ion trap (including Orbitrap), quadrupole, time of flight (TOF) and Fourier transform ion cyclotron resonance (FTICR) analysers. Variants of each of these exists, as well as hybrids and systems which combine their usage in tandem (Aebersold and Mann, 2003). Ion trap and FTICR utilize m/z resonance frequency to separate and analyse ions, whereas quadrupoles use m/z stability and TOF analysers use flight time to separate ions (Yates et al., 2009).

Ion trap mass analysers vary in their design but essentially “trap” ions using their m/z properties before using a detector to measure the number of ions present. They are reasonably cheap, sensitive and durable and as a result are frequently seen in mass spectrometry based proteomic studies. Their disadvantage however is their low mass accuracy due to the finite number of ions that can be held before charge disturbs their spacing and ultimately their mass measurement (Aebersold and Mann, 2003). FTICR

analysers use similar principals to the ion trap in respect to trapping ions and holding them before analysis using their m/z. FTICRs can actually be correctly referred to as a Penning ion trap as it utilizes a high vacuum and strong magnetic fields to achieve a similar result. FTICR analysers have good sensitivity, accuracy, resolution and dynamic range but are expensive and require experience and training in their use (Aebersold and Mann, 2003).

Quadrupoles also utilise ion m/z to select their ions and analyse them. Quadrupoles consist of four rods connected together electrically in which two opposing rods share a direct current (DC) and the remaining two opposing rods share an alternating current (AC). A radio frequency voltage originating from the AC rods is emitted along with an offset DC voltage from the DC rods. These opposing voltages influence the path of passing ions in relation to their m/z value. Ions which have an unstable (critical) m/z value don't acquire a stable trajectory as they are repelled and attracted to each rod unevenly, causing them to either hit an electrode or exit the quadruple structure. Those that have m/z values which oscillate between the DC and radio frequency voltages evenly have a stable trajectory and make it through the quadrupole to a connected detector (Miller and Denton, 1986) (**Figure 1.6**). These radio frequency and DC voltages can be manipulated to allow only certain m/z values to pass through the quadrupole at any one time. This means precursor ions can be specifically selected and often fragmented before detection to obtain MS/MS spectra. Quadrupoles are simple, easy and cheap to produce and as a result are often used in proteomic mass spectrometry.

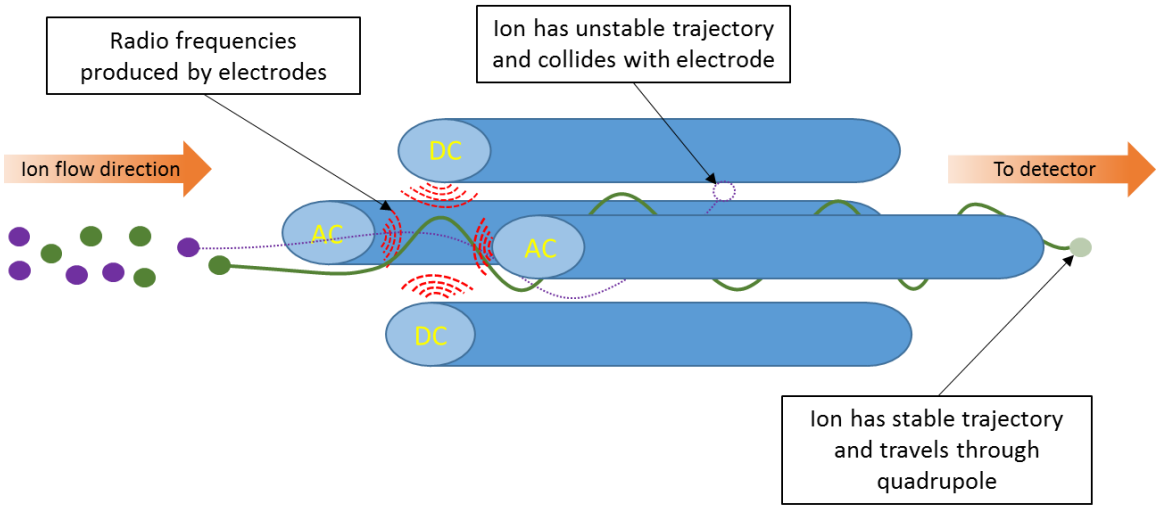


Figure 1:7 Diagram of a quadrupole mass analyser.

Two opposing DC and AC electrode rods create a radio frequency that guides ions of a certain mass to charge ratio through on a stable trajectory. Those with mass to charge ratios in the unstable, critical range have an unstable trajectory and do not make it through the quadrupole

TOF analysers determine the m/z value by measuring the time it takes for an ion to get from one point to another through a vacuum. Ions are accelerated through an electric field at a known strength toward a detector. The time it takes for an ion to reach the detector depends on the velocity of the ion, which is dependent on the weight of the ion (as heavier ones will move slower than smaller ones) and the charge of the ion as it receives more or less kinetic energy from the initial acceleration (Domon and Aebersold, 2006, Guilhaus et al., 1997). The time it takes the ion to travel from the starting point to the detector can be used to calculate the m/z value with exceptional resolution and accuracy.

1.3.3 Detector

The final part of the mass spectrometer is the detector. As ions hit or pass by the detector, it measures the charge induced or the current produced, respectively. This charge or current excites an electron in the detector, which is then recorded on a computer. As ions are typically in very low abundance in mass spectrometry, the numbers of electrons excited in the detector are similarly low as a result of this. For this reason, many detectors use an electron multiplier where one excited electron can stimulate multiple other electrons which can subsequently stimulate additional electrons, thus amplifying the signal which can then be read on a computer and converted to a mass spectrum.

1.3.4 Tandem mass spectrometry

Early proteomics methods involving mass spectrometry (MS) had one level of measurement whereby analytes in a sample are measured at the state in which they were introduced into the mass spectrometer (i.e. peptides), giving an MS spectra. In tandem mass spectrometry (which the majority of mass spectrometers are nowadays), analytes are measured in a first scan as afore mentioned, but subsequent to this, analyte ions are fragmented (giving fragment ions) and also measured, giving MS/MS spectra. This enables the user to determine what the constituents of a precursor ion (i.e. peptides) are, aiding in protein identification (McLafferty, 1981). Fragmentation of ions can be performed in a variety of ways, the most common of which in proteomic studies is through collision-induced dissociation (CID). CID is achieved when an inert gas such as argon, helium or nitrogen is accelerated, giving them kinetic energy, towards the target ion. When the inert gas collides with the ion, some of the energy is transferred to the ion causing bonds to break, resulting in smaller fragments of the initial ion (Mitchell Wells and McLuckey, 2005). These fragment ions are then detected and produce MS/MS spectra.

1.3.5 Data acquisition

During the mass spectrometry process there are two main types of acquisition modes; data dependant acquisition (DDA) and data independent acquisition (DIA). During DDA experiments, precursor ions are measured in an initial scan, giving an MS spectra and then subjected to fragmentation to acquire MS/MS spectra. Once an ion is analysed, it can be dynamically excluded from future scans to ensure that lower abundant peptides are also measured (Peng and Gygi, 2001). The issue with this type of acquisition is that it favours those ions with the highest abundance, so low abundance ions can be missed. Up to 84% of proteins in a complex protein mixture can remain unsampled for this reason (Egertson et al., 2013). In addition to this, because the selectivity is random there can be up to 30% variability between replicates of the same sample (Egertson et al., 2013). However, it has been found that because of this randomness, replicates of the same sample have the ability to identify additional (including lower abundance) ions (Liu et al., 2004).

A more recent type of acquisition mode is DIA which, in brief, fragments all ions to obtain MS/MS spectra for the whole sample. There are currently a few ways of doing this, one of which is named MS^E, created by the Waters corporation. MS^E does this by performing a full scan at low energy and then quickly switching to a high energy state and fragmenting all ions (Doerr, 2015, Shliaha et al., 2013). As this method fragments everything, it produces a vast array of MS/MS data with no clear indication of what the parent ions are to each fragment ion. It is for this reason that MS^E uses extra hardware, known as traveling wave ion mobility separation, within the mass spectrometer to produce more information about ions to correctly assign them to parent molecules. Traveling wave ion mobility separation separates ions by their size and shape, meaning that those with the same m/z value can be separated, thus giving more parameters to enable better identification when searching huge MS/MS data against databases (Shliaha et al., 2013). Due to the complex data output from DIA and specifically MS^E, specialised pieces of software are needed to decipher MS/MS spectra, one such program also created by the Water Corporation is Protein Lynx Global Server (PLGS). PLGS works by gathering data on MS and MS/MS spectra and identifying a single matched protein, after which all associated peptides and fragments are removed from the search; the process is then repeated for the next peptide and subsequent peptides (Shliaha et al., 2013).

1.3.6 Qualitative and quantitative proteomics

The type of data acquisition mode is largely dependent on the output desired, be it qualitative or quantitative. Qualitative proteomics aims to identify the proteins present in a sample with no quantitation. DDA experiments are often better in this scenario as when they detect a peptide of a certain m/z, they exclude it from future selection to enable better coverage of the sample, in an attempt to discover all peptides in a sample. The issue with this type of acquisition is that quantification cannot be achieved as the number of peptides identified is not proportional to the true amount of peptide in a sample. It is for this reason that DDA experiments are optimal for qualitative experiments but lack the ability to perform absolute quantitative experiments.

DIA detects a peptide each time it is encountered and so is proportional to the amount within the sample. This can be taken one step further to achieve absolute quantification by spiking in a known amount of peptide into each sample (internal standard) as a reference and calculating the amount of an unknown peptide against it using the Hi3 method (Silva et al., 2006). The Hi3 method uses the three most intense tryptic peptide ions from an internal standard to create a universal signal response factor which in turn is compared to the three most intense tryptic peptides of each protein identified to gain quantification of the identified proteins (Doneanu et al., 2012). Using DIA in this manor is a useful approach for quantitative biomarker discovery. However, due to the assumption that the amount of known peptide spiked into a sample and detection of the total amount of the known peptide using DIA are accurate, which in some cases may not be entirely correct, validation using a more specific form of quantitation is beneficial; this often comes in the form of targeted proteomics.

1.3.6.1 Targeted and untargeted proteomics

Untargeted proteomics is the method of acquiring proteomic data without targeting specific proteins. It is often carried out for biomarker discovery as it is not limited to looking at specific proteins of interest but as a means to identify all proteins in a sample. Targeted proteomics however is an approach that is used to measure specific proteins (in a sample) which have usually been determined by previous biomarker discovery methods. The main type of targeted proteomics is selective reaction monitoring (SRM)/multiple reaction monitoring (MRM) (Lange et al., 2008). The basic principle of MRM is to select peptides from a protein of interest and have an isotopically heavy labelled version of each of these peptides synthesised. Then, using a known amount of heavy labelled peptide in a sample, it is possible for one to calculate the true concentration of the native peptide (and thus the native protein) in the sample. As MRM is a targeted approach it benefits from higher sensitivity than non-targeted approaches, because the instrument is set to detect and fragment only those ions of specific m/z's dependant on the peptide being investigated.

1.3.7 Data analysis

MS spectra obtained from mass spectrometry must be processed and analysed to establish which m/z values belong to which peptides and therefore what proteins were present in the original sample. In the early days of mass spectrometry this was usually done through peptide mass fingerprinting, which is where the MS peak value is compared to theoretical mass values that could be obtained from known proteins being cleaved by the specific proteases (e.g. trypsin). In that early period of mass spectrometry, genomic databases were much smaller than they are currently and, as such, just three or four peptide matches were needed to correctly identify a protein. Nowadays however, databases are huge and ever growing and therefore the criteria for correct protein identification has become more stringent, meaning that more peptide matches and more coverage of the protein sequence are needed in order to have confidence in identifying the correct protein (Baldwin, 2004). By using CID MS/MS data, current search engines and algorithms can match real fragments in MS/MS data to a database of hypothetical fragments which could be obtained from protease cleavage from all known/hypothetical peptides. This approach enables an extra dimension of sequence coverage and specificity to protein matches, thus increasing the chances of correct protein matching and increasing confidence in the results generated by this process (Baldwin, 2004). However, the more attempts searches make at matching a fragment to a database sequence, the higher the probability of getting a false “match” by random chance. To counteract this, most search engines use a factor known as false discovery rate (FDR) whereby the data is also searched against a dummy database of either random or reversed genomic sequences; those fragments that match readily with the decoy database can be excluded at a pre-set threshold, increasing the confidence in peptide and protein matches that are genuine (Choi and Nesvizhskii, 2008).

1.3.8 Fractionation

Proteomic mass spectrometry is frequently used for complex protein/peptide mixtures, therefore samples are fractionated to produce several, less complex sample fractions prior to introduction into the mass spectrometer. Most fractionation techniques utilise characteristics specific to peptides such as their size, their isoelectric point, hydrophobicity

and polarity to separate protein/peptide mixtures. Two of the most frequently employed fractionation methods are gel fractionation and liquid chromatography (LC).

Polyacrylamide gel electrophoresis (PAGE) is a common technique used throughout biological sciences whereby samples are separated as a result of different speeds of movement through a polyacrylamide gel. Gel electrophoresis can be done in one dimension and two dimensions. In 1D gels, proteins are separated according to their size by being pulled through the polyacrylamide gel by electric current, with smaller proteins travelling further in the gel than larger proteins. In 2D gel electrophoresis, proteins are first separated by their isoelectric points by increasing pH through a polyacrylamide gel and the proteins are then separated according to their size using gel electrophoresis in the same way that they are separated in 1D gels (Hames, 1998). Once the proteins have been separated, specific bands and areas of interest can be dissected out and subjected to proteomic analysis, reducing the complexity of the input sample for proteomics.

Another method of fractionating complex samples is through liquid chromatography. There are various different types of liquid chromatography methods that can be used to separate proteins by their characteristics, these include anion exchange, cation exchange, hydrophobicity and polarity. Liquid chromatography consists of two main components, a stationary phase, most commonly an immobilised particle on the lining of a column, and a mobile phase, the liquid containing analytes which passes by the stationary phase. Anion exchange and cation exchange utilise charge of the analyte whereby negatively charged and positively charged analytes, respectively, are adsorbed to stationary phase (Niessen, 2006). The most common type of LC used in mass spectrometry is reverse phase chromatography as it couples well with ESI and is easy to achieve. Reverse phase chromatography utilises a hydrophobic stationary phase, often a silica molecule with a carbon chain (C18) attached. As hydrophobic analytes in the mobile phase pass by the stationary phase, they are adsorbed to the column and are then later eluted by a gradual increase in an organic solvent, causing peptides to elute off gradually, essentially fractionating the sample. Reverse phase chromatography can be repeated in tandem to achieve even better fractionation whereby the first column is at high pH and the second is at low pH (Yang et al., 2012). In this process, the analytes bind to the first high pH column and are eluted off in “slugs” through a stepwise increment in organic buffer. Each

incremental aliquot released from the high pH column is then further separated on a low pH column using a gradual increase around the concentration used to elute from the first column. For example, the first aliquot from the high pH column will be eluted off at 11% acetonitrile, the analytes released by this process are them captured on a second low pH column where a gradient of 9% to 13% acetonitrile is used, and thus separating that "slug" more.

1.3.9 Proteomics in cancer

Proteomics encompasses a wide variety of techniques, ranging from simple western blots to complex targeted mass spectrometry based proteomics. Before the usage of mass spectrometers in proteomics in the late 20th century, laboratories wishing to identify protein biomarkers would use a more directed single hypothesis approach which investigated one or a few proteins of interest in the relevant tissue/sample. Nowadays, many proteomics studies, especially for biomarker discovery, work on more general hypotheses (e.g. many proteins may be different) and their results are frequently hypothesis generating. Typically, mass spectrometry based proteomics identifies a number of biomarkers, which can then be validated using more targeted methods, such as MRM mass spectrometry.

Mass spectrometry based proteomics has led the way in protein biomarker discovery in recent decades (Srinivas et al., 2002). Many biomarker discoveries can be attributed to mass spectrometry based proteomics and techniques are constantly being adapted and improved (Sallam, 2015). The past few decades has seen an increase in serum biomarker discovery (Jacobs et al., 2005) in the hope of identifying markers that can be tested easily in serum and provide early diagnosis of disease. Much of the tissue-based proteomics focuses on fresh tissue because the fresher the tissue, generally the more intact the proteins are and therefore the better the results are. Nonetheless, advances in methodology has led to several proteomics studies investigating preserved tissue, including tissue preserved with formalin fixation (**Appendix 1**). It is recognised that samples that have been preserved can be problematic for downstream mass spectrometry analysis, due to protein destruction, protein modifications and/or the presence of certain "contaminating" reagents leading to reduced protein yield and accuracy.

1.4 Formalin fixed paraffin embedded proteomic studies

Many cancers identified in clinical practice are excised for diagnostic and/or therapeutic purposes. In most cases, the excised tumour is formalin fixed and then paraffin embedded (FFPE) so that it becomes ready for tissue sections to be cut with a microtome and histologically stained to confirm the diagnosis and assess any known characteristics that aid accurate prognosis. As this method of tissue processing is generally standard throughout the NHS and globally, there is a huge bank of FFPE samples available for use in clinical research studies. Coupled with detailed clinical notes and clinical outcome, these FFPE samples represent a useful biorepository of samples (Wisniewski, 2013).

Unfortunately, formalin fixation causes multiple cross-linking between proteins, limiting their analysis using standard proteomic protocols. Many research studies have attempted to perform proteomics on FFPE samples by using various methods to extract and prepare the proteins/peptides for MS (**appendix 1**). The total number of publications relating to "FFPE" and "Proteomics" has increased 54 fold since the first 4 publications in 2005 (**Figure 1.7**). **Appendix 1** is a table of a large proportion of published proteomic studies aimed at testing different methods for the use of FFPE samples in proteomics.

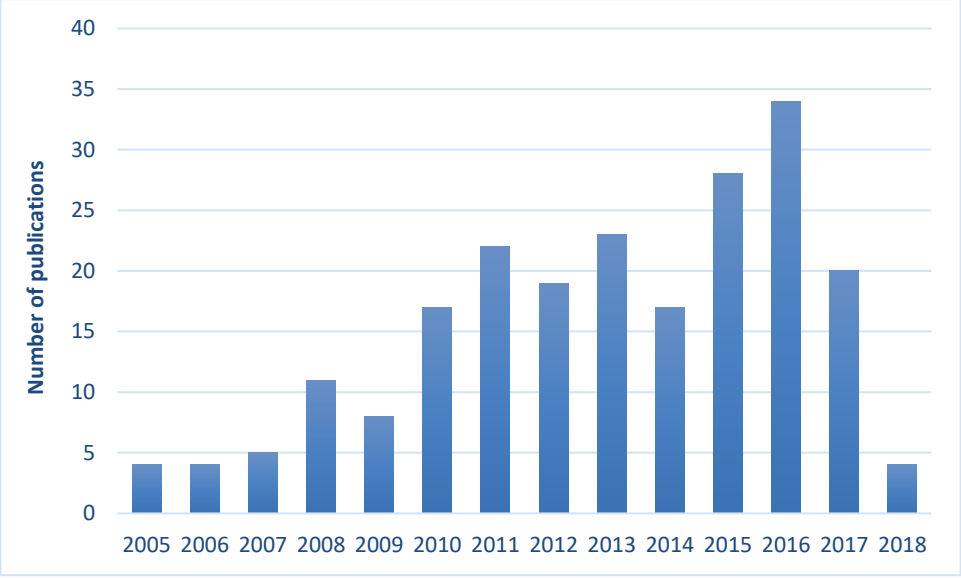


Figure 1:8: PubMed publications using key words "FFPE" and "Proteomics". Search carried out in early 2018.

FFPE, formalin fixed paraffin embedded

Chapter 1

The highest number of proteins identified in a novel FFPE proteomic methodology study in appendix 1, was 9,437 (Bateman et al., 2011). This method, like many others, utilised a FFPE protein extraction kit, specifically "Liquid Tissue". Many other studies have used this kit, albeit with slight modifications, but have achieved lower protein IDs. For instance, the highest number of protein IDs achieved by Naidoo *et al* (2012) was 1,504, Takadate *et al* (2013) 1,229, Byrum *et al* (2011) 888, Kawamura *et al* (2010) 449 and several other studies achieved less than this. The second and third highest protein yield were both from Wisniewski *et al* (2013, 2011) attaining 8,481 and 5985 IDs respectively, utilising their filter aided separation (FASP)-strong anion exchange (SAX) protocol. Wisniewski *et al*, amongst many others, utilised heat induced antigen retrieval, similar to the method used in immunohistochemistry to reveal antigen binding sites (Yamashita and Katsumata, 2017). Using heat in the presence of water and a reducing agent, for instant dithiothreitol (DTT), facilitates the hydrolysis of cross-linked bonds, thus freeing covalently bound proteins and peptides. Selected methods also attempt to utilise pressure to achieve antigen retrieval, however studies using this approach are less common and are generally less successful. An issue which has been focused on in numerous methods within studies is solubility (or lack of solubiliy, which varies between the types of tissue used in different studies) of proteins because if the sample is insoluble, the processes of reducing, alkylating and especially digestion, becomes inefficient. SDS, glycerol, polyetheleneglycol and a few other substances have been used with a degree of success. However, the use of these reagents is unfavourable as subsequent steps are required to ensure the complete removal of them because they often contaminate LC systems and saturate signals leading to poor MS spectra. Several studies have utilised RapiGest (Waters, Massachusetts), a surfactant that increases solubility of proteins and peptides as well as catalysing digestion, while the resultant solution remains suitable for analysis by LC-MS (Yu et al., 2003).

Although the studies in appendix 1 are ranked according to protein yield, it cannot be accredited solely to the methodology used, the type of tissue used as well as instrumentation play a pivotal role in the number of protein IDs. Of the different types of tissue employed, colon, liver, renal and brain were the most widely used. Protein yield seems to be indiscriminate of tissue types used as there is no clear tissue achieving more protein yields. However, because it is a fairly new technique, the number of studies to asses this is limited and as such no firm conclusions can be made. In addition to tissue type, it

seems that the LC-MS system used is fairly independent of protein yield also, but this could be due to the majority of studies utilising very similar, if not identical system set-ups.

Of the studies listed in Appendix 1, five papers used skin-based tissues, three on cSCC (Azimi et al., 2016, Foll et al., 2017, Azimi et al., 2019), and two on melanoma (Byrum et al., 2011, Byrum et al., 2013). All three papers investigating cSCC used RapiGest in their protocol as well as heat-induced antigen retrieval. Furthermore, both melanoma studies also utilised heat-induced antigen retrieval but with different buffers and reagents. 3 out of 5 of these experiments used RP-HPLC-Orbitrap setups. Due to the sensitivity and resolution of orbitraps, they are much better at achieving higher numbers of protein identifications. However, they are not well suited to perform true absolute quantification and as such often rely on targeted quantification methods, making biomarkers discovery less possible.

1.5 Bioinformatics

The vast quantity of data that proteomic studies generate requires a number of different analysis techniques to utilise all data available. Analysing biological data with the use of computational power, or *informatic*s, is widely referred to as bioinformatics and can entail a host of different methods.

Bioinformatics encompasses a number of different analysis techniques, including classical statistics, pathway analysis, machine learning and modelling and topological data analysis. Classical statistics generally comprises of parametric and non-parametric comparative tests such as T test or Mann Whitney-U test respectively. These tests are used to calculate whether there is a difference between two groups (or multiple groups in the case of ANOVA and Kruskal Wallis) by determining whether a null hypothesis is true or false through calculation of a P-value and whether the P-value is below a pre-defined threshold (usually $P<0.05$) (Wasserstein and Lazar, 2016).

Pathway analysis is a generic term for bioinformatics tools that aim to establish which pathways are likely to be involved in a biological process based on how many of the identified proteins or expressed genes involved in that pathway are increased or decreased in the biological samples being investigated. It is inclusive of, but not limited to, gene ontology analysis, protein-protein interactions and KEGG pathway analysis (Khatri et al.,

2012). Gene ontology analysis uses inputted protein or gene IDs to establish which areas of biological processes, molecular functions and components of a user defined background are over or under-represented (Khatri and Draghici, 2005). The differences identified are then given scores dependant on the measuring metric used (often a Z score), which can then be interpreted to identify important groups of proteins in a given set of gene/protein IDs. KEGG is an acronym for "Kyoto Encyclopaedia of Genes and Genomes" and includes three features; the genes database, the pathway database and the ligand database. The genes and pathway database can be used collectively by the input of several genes of interest to aid in predicting pathway enrichment (defined in their pathway database) (Kanehisa and Goto, 2000).

The term "model" can mean different things dependant on the discipline and context it is used. In terms of biological systems, it is generally used to refer to an organism that holds potential to investigate a disease or organism of interest with the intent to learn about the disease or another organism from that organism (Fields and Johnston, 2005). A mathematical model is any system utilising mathematical theories and language to calculate an unknown output. Modelling in computational biology combines these two definitions so that biological data is used to create a mathematical model using machine learning techniques. This is mainly done by one of two methods, either through supervised or non-supervised learning. Supervised learning is the creating of models using biological data with known outcome. Using the known outcome in supervised learning strategies enables algorithms to assess what features of the biological data can be used to predict the outcome on future data, where the outcome is unknown. This is known as predictive modelling and has many computational methods designed for/attributed to it, including generalised linear modelling (GLM), support vector modelling (SVM), decision trees (a full list of machine learning algorithms used in this thesis can be seen in Appendix 4). Each of these modelling techniques has its own benefits. A common way of assessing and visualising the output of data from these models is by using receiver operating characteristics (ROC) curve and the area under the curve (AUC). ROC curves investigate the true positive and true negative rates of a given model and the AUC can be used as a metric to compare models (Larranaga et al., 2006). Non-supervised learning is the creating of models using only biological data, without the input of outcome. This method requires the chosen algorithm to infer a function that exploits hidden correlations in the data. There are

several types of models which utilise non-supervised learning, including neural networks and topological data analysis. Topological data analysis is the process of analysing data using topology, which is the mathematical study of continuous space and shape. Application of topological data analysis to biological data allows investigators to look at the "space and shape of the biological data" to try to identify information in the data that classical statistics may overlook, such as the presence of different subgroups.

Proteomics offers great potential to examine for biomarkers which indicate whether metastases will develop from a primary cancer which has been recently excised. cSCC and melanoma are two important types of skin cancer which can metastasise and for which there are limited biomarkers available to assist with determining prognosis.

1.6 Hypothesis

The hypothesis for this project on cSCC and melanoma is that there is a significant difference in the protein profile between primary skin tumours that, despite excision of the primary cancer, have gone on to metastasise and primary skin tumours which have not metastasised by 5 years following excision of the neoplasm.

1.7 Aims

The aims of this PhD project are:

- To perform proteomic analysis on FFPE cSCC samples, and separately on FFPE melanoma samples, in order to examine for differences between primary tumours which metastasised and primary tumours which had not metastasised after excision of the primary tumour.
- To use the proteomics data to identifying key pathways and processes involved in metastasis of skin cancer.
- To validated discovery proteomics using multiple reaction monitoring on selected proteins.

- To use predictive modelling and machine learning to try to develop a model capable of predicting metastasis using protein biomarkers in cases where successful validation with MRM has been conducted.
- To develop a mathematical model which uses clinical and/or histological data from cSCC to predict metastasis.

Chapter 2: MATERIALS AND METHODS

2.1 Tissue samples

Tissue blocks were identified and obtained from the Histopathology Department, University Hospitals Southampton NHS Foundation Trust. Primary metastatic (P-M) tumours were selected on the criteria that they had metastasised, confirmed by histological evidence of metastasis in the Histopathology Department. Primary non-metastatic (P-NM) tumours were selected based on the fact that the patient had been seen in the Dermatology Centre, University Hospitals Southampton NHS Foundation Trust at least 5 years after their tumour excision and there was no documented evidence of metastases at that stage. The research was approved by the South Central Hampshire B National Research Ethics Service Committee (reference number 07/H0504/187).

2.2 Haematoxylin and Eosin staining (H&E)

4µm tissue sections were cut using a microtome and mounted on 3-aminopropyltriethoxysilane (APES) coated slides. Slides were deparaffinised in two washes of xylene, each for 5 minutes and then rehydrated in 100% ethanol for 10 minutes and 70% ethanol for 5 minutes followed by 3 minutes in distilled water. Slides were then stained with Mayer's Haematoxylin (MHS32 - Sigma) for one minute and then washed in running tap water for five minutes. Next, sections were stained in eosin (E4009 - Sigma) for one minute and again, washed in running tap water before being dehydrated in 70% ethanol for five minutes and 100% ethanol for 10 minutes. Sections were cleared through two washes of xylene (X/0250/17 - Fischer), each for five minutes, before cover slips were mounted, using DPX (06522 - Sigma).

2.3 Immunohistochemistry (IHC)

4µm tissue sections were cut with a microtome, deparaffinised and rehydrated, then endogenous peroxidase was inhibited by incubating slides with 0.5% hydrogen peroxide (H1009 - Sigma) in methanol for 10 minutes. Slides were washed with TBS three times, each for two minutes. Citrate buffer heat antigen retrieval was carried out by boiling slides in a

microwave at medium-high power for 25 minutes in 10mM citric acid monohydrate (C/6200/53 - Fisher), pH 6. Sections were then washed under running tap water for three minutes before being washed with TBS three times, each for two minutes. Avidin and biotin (SP-2001 - Vector) were applied separately to the slides for 20 minutes each, with three TBS washes, each two minutes, after the avidin and again after the biotin. Slides were then immersed in culture medium containing 20% FBS and 1% BSA in Dulbecco's Modified Eagle Medium (DMEM) for 20 minutes. Slides were incubated with primary antibody in culture medium overnight at 4°C. Langerhan's cells were immunostained using an anti-CD1a monoclonal antibody (M3571, clone 010 - Dako) at a dilution of 1:50. B cells were immunostained using an anti-CD20 monoclonal antibody (M0755, clone L26 - Dako) at a dilution of 1:100. After overnight incubation, slides were washed with TBS for five minutes each, then anti-mouse, biotin conjugated, secondary antibody (315-066-045-JIR) used at a dilution of 1:400 in TBS was added to the slides and left to incubate for one hour at room temperature. Slides were then washed three times in TBS, each for five minutes, before applying 3,3'-Diaminobenzidine (DAB) chromogen (K3468 - Dako) for five minutes. Slides were immediately washed with TBS and then rinsed under running cold tap water for three minutes. Slides were then counterstained with Mayer's haematoxylin (MHS32 - Sigma) for 1.5 minutes. Once counterstained, slides were washed in cold running tap water for four minutes before being dehydrated for five minutes in 70% ethanol, 10 minutes in 100% ethanol, followed by two subsequent washes in xylene, each for five minutes. DPX was then used to mount coverslips on slides.

Whole slides were imaged using an Olympus DotSlide at 20x magnification. Representative high power fields of view (at 20x) were then captured and used for analysis. Five fields of view were selected for CD1a stained sections and 10 fields of view for CD20 stained sections; this higher number of fields of view were taken for CD20 because staining was often concentrated in certain areas of the tissue, and therefore 10 fields of view enabled a more representative analysis.

2.4 Image analysis

Image analysis was conducted using either Image J (Jensen, 2013) or TMarker (Schuffler et al., 2013). Image J was used to count and calculate the percentage of immunopositive cells in the tissue on the immunohistochemistry (IHC) stained slides. This was achieved by separating the red, green, blue filters and altering the intensity to show only blue (haematoxylin) and brown (DAB) staining. Images were then converted to pixels and automatically counted to obtain the number of cells in each group. TMarker is a piece of software designed for the counting of IHC staining and is somewhat similar to Image J, but TMarker automates a lot of the process, allowing batch analysis of images. Within the fields of view, all immunopositive and relevant immunonegative cells were counted, in order to obtain the percentage of cells which were immunopositive.

2.5 Tissue microdissection

FFPE tissue samples were cut with a microtome, mounted, deparaffinised and rehydrated. 10µl of tissue lysis buffer was added to the slides and then the tumour and immune infiltrate were microdisected away from the surrounding skin tissue using a sterile hypodermic needle, and then placed into an microcentrifuge tube. The entire tumour and adjacent immune infiltrate were removed together into the microcentrifuge tube because it was considered that both these components were likely to be important in determining metastatic spread (**Figure 2.1**).

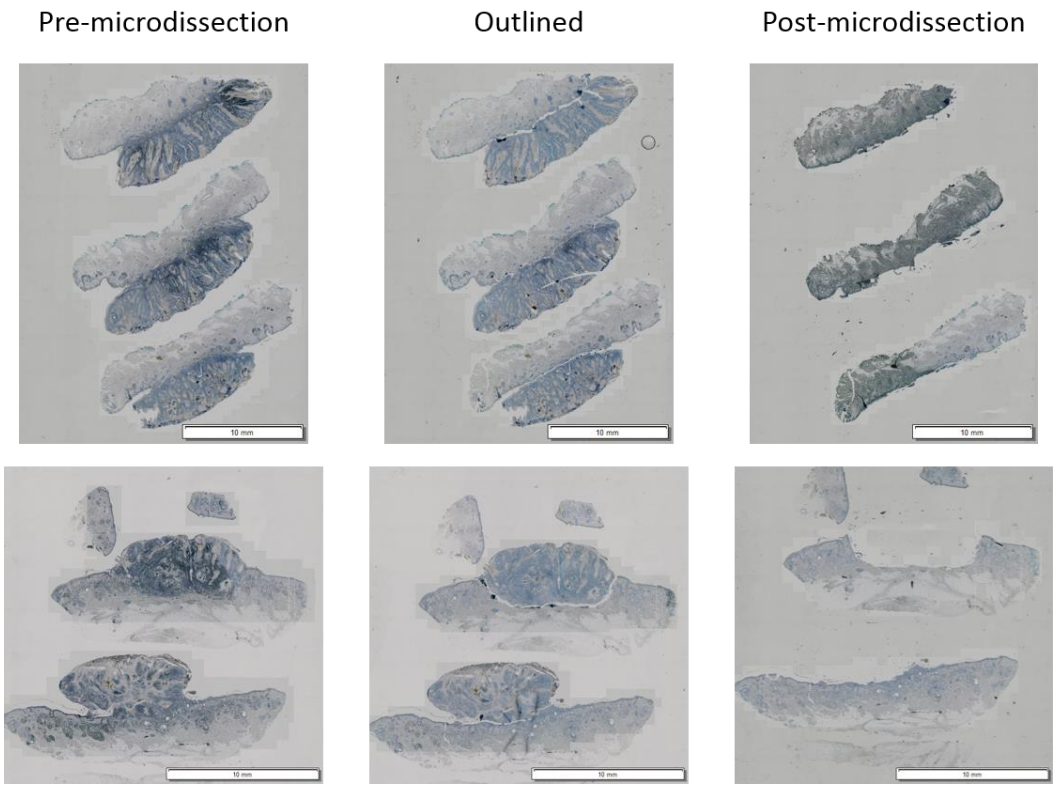


Figure 2:1 Example of microdissection of cutaneous Squamous Cell Carcinoma (cSCC) for proteomic analysis.

Left images are tissue sections of cSCC samples before microdissection. Middle images are after the tumour and relevant peritumoral immune infiltrate has been microdissected from the surrounding skin using a sterile hypodermic needle. Right images are after the microdissected tumour and relevant peritumoral immune infiltrate has been removed from the glass slide into an microcentrifuge tube.

2.6 Protein extraction from Formalin Fixed Paraffin Embedded (FFPE) samples

A range of approaches were evaluated whilst optimising the method for protein extraction but the method used for the proteomic analysis of cSCCs in this project was adapted from the technique described by Nirmalan *et al* (2011). Three 10µm sections of each tissue sample were cut with a microtome, deparaffinised and rehydrated before being counter stained in Mayer's haematoxylin. Slides were subsequently washed under running water for 3 minutes and then microdissected into microcentrifuge tubes containing 100µl lysis buffer, consisting of 0.2% RapiGest SF (186001861 - Waters), 50mM ammonium bicarbonate (09830 - Sigma) and 5mM DTT (D9779 – Sigma). Samples were then kept on

ice whilst being transferred from Dermatopharmacology to the Centre for Proteomic Research where the remainder of the protein extraction was performed. Samples were boiled in a heatblock for 30 minutes at 105°C; during boiling, the microcentrifuge tubes were weighted shut to avoid evaporation and periodically tapped to keep the liquid at the bottom of the tubes. The samples in the microcentrifuge tubes were subsequently cooled on ice for five minutes, briefly vortexed and then put into a heatblock at 80°C for two hours. After heating, samples were cooled on ice and reduced through the addition of 5mM dithioerythritol (DTE) (D8255 - Sigma) in 100µl molecular grade water for 30 minutes at 60°C. Samples were then alkylated through the addition of 15mM iodacetamide (I6125 - Sigma) in 200µl molecular grade water for 30 minutes at room temperature in the dark. Samples were subsequently digested by addition of 1µg trypsin (V5111 - Promega) (which cleaves at Lysine or Arginine) in 2µl molecular grade water and left at 37°C overnight to facilitate digestion. After digestion, 0.5% triflouracetic acid (TFA) (91700 - Sigma) was added to the microcentrifuge tubes for 30 minutes at 37°C to enable the hydrolysis of the RapiGest SF surfactant. Samples were then centrifuged at 15,000xg for 15 minutes to pellet insoluble material and precipitate. Supernatant containing the tryptic peptides were transferred to a new microcentrifuge tube and then lyophilised in vacuum using an Eppendorf Concentrator-5301 before being reconstituted in 106µl of 200mM ammonium formate.

2.7 Direct detect spectrometry for measurement of peptide concentration

Peptide concentration was measured using a direct detect spectrometer (Merck) according to the manufacturer's recommendations. Briefly, 2µl of control (200mM ammonium formate) was spotted onto the control segment of the direct detect membrane card. Three x 2µl of sample were then spotted onto the three sample segments on the direct detect membrane. A laser emitting infrared light is then pointed at each sample/control position and absorbance against wavenumber (cm^{-1}) is measured. This is used by the software to calculate the number of amine bonds and the total protein/peptide concentration.

Chapter 2

2.8 C18 peptide clean up

During the protein extraction step, samples are often exposed to various salts and other reagents to aid in solubilising and digesting protein. Using a reverse phase separation by running the sample through a C18 column, it is possible to remove a lot of this contaminating matter, which otherwise could interfere with downstream mass spectrometry analysis. This process is standard in most proteomic laboratories and has very little risk associated with it. The only potential risk is that some peptides don't adhere to the C18 column and as such are lost as they flow straight through. However, if this were to happen, it would not likely change the results as LC systems coupled to mass spectrometers use reverse phase to separate samples and therefore such peptides would be lost anyway. Samples were desalted prior to MS analysis using an Empore^{TM} C18 plate (Sigma, 66875-U). Before C18 clean up, samples were acidified, either by lyophilising and reconstituting in C18 wash buffer, consisting of 0.5% acetic acid in water, or by adding 0.2µl 100% TFA. The C18 plate was equilibrated with 100µl methanol and centrifuged at 250xg until the methanol had passed through the filter. A further 50µl methanol was then added to wells and allowed to drip through the filter, after which the sample solutions were added to the wells and centrifuged at 250xg for 10 minutes in order for the samples to bind to the membrane in their respective wells. The wells were then washed with 200µl wash buffer and centrifuged at 250xg and the filtrate discarded. This wash and centrifugation step was repeated, and the collection plate emptied again. After the 2nd wash, the collection plate was removed and a new sterile collection plate was used for collection of the "cleaned up" peptide samples. 150µl elution buffer (consisting of 80% acetonitrile, 0.5% acetic acid in water) was added to the wells and centrifuged for 10 minutes until all the solution had passed through the membrane and into the collection plate. This solution in the collection plates contained the cleaned-up peptides, which were now ready for introduction into the LC/MS system.

2.9 Mass spectrometry

2.9.1 Discovery proteomics with LC/MS^{e}

2.9.1.1 1 dimensional liquid chromatography (1D)

Prior to introduction to the LC system, 3.75µg of sample was lyophilised and reconstituted in 6µl of the buffer A (0.1% formic acid in water), specific to that method. 1D reverse phase liquid chromatography was performed using a nanoAcquity UPLC system (Waters) whereby peptides were injected and trapped onto a Symmetry C18 180µm x 20mm trap column (Waters, 186006527) and washed for 5 minutes in buffer A. Peptides were separated on a 75µm I.D x 250mm, 1.7µm particle size C18 analytical column (Waters, 186007474) using flow rate of 300nl/min and a linear gradient of 1 to 50% organic buffer B (with buffer A), (buffer A = 0.1% formic acid in water, buffer B = 0.1% formic acid in acetonitrile) over 150 minutes with a wash at 60% buffer B (with buffer A) at the end. The separation was performed on the LC system, coupled to the mass spectrometer (in a set-up known as "online separation") and sprayed directly into the nanospray source of a Waters G2-S Synapt HDMS mass spectrometer operating in MS^{e} mode with ion mobility enabled (HDMS^{E}). Alternating low (5v) and high (20v-40v) collision energy scans were enabled in ion mobility mode and data was acquired between 50 to 2000 m/z. Glu-fibrinopeptide, (m/z = 785.8426) was used as an internal calibration standard known as "lock-mass". Three blank runs were performed between each sample to ensure no carry over between samples occurred. Samples were randomly batched into groups of 12 and calibrations were performed at the beginning of each batch. At the start, middle and end of each batch an enolase standard was used to assess the performance of the machine in terms of resolution, peak width and sensitivity. Before starting each batch, the system was operated using 50% buffer A, 50% buffer B for several hours, in addition to 3 blanks, to ensure a clean column.

2.9.1.2 2 dimensional liquid chromatography (2D)

Online 2D reverse phase liquid chromatography was also performed using a nanoAcquity UPLC system (Waters), injected into a 5µl loop but then first adsorbed to a high pH column (XBridge, BEH130 C18 5µm 300x50 nano – 186003682). The first column was then eluted into 6 fractions, using 6 different compositions (11.1%, 14.5%, 17.4%, 20.8%, 45% and 65%) of buffer B. After each fraction, the eluted sample was trapped onto a low pH column, where it was subsequently eluted via a buffer B gradient directly into the nano spray source of the mass spectrometer (as stated in the 1D LC/MS^e method). The high pH pump was set at a constant flow rate of 1µl/min and had a buffer A composition of 20mM ammonium formate in water.

2.9.2 Mass spectrometry quantification

Throughout the discovery phase, the method of absolute quantification was used during LC/MS whereby 100fmol of digested enolase standard (Waters) was spiked into samples prior to LC/MS. As MS^e fragments all ions, using the Hi3 approach whereby the top three tryptic peptides of an internal standard (enolase) are correlated to the three most abundant tryptic peptides of each protein ID, absolute quantification can be achieved (Silva et al., 2006).

2.9.3 Targeted mass spectrometry

To confirm the results that were found in the discovery phase of this project, the same samples were subjected to targeted mass spectrometry in the form of multiple reaction monitoring (MRM). Furthermore, validation on previously untested samples was also carried out using MRM analysis. MRM utilises the targeted quantification strategy to accurately measure the amounts of specific peptides in a sample. Using the chromatography and mass spectra from the SCC discovery proteomics, a spectral library was created in Skyline.

2.9.3.1 Creation of spectral library

Discovery proteomic chromatograms and accompanying spectra were imported into skyline to create a spectral library. Mass spectra were searched against a protein/peptide database (in this case the human proteome) to score spectra to peptides. The result was a spectral library that consisted of tens of thousands of spectra which are associated with peptides. Spectral libraries were used to identify unique peptides for each protein of interest.

2.9.3.2 Multiple reaction monitoring (MRM)

The unique peptides identified from spectral libraries were synthesised to incorporate a >6 Dalton shift by isotopic labelling of one of the amino acids ($^{13}C_6{}^{15}N_4$ or $^{13}C_6{}^{15}N_2$) (Cambridge Research Biochemicals).

A two-fold dilution series of each heavy peptide from 400fmol down to 0.78fmol was created in buffer A to form a calibration curve. 1µg of peptides extracted from SCC samples was spiked into each dilution to serve as a background matrix. Dilution series were performed using the 1D LC/MS method previously mentioned, however, instead of MS^e, a targeted acquisition method was applied. Transition ions (that is ions that are created during the fragmentation process of precursor ions in the mass spectrometer) were selected from the spectral library created from the discovery proteomics. A targeted method was created for each peptide, using the transition ions identified in Skyline. In doing this, the mass spectrometer focusses on the light (native) and heavy (synthesised) peptides in a run. Performing a dilution series gave a calibration curve which could later be used to determine the amount of heavy and subsequently light (native) peptides in a sample.

After the calibration curves had been created, the samples used in the discovery phase were examined using the MRM method containing 100fmol of each heavy peptide. Results were imported into Skyline for analysis. Using the slope of the calibration curves, skyline calculates the actual amount of heavy peptide present in each sample as well as the ratio between heavy and light peptides. The amount of light peptide is then calculated by dividing the calculated heavy amount by the ratio of light to heavy. In addition to the

discovery set, MRM was also carried out on a validation set of samples which were previously untested.

2.9.4 Data processing

Five minutes of extracted ion chromatogram was inputted into threshold inspector (Waters) to determine correct high and low collisional energy thresholds to filter out as much noise as possible and allow the highest number of peptide identifications. Once established for each batch, thresholds were set and samples were processed using Waters Protein Lynx Global Server (PLGS) ver 3.0.3. The 6 fractions that 2D LC produced were all processed individually and then subsequently merged in PLGS. Processed files were then searched against the human UniProt – SwissProt protein database (November 2016). During searching, using PLGS, a workflow was set to allow only those peptides which acquired 3 or more ions, proteins identified from 1 or more peptides and proteins that had 7 matched products for identification. PLGS Primary digest reagent was set to trypsin and 1 missed cleavages were allowed. Peptides can often become modified during the extraction process through oxidation and deamidation. It is therefore necessary to add these modification to a variable modifications list so that if they are indeed modified, they still get identified. Variable modifier reagents were deamidation of asparagine and glutamine along with oxidation of methionine. It has also been found that methylol groups (hydroxymethylation) of cysteine is often present in FFPE tissue (Metz et al., 2004) and therefore was also included as a variable modification. Fixed modifier reagents were carbamidomethylation of cysteine residues.

2.10 Data pre-processing and Statistical analysis

A matrix consisting of sample ID in rows and protein identifications, with abundancy values, in columns was created. This matrix was then imported into Inferno, an R package created for analysis of proteomic data.

2.10.1 Missing values

Many 'omics studies results contain missing values due to various reasons. There have been several methods developed to attempt to counter this, some of which rely on creating random numbers within standard deviations and some more elaborate that attempt to encompass all available data to create and impute numbers. However, due to the complexity of imputing data and allowing consideration for why data was originally missing (too low to detect, not present, or by random chance wasn't sampled), it has been suggested that ultimately, no imputation results in higher statistical power and confidence (Bantscheff et al., 2012, Webb-Robertson et al., 2015). Nonetheless it is impractical to use protein IDs which only have one value and therefore a threshold is still required. It has been reported that up to 50% of data is missing in 2D gel electrophoresis proteomics and that this has no effect on statistical analysis (except on correlational studies) (Jung et al., 2005). Furthermore, in an FFPE proteomic study looking at cSCC, the authors used only protein data which appeared in 50% or more of samples (Foll et al., 2017). The current study therefore allowed up to 49% missing values per protein ID, any protein which appeared in less than 50% of samples was not included in statistical analysis, thus ensuring results in which one can have high confidence in the analysed data.

2.10.2 Normalisation

Although the technique of absolute quantification allows direct comparison between samples, normalisation of data is first required to ensure an equal comparison. In this study, a total protein concentration normalisation strategy was carried out. This was performed by calculating the median of the values within a sample and dividing each value by this median. Median rather than mean was chosen because proteomics data often has a 'floor effect' whereby values are only given above a certain abundance (dependant on how sensitive the machine is), therefore creating a non-normal distribution (**Figure 2.2**). Histograms were created for all samples to assess the Gaussian distribution of normalised data using Inferno, manually setting the bins to 25.

2.10.3 Histograms of p-values

The non-parametric Mann Whitney U test for significance was used to test differences between the primary metastatic (P-M) and primary non-metastatic (P-NM) groups. Statistical advice (by Research, Design and Methodology, University of Southampton) recommended plotting all p-values obtained in a histogram to assess the confidence and false positive rate. This was performed using Prism software.

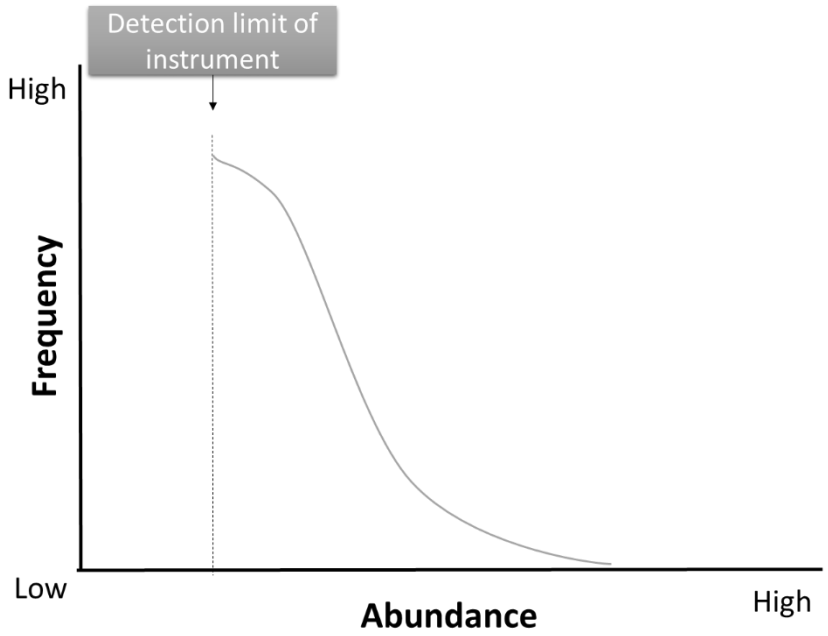


Figure 2:1: 'Floor effect' often produced by proteomic data.

Much of proteomic data suffers from the 'floor effect' whereby the lower values are at higher abundance due to instrumentation measurement limitations

2.11 Bioinformatics

As proteomic results yield a lot of data, many different bioinformatics approaches are required to fully interrogate the data. Most of the bioinformatics analysis was carried out using R unless otherwise stated.

2.11.1 Time to metastasis plots

Time to metastasis plots were created using clinical data where time zero is the date of excision and time is the number of days until metastasis was identified (up to a five year period) where a binary operator was used, i.e did metastasis occur, yes/no. Where

metastasis was present at the day of excision (i.e through nodal biopsies or CT scan), time to metastasis is 0. As P-NM samples never metastasised (during the 5 year period which was the criteria for P-NM samples), they had a set time of 1,825. Plots were created using the survminer and survival packages in R.

2.11.2 Volcano plots

Volcano plots were plotted within R using log10 p-values and log2 fold changes. Log10 p-values were calculated from p-values obtained using the Mann Whitney U test for significance comparing P-M and P-NM abundancies for individual proteins. Log2 fold change values were calculated by subtracting the specific protein value medians between P-M and P-NM, and then log2 transforming these data. Coordinates that had p-values greater than 0.05 and log2 fold changes less than 1 were coloured red. Coordinates that had p-values greater than 0.05 and log2 fold changes greater than 1 were coloured black. Coordinates that had p-values less than 0.05 and log2 fold changes less than 1 were coloured orange. Coordinates that had p-values less than 0.05 and log2 fold changes greater than 1 were coloured green. Coordinates that had p-values less than 0.01 and log2 fold changes greater than 0.5 were coloured blue. Coordinates that had p-values less than 0.001 were coloured pink. Coordinates were labelled using Uniprot accession numbers

2.11.3 Search tool for the retrieval of interacting genes/proteins (STRING) analysis

Interactions between genes/proteins were assessed using STRING analysis (Szklarczyk et al., 2015). Significantly differentially expressed proteins ($p<0.05$) between P-M and P-NM were inputted into STRING analysis software. A medium confidence score of 0.4 was set as an allowance parameter for associations, as suggested by the software. Individual proteins were mapped as nodes with lines representing a contribution to a shared function; thicker lines indicate a stronger confidence in the interaction. KEGG pathway analysis was then mapped on top of these created structures to identify significantly enriched areas.

2.11.4 Gene ontology analysis

Gene ontology analysis was carried out using GoGorilla gene enrichment analysis (Eden et al., 2009), whereby the list of significantly differentially expressed proteins were imported into this software programme. A two unranked list of genes approach was used, with the human Uniprot-SwissProt database set as the background proteome. Output was shown in reduced and visualised gene ontology (REViGO) format (Supek et al., 2011). The REViGO R script for generating treemaps was downloaded and adapted.

2.11.5 Weighted gene co-expression network analysis (WGCNA)

Weighted gene co-expression network analysis (WGCNA) was carried out on the proteomic data using the WGCNA package in R. The mean connectivity and scale free independence of the data was assessed to identify a suitable soft threshold to use when creating the similarity and adjacency matrix. A soft threshold is a value used to power the correlations between genes to highlight more significant connections and reduce noise. A topological overlap matrix (TOM) was created and used to carry out hierarchical clustering and module identification. Modules were then correlated to clinical/histopathological traits in addition to analysis through KEGG pathway enrichment. A brief overview of the pre-processing steps involved in WGCNA can be seen in **Figure 2.3**.

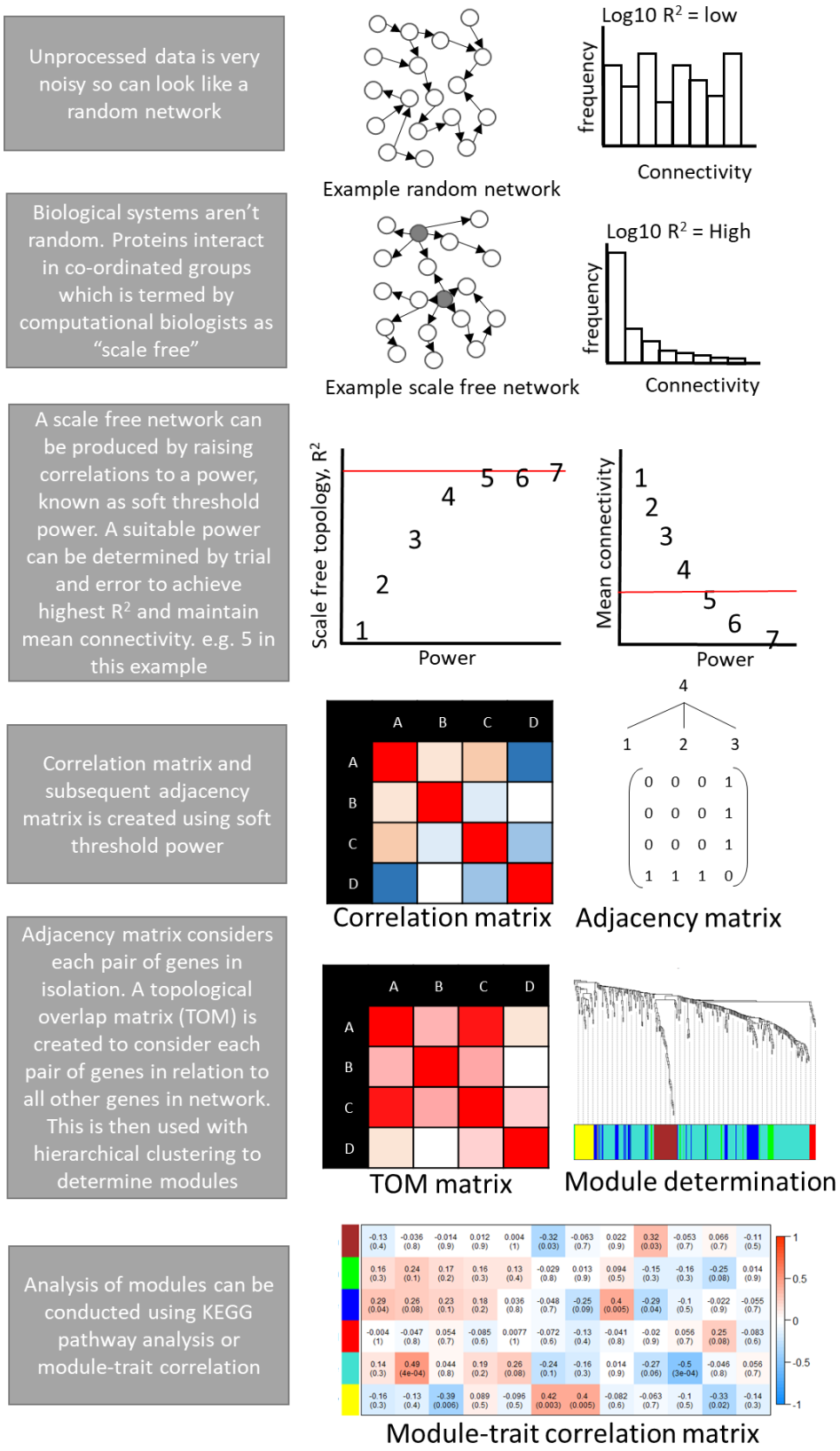


Figure 2:3: Overview of weighted gene co-expression network analysis

2.11.6 Topological data analysis

Topological data analysis (TDA) is an alternative approach to exploring data by looking at the shape of the data as opposed to direct comparisons. To create a topological structure, sample proteomes are correlated to other sample proteomes using any one of a number of algorithms (i.e. regression, hamming, etc) and then clustered in accordance to their similarities. Ayasdi is a machine intelligence software that allows topological modelling of large datasets. Differentiation, depth and diameter of samples (recorded during tumour excision) were used to generate a topological structure in Ayasdi, using a hamming metric and neighbourhood lenses. Gain and resolution was set at 35 and 5.5 respectively. Outcome was then mapped on top of the structure created from differentiation, depth and diameter.

Normalised proteomic abundancy data with a 50% missing value threshold was used to generate a topological structure in Ayasdi. This was repeated for 1D data and 2D data. Subsequent structures were mapped for outcome, differentiation, depth and diameter. Outcome revealed separate P-M and P-NM groups within the structure. These groups were analysed using Kolmogorov-Smirnov test for significance to identify proteins that differed in the P-M and P-NM groups.

2.11.7 Predictive modelling

Predictive modelling was carried out using the statistical programming language R. Packages caret, caret ensemble, pROC and doParallel were used in the production of predictive models. Data was split into training and test sets with models trained on the training set using 10 fold cross validation repeated 3 times. The model was then used to predict the outcome in the test set. A full list of the algorithms used in this thesis can be found in **Appendix 4.**

Chapter 3: Proteomic characterisation of cutaneous squamous cell carcinomas (cSCC)

3.1 Introduction

Cancer results from the dysplastic growth of mutated cells that form tumours which subsequently invade into surrounding tissue (Hanahan and Weinberg, 2000). In healthy cells, transcription and translation are meticulously maintained to ensure homeostasis within the cell and to aid in its ability to uphold its functions. DNA damage and mutations, as occurs in the early stages of cancer development, can cause alterations in protein expression as well as alterations in protein functions, both of which can have critical effects on the cell (Hanahan and Weinberg, 2011). Dysregulation in protein abundancies within a cell can lead to dysplastic behaviour and an inability to perform its intended function (Le Quesne et al., 2010). This abnormality in protein expression can contribute to the development and progression of cancer (Hanahan and Weinberg, 2000, Hanahan and Weinberg, 2011). Cancerous cells possess the ability to rapidly proliferate and invade, but in the early stages of cancer development they usually lack the ability to metastasise. This ability comes through clonal evolution within the tumour whereby additional genetic alterations and selective pressures cause genetically diverse subclonal populations to form (Greaves and Maley, 2012). Cells from some of these genetically diverse subpopulations can then spread to other organs via the lymphatics or blood vessels and form metastases (Brodland and Zitelli, 1992, Greaves and Maley, 2012).

Although many studies have been performed on the genetics of cSCC development (Li et al., 2015, Pickering et al., 2014, South et al., 2014, Durinck et al., 2011) few studies have looked at identifying differentially expressed proteins in cSCC (Dang et al., 2006). The vast numbers of mutations in cSCC (i.e. approximately 50 mutations per megabase) (South et al., 2014) means that it may be challenging to identify genetic markers predictive of the development of cSCC metastases. However, the fact that genes “instruct” cells how to behave, whereas proteins “carry out” those instructions, it is possible that it may be easier to find markers which predict cSCC metastases using proteomics. One proteomics study that used FFPE cSCC samples identified a number of proteins differentially expressed in

cSCC compared to normal skin; these included tenascin, vinculin, calmodulin like protein 5, IQGAP1 and transgelin (Azimi et al., 2016). However, this study by Azimi *et al* (2016) was limited in its sample size and it did not investigate the relation to metastasis. Very recently however, a larger study by the same group (Azimi et al., 2019) did investigate the relation between precancerous legions and cancer by examining Bowen's disease, actinic keratosis and cSCC. Nonetheless, again, they do not investigate cSCC metastasis or relate their findings to it.

Determining what proteins are involved in the progression from the primary cancer to metastatic cancer is vital for identifying patient prognosis. At present, cSCC patients are separated into high and low risk categories according to the size, depth, differentiation, and site of the tumour, whether the tumour shows perineural or perivascular invasion, and the immunocompetence of the patient. The identification of protein biomarkers which are involved in cSCC metastasis would lead to a better ability to identify those patients likely to develop metastases from cSCC and might also lead to improved treatment strategies.

Mass spectrometry based proteomics is at the forefront of biomarker discovery (Reymond and Schlegel, 2007) and has been used in the investigation of breast (Gast et al., 2009), colon (Ward et al., 2006), pancreatic (Koopmann et al., 2004) cancers and indeed many more. Studies have identified the potential to undertake mass spectrometry based proteomics in FFPE tissues (see table in Appendix 1), which opened up opportunities to undertake cancer studies on FFPE samples. This current study aimed to identify biomarkers relevant to metastasis in cSCC using a proteomic approach by looking for differentially expressed proteins between primary cSCC which subsequently metastasised (P-Ms) and primary cSCCs which had not metastasised at 5 years after excision (P-NMs).

3.2 Materials and Methods

A total of 89 samples were used for this part of the study, consisting of 44 P-M samples (24 P-M samples for proteomics) and 45 P-NM samples (24 P-NM samples for proteomics). Immunohistochemical staining and proteomic analysis was performed to identify factors within cSCC that contribute to metastasis

3.2.1 Immunohistochemistry and image analysis

P-M and P-NM samples were immunohistochemically stained according to methods outlined in chapter 2.3. Briefly, sections were cut using a microtome before deparaffinization and rehydration in xylene and ethanol, respectively. Sections were then subjected to heat induced antigen retrieval and blocked, probed with a primary antibody before being washed and probed with a secondary antibody. Images of sections were taken on an Olympus DotSlide microscope. ImageJ and a software developed for quantifying IHC (TMarker) were compared to assess which is more effective for future quantification, of which the superior one was used for all future IHC analysis (full materials and methods outlined in chapter 2.4).

3.2.2 Proteomic analysis of cutaneous Squamous Cell Carcinoma (cSCC) samples

Two cSCC samples with undetermined outcome (P-M or P-NM) were used when optimising the protein extraction method. The best performing method was used for protein extraction of 24 P-M samples and 24 P-NM samples of which full materials and methods can be found in chapter 2.6. Samples were quantified using a Direct Detect infrared spectrometer outlined in chapter 2.7 and cleaned up using a C18 reverse phase technique (full material and methods can be found in chapter 2.8). Samples were then analysed using a Waters Synapt G2-Si high resolution mass spectrometer using the methods described in chapter 2.9.

3.2.3 Bioinformatics and data analysis

Protein concentrations were normalised as described in chapter 2.10.2. Statistical analysis was performed on the results by comparing P-M data to P-NM data. Whole proteome analysis was carried out through the use of volcano plots (methods in chapter 2.11.1), topological data analysis (methods in chapter 2.11.5) and predictive modelling (methods in chapter 2.11.6). Significantly differentially expressed proteins were further analysed using STRING, gene ontology and WGCNA as outlined in chapters; 2.11.2, 2.11.3, 2.11.4, respectively.

"Time to metastasis" plots were created using R and the packages survminer and survival. Time to metastasis was deduced from patient records where the start point was the day of

excision of the tumour except in cases where the patient presented with metastases. If metastasis was present from initial presentation then time to metastasis was 0. P-NM samples were set a consistent time to metastasis at 1,825 days (5 years), which was the cut-off used to define P-NM samples. High and low expression was defined as either above or below the median, respectively. P values were obtained by log-rank test.

3.3 Results

3.3.1 Clinical and histological characterisation of samples used for immunohistochemistry

44 primary metastatic (P-M) and 45 primary non-metastatic (P-NM) tumours were used in the IHC staining of cSCC. A summary of the samples used can be seen in **Table 3.1**. Briefly, there were more samples from male patients, P-M tumours were more poorly differentiated than P-NM tumours, and P-M tumours consisted of more samples reporting perivascular invasion, perineural invasion and/or in immunosuppressed subjects. The average depth and diameter were larger in P-M than P-NM samples. Information on geographic ancestry was not collected and therefore was not available

Table 3.1: A table showing clinical and histological details of cutaneous Squamous Cell Carcinoma (cSCC) samples used for immunohistochemistry staining.

	P-M	P-NM
number of Samples	44	45
Male	33 (75%)	32 (71.11%)
Female	11 (25%)	13 (28.89%)
Well differentiated	1 (2.27%)	14 (31.11%)
Moderately differentiated	13 (29.55%)	27 (60%)
Poorly differentiated	30 (68.18%)	4 (8.89%)
Perivascular invasion	9 (20.45%)	1 (2.22%)
Perineural invasion	9 (20.45%)	2 (4.44%)
Immunosuppressed	7 (15.91%)	3 (6.67%)
Avg tumour depth (mm)	7.70 ± 5.34	4.30 ± 2.79
Avg Tumour diameter (mm)	29.35 ± 31.39	12.00 ± 8.27

P-M, Primary metastatic. P-NM, Primary non-metastatic.

Initially, several cSCCs were stained with H&E to help gain histological experience, using microscopy, in recognising the relevant parts of cSCC samples. This included recognition of the tumour cells and the peritumoral immune infiltrate and being able to distinguish these from the surrounding normal skin tissue (**Figure 3.1**). This skill was important for subsequent analysis of immunohistochemical staining and for the acquisition of the relevant cSCC tissue for proteomic analysis.

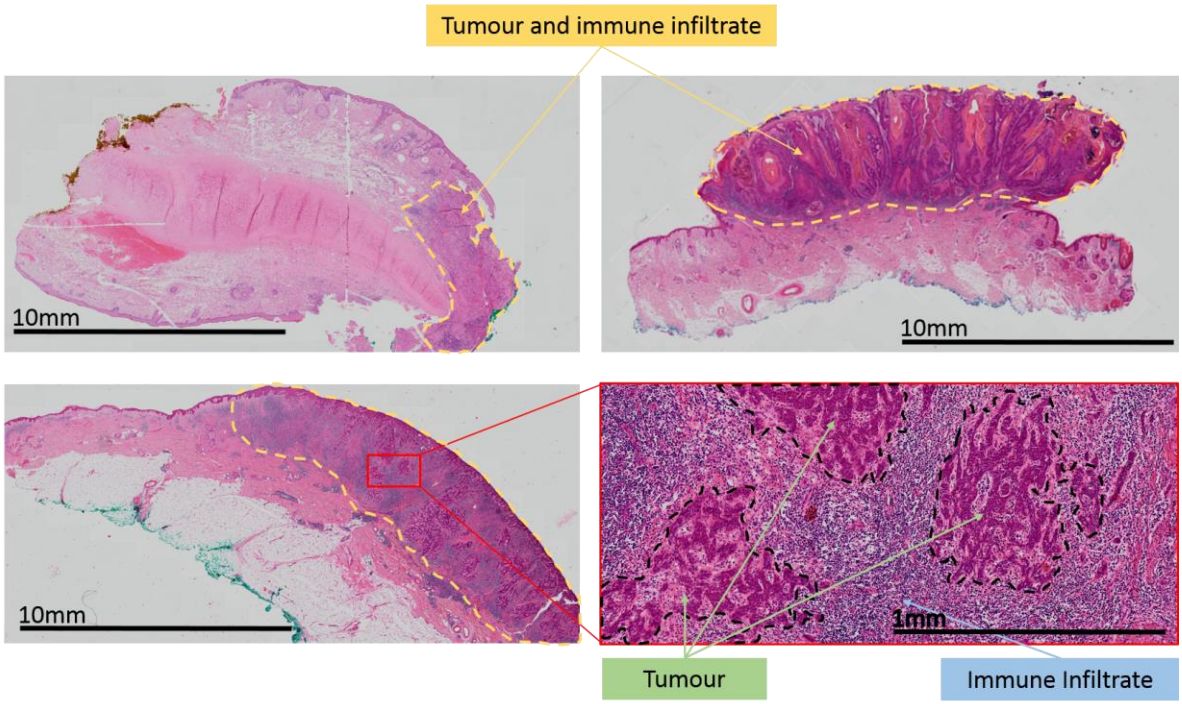


Figure 3:1: Haematoxylin and eosin staining of SCCs which allows identification of the tumour and peritumoural immune infiltrate

3.3.2 Image analysis and quantification of CD20+ cells in cutaneous Squamous Cell Carcinoma (cSCC)

Previously in Dermatopharmacology, P-M and P-NM cSCCs were immunostained for CD8 and FOXP3, and it was found that lower numbers of CD8+ cells and higher numbers of FOXP3+ cells (i.e. regulatory T cells) were present in P-Ms than in P-NMs. This previous work was carried out using ImageJ to aid in counting and analysis of images. However, TMarker is an alternative image analysis programme which enables batch analysis and so speeds up image processing time dramatically as well as removing some observatory bias. Following immunohistochemical staining of 20 cSCCs, quantification by TMarker and ImageJ were compared, and TMarker seemed superior to ImageJ in relation to the values obtained by manual counting of immunopositive cells, achieving; 0.87, 0.99 and 0.99 R values in total number of cells, number of immunopositive cells and percentage of cells which were immunopositive, respectively (**Figure 3.2**). Therefore, TMarker was used in this project for the counting of cells in P-M and P-NM tissue sections which had undergone immunohistochemical staining.

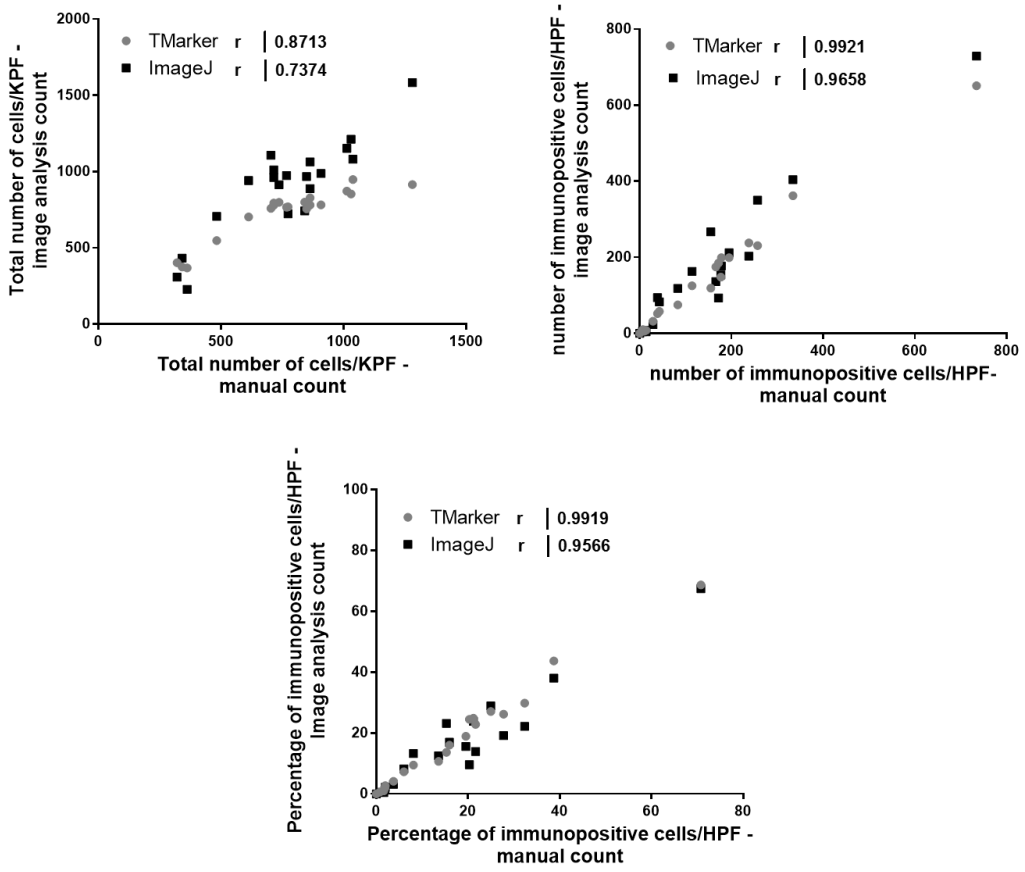


Figure 3:2: Comparison between TMarker and ImageJ in relation to manual counting.

20 high power field (20x magnification) images from 20 cSCCs that had been immunostained for CD20 were counted using TMarker and ImageJ software and compared to manual counting (R values). The results suggest that TMarker is an appropriate software for counting the immunopositive cells and total cells in cSCC sections which have undergone immunohistochemistry.

3.3.3 CD20+ and CD1a+ cells in P-M and P-NM cutaneous squamous cell carcinoma (cSCC)

P-M and P-NM cSCCs were immunostained separately with anti-CD20 and anti-CD1a antibodies to calculate the percentages of B cells and Langerhans cells respectively in these samples. Previous work in Dermatopharmacology investigating CD8+ and FOXP3+ cells in cSCCs had quantified the numbers of immunopositive cells in 5 high power fields (HPFs), however, in the current project CD20+ cells were concentrated in certain areas of the peritumoral infiltrate, whereas CD1a+ cells were scattered throughout the tumour and peritumoral infiltrate. For this reason, in order to obtain a representative analysis of CD20+,

and separately CD1a+, cells in P-M and P-NM cSCCs, quantitation of immunopositive cells was conducted on 10 HPFs for CD20 and on 5 HPFs for CD1a.

There was no significant difference in the percentage of CD20+ cells between the P-Ms and the P-NMs (P=0.1238, Mann-Whitney U). The P-M group was found to show a median of 7.42% of peritumoural cells staining CD20+ (IQR= 1.603 to 21.55, n=44) and the P-NM group showed a median of 15.32% of peritumoural cells staining CD20+ (IQR= 3.28 to 26.76, n=45) (**Figure 3.3**).

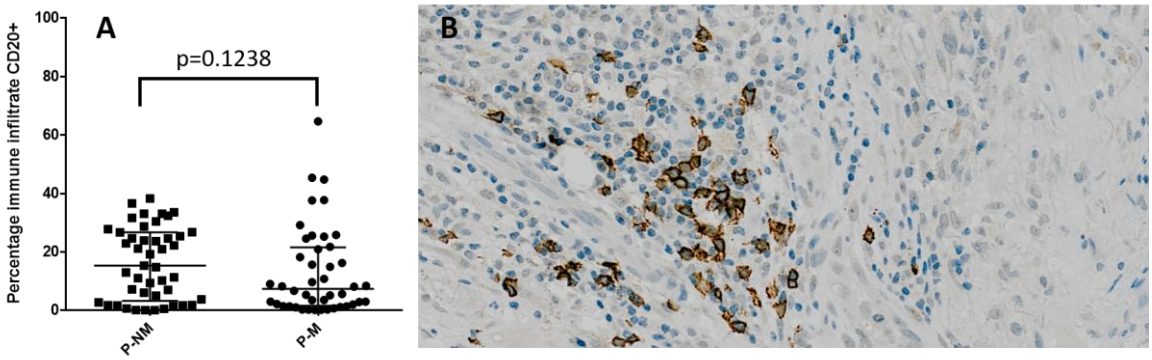


Figure 3:3: immunohisochemical staining of P-M and P-NM tumours for CD20.

P-M and P-NM tumours were immunostained for CD20+ cells, and quantification of staining was calculated using TMarker (10 high power fields of view at 20x magnification for each tumour). **A**. Graph showing results of CD20+% staining of peritumoural cells r (P-NM, n=44; P-M, n=45; P=0.1238). Each dot represents single tumour. Error bars show interquartile range with median plotted on them. A Mann-Whitney U test was performed for statistical significance **B**. Representative image of CD20+ staining in P-M group.

Immunohistochemical staining demonstrated CD1a expression within (intratumoural) and adjacent to the malignant keratinocytes (peritumoural) and therefore both were quantified independently. P-NM and P-M had a median of 2.26% (IQR=0.865 to 3.43, n=45) and 0.645% (IQR= 0.1025 to 1.583, n=44) CD1a+ cells intratumorally respectively (P = 0.0003, Mann Whitney U test). Medians of 0.4400% (IQR= 0.220 to 0.9050, n= 45) and 0.16% (IQR=0.04 to 0.6725, n=44) of CD1a+ cells were present peritumorally in P-NM and P-M respectively (P=0.0045, Mann-Whitney U test) (**Figure 3.4**).

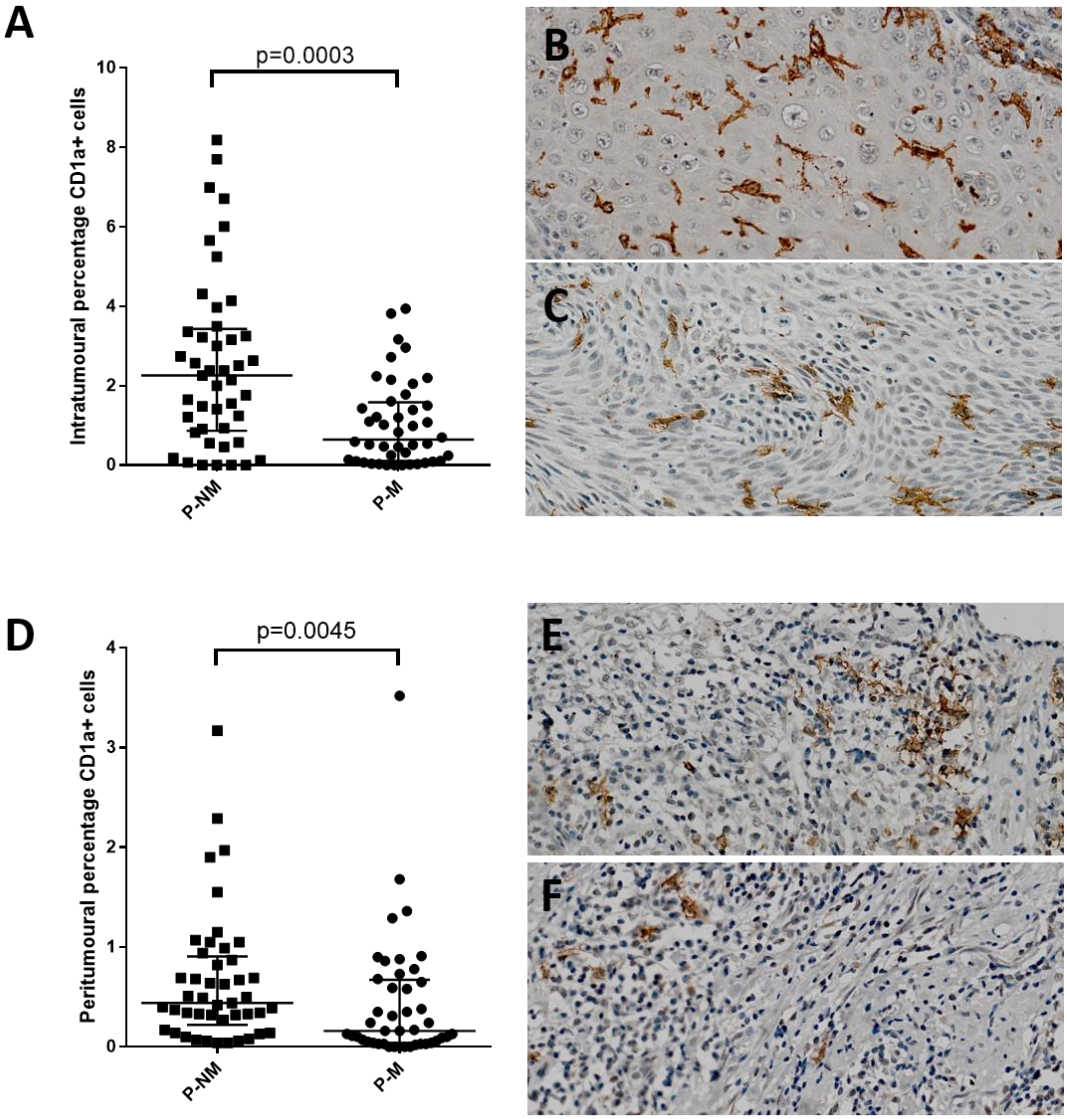


Figure 3:4: Immunohistochemical staining of P-M and P-NM tumours for CD1a

P-M and P-NM were stained using an anti-CD1a antibody. Positive immunostaining in images was quantified using TMarker (5 high power fields of view at 20x magnification for each tumour). **A**. Graph displaying percentage of intratumoural CD1a+ cells (P-NM, n=45; P-M, n=44; P=0.0003). **B** and **C**. Representative images of intratumoural CD1a+ staining in P-NMs and P-Ms respectively. **D**. Graph displaying percentage of peritumoural CD1a+ cells (P-NM, n=45; P-M, n=44; P=0.0045). **E** and **F**. Representative images of peritumoural CD1a+ staining in P-NMs and P-Ms respectively. In A and D, each dot represents single tumour, and error bars show interquartile range with median plotted on. A Mann-Whitney U test was performed for statistical significance. Results are shown as percentage of total number of immune cells which stained positive for CD1a.

3.3.4 Association of CD20+ cells and CD1a+ with time to metastasis in cutaneous Squamous Cell Carcinoma (cSCC)

A “time to metastasis” plot was developed to determine if there is a relationship between the numbers of CD20+ and/or CD1a+ cells and the time taken for a cSCC to metastasise. Higher amounts of CD20 were significantly associated with less chance of metastasis (p=0.027, Log-rank test) **Figure 3.5**.

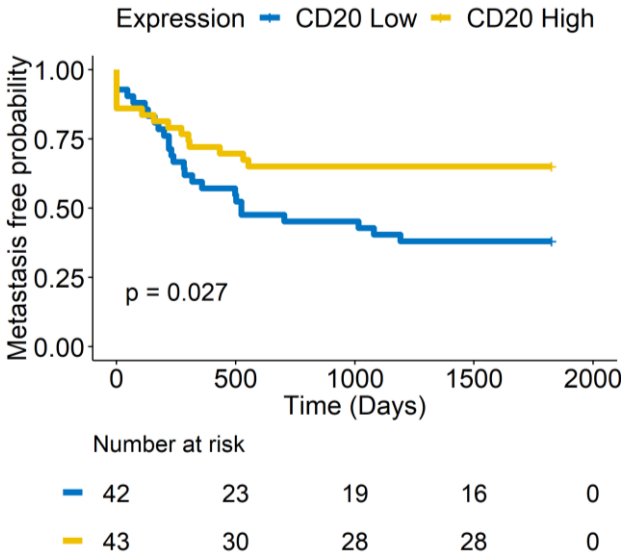


Figure 3:5: The effect CD20 expression has on time to metastasis

“Time to metastasis” plots were created using date of excision as start time. CD20 high and low expression was determined by above or below the median, respectively. P value obtained by Log-rank test.

Furthermore, higher levels of intratumor CD1a staining was significantly associated with a decreased risk of developing cancer (p=0.011). However, there was no association between CD1a+ peri tumoural cells and time to metastasis (p=0.17) **Figure 3.6**.

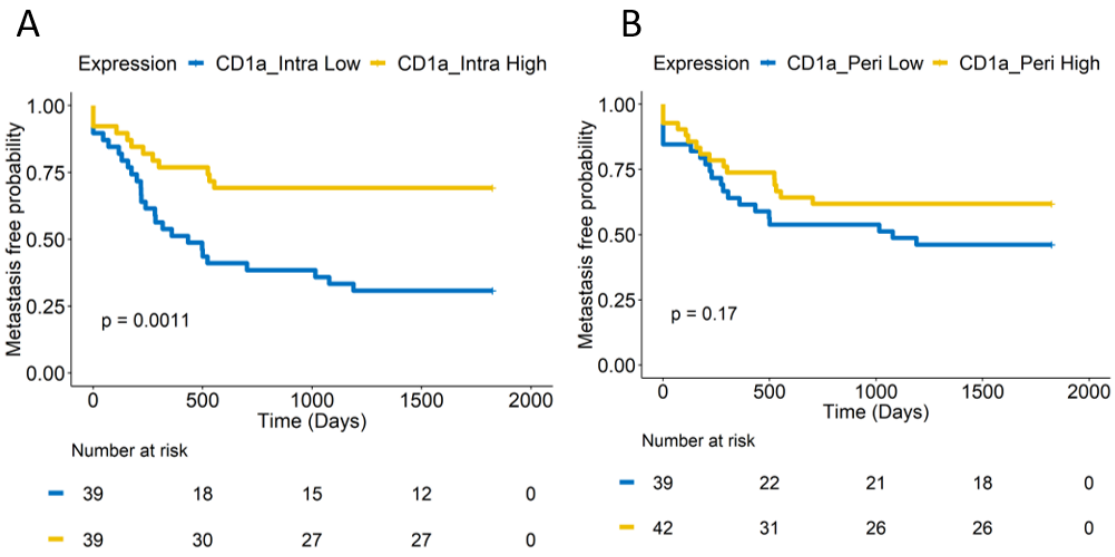


Figure 3:6: The effect CD1a expression within and around the tumour has on time to metastasis.

"Time to metastasis" plots were created using date of excision as start time. CD1a expression was determined as high or low according to the value being above or below the median, respectively. P value obtained by Log-rank test.

3.3.5 Optimisation of tissue sample preparation and fractionation technique for subsequent proteomic investigation.

The next aim of the project was to determine whether proteins could be successfully extracted and fractionated from the cSCC samples. The extraction and processing of proteins from cSCCs proved difficult as a result of the formalin fixation of these FFPE samples. For this reason, different techniques were tested in combination with fractionation methods to establish the most suitable approach (**Figure 3.7**). Two different fractionation techniques were tested on pre-digested cell lines; strong anion exchange (SAX) (Wisniewski, 2013) and online 2D fractionation (**Figure 3.7**). SAX performed poorly and identified fewer than 1,000 unique proteins. 2D fractionation produced higher protein yields (≥2,000 proteins) than SAX, with protein numbers similar to those often achieved from pre-digested cell lines by other researchers in the Centre for Proteomics, University of Southampton and was therefore considered superior to SAX. Although only two fractionation methods were used, a variety of extraction methods were assessed. Based on studies by Wisniewski (Wisniewski, 2013, Wisniewski et al., 2011), a filter aided separation protocol (FASP) was attempted. FASP was tested in conjunction with SAX fractionation and

varied widely in terms of protein IDs and was therefore considered not adequately reproducible. The FASP/2D method was tested on frozen and FFPE cSCC samples as well as on fresh and frozen normal skin. The data indicate that this method was more suitable for FFPE protein extraction. Furthermore, the FASP/2D method achieved higher protein yields from cSCC than from normal skin. Nonetheless, although this method produced protein yields higher than 1,000 proteins, its variability was very large and therefore was not considered reliable. The FASP/2D technique was also tested utilising different boiling times but these changes seemed detrimental and did not improve protein yield.

The final method tested was one based on Nirmalan et al. (2011) which utilised RapiGest surfactant, coupled to 2D fractionation. This method produced consistent results, resulting in a median of 772 protein identifications. In addition to yielding higher and more reproducible protein ID numbers, the RapiGest method was superior to others as it was much cheaper, faster and easier to undertake. Furthermore, FASP and SAX methods resulted in a large amount of insoluble material being discarded, whereas the RapiGest method produced minimal insoluble material.

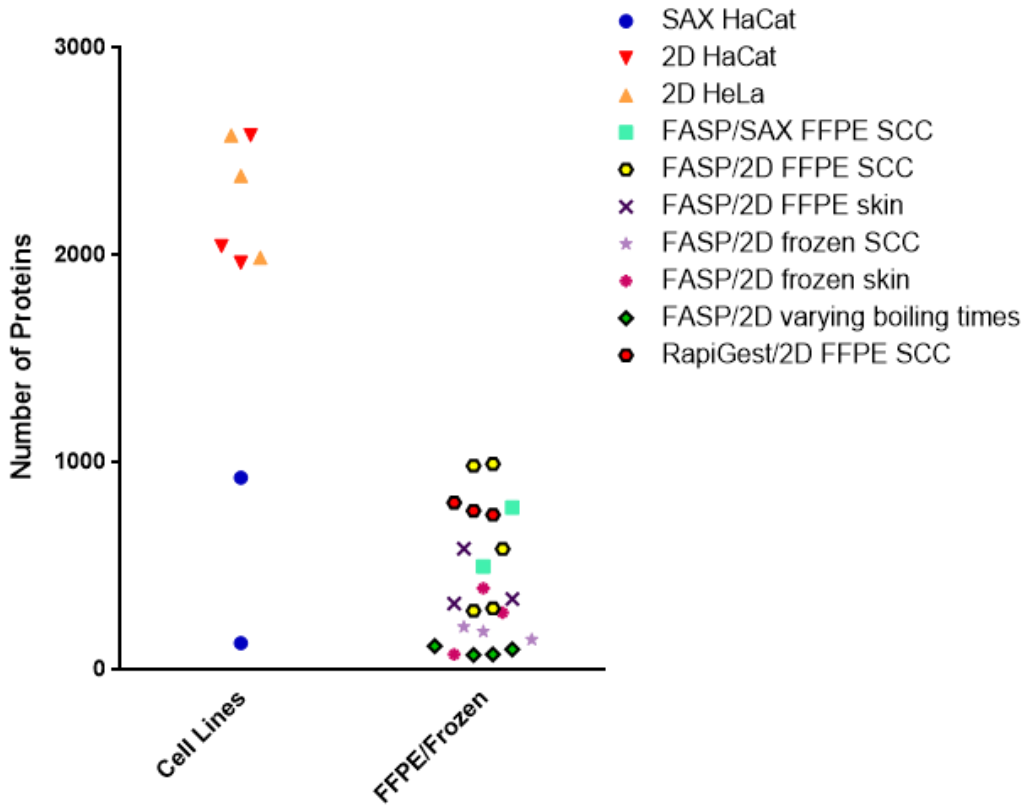


Figure 3:7: Varying combinations of extracting and fractionating samples were tested to identify a suitable methodology

Two cSCC samples with undetermined clinical outcome, in addition to matched skin samples, were used to optimise the protein extraction method from FFPE tissue. Cell lines (HaCaT and HeLa) which had not been fixed in formalin nor paraffin embedded were used for comparison purposes. Analysis was conducted using a nano aquity UPLC system (Waters) and a Waters Synapt G2-si high resolution mass spectrometer. 2D fractionation gave a higher number of protein IDs than SAX in cell lines. FASP methods gave varying results in conjunction with both SAX and 2D fractionation in cSCC and skin samples. SAX, Strong anion exchange. FASP, Filter aided separation protocol.

3.3.6 Verification of RapiGest method

MS technical repeats (i.e. 3 experiments using the same protein extraction sample) and biological repeats from the same sample (i.e. 3 independent protein extractions using the RapiGest method from the same tissue samples) were carried out to examine the reproducibility. Two SCC samples were used and named A and B. A1, A2, A3 and B1, B2, B3 refer to independent days of extraction, where A1a/A2a/A3a, A1b/A2b/A3b and A3a/A3b/A3c are reference to triplicate repeats within an independent extraction. The

results of the three RapiGest/2D experiments in **Figure 3.7** were compared to assess the technical reproducibility of the MS method. **Figure 3.8** shows a Venn diagram visualising the proteins' IDs that were shared between the three experiments. 48.2% of the protein IDs identified were similar in all three cases and 66.2% of protein IDs identified were detected in at least 2 experiments. A coefficient of correlation analysis of the protein ID abundancies between experiments produced high r values (r>0.85), indicating good reproducibility between samples.

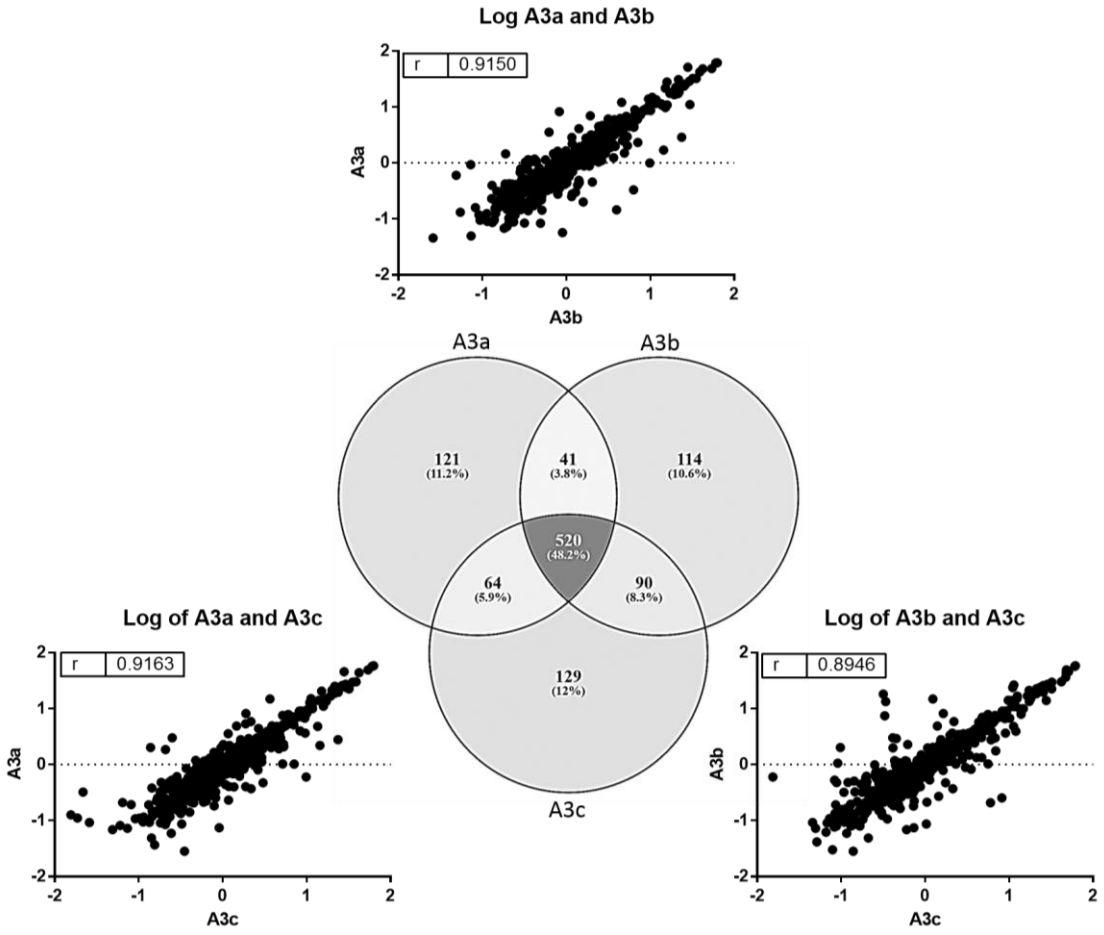


Figure 3:8: Investigating the technical reproducibility of the RapiGest method.

Protein identifications from the three MS experiments which used RapiGest extraction (data points from **Figure 3.7**) were analysed for reproducibility. Identical protein IDs were found in 48.2% of cases when comparing all 3 experiments and 66.2% of cases when comparing 2 of the 3 experiments. Coefficient of correlation analysis of protein abundancies revealed high similarity in protein IDs between all 3 experiments, and thus reproducibility of this method.

In the biological repeat experiment, 46% of proteins identified were shared between all 3 replicates in samples A1-3 and 65.6% of protein IDs were found in at least 2 replicates (**Figure 3.9**). 46.9% of protein IDs in samples B1-3 were shared between all 3 replicates and 65.7% of IDs were found in at least 2 replicates.

The coefficient of correlation analysis of protein ID abundancies between A1-3 and B1-3 reveal high r values, suggesting the data obtained from RapiGest extraction was of high similarity, further highlighting the reproducibility of the method.

The number of IDs achieved in biological repeats (median A = 870, median B =788) was consistent with what has been achieved during the earlier optimisation experiments (722). These data combined indicate that the RapiGest method achieved high protein yield consistently, with good protein ID coverage between biological and technical repeats.

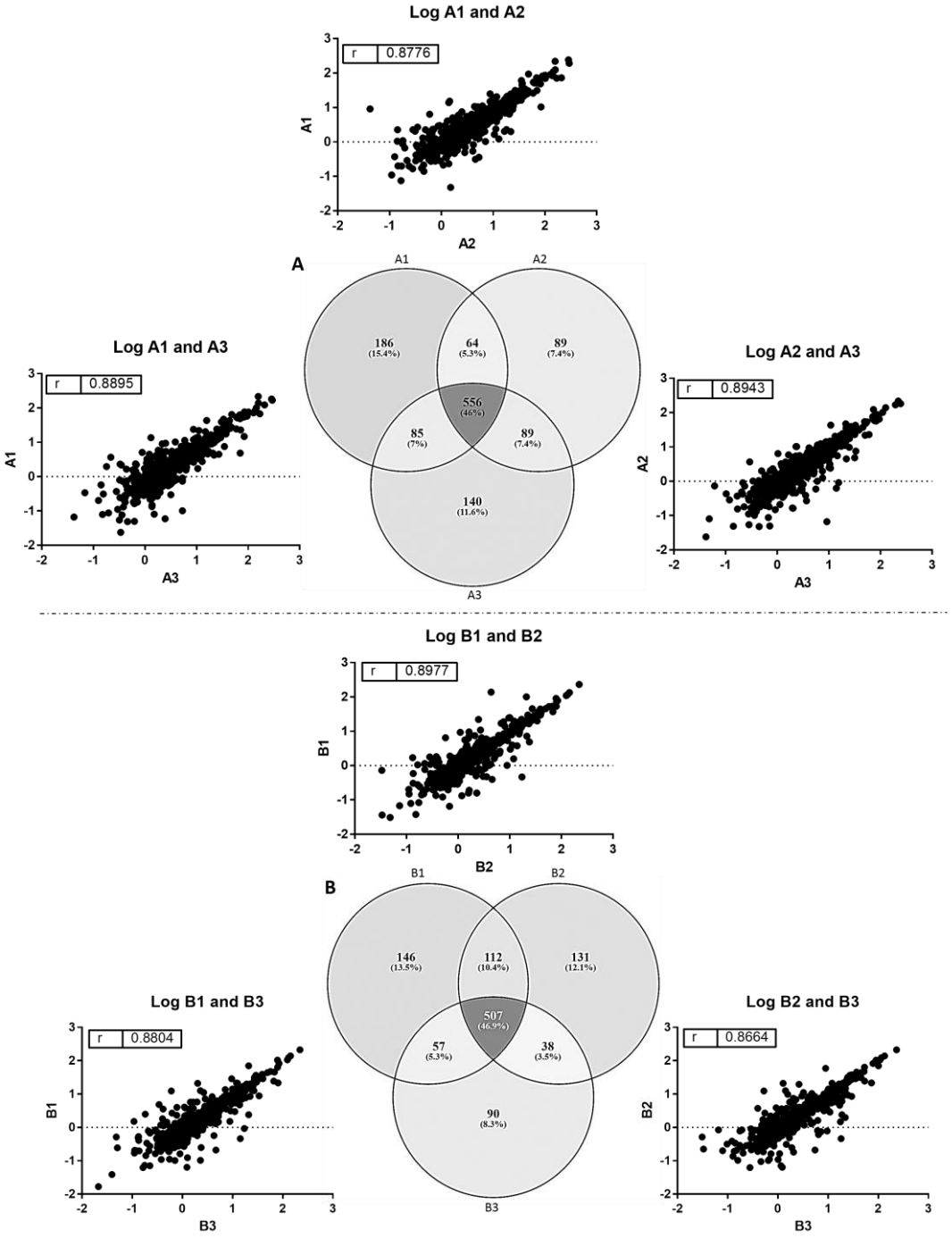


Figure 3:9: Investigating the reproducibility between RapiGest biological repeats.

Protein extraction of cSCC samples A and B was conducted three separate times each. Protein IDs were expressed in a Venn diagram. Identical protein IDs were noted in 46% and 46.9% of cases in A and B respectively following MS of all three extracted samples and almost 2/3rds of the protein IDs were identical following MS of 2 extracted samples. Coefficient of correlation of protein abundancies between the three repeats was analysed and displayed high positive correlation. A, biological replicates of cSCC "A". B, biological replicates of cSCC "B".

Initial experiments using the RapiGest method utilised three histological sections of FFPE tissue at 10µm. It was investigated whether using more sections would increase the protein ID yield, however, it was found that increasing the number of sections to 10 had a detrimental effect on the protein yields (median = 389 protein identifications). After these analyses, it was concluded that the RapiGest method on three histological tissue sections from cSCC was reproducible and achieved high protein identification yields, therefore proteomic analysis of P-M and P-NM cSCCs was commenced.

3.3.7 Clinical and histological characteristics of discovery proteomic samples

A total of 48 samples were used for the discovery proteomics consisting of 24 P-Ms and 24 P-NMs. The clinical and histological characteristics of these samples can be seen in **Table 3.2**. There was an equal ratio of males: females between P-M and P-NM samples. There were similar numbers of moderately differentiated samples in the P-M and P-NM groups, however, there were a greater number of well differentiated cSCCs in the P-NM than the P-M group as well as a larger number of poorly differentiated samples in the P-M than in the P-NM group. There were slightly more cSCCs that had perivascular and perineural invasion in the P-Ms, and more immunosuppressed patients in this group. P-M samples were, on average, deeper (in terms of depth of invasion of the tumour) and larger than P-NM samples.

Table 3.2: Clinical and histological details of cSCC samples used for discovery proteomics.

	P-M cSCCs	P-NM cSCCs
Number of samples	24	24
Male	18 (75%)	18 (75%)
Female	6 (25%)	6 (25%)
Well differentiated	1 (4.17%)	8 (33.33%)
Moderately differentiated	11 (45.83%)	14 (58.33%)
Poorly differentiated	12 (50%)	2 (8.33%)
Perivascular invasion	3 (12.5%)	0 (0%)
Perineural invasion	4 16.67%)	2 (8.33%)
Immunosuppressed	2 (8.33%)	0 (0%)
Average tumour depth (mm)	6.94 ± 4.01	3.88 ± 2.08
Average tumour diameter (mm)	27.04 ± 14.70	12.82 ± 7.77

3.3.8 Protein extraction quantification

After extraction and digestion of proteins from P-M and P-NM samples, peptides were quantified using the direct detect spectrometer from Merck using the method outline sin Chapter 2. The results of this quantification can be seen in **Table 3.3**. The mean total protein extracted was 144.05µg in the P-M group and 84.55µg in the P-NM group.

Table 3.3: Concentrations and total amounts of proteins extracted from P-M and P-NM cSCCs.

P-M cSCCs			*P-NM cSCCs*		
Sample ID	Concentration (µg/µl)	Total peptide (µg)	**Sample ID**	Concentration (µg/µl)	Total peptide (µg)
P-M1	1.198	119.8	*P-NM2*	2.489	248.9
P-M2	0.703	70.3	*P-NM5*	1.349	134.9
P-M3	1.450	145.0	*P-NM6*	0.860	86.0
P-M4	3.103	310.3	*P-NM11*	0.226	22.6
P-M5	0.866	86.6	*P-NM13*	0.527	52.7
P-M7	1.894	189.4	*P-NM16*	0.366	36.6
P-M10	1.155	115.5	*P-NM20*	0.539	53.9
P-M11	1.516	151.6	*P-NM22*	1.433	143.3
P-M13	2.328	232.8	*P-NM25*	2.155	215.5
P-M14	1.700	170.0	*P-NM28*	2.005	200.5
P-M15	2.144	214.4	*P-NM29*	0.810	81.0
P-M16	0.870	87.0	*P-NM31*	2.970	297.0
P-M22	1.059	105.9	*P-NM32*	0.426	42.6
P-M25	2.273	227.3	*P-NM35*	0.951	95.1
P-M26	0.915	91.5	*P-NM38*	1.196	119.6
P-M27	0.537	53.7	*P-NM39*	1.594	159.4
P-M28	0.617	61.7	*P-NM40*	0.831	83.1
P-M39	0.421	42.1	*P-NM41*	0.705	70.5
P-M41	1.705	170.5	*P-NM42*	2.181	218.1
P-M43	2.317	231.7	*P-NM43*	0.478	47.8
P-M45	1.431	143.1	*P-NM46*	0.465	46.5
P-M47	2.054	205.4	*P-NM47*	0.576	57.6
P-M48	0.441	44.1	*P-NM48*	0.731	73.1
P-M49	1.725	172.5	*P-NM51*	2.102	210.2

3.3.9 Protein ID yields from 1D and 2D fractionation

The above described RapiGest method utilised 2D liquid chromatography separation, however, much research undertaken in the Centre for Proteomics, University of Southampton utilises 1D separation for other proteomic studies. As this method was well established in our laboratory, it was decided to use both 1D and 2D liquid chromatography separation as two independent methods to research and investigate biomarkers of metastasis in cSCC. All MS was therefore carried out utilising the RapiGest protein extraction method with 1D fractionation and, separately, 2D fractionation.

The numbers of proteins identified in each sample was similar using 1D and 2D fractionation, with 2D identifying marginally more protein IDs in each sample (**Figure 3.10**).

The highest number of protein IDs achieved was 960 and the lowest 58. The majority of samples identified between 400 and 600 proteins with a mean of 614 IDs in the P-M group and a mean of 509 in the P-NM group.

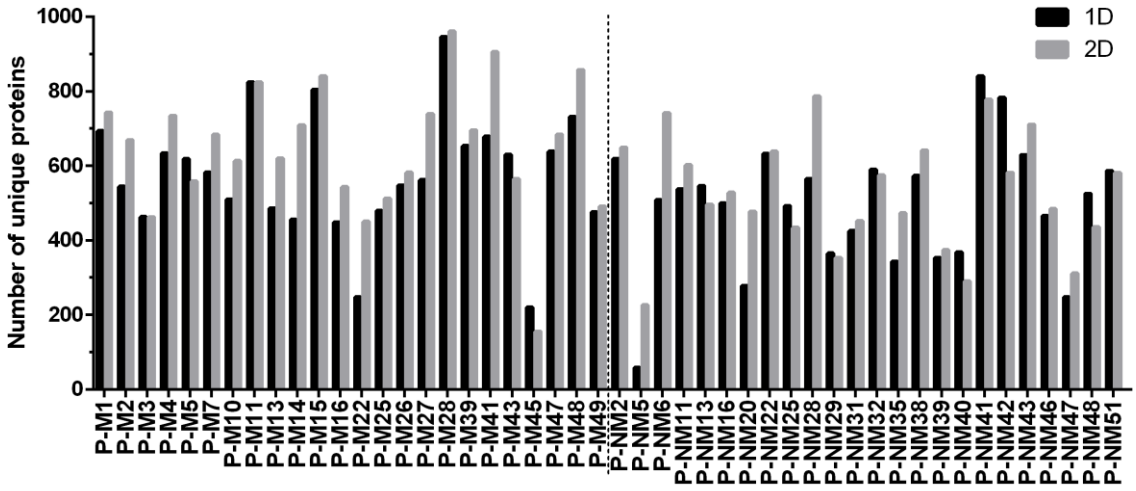


Figure 3:10: Number of protein IDs in all samples using 1D and 2D fractionation.

3.75µg of protein from each sample was subjected to LCMS and mass spectra were processed into protein IDs using Protein Lynx Global Server (PLGS).

The number of unique proteins identified in all samples was 2,986 following 1D and 2,848 following 2D LC fractionation (total 4,018 combining 1D and 2D). 45% of proteins (1,817 of 4,018) were identical following 1D and 2D LC (**Figure 3.11**).

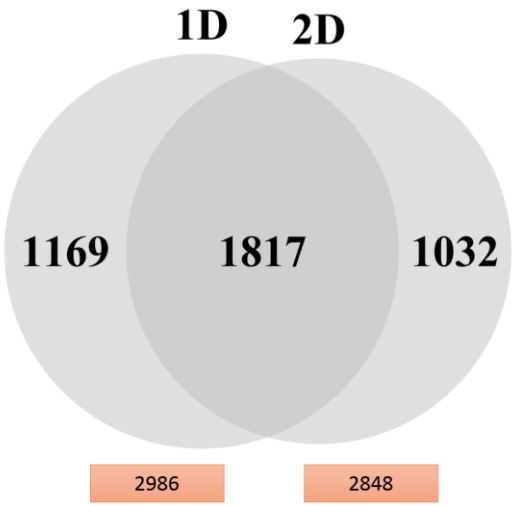


Figure 3:11: Number of unique proteins identified by MS following 1D and 2D LC fractionation.

Total number of unique proteins identified in all P-M and P-NM samples using MS following 1D and 2D LC.

3.3.10 Establishing the distribution of the mass spectrometry protein results

To establish the best statistical analysis test for the data, histograms of each cSCC sample were created to determine if the data was parametric or non-parametric (**Appendix 2**). Alike other proteomics studies, the data suffered from the floor effect because the instrument can only detect down to a certain abundance of protein **(Figure 3.12**). Log10 transformation often results in a more normal distribution of data (because the scale is reduced), nonetheless, with the data Log10 transformed, there were still several samples that had a non-normal distribution (**Appendix 2**). As non-parametric tests are more conservative, it was decided that the non-parametric Mann Whitney U test for significance with non-log10 transformed data would be used to determine significantly differentially expressed proteins between P-Ms and P-NMs.

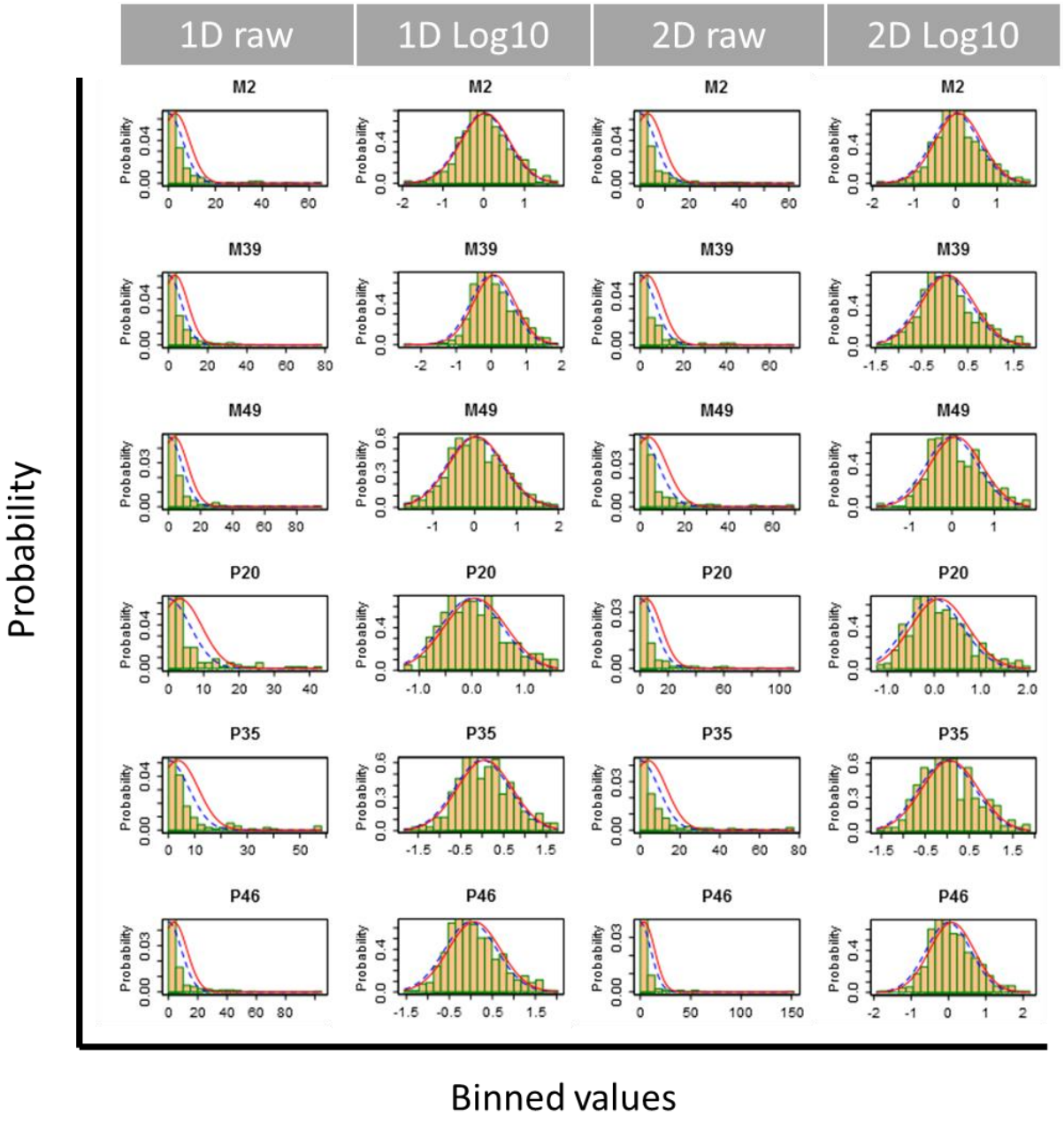


Figure 3:12: Histograms of proteomic quantification data revealed a non-normal distribution.

Protein abundancies of each sample were plotted in histograms using the R package "Inferno", to analyse the normality of the distribution to determine whether parametric or non-parametric statistics should be used. Abundancy data in samples were 'binned' (i.e. separated into a series of intervals) and plotted against frequency to create histograms (yellow bars). Data was also log10 transformed and plotted as histograms. The majority of the data is normally distributed but some of the data is not and therefore a conservative non-parametric test was used (**all histograms Appendix 2**)

3.3.11 Investigating confidence in significantly differentially expressed proteins

Statistical analysis of 'omics data can be difficult because hundreds, and even thousands, of comparisons between variables creates a very high chance of false positives (Franceschi et al., 2013).

One way to assess the false discovery rate is to plot all p-values obtained through statistical analysis into a histogram (**Figure 3.13**). A normal distribution of p-values is indicated by a higher frequency of p-values closer to 0, with a sharp decline down to 0.5, followed by a level frequency thereafter.

p-values obtained through comparison of protein abundancies between P-M and P-NM tumours were plotted in a histogram. This revealed a trend that would be expected of data with a low false positive rate and true significant differences. As previously stated, 'omics data also often suffers from missing values. To accommodate for this, multiple amounts of missing data were analysed and it was found that the higher the allowance of missing data in the analyses, the less confidence there was in the data in terms of false positive rate. Allowing a missing value percentage of 50 produced a high confidence p-value histogram in addition to maintaining high n numbers.

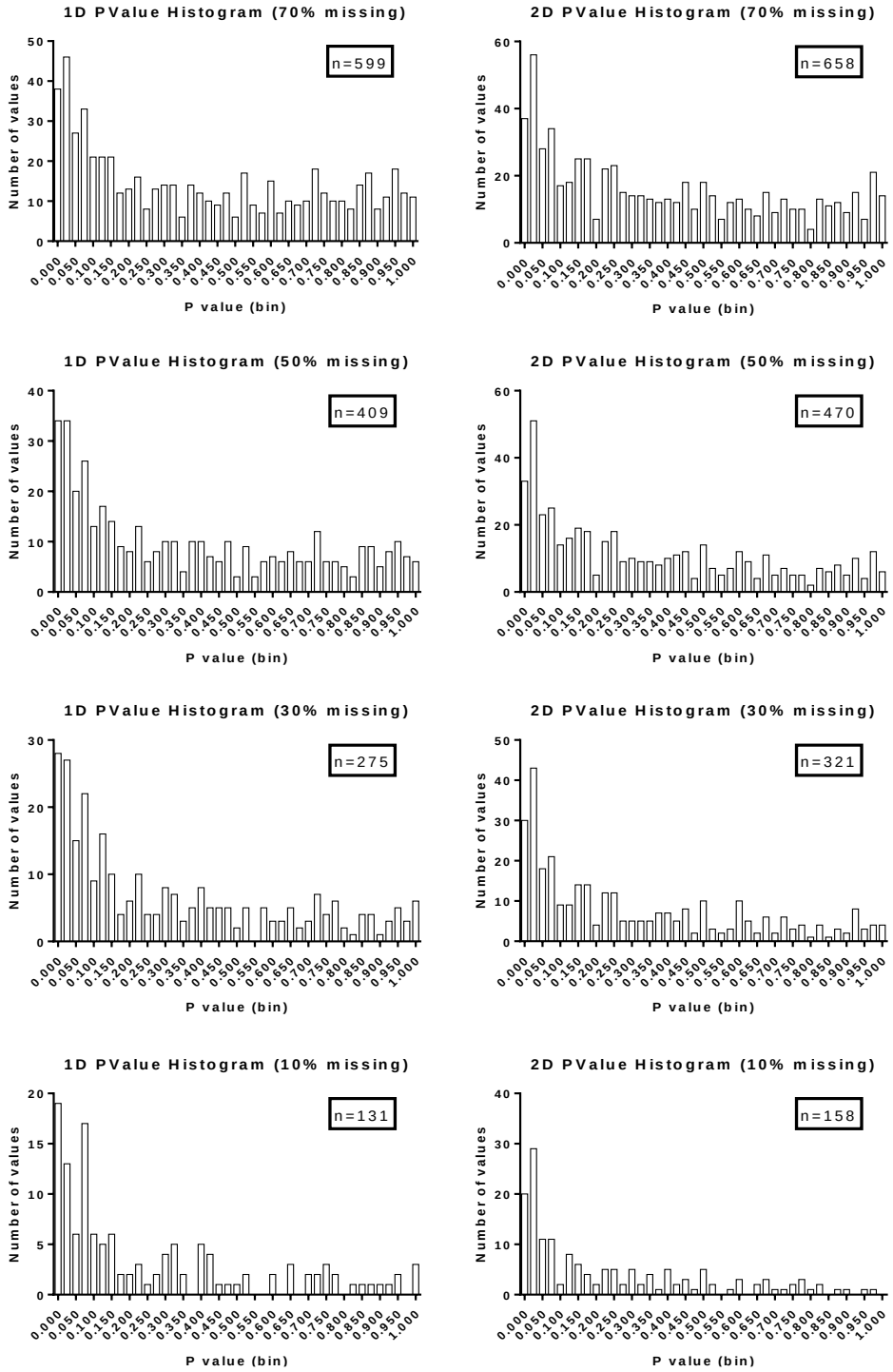


Figure 3:13: p-value histograms of the comparison of differences in the abundancy of unique proteins between P-M and P-NM cSCCs.

p-values obtained through comparing P-M and P-NM protein abundancies (allowing four different percentages of missing values) using Mann Whitney U test were plotted as histograms. The higher the amount of missing data, the less confidence there was in the data. 50% missing data displayed an appropriate balance between confidence in the data and number of significant results.

3.3.12 Differentially expressed proteins

There was a total of 79 significantly differentially expressed proteins identified between the P-M and P-NM groups in the 1D data and 98 in the 2D data (P<0.05). 33 of these were identified in the data obtained both following 1D, and separately, 2D fractionation, equating to a total of 144 significantly differentially expressed proteins (P<0.05) identified in the combination of 1D and 2D data (**Figure 3.14**).

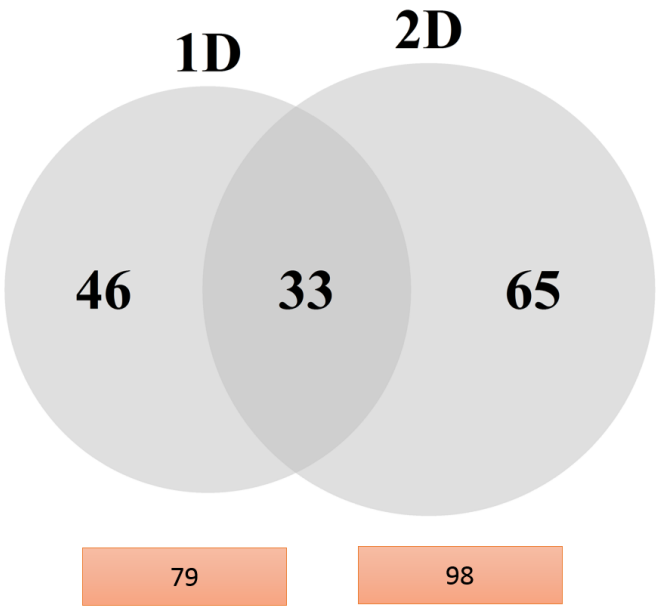


Figure 3:14 : Venn diagram displaying the number of significantly differentially expressed proteins identified in 1D and 2D data.

p-values were obtained through Man Whitney U test for significance of differential expression of proteins between the P-M and P-NM groups. 79 proteins were differentially expressed following 1D and 98 following 2D LC (P<0.05). 33 (22.9%) proteins were found to be differentially expressed in both the 1D and 2D data.

3.3.13 Volcano plots

In addition to whether proteins are significantly differentially expressed, the fold change in expression between P-M and P-NM is also important. One method to compare the p-values and fold changes of protein abundancies is to visualise them in the form of a volcano plot.

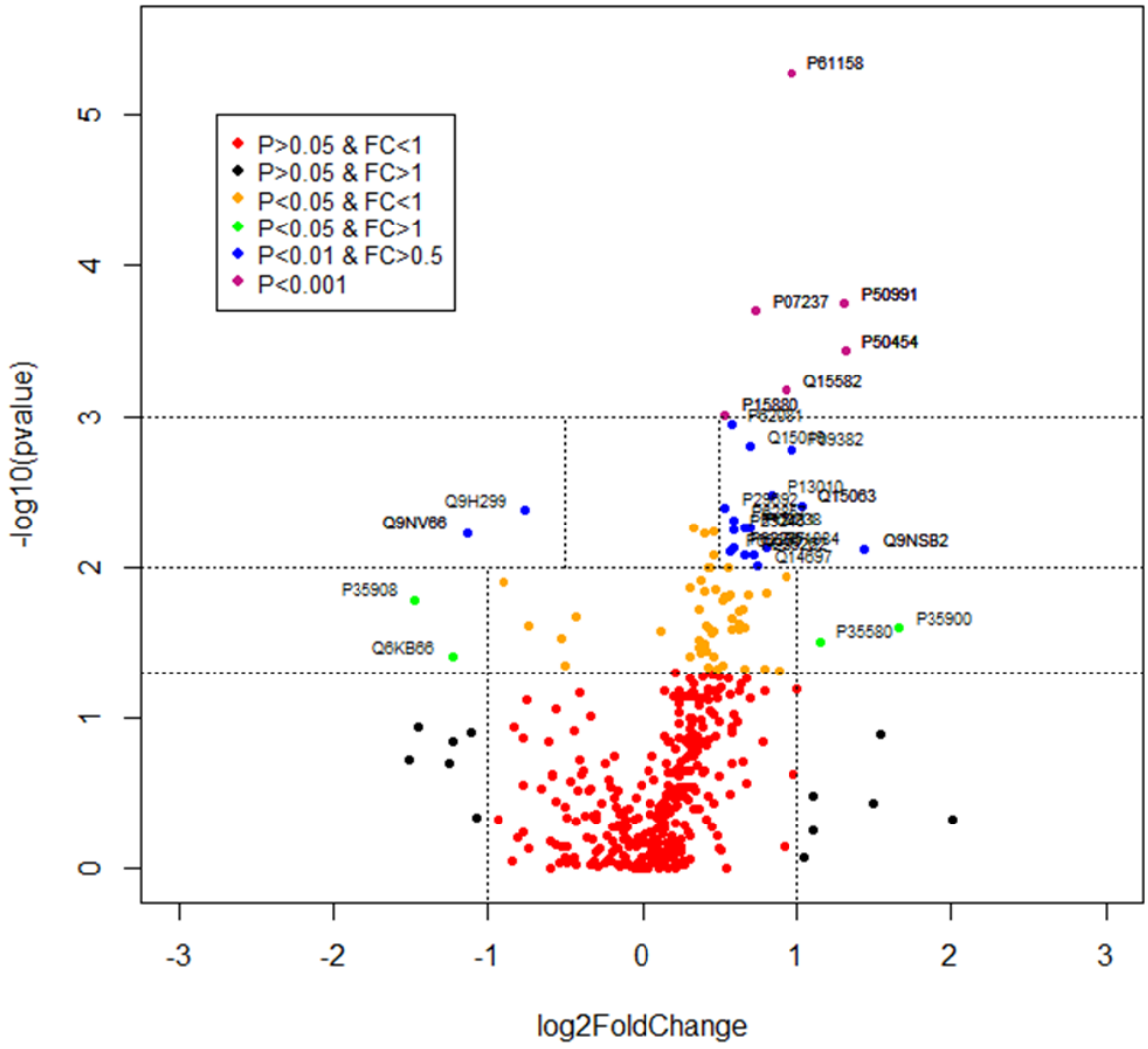


Figure 3:15: Volcano plot of 1D data highlights proteins of interest.

Protein p-values were obtained through comparing protein abundancies in P-M and P-NM tumours using the Mann Whitney U test. Fold change was calculated by subtracting the mean of the protein abundancies in the P-NM group from the P-M group. p-values were log10 transformed and fold change was log2 transformed to generate the volcano plot of the data. Red points indicate non-significant p-value (P>0.05) and fold change <1 log2. Black points indicate non-significant p-value (P>0.05) but fold change >1 log2. Orange points indicate significant p-value (P<0.05) and fold change <1 log2. Green points indicate significant p-value (P<0.05) and fold change >1 log2. Blue points indicate a higher significant p-values (P<0.01) with fold change >0.5 log2. Purple points represent points with highest significance (P<0.001). Labels are Uniprot protein accession numbers. Volcano plot created in R.

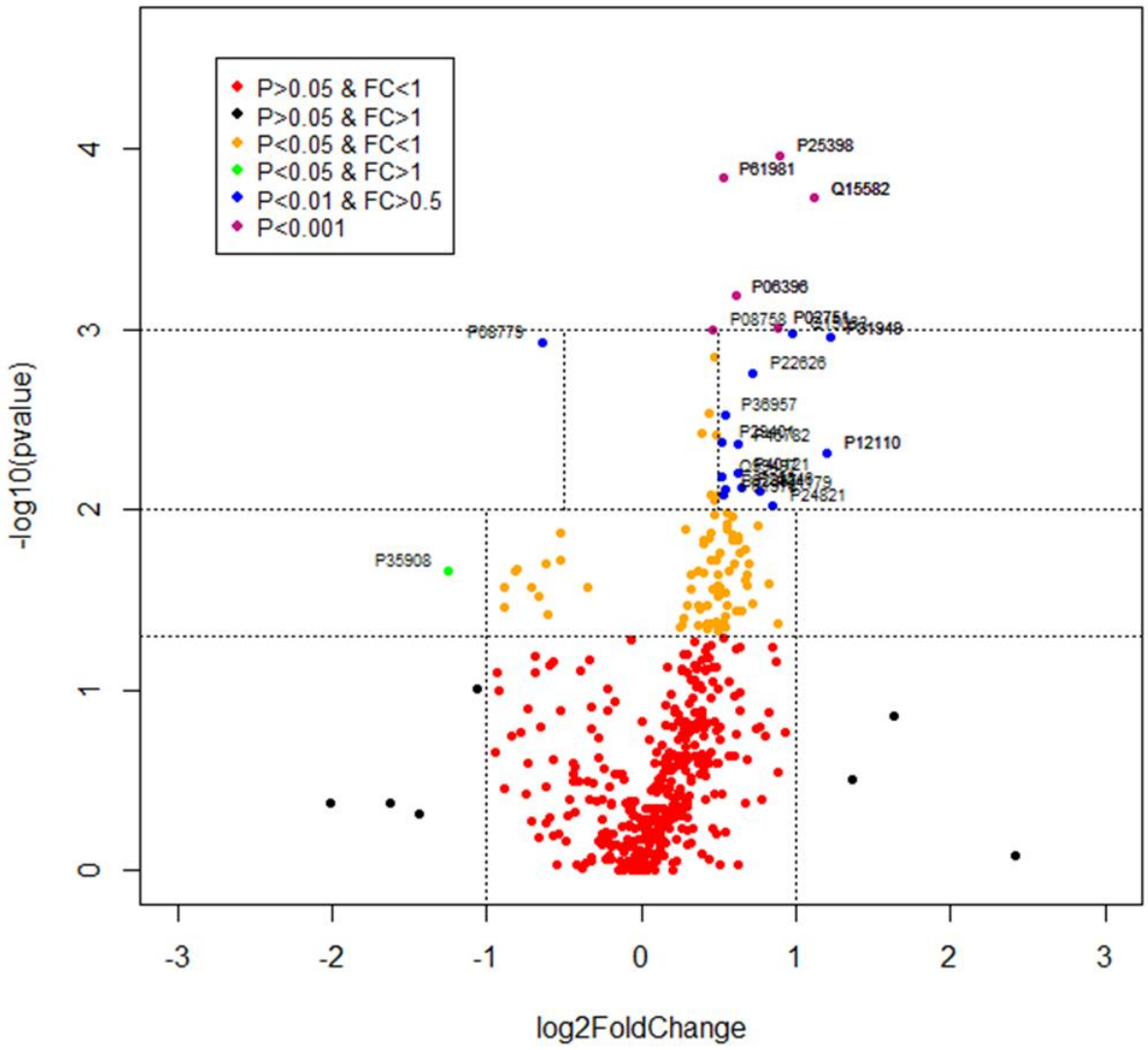


Figure 3:16: Volcano plot of 2D data highlights proteins of interest.

P-values of differential protein expression between P-M and P-NMs were obtained using Mann Whitney U test for significance. Fold changes were calculated by determining the differences between mean expressions of proteins in P-NM to P-M tumours. P-values were log10 transformed and fold changes were log2 transformed for the purpose of producing the volcano plot. Red points indicate non-significant p-value (P>0.05) and fold change <1 log2. Black points indicate non-significant p-value (P>0.05) but fold change >1 log2. Orange points indicate significant p-value (P<0.05) and fold change <1 log2. Green points indicate significant p-value (P<0.05) and fold change >1 log2. Blue points indicate a higher significant p-values (P<0.01) with fold change >0.5 log2. Purple points represent points with highest significance (P<0.001). Labels are Uniprot protein accession numbers. Volcano plot created in R.

The volcano plot of the 1D data shows that 124 proteins (30.32%) were down regulated (<0 fold change) and 285 proteins (69.68%) were upregulated (>0 fold change) between the P-M and the P-NM groups. The plot also highlights a number of potential biomarkers suitable for further investigation (**Figure 3.15**), including 29 proteins which were significantly differentially expressed (P<0.05) with a "significant" fold change (dependant on threshold). Protein accession numbers (and their corresponding gene IDs) of these potential biomarkers are P61158 (*ACTR3*), P50991 (*CCT4*), P07237 (*P4HB*), P50454 (*SERPINH1*), Q15582 (*TGFBI*), P15880 (*RPS2*) , Q9NSB2 (*KRT84*), P35900 (*KRT20*), Q9H299 (*SH3BGRL3*), Q9NV66 (*TYW1*), Q15019 (*SEPT2*), P09382 (*LGALS1*), Q15063 (*POSTN*), P29692 (*EEF1D*), P35908 (*KRT2*), and Q6KB66 (*KRT80*) (full list in **Table 3.4**). Several blue data point accession numbers clustered in the top right of the volcano plot correspond to ribosomal proteins.

In the 2D volcano plot, 139 proteins (29.57%) were down regulated (<0 fold change) and 331 proteins (70.43%) were upregulated (>0 fold change). Exploring the data from this volcano plot revealed 22 proteins that were significantly differentially expressed (P<0.05) with a "significant fold" change (**Figure 3.16**) including several that have been highlighted in the 1D plot. Accession numbers (and corresponding gene IDs) of potential biomarkers in the 2D volcano plot are P25398 (*RPS12*), P61981 (*YWHAG*), Q15582 (*TGFBI*), P06396 (*GSN*), P08758 (*ANXA5*), P02751 (*FN1*), P31949 (*S100A11*), P22626 (*HNRNPA2B1*), P24821 (*TNC*), P29401 (*TKT*), P40121 (*CAPG*), Q15063 (*POSTN*), P35908 (*KRT2*), P12110 (*COL6A2*), P08779 (*KRT16*) and P62937 (*PPIA*) (full list in **Table 3.5**). Similarly to the 1D volcano plot, several of the blue data point's accession numbers in the 2D plot represent ribosomal proteins.

3.3.14 Significantly differentially expressed proteins and their respective fold changes

The significantly differentially expressed proteins in the proteomic profiling results, following 1D LC, and their respective fold changes between P-M and P-NM groups can be seen in **Table 3.4**. Actin related protein 3 (*ACTR3*), T-complex protein 1 subunit delta (*CCT4*), Protein disulphide-isomerase (*P4HB*), Serpin H1, TGFB induced protein (*TGFBI*) and 40S ribosomal protein S2 (*RPS2*) all have high significance (P<0.001). Of the differentially expressed proteins presented in **Table 3.4**, the fold change ranges from the most downregulated protein, *KRT2* (-1.47336 log2 fold change) to the most upregulated protein, *KRT20* (1.661553 log2 fold change).

Table 3.4: A table of significantly differentially expressed proteins with fold change between P-Ms and P-NMs during MS following 1D fractionation.

Uniprot ID	Gene ID	log2 Fold Change	p-value
P61158	ACTR3	0.961506898	5.21E-06
P50991	CCT4	1.30927292	0.000178
P07237	P4HB	0.733735226	0.000197
P50454	SERPINH1	1.317776969	0.000362
Q15582	TGFBI	0.928341791	0.000667
P15880	RPS2	0.535540549	0.000977
P62081	RPS7	0.577872021	0.001138
Q15019	SEPT2	0.699315588	0.00159
P09382	LGALS1	0.966564061	0.001659
P13010	XRCC5	0.831577845	0.003321
Q15063	POSTN	1.032788051	0.003898
P29692	EEF1D	0.535534134	0.004054
Q9H299	SH3BGRL3	-0.761111358	0.004194
P62857	RPS28	0.594810883	0.004916
P08238	HSP90AB1	0.696523855	0.005466
P12111	COL6A3	0.6616087	0.00553
P60709	ACTB	0.331314871	0.00553
P23246	SFPQ	0.58683941	0.005641
P08133	ANXA6	0.457416134	0.00585
O43707	ACTN4	0.39691863	0.005863
Q9NV66	TYW1	-1.134978807	0.005908
P51884	LUM	0.80396623	0.007435
P62277	RPS13	0.589690303	0.007468
Q9NSB2	KRT84	1.438285492	0.007721
P60660	MYL6	0.565832754	0.007811
P22626	HNRNPA2B1	0.662690619	0.008254
P61978	HNRNPK	0.454874485	0.008254
P35222	CTNNB1	0.721477453	0.008257
Q14697	GANAB	0.746390079	0.0099
P07437	TUBB	0.426709653	0.009969
P08758	ANXA5	0.441600309	0.010052
P04844	RPN2	0.554418246	0.010152
P24821	TNC	0.929523896	0.011412
P11142	HSPA8	0.383725658	0.012193
Q9NZT1	CALML5	-0.901133441	0.012585
P62805	HIST1H	0.312532197	0.013583
P59998	ARPC4	0.466567609	0.013831
P36578	RPL4	0.402361344	0.014564
P16403	HIST1H1C	0.79468494	0.014741
(Continued)			
P19338	NCL	0.570498	0.015221
P39656	DDOST	0.686281	0.015341
O00571	DDX3X	0.55198	0.015599
P46783	RPS10	0.533801	0.015599
P35908	KRT2	-1.47336	0.016373
P62937	PPIA	0.522324	0.01663
P26038	MSN	0.372223	0.018779
P50990	CCT8	0.643611	0.018874
Q99623	PHB2	0.626632	0.019505
P04259	KRT6B	-0.43177	0.021146
P62140	PPP1CB	0.582665	0.021844
P13796	LCP1	0.626618	0.023763
P50395	GDI2	0.417755	0.024282
P21810	BGN	-0.73761	0.024587
P02675	FGB	0.662066	0.025039
P35900	KRT20	1.661553	0.025217
P04792	HSPB1	0.43912	0.025692
P10599	TXN	0.578065	0.025747
Q07065	CKAP4	0.620822	0.025817
P29401	TKT	0.465121	0.026105
P09651	HNRNPA1	0.119659	0.026825
P52597	HNRNPF	0.450961	0.027294
Q9HCY8	S100A14	-0.51908	0.029751
P13639	EEF2	0.36846	0.030267
P35580	MYH10	1.153696	0.030978
P07741	APRT	0.406844	0.031729
Q02878	RPL6	0.365138	0.033587
O00148	DDX39A	0.408352	0.03601
P46940	IQGAP1	0.379869	0.036832
Q6KB66	KRT80	-1.22693	0.039027
P12109	COL6A1	0.461111	0.039192
P60866	RPS20	0.312545	0.039198
O75369	FLNB	0.524255	0.044172
P15088	CPA3	-0.49888	0.044723
P07900	HSP90AA1	0.424161	0.046572
P16144	ITGB4	0.791541	0.047448
P62318	SNRPD3	0.484703	0.047584
P30044	PRDX5	0.661837	0.047683
P42224	STAT1	0.882976	0.048652

p-values were obtained through Mann Whitney U test for significance between P-Ms and P-NMs. Fold change calculated from mean of protein abundancies between each group. Green shading indicates proteins which were significantly differentially expressed in both the 1D and 2D data.

Table 3.5: Details of proteins that were significantly differentially expressed between P-Ms and P-NM groups in the MS data following 2D LC.

Uniprot ID	Gene ID	log2 Fold Change	p-value
P25398	RPS12	0.899331	1.08E-04
P61981	YWHAG	0.529794	0.000145
Q15582	TGFBI	1.116691	0.000187
P06396	GSN	0.609908	0.000655
P02751	FN1	0.883513	0.00098
P08758	ANXA5	0.463055	0.001
Q15063	POSTN	0.981432	0.001047
P31949	S100A11	1.220557	0.001104
P08779	KRT16	-0.64341	0.001175
P63000	RAC1	0.475662	0.001428
P22626	HNRNPA2B1	0.723681	0.001739
P02545	LMNA	0.436365	0.002937
P36957	DLST	0.541366	0.002971
P18206	VCL	0.391585	0.003743
P62277	RPS13	0.479793	0.003863
P29401	TKT	0.521632	0.004185
P46782	RPS5	0.62297	0.004337
P12110	COL6A2	1.202369	0.004823
P40121	CAPG	0.6193	0.006275
Q99497	PARK7	0.521044	0.006502
P23246	SFPQ	0.643473	0.007564
P62937	PPIA	0.53968	0.007713
P04179	SOD2	0.762658	0.007975
P08123	COL1A2	0.450851	0.00823
P08238	HSP90AB1	0.497668	0.00823
P61978	HNRNPK	0.536419	0.008247
P62158	CALM	0.477913	0.008814
P24821	TNC	0.850068	0.009468
Q07960	ARHGAP1	0.559179	0.010409
P07437	TUBB	0.471989	0.010616
P62314	SNRPD1	0.592314	0.010933
P60174	TPI1	0.556856	0.012021
P31146	CORO1A	0.759081	0.012238
P68104	EEF1A1	0.283546	0.012781
P09525	ANXA4	0.553413	0.01287
P04259	KRT6B	-0.52402	0.013583
P08670	VIM	0.454294	0.013583
P14625	HSP90B1	0.59337	0.013831
P02671	FGA	0.62589	0.014133
P60660	MYL6	0.436868	0.014429
Q03252	LMNB2	0.590182	0.014685
Q99878	HIST1H2AJ	0.407634	0.014719
P29590	PML	0.628868	0.014724
P23396	RPS3	0.399818	0.015319
Q99623	PHB2	0.67447	0.016553
P07237	P4HB	0.631181	0.017242
P12111	COL6A3	0.501978	0.017242
P27482	CALML3	-0.51899	0.018889
P50395	GDI2	0.444733	0.018889

(Continued)

Q14697	GANAB	0.481031	0.018922
P27824	CANX	0.5971	0.019724
P16615	ATP2A2	-0.62138	0.019818
P02675	FGB	0.694736	0.019933
Q9NZT1	CALML5	-0.80848	0.021169
P04264	KRT1	-0.81594	0.021718
P35908	KRT2	-1.24764	0.021718
P62249	RPS16	0.365977	0.021847
P50990	CCT8	0.569868	0.022002
Q96FW1	OTUB1	0.406082	0.022399
P30044	PRDX5	0.683793	0.022719
P26038	MSN	0.494652	0.02298
P63104	YWHAZ	0.323326	0.02298
P20700	LMNB1	0.676225	0.024282
P05141	SLC25A5	0.823864	0.025666
Q562R1	ACTBL2	0.495019	0.026073
P13796	LCP1	0.685076	0.026452
P01871	IGHM	-0.88295	0.026604
P35555	FBN1	0.508086	0.027037
P48668	KRT6C	-0.71584	0.027113
P02538	KRT6A	-0.34777	0.027147
P37802	TAGLN2	0.455563	0.027434
P60866	RPS20	0.319156	0.027496
P39656	DDOST	0.541621	0.029103
P01011	SERPINA3	0.490469	0.030512
P29508	SERPINB3	-0.66772	0.030512
Q99715	COL12A1	0.722631	0.032952
P00338	LDHA	0.298966	0.033683
O43390	HNRNPR	0.430315	0.034032
P01009	SERPINA1	0.554795	0.034206
P62081	RPS7	0.363127	0.034206
Q02388	COL7A1	-0.88471	0.034513
P11021	HSPA5	0.377159	0.035508
P07195	LDHB	0.6105	0.036002
Q05707	COL14A1	0.650654	0.03669
P55795	HNRNPH2	-0.60532	0.038253
O00299	CLIC1	0.539654	0.039171
P21333	FLNA	0.267202	0.039408
P00558	PGK1	0.485908	0.041488
P62899	RPL31	0.458861	0.042321
P30041	PRDX6	0.877492	0.042339
P07900	HSP90AA1	0.419973	0.042481
P51884	LUM	0.532743	0.042486
P19338	NCL	0.366465	0.043658
P62805	HIST4H	0.265299	0.043658
Q71UI9	H2AFV	0.546786	0.044845
P62269	RPS18	0.250738	0.045023
P30101	PDIA3	0.421946	0.045921
P27816	MAP4	0.496461	0.046927

Significance was calculated using Mann Whitney U test. Fold change was calculated from the ratio of the mean expression of the protein in the P-M group relative to the P-NM group and log 2 transforming the data. Green shading indicates proteins that were significantly differentially expressed in both the 1D and 2D data.

Similarly to the 1D results, the fold changes seen within the differentially expressed proteins following 2D proteomic experiments vary greatly (**Table 3.5**). 40S ribosomal protein S12 (RPS12), protein 14-3-3 gamma (YWHAG), TGFβ induced protein (TGFBI), gelsolin (GSN) fibronectin (FN1) and annexin A5 (ANXA5) are all highly differentially expressed ($P \leq 0.001$). The most downregulated protein is KRT2 (-1.24764) and the highest upregulated protein is S100A11 (1.220557).

Many of the significantly differentially expressed proteins identified in the 1D data (**Table 3.4**) can also be seen in 2D data (**Table 3.5**) (highlighted in green). Example of these are TGFB induced protein (TGFBI), periostin (POSTN), heat shock protein 90-beta (HSP90AB1), calmodulin-like protein 5 (CALML5), collagen alpha-3(VI) chain (COL6A3), fibrinogen beta chain (FGB), lumican (LUM), nucleolin (NCL) and tenascin (TNC). Although the p-values for comparison of many of these proteins between P-M and P-NM cSCCs vary between the 1D and 2D data, their fold change is relatively consistent between both sets of data.

Some examples of significantly differentially expressed proteins from the 1D data can be seen in **Figure 3.17**. Of the proteins in this figure, the lowest number of samples that detected a specific protein was 15 and the highest was 24 (all of them). The median abundancies of these proteins varies, ranging from the lowest, SEPT2 (0.4352ng), to the highest, COL6A3 (75.22ng). All of the proteins presented in **Figure 3.17** express an increase in abundancy in P-M samples compared to P-NM samples.

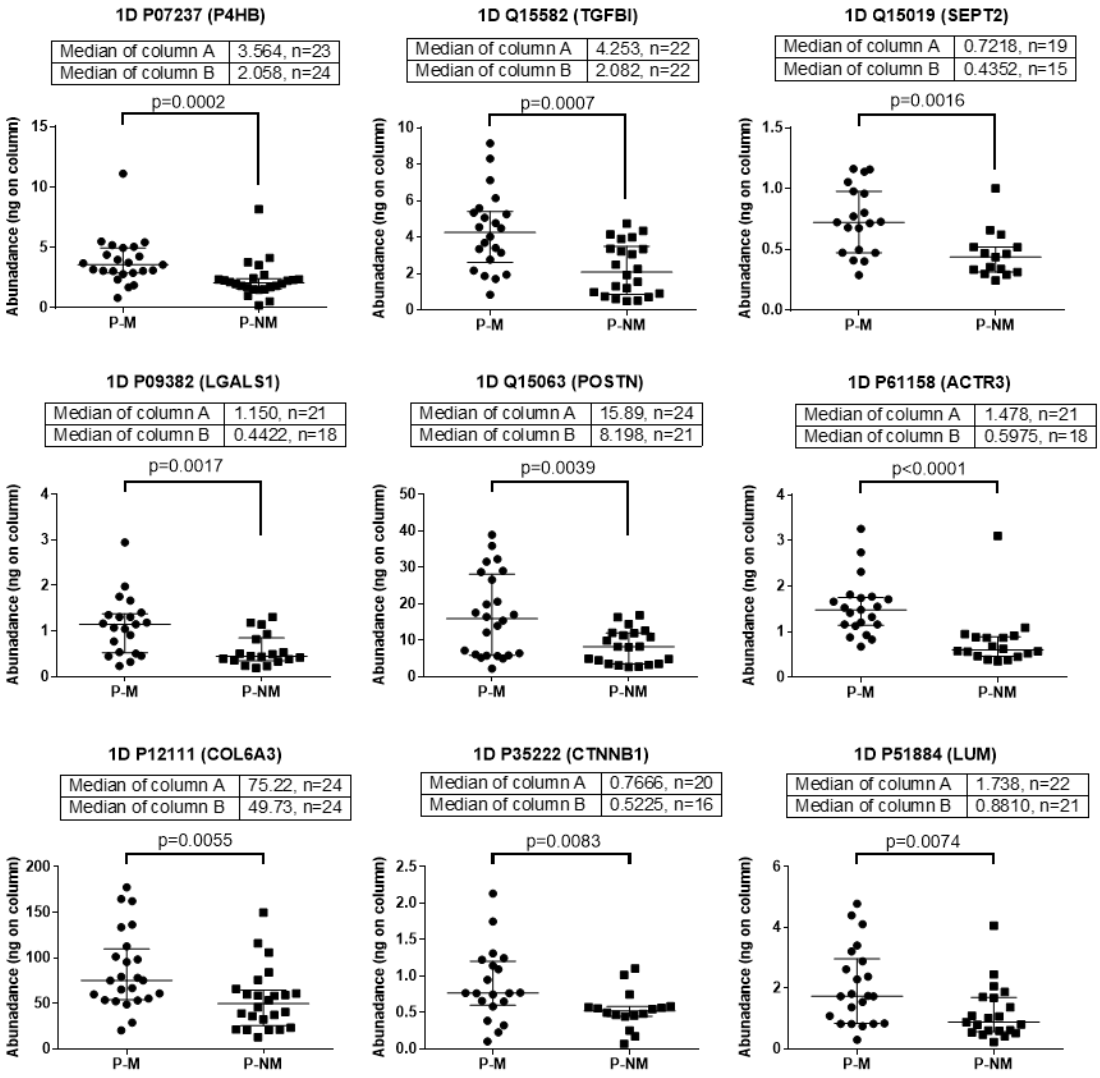


Figure 3:17: Examples of significantly differentially expressed proteins from 1D proteomic profiling experiments.

Protein abundancies from P-M samples were compared to P-NM samples using Mann Whitney U test for significance and plotted using Prism. Median +/- interquartile range shown.

Examples of significantly differentially expressed proteins from the 2D data can be seen in **Figure 3.18**. The lowest number of samples where the protein was identified is 17 and the highest number of cSCCs in which the protein was detected is 24. The lowest median abundancy is S100A11 (0.2615ng) and the highest is POSTN (15.51ng). Similarly to the 1D data (**Figure 3.17**), all of the proteins in Figure 3.18 are upregulated in P-Ms compared to P-NMs.

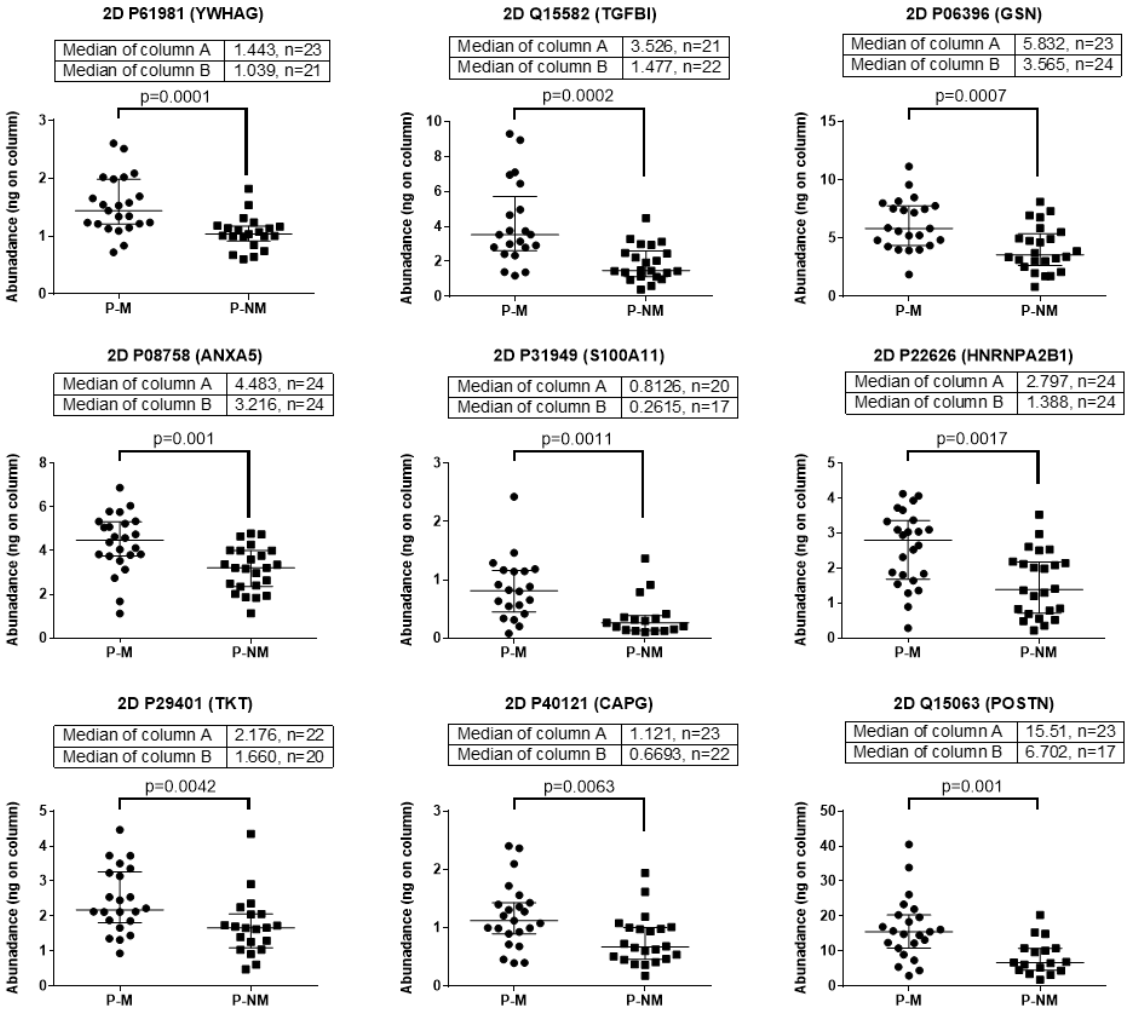


Figure 3:18: Examples of significantly differentially expressed proteins from 2D proteomic profiling experiments.

Protein abundancies were compared between P-M and P-NM samples using Mann Whitney U test for significance. Median +/- interquartile range shown here.

3.3.15 Search tool for the retrieval of interacting genes/proteins (STRING) analysis

STRING is a database of known and predicted protein-protein interactions, gathering its information from various sources including experimental data, computational modelling and text mining. The data from the MS experiments following 1D and 2D separation was analysed using STRING to generate a structure of protein interactions that are likely to play a role in metastases of cSCC.

The structure created using STRING for 1D can be seen in **Figure 3.19**. Nodes represent proteins and the lines connecting the nodes indicate an interaction between these

proteins, which may include physical interaction (e.g. direct binding), signalling pathways, or common biological effects. The density of each individual line in the structure corresponds to the confidence of the interaction between nodes. Within the structure produced from the 1D proteomic profiling data, many proteins are interacting. For example, 98 interactions would be expected by chance from the 77 proteins analysed, but the created structure resulted in 246 interactions which therefore suggests the inputted proteins are acting in a combined manner to promote metastasis. The structure showed clusters of interacting proteins, which included ribosomal proteins, extracellular proteins and proteins involved in protein folding.

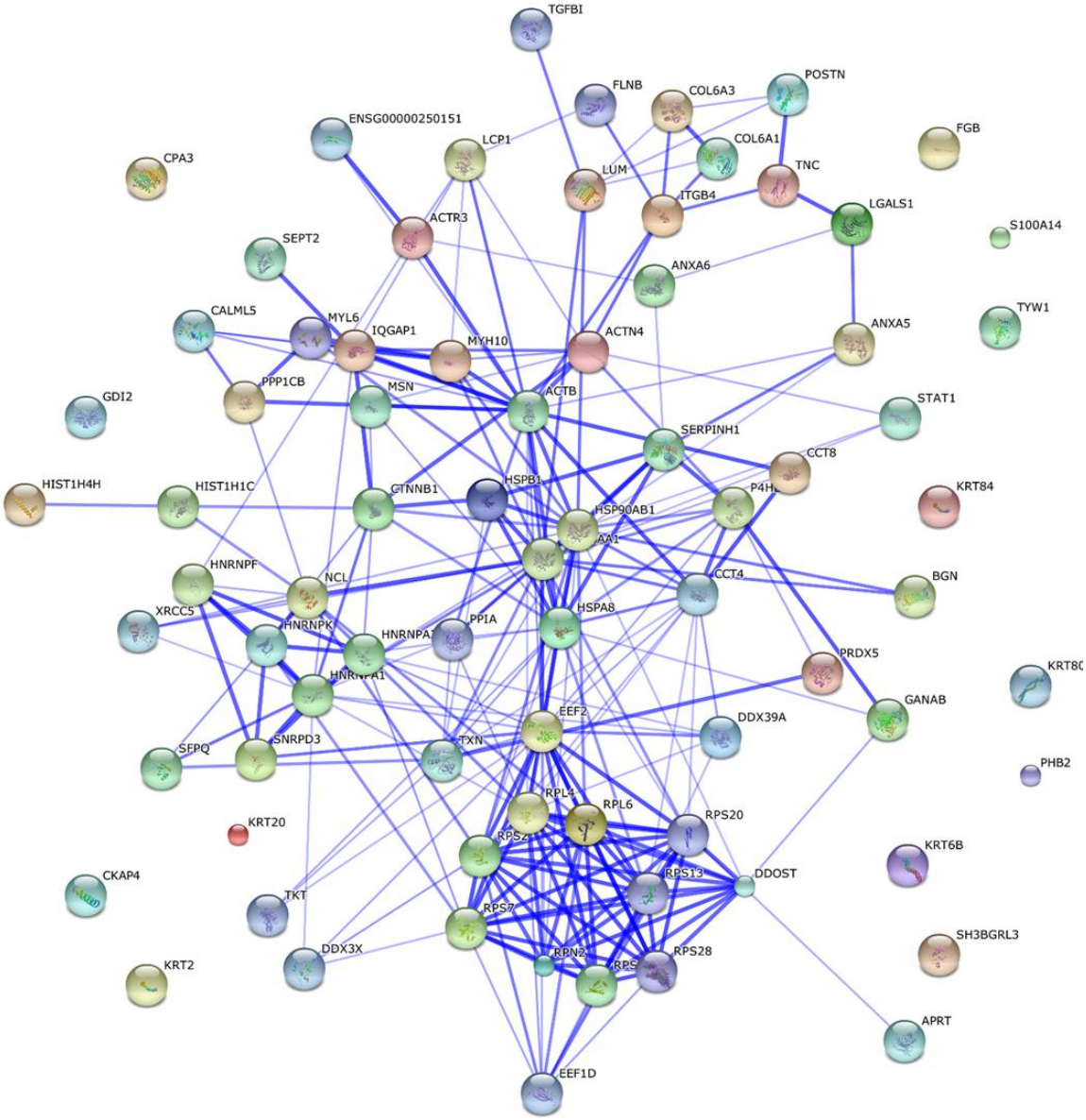


Figure 3:19: STRING structure of significantly differentially expressed proteins from the 1D data.

Significantly differentially expressed proteins identified from Mann-Whitney U test between P-M and P-NMs were analysed using STRING software to create a structure of known interactions. A medium confidence score of 0.4 was allowed for interaction certainty. Nodes represent proteins and lines represent known interactions between proteins. The thicker the line, the higher confidence in the interaction data. Total number of nodes is 77. Total number of interactions is 246.

After this STRING structure was created, other areas of interest such as KEGG pathway enrichment could be mapped onto this. KEGG pathway analysis identifies proteins involved in biological systems through a manually curated database. Proteins are scored through well published algorithms (Franceschini et al., 2013, Von Mering et al., 2003, Szklarczyk et al., 2015) and resulting p-values are corrected for false discovery rate (FDR). **Figure 3.20** utilises the same structure produced in **Figure 3.19**, mapping on significantly enriched KEGG pathways by highlighting nodes red if they are involved. Ribosomal proteins were identified as being the most significantly enriched KEGG pathway with an adjusted p-value of 0.00000792. Focal adhesion was significantly enriched, obtaining an adjusted p-value of 0.00000968, in addition to protein processing in the endoplasmic reticulum (P=0.000165) and regulation of cytoskeleton (P=0.000959).

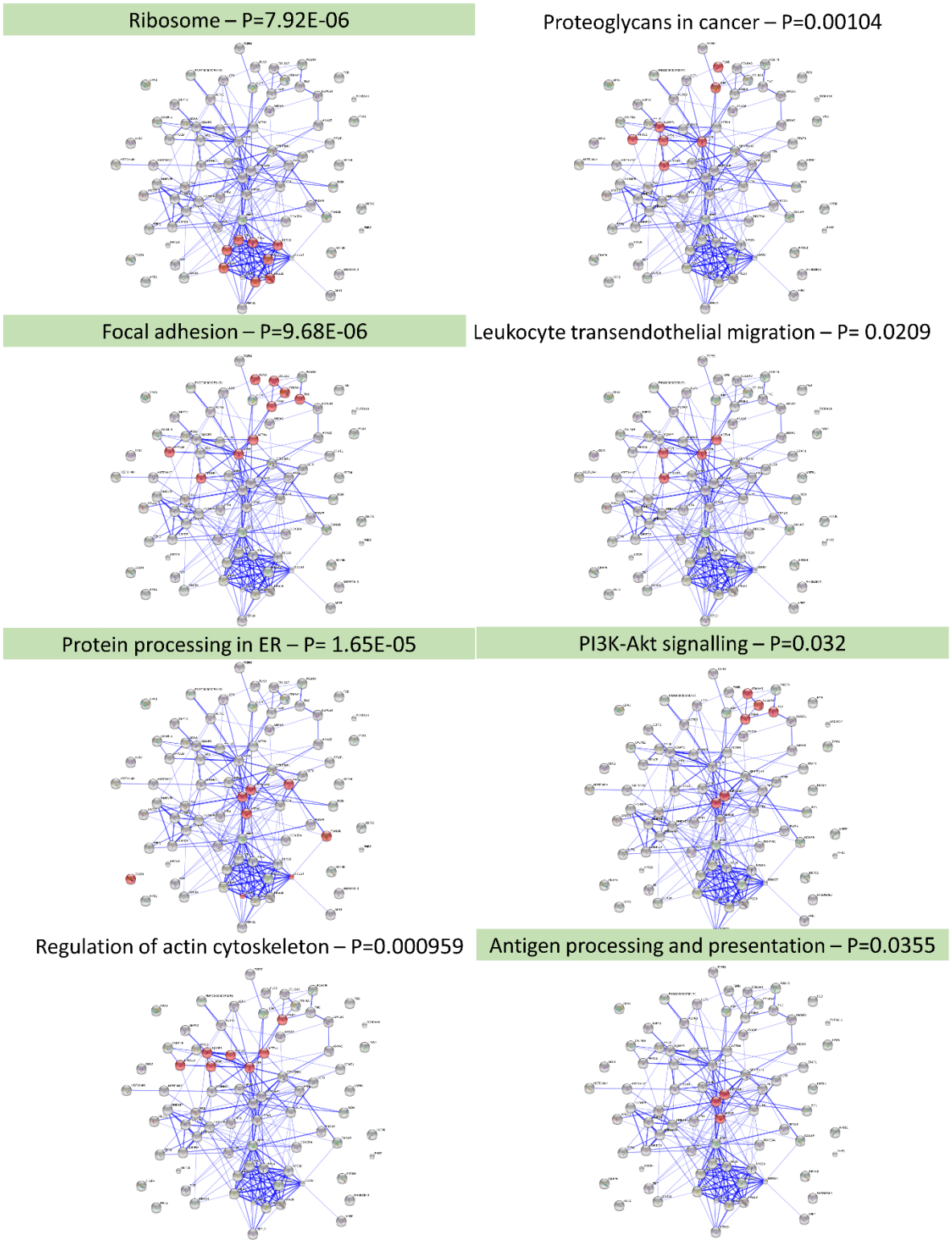


Figure 3:20: STRING structure with significantly enriched KEGG pathways from 1D data. KEGG pathway analysis was carried out within STRING software to give FDR adjusted p-values of enriched pathways. Proteins involved are highlighted in red. Green highlighted text indicates KEGG pathway enriched in both the 1D and 2D data. ER, endoplasmic reticulum.

The STRING structure for the 2D data can be seen in **Figure 3.21**. Similar to the 1D results, the amount of interactions expected (128) was far less than actually observed (340) from the 94 proteins input into the software and this therefore suggests that these proteins are interacting in a manner which is greater than that expected to be seen by chance. Furthermore, similar clusters are seen in the KEGG analyses of the STRING structures of the 2D and the 1D data, such as ribosomal proteins, and proteins involved in the extracellular matrix and protein folding.

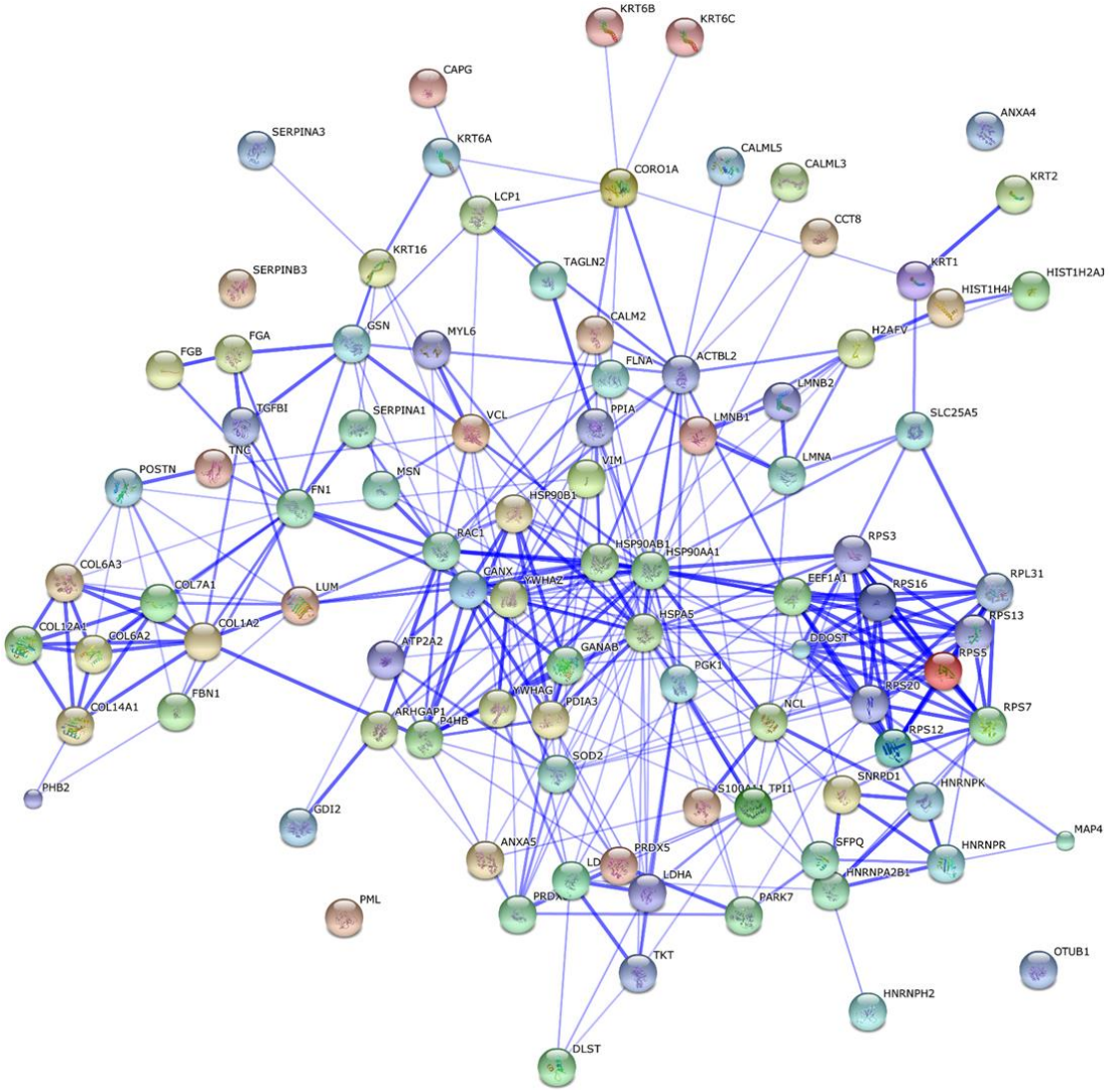


Figure 3:21: STRING structure of significantly differentially expressed proteins from the 2D data.

This structure was created in STRING using the differentially expressed proteins between P-M and P-NMs found in the 2D data. A medium allowance of 0.4 was set for similarity. The total number of nodes was 94. The expected number of interactions was 128, the actual number of interactions was 340.

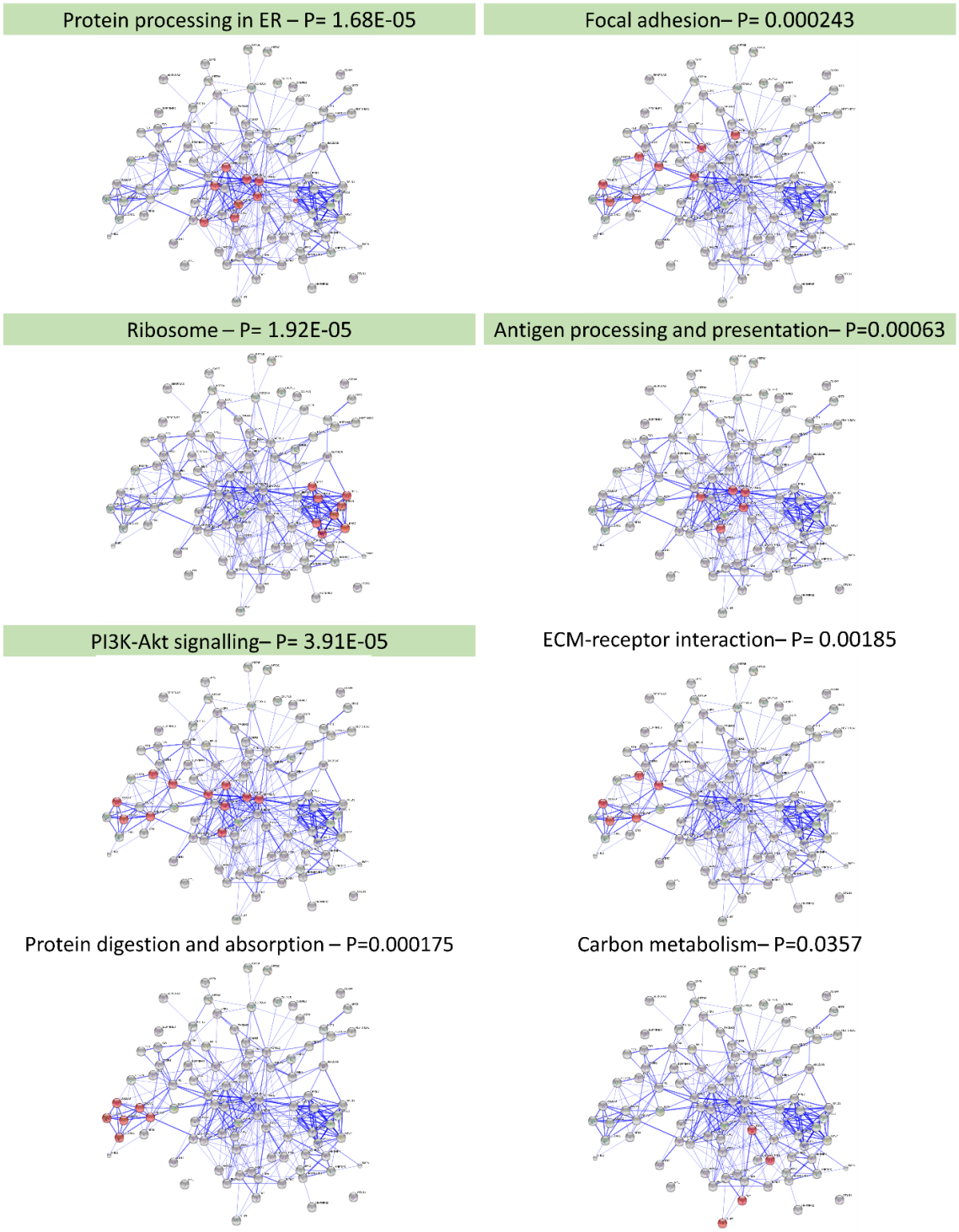


Figure 3:22: KEGG pathway enrichment of 2D STRING structure.

STRING software was used to map KEGG pathway involvement within the STRING structure from the 2D data to give an FDR adjusted p-value. Proteins involved in pathways are indicated in red. Green highlighted text indicates KEGG pathway enrichment in the STRING structures from both the 1D and 2D data. ER, endoplasmic reticulum.

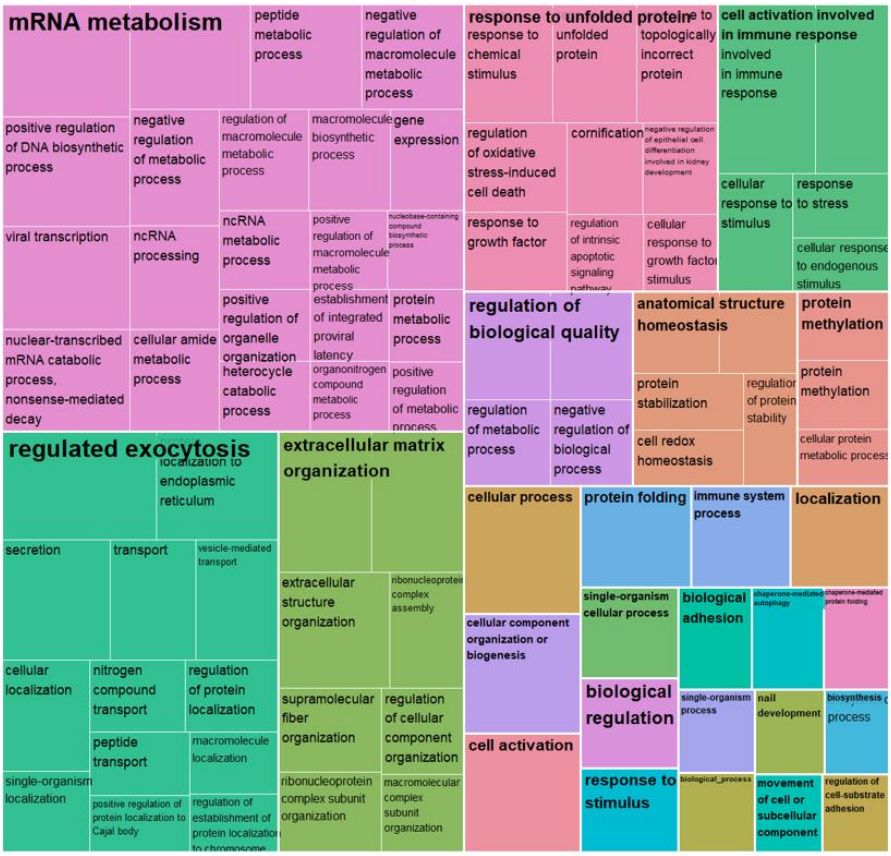
KEGG pathway enrichment analysis of the STRING structure from the 2D data revealed several enriched pathways (**Figure 3.22**). The highest enriched KEGG pathway was protein processing in the endoplasmic reticulum with an adjusted p-value of 0.0000168. The ribosome pathway was enriched with a p-value of 0.0000192, in addition to PI3K-AKt signalling (P= 0.0000391) and protein digestion and absorption (P=0.000175).

Five of the KEGG pathways identified were enriched in both 1D and 2D proteomic profiling data. Ribosomal pathway was the most enriched in the 1D data and the 2nd most enriched in the 2D data. Protein processing in the endoplasmic reticulum was the highest enriched in the 2D data and the 3rd most enriched in the 1D data. Focal adhesion was highly enriched in both the 1D and 2D data, along with PI3K-AKt signalling and antigen processing/presentation.

3.3.16 Gene ontology analysis

Results were further analysed with GoGorilla gene ontology analysis which comprises a database of all known genes, classified according to cell biological processes, cell molecular functions and cellular component (Ashburner et al., 2000). The results obtained from gene ontology enrichment analysis in GoGorilla from significantly differentially expressed proteins (between P-Ms and P-NMs) is too large to display graphically. REViGO (i.e. reduce, visualise gene ontology) was subsequently employed as it condenses the results from gene ontology enrichment analyses, such as Go Gorilla, into simpler graphics for visualisation as seen in **Figure 3.23**. Modified R code was used to create a REViGO tree map for differentially expressed proteins in the 1D data, with the proportion of the size of individual squares and rectangles, and thus the area coverage, in the figure indicating the amount of enrichment of the relevant biological process, molecular function and/or cellular component.

1D Biological Processes



1D Molecular Function

Chapter 3

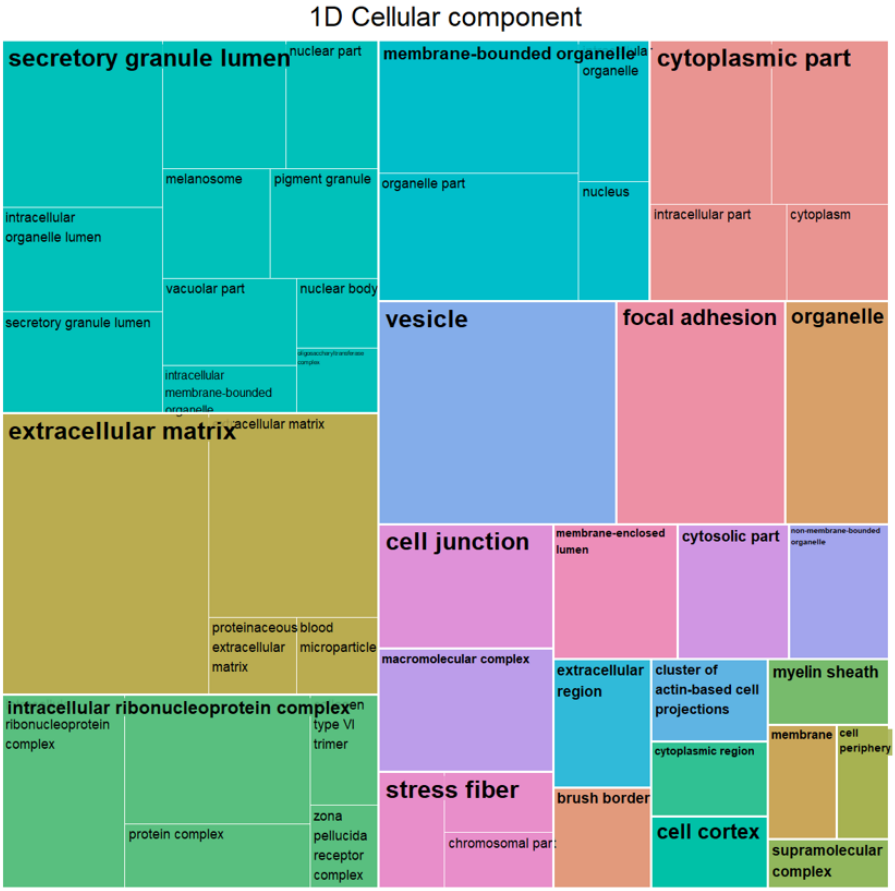
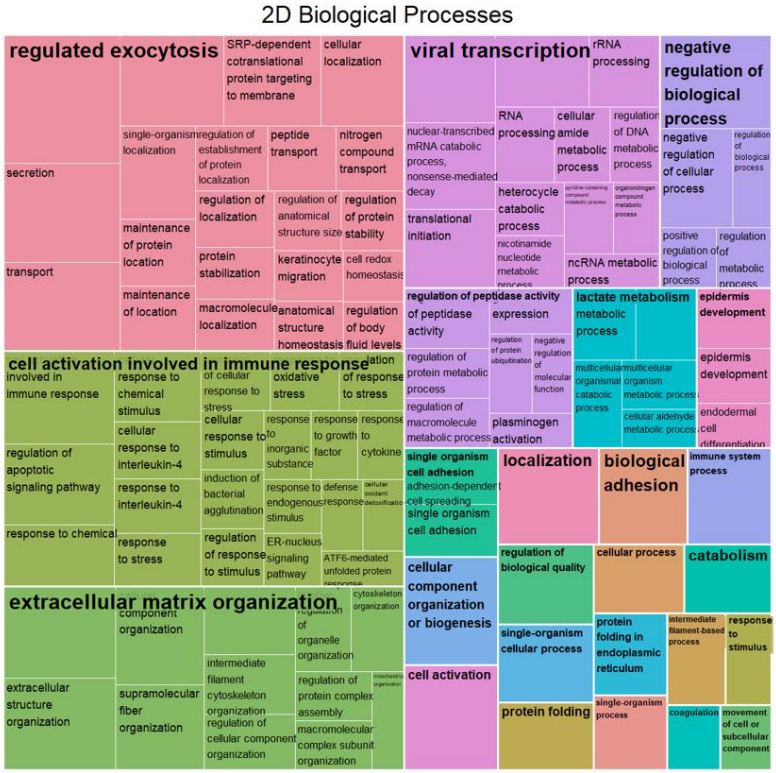


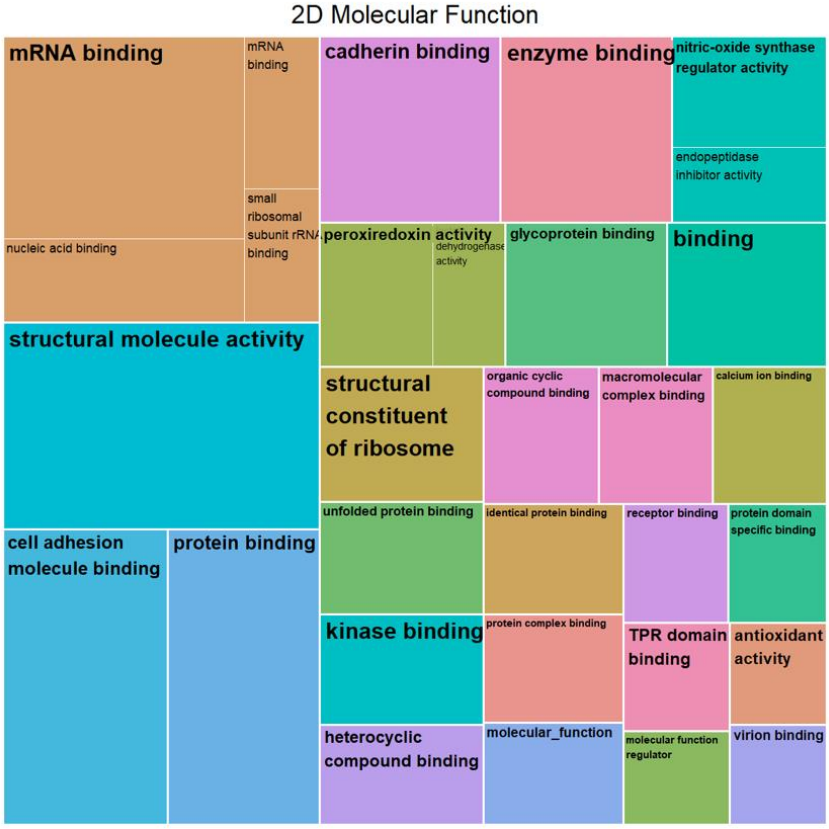
Figure 3:23: REViGO gene ontology analysis of significantly differentially expressed proteins in the 1D data.

Significantly differentially expressed proteins in the 1D data were inputted into GoGorilla gene ontology software. Gene ontology terms were then inputted into REViGO to reduce redundancies. The area of each component in the figure is representative of its enrichment.

Using the differentially expressed proteins from the 1D data, several areas of gene ontology were found to be highly enriched. The largest, and therefore the highest, enriched area within biological processes was ‘mRNA metabolism’, followed by ‘regulated exocytosis’. ‘Extracellular matrix organisation’, ‘response to unfolded proteins’ and ‘cell activation involved immune response’ were also areas of enrichment within the significantly differentially expressed proteins from the 1D data. Within the molecular function gene ontology area, the largest proportion of enrichment is in ‘telomeric DNA binding’. Second to telomeric DNA binding are several other areas of enrichment, sharing similar proportions, including ‘cell adhesion’, ‘structural molecular activity’, ‘cadherin binding’, ‘protein binding’ and several other ontologies involved in various aspects of binding.

'Secretory granule lumen', 'extracellular matrix (ECM)' and 'intracellular ribonucleoprotein complex' were all areas of enrichment within the cellular component area of gene ontology.





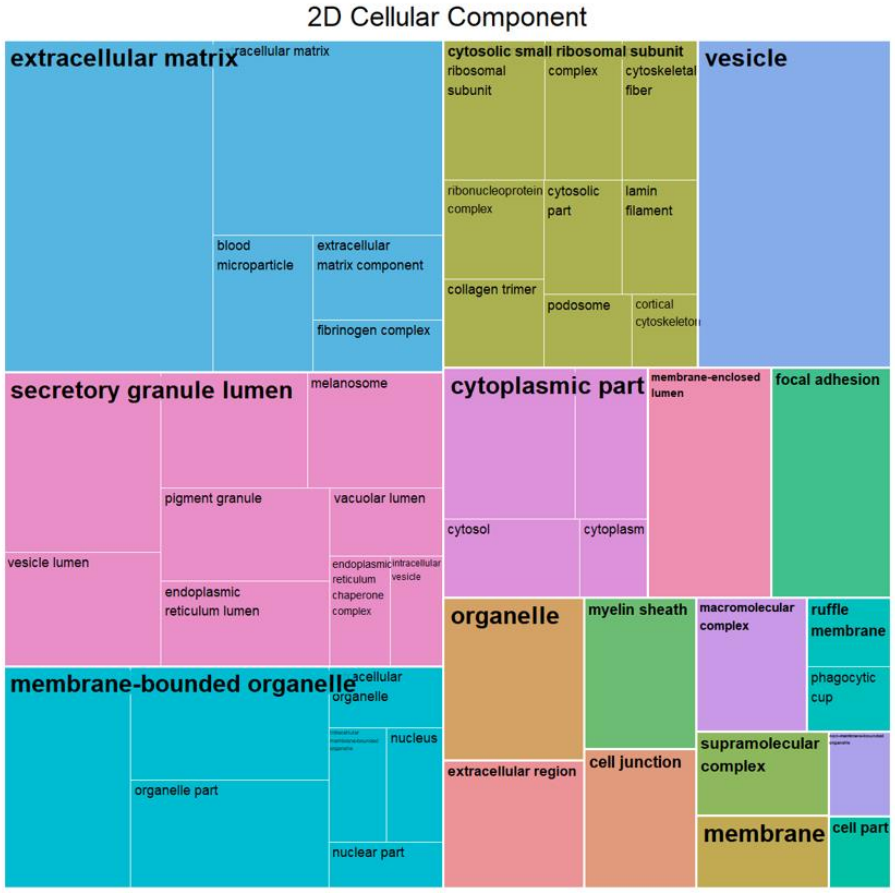


Figure 3:24: REViGO gene ontology analysis of significantly differentially expressed proteins in the 2D data.

Significantly differentially expressed proteins from the 2D data were subjected to gene ontology analysis using GoGorilla. Gene ontology terms and p-values were reduced of redundancies (keeping most enriched terms) by REViGO and presented in tree map format, with the size of each area representative of the amount of enrichment of this process, function and/or component.

Using the differentially expressed proteins identified in the 2D data, the highest area of enrichment in biological processes was 'regulated exocytosis' and 'cell activation involved in immune response' (**Figure 3.24**). 'Extracellular matrix organisation' and 'viral transcription' was also highly enriched in the biological process area. 'mRNA binding' and 'structural molecule activity' was highly enriched in molecular function gene ontology areas in addition to various other binding ontologies. Within the cellular component area, 'extracellular matrix' and 'secretory granule lumen' are highly enriched. 'Membrane bound organelle', 'cytosolic small ribosomal subunit' and 'vesicle' gene ontologies were also enriched.

'Regulated exocytosis' and 'cell activation involved in immune response' were both highly enriched biological processes in the REViGO results from both the 1D and 2D data. Within molecular function, many binding ontologies were enriched in both the 1D and 2D data. 'Extracellular matrix' and 'secretory granule lumen' gene ontologies were highly enriched within cellular component in both the 1D and 2D data.

3.3.17 Ingenuity pathway analysis

In addition to STRING/KEGG pathway enrichment analysis and GO/REViGO, ingenuity pathway analysis (IPA) was performed on all of the 1D and 2D proteomic data as well as on the significantly differentially expressed proteins from the 1D and 2D data (**Figure 3.25**). The results were combined into one graph and ranked by the sum of the Log10 P-values for each data set. A –log10 P-value cutoff of >5 (P< 0.00001) was employed on the sum of the –log10 P-values for each pathway. IPA revealed a number of significantly enriched pathways, the most significant being 'EIF2 signalling'. Some data sets had insufficient data to report an activation state (Zscore), for instance 'EIF2 signalling' in the significantly differentially expressed proteins data sets following 1D and 2D fractionation. There were several immune related pathways which were significantly enriched, including 'leukocyte extravasation signalling', 'FC receptor-mediated phagocytosis in macrophages and monocytes', 'production of nitric oxide and ROS in macrophages' and 'CD28 signalling in T helper cells'. Furthermore, there were a number of significantly enriched pathways associated with integrin signalling and intracellular signalling pathways, including 'ILK signalling', 'integrin signalling', 'PI3K/AKT signalling' and 'ERK/MAPK signalling'. Several Rac/Rho signalling pathways were also identified, including 'RhoGDI signalling', 'RhoA signalling' and 'signalling by Rho family GTPases' and 'Rac signalling'. Furthermore, 'signalling by Rho family GTPases' was amongst the most activated pathways (mean Z score = 2.377). Conversely, RhoGDI was one of the few pathways identified as inhibited (mean Z score = -1.741).

A unique function of IPA is to identify upstream regulators based on the data provided. The four data sets used for pathway analysis were also used for this function (**Figure 3.26**). It was predicted that a number of upstream regulators were significantly enriched, several of which were immune related, including TCR (mean Z score = 2.125), IgG (mean Z score = -

3.769), IL15 (mean Z score = 3.024), IL6 (mean Z score = 2.556), IL1a (mean Z score = 2.498) and TGFB1 (mean Z score = 3.029); some of these were associated with a positive Z score whereas others had a negative Z score. MicroRNA 122, and more specifically miRNA-122-5p, were also denoted by a negative Z score, thus predicted to be inhibited. Several other noteworthy predicted upstream regulators were EGFR, TP63, PI3K, CTNNB1 and CD44.

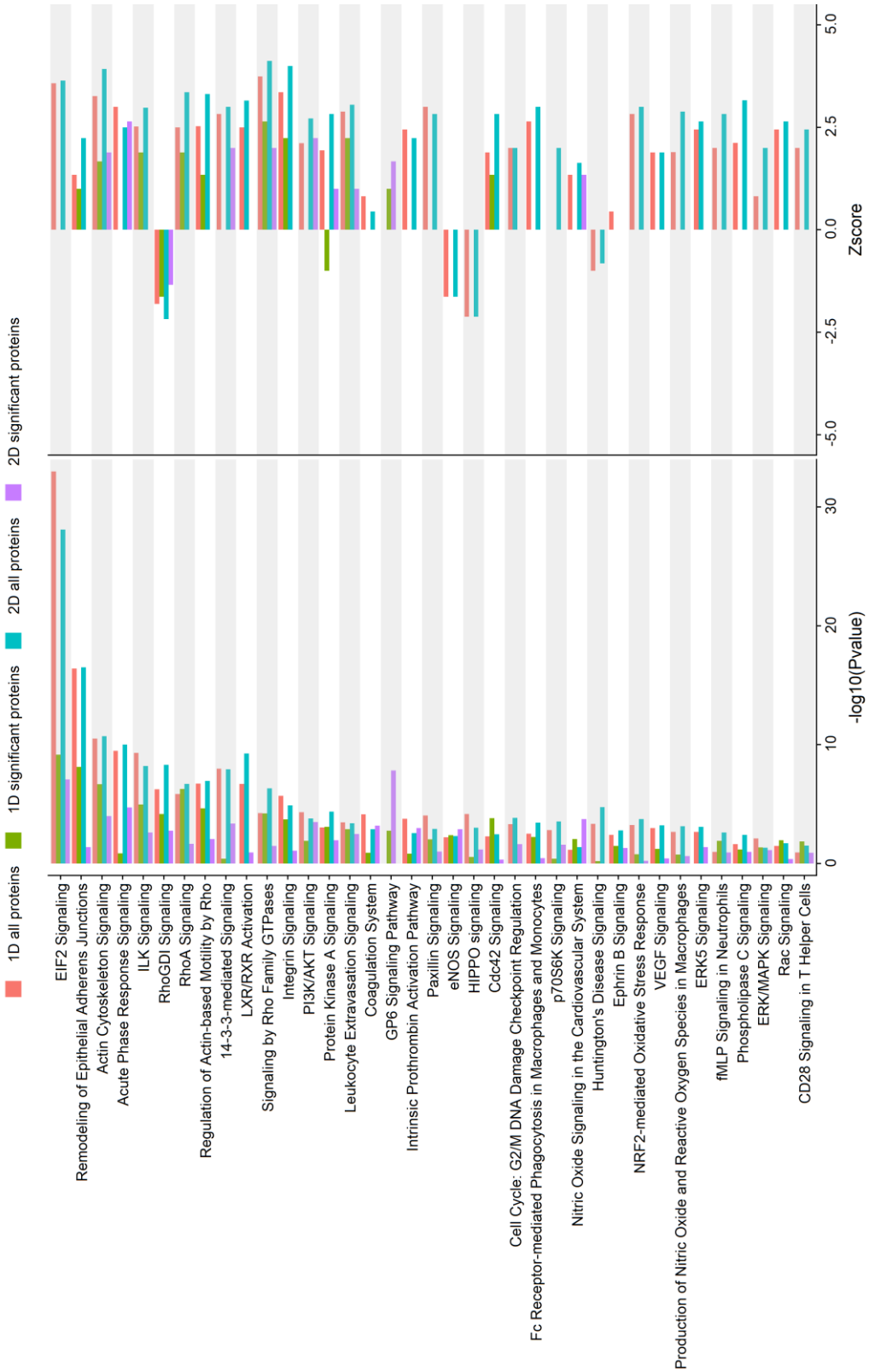


Figure 3:1: Ingenuity pathway analysis of cSCC proteomic data.

Four datasets were input into IPA individually, including all proteomic data from 1D and, separately, 2D experiments, as well as the significantly differentially expressed proteins in the 1D and, separately, 2D data. Results were stacked and ranked according to the sum of the log10 pvalues for each pathway. Activation state given as Zscore where positive numbers represent an activation and negative numbers an inhibition.

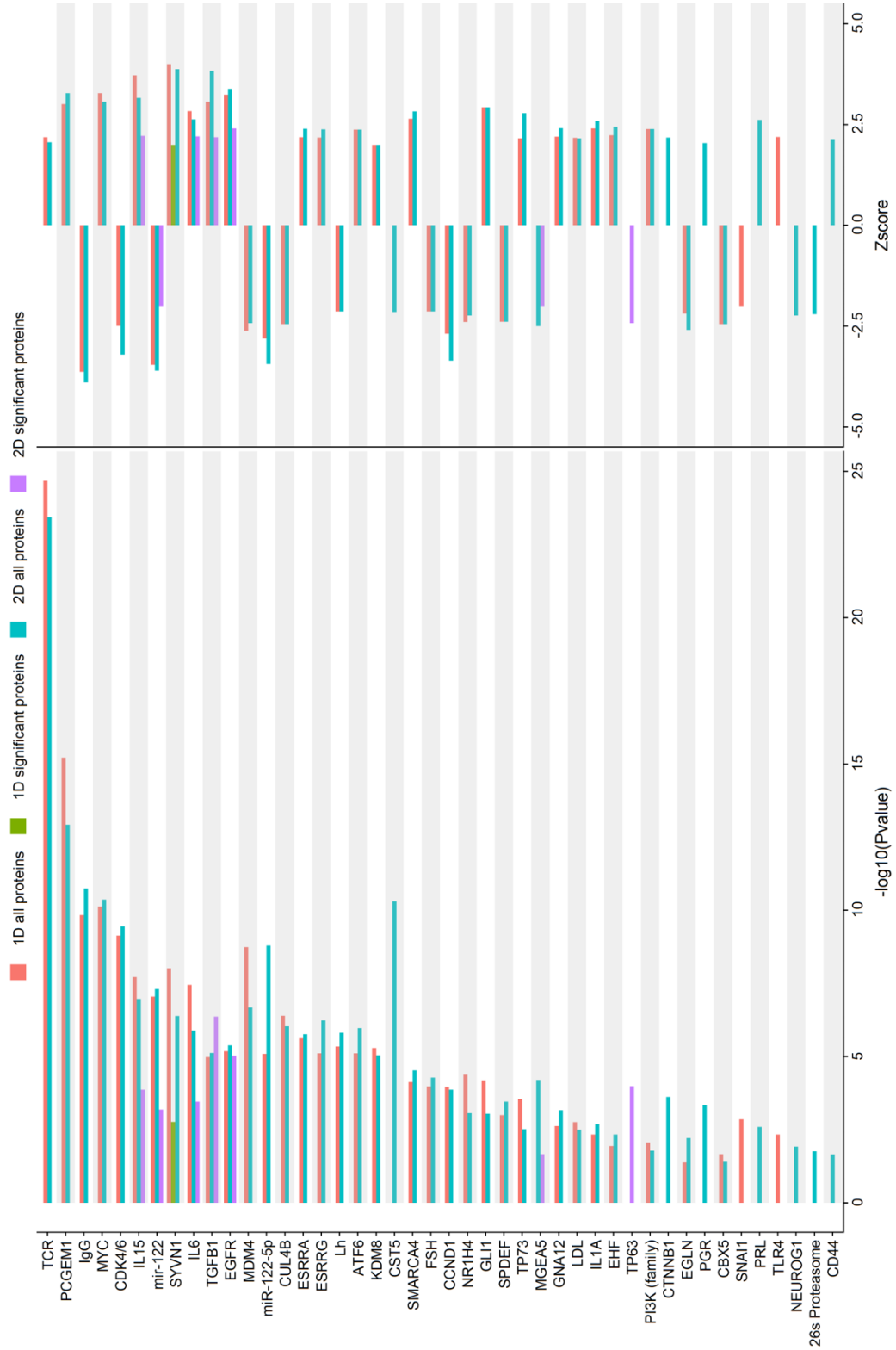


Figure 3:2: Upstream analysis of cSCC proteomic data using ingenuity pathway analysis

Four datasets were input into IPA individually, including all proteomic data from 1D and separately 2D as well as the y significantly differentially expressed proteins in the 1D and, separately, 2D data. The results were stacked and ranked according to the sum of the log10 pvalues for each pathway. Activation state is given as a Zscore where positive numbers represent an activation and negative numbers an inhibition.

3.3.18 Weighted gene co-expression network analysis

Weighted gene co-expression network analysis (WGCNA) is a bioinformatics approach which aims to highlight groups of genes (or proteins) which are heavily correlated (either positively or negatively) with each other. Several pre-processing steps are required before such a network can be derived (**Figure 2.3**). Correlated proteins were clustered into a total of 6 modules and subjected to pathway analysis (**Figure 3.27**) and correlated to clinical/histological characteristics (**Figure 3.28**). The 'turquoise' module had the greatest number of significantly enriched pathways and protein complexes with the most significantly being neutrophil degranulation. Other noteworthy significant results in the 'turquoise' module are protein processing in the ER, platelet degranulation and IL-12 family signalling. The 'blue' module had the second most number of pathway enrichments, including; ribosome, peptide chain elongation, and selenocysteine synthesis. The 'yellow' module had significant enrichment in neutrophil degranulation and keratinization. Module-trait analysis identified several relationships between protein modules and clinical/histological characteristics. The 'brown' module correlated positively with CD1 intratumoural stain (r=0.32, P=0.03). The 'blue' module was positively correlated with metastasis (r=0.29, P=0.04) and inversely with CD1a intratumoural stain (r=-0.29, P=0.04). The 'turquoise' module heavily correlated positively with both Clarks level (r=0.49, P=0.0004) and inversely with CD1a peritumoural stains (r=-0.5, P=0.0003). The 'yellow' module had an inverse correlation with differentiation (r=-0.39, P=0.006) and CD20 (r=-0.33, P=0.02).

source	term name Protein databases (CORUM protein complexes)	term ID	n. of term genes	corrected p-value	BLUE	TURQUOISE	YELLOW	RED	BROWN	GREEN
cor	40S ribosomal subunit, cytoplasmic	CORUM:305	33	6.13e-09	9	8				
cor	iNOS-S100A8/A9 complex	CORUM:6827	3	1.86e-04		2				
cor	CCT complex (chaperonin containing TCP1 complex)	CORUM:126	8	1.96e-07		7				
cor	BBS-chaperonin complex	CORUM:6247	10	1.26e-04		6				
cor	Ribosome, cytoplasmic	CORUM:306	80	7.81e-14	15	17				
cor	Nop56p-associated pre-rRNA complex	CORUM:3055	104	3.22e-09	13	18				
cor	(S100A8/S100A9)2 heterotetramer	CORUM:6826	2	6.19e-05		2				
cor	40S ribosomal subunit, cytoplasmic	CORUM:338	31	3.25e-09	9	8				
cor	Emerin complex 52	CORUM:5615	23	2.79e-04		8				

source	term name Biological pathways (KEGG)	term ID	n. of term genes	corrected p-value	BLUE	TURQUOISE	YELLOW	RED	BROWN	GREEN
keg	Protein processing in endoplasmic reticulum	KEGG:04141	165	2.22e-05		17				
keg	Regulation of actin cytoskeleton	KEGG:04810	213	7.84e-04		17				
keg	Ribosome	KEGG:03010	136	1.03e-15	15	17				
keg	Metabolic pathways	KEGG:01100	1297	3.62e-05					17	
keg	Glycolysis / Gluconeogenesis	KEGG:00010	69	4.16e-05		11				
keg	Legionellosis	KEGG:05134	55	3.95e-05		10				
keg	Biosynthesis of amino acids	KEGG:01230	76	8.44e-04		10			4	
keg	Pathogenic Escherichia coli infection	KEGG:05130	55	2.18e-08		13				
keg	Salmonella infection	KEGG:05132	84	3.13e-04		11				
keg	Tight junction	KEGG:04530	172	2.02e-04		16				

source	term name Biological pathways (Reactome)	term ID	n. of term genes	corrected p-value	BLUE	TURQUOISE	YELLOW	RED	BROWN	GREEN
rea	Platelet degranulation	R-HSA-114608	127	1.51e-13		23				
rea	Apoptosis	R-HSA-109581	169	8.85e-05		16				
rea	Glycolysis	R-HSA-70171	72	5.41e-04		10				
rea	Interleukin-12 family signaling	R-HSA-447115	59	5.19e-07		12				3
rea	Peptide chain elongation	R-HSA-156902	90	1.47e-18	15	19				
rea	Neutrophil degranulation	R-HSA-6798695	476	3.30e-23		53	10			
rea	Viral mRNA Translation	R-HSA-192823	90	1.47e-18	15	17				
rea	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	R-HSA-975956	96	4.13e-18	15	18				
rea	Host Interactions of HIV factors	R-HSA-162909	130	9.38e-04						5
rea	Selenocysteine synthesis	R-HSA-2408557	94	2.95e-18	15	17				
rea	Keratinization	R-HSA-6805567	216	2.07e-17			15			
rea	HSF1 activation	R-HSA-3371511	12	8.00e-04		5				
rea	Post-translational protein phosphorylation	R-HSA-8957275	108	5.51e-04		12				
rea	Beta oxidation of palmitoyl-CoA to myristoyl-CoA	R-HSA-77305	3	9.66e-06					3	
rea	Apoptotic cleavage of cell adhesion proteins	R-HSA-351906	11	5.25e-04			3			
rea	RHO GTPases activate PAKs	R-HSA-5627123	21	3.88e-05		7				

Figure 3:27: Protein complex and pathway analysis of WGCNA modules.

Modules identified by WGCNA were subjected to KEGG and Reactome pathway analysis in addition to overlay of the CORUM database. Strong hierarchical filtering was employed to reduce the number of terms and ease interpretation. Only results with P<0.001 are shown.

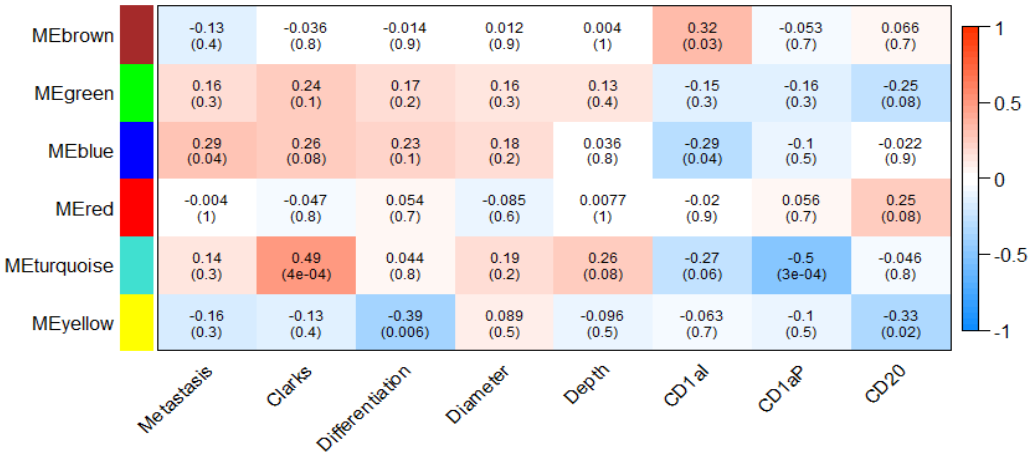


Figure 3:28: Module-trait analysis of WCGNA.

WGCNA modules were correlated to clinical and histological characteristics using Pearson correlation to identify relationships. In each cell, the upper value is the correlation coefficient (r) and the lower value is the P-value. ME, module eigengene; CD1aI, CD1a intratumoural stain; CD1aP, CD1a peritumoural stain.

3.3.19 Topological data analysis (TDA)

Topological data analysis is a higher dimensional analysis strategy aimed at identifying patterns in data and represents an alternative way of exploring data sets in addition to classical statistical analysis. Using the non-parametric Kolmogorov-Smirnov (KS) test for significance and various metrics (e.g. hamming, regression) for clustering, it enables an alternative method of investigating large data sets. Structures are generated using 2 or more variables, allowing colour mapping to reveal location of variables within sample nodes. Nodes can consist of one or more samples, depending on the similarity between them.

Currently, prognosis of cSCCs regarding whether or not they will metastasise is typically calculated by a number of factors, including differentiation, diameter and depth. For this reason, diameter, depth and differentiation of samples were inputted into Ayasdi to create a topological model structure to assess how these factors relate to outcome (**Figure 3.29**). Colour mapping of outcome onto the structure from the diameter, depth and differentiation revealed groups of blue (P-NMs) and groups of red (P-Ms) with some intermediate groups of yellow and green indicating nodes with both P-Ms and P-NMs within them.

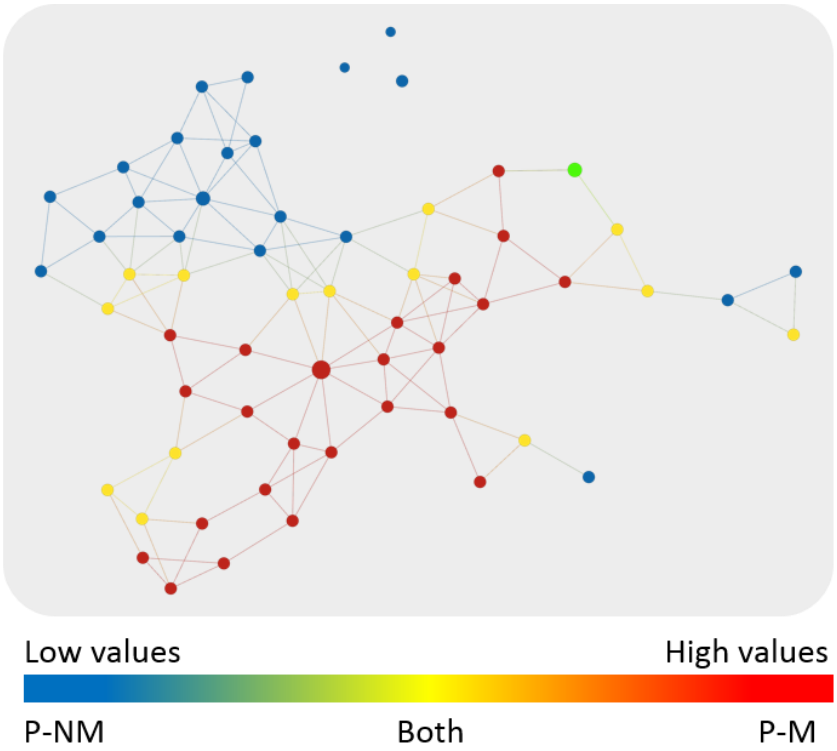


Figure 3:29: Topological model of cSCC differentiation, diameter and depth against subsequent development of metastases.

A topological model was created from samples' diameter, depth and differentiation values. A hamming metric was used with two neighbourhood lenses; resolution 40, gain 7.5. Differentiation was categorised into 1 (well differentiated) 2, (moderately differentiated) and 3 (poorly differentiated). Outcome (i.e. subsequently metastasised or did not metastasise) was colour mapped onto the resulting structure. Blue indicates P-NMs, red indicates PMs and yellow and green represent a combination of P-NMs and PMs.

To determine if the same separation of P-NMs and P-Ms could be obtained using proteomic abundancy data, a structure using just the identified proteins and their abundancies was created in Ayasdi (**Figure 3.30**). Protein data from both 1D and 2D separation experiments were capable of producing topological structures that separated P-NMs (blue) from P-Ms (red). Intermediate groups (yellow) containing both P-Ms and P-NMs could be found interlinking blue and red nodes.

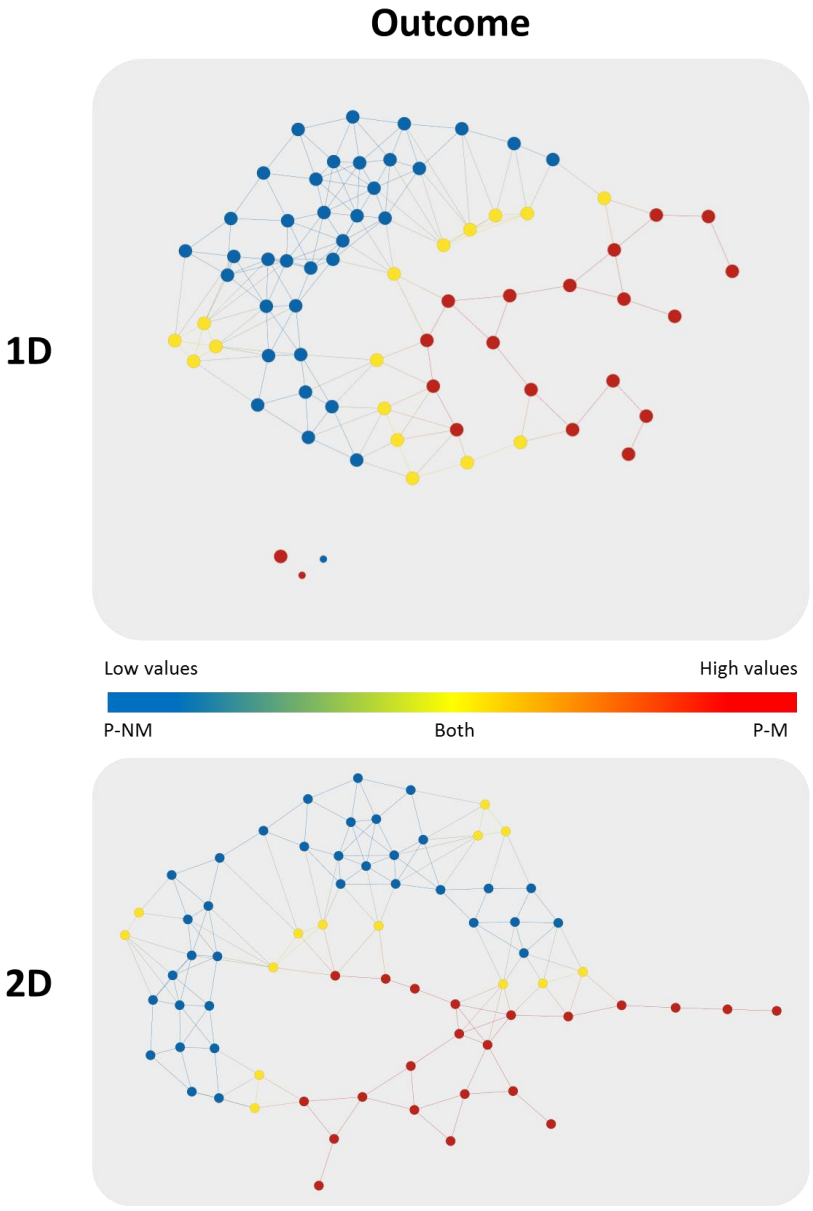


Figure 3:30: Topological model of 1D and 2D protein data against subsequent development of metastases.

Using proteomic data from 1D and 2D LC experiments, a topological model was created in Ayasdi using a hamming metric with two neighbourhood lenses. Lens resolution = 40, gain = 7.5 for 1D data. Lens resolution = 33, gain = 7 for 2D data. Outcome (i.e. subsequently metastasised or did not metastasise) was colour mapped on top of the structure. Blue represents P-NMs, red represents P-Ms and yellows indicate nodes containing both P-Ms and P-NMs. Protein data from 1D and 2D data resulted in topological structures which largely separated the P-M and P-NM groups.

To determine if protein driven topological structures have a correlation with differentiation, depth or diameter, the structures created in **Figure 3.30** were colour mapped accordingly (**Figure 3.31**). Colour mapping of differentiation on the 1D and 2D protein data topological structures revealed some equivocal separation of the clinical groups. Depth and diameter in **Figure 3.31** show less pronounce groups than differentiation but nonetheless show a general increase in higher values towards the P-M outcome section (shown in **Figure 3.30**).

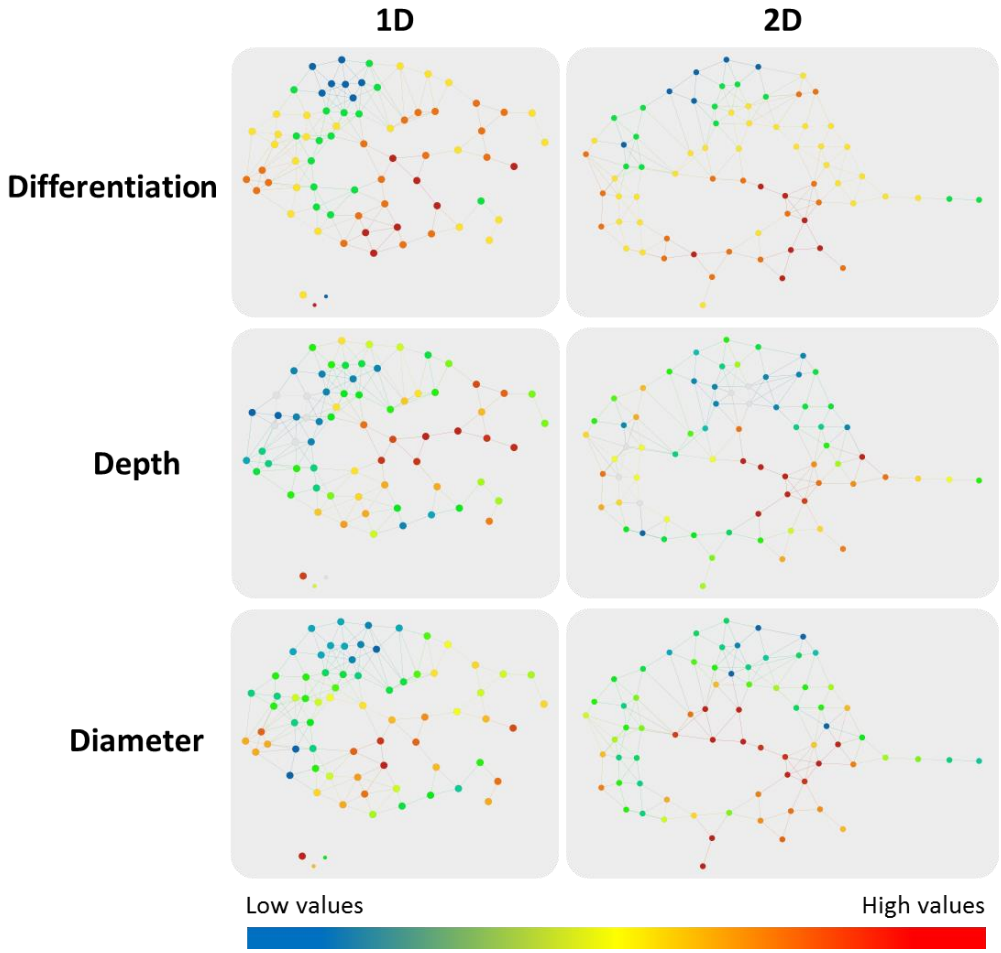


Figure 3:31: Topological structures of 1D and 2D protein data in relation to differentiation, depth and diameter of the cSCCs.

The topological structures for the 1D and 2D protein data were colour mapped for differentiation, depth and diameter of the cSCCs. Blue and green represent lower values, red and orange represent higher values (i.e. Larger in diameter or depth and less differentiated). Colour mapping revealed some clustering of poor, medium and well differentiated nodes, whereas the structures related to diameter and depth revealed several loose clusters of nodes. Higher values of differentiation were seen in P-M outcome regions identified in **Figure 3.30**.

Creating a structure that can differentiate between P-M and P-NM clinical outcome from the samples' proteomes (**Figure 3.30**) enables the comparison between the groups to establish driving variables that could be of importance in the transition between P-NMs and P-Ms. A comparison of exclusively P-NM nodes (completely blue nodes) was compared to exclusively P-M nodes (completely red nodes) from the topological structure from the 1D data in **Figure 3.30**. In doing so, a number of variables were seen to drive the different groups with 40 of these variables being significant (**Table 3.6**). 80% of significant differences (32 variables) were identified in previously performed analyses. The 8 newly identified differences between the groups consisted of 6 proteins and 2 clinical parameters. The clinical parameters found to drive the differences between the groups were diameter (P= 0.00808) and differentiation (P= 0.04685). The 6 newly identified significant proteins were ACTN3 (P= 0.0984), SFN (P= 0.0239), OTUB1 (P= 0.0335), RPL18 (P= 0.0355), NAGK (P= 0.0397) and KRT74 (P= 0.0443).

P-M and P-NM groups from the TDA structure of the 2D data in **Figure 3.30** were compared using Kolmogorov-Smirnov test for significance and revealed 59 variables significantly different between the two groups (**Table 3.7**). 80% of these (47 variables) had been previously identified in preceding analyses. Of the 12 newly identified variables, 2 were clinical parameters (diameter and differentiation of the cSCCs) and the remaining 10 were proteins. The 10 newly identified significantly different proteins were HSPB1 (P= 0.0054), UQCRC1 (P= 0.0197), APOE (P=0.0236), DMKN (P= 0.0236), APRT (P= 0.0250), SARNP (P= 0.0298), KRT14 (P= 0.03349), IGKV3-20 (P= 0.03968), COL1A1 (P= 0.04435) and USP5 (P= 0.04435). 15% of all significantly different variables identified were found in both TDA analyses of the 1D and 2D data.

Table 3.6: Group analysis for TDA structure of 1D data comparing P-M and P-NM groups.

Variable	Gene ID	p-value	KS Score
Outcome (metastasis)	N/A	2.55E-09	1
P61158	ACTR3	3.19336E-05	0.816176471
P13010	XRCC5	0.001163116	0.769230769
P50991	CCT4	0.001702508	0.75
P07237	P4HB	0.001702508	0.595959596
Q9NSB2	KRT84	0.002467246	0.703296703
P50454	SERPINH1	0.003800385	0.588235294
P08133	ANXA6	0.003800385	0.582352941
P15880	RPS2	0.006610785	0.555882353
Q15019	SEPT2	0.008082509	0.637362637
P35222	CTNNB1	0.008082509	0.607142857
Diameter of cSCC	N/A	0.008082509	0.528947368
O00571	DDX3X	0.009223018	0.55
Q08043	ACTN3	0.009846365	0.615384615
Q02878	RPL6	0.012739434	0.539473684
O43707	ACTN4	0.013573266	0.498746867
Q14697	GANAB	0.016377393	0.529411765
P62081	RPS7	0.017421443	0.50140056
P29692	EEF1D	0.018524639	0.516447368
Q6KB66	KRT80	0.02358712	0.634920635
P31947	SFN	0.02358712	0.476315789
P62857	RPS28	0.026551868	0.482954545
Q9NZT1	CALML5	0.026551868	0.482352941
P62277	RPS13	0.028154288	0.56043956
Q15063	POSTN	0.028154288	0.473684211
P62318	SNRPD3	0.031617112	0.623931624
P52597	HNRNPF	0.033485006	0.549450549
Q96FW1	OTUB1	0.033485006	0.549450549
P08758	ANXA5	0.033485006	0.45112782
Q9NV66	TYW1	0.035449065	0.659090909
Q07020	RPL18	0.035449065	0.461111111
Q9UJ70	NAGK	0.03968188	0.573426573
P50395	GDI2	0.03968188	0.456582633
P36578	RPL4	0.041959012	0.566666667
P30044	PRDX5	0.041959012	0.456140351
P62140	PPP1CB	0.044349062	0.575757576
P16403	HIST1H1C	0.044349062	0.538461538
Q7RTS7	KRT74	0.044349062	0.538461538
P24821	TNC	0.044349062	0.470588235
Differentiation of cSCC	N/A	0.046856493	0.428229665
P15088	CPA3	0.049485877	0.504807692

*Blue nodes (P-NM) and red nodes (P-M) from the topological structure for the 1D data in **Figure 3.30** were compared using Kolmogorov-Smirnov's test. Orange shading in the table indicates proteins already identified in previous analyses. Blue text indicates variables found in analysis of TDA structures for both 1D and 2D data comparing P-Ms and P-NMs.*

Table 3.7: Significantly different variables between P-M and P-NM groups in 2D TDA.

Variable	Gene ID	p-value	KS Score
Outcome (metastasis)	N/A	2.09E-10	1
Diameter of cSCC	N/A	0.000920998	0.590909091
Q15063	POSTN	0.000995911	0.633053221
P08779	KRT16	0.001256214	0.564393939
P61981	YWHAG	0.001977241	0.571428571
P06396	GSN	0.003067621	0.535714286
P25398	RPS12	0.003295976	0.709090909
P31949	S100A11	0.003539914	0.60130719
P50395	GDI2	0.004691266	0.528138528
P04792	HSPB1	0.005387613	0.507575758
P02751	FN1	0.005770175	0.513457557
Q15582	TGFBI	0.007071711	0.523923445
P02545	LMNA	0.010507637	0.477272727
Q99878	HIST1H2AJ	0.014455892	0.466403162
P22626	HNRNPA2B1	0.014455892	0.462121212
P02675	FGB	0.015389755	0.464426877
P18206	VCL	0.016377393	0.476190476
P51884	LUM	0.016377393	0.476190476
P16615	ATP2A2	0.017421443	0.492105263
Q96FW1	OTUB1	0.019689815	0.529411765
P31930	UQCRC1	0.019689815	0.65
P04179	SOD2	0.019689815	0.567307692
P08758	ANXA5	0.02091991	0.443181818
P02649	APOE	0.02358712	0.522556391
Q6E0U4	DMKN	0.02358712	0.578947368
P48668	KRT6C	0.02358712	0.45
P07741	APRT	0.025030636	0.45021645
P02538	KRT6A	0.028154288	0.428030303
P62158	Calm1	0.028154288	0.43452381
P27816	MAP4	0.028154288	0.454761905
P63000	RAC1	0.028154288	0.454761905
P01871	IGHM	0.029841473	0.625
P82979	SARNP	0.029841473	0.554945055
P40121	CAPG	0.029841473	0.441558442
P11021	HSPA5	0.031617112	0.424242424
P35908	KRT2	0.031617112	0.424242424
P62314	SNRPD1	0.033485006	0.507936508
P02533	KRT14	0.033485006	0.420454545
P29401	TKT	0.033485006	0.45
O43390	HNRNPR	0.033485006	0.430641822
P46782	RPS5	0.035449065	0.543956044
P14625	HSP90B1	0.035449065	0.420948617
Q07960	ARHGAP1	0.035449065	0.510121457
P60660	MYL6	0.03968188	0.412878788
Differentiation of cSCC	N/A	0.03968188	0.412878788
P12111	COL6A3	0.03968188	0.412878788
P09525	ANXA4	0.03968188	0.422360248
P60174	TPI1	0.03968188	0.412878788
P01623	IGKV3-20	0.03968188	0.504201681
Q562R1	ACTBL2	0.041959012	0.492063492
P36957	DLST	0.041959012	0.444444444
Q14697	GANAB	0.041959012	0.428571429
P62937	PPIA	0.041959012	0.409090909
P02452	COL1A1	0.044349062	0.40530303
P08123	COL1A2	0.044349062	0.40530303
P45974	USP5	0.044349062	0.575757576
P12110	COL6A2	0.046856493	0.401515152
P68104	EEF1A1	0.046856493	0.401515152
P24821	TNC	0.049485877	0.409090909
P62277	RPS13	0.049485877	0.436507937

Blue nodes (P-NM) were compared to red nodes (P-M) from the topological structure for the 2D data in ***Figure 3.30*** *using Kolmogorov-Smirnov's test to investigate driving variables. Orange shading in the table indicates proteins previously identified in prior analyses. Blue text indicates variables identified in analysis of TDA structures for both 1D and 2D data comparing P-Ms and P-NMs.*

Chapter 3

3.4 Discussion

This study aimed to identify biomarkers of metastasis in cSCC using FFPE samples. The initial direction of the study focused on IHC staining of P-M and P-NM samples to look at certain immunological parameters which might have relevance to development of cSCC metastases while also becoming familiar with histological parameters of cSCC.

It has previously been reported that there are higher numbers of tumour associated B cells in primary melanomas that did not metastasise than in those that did metastasise and that a higher number of tumour associated B cells is associated with significantly better overall survival in cutaneous melanoma (Garg et al., 2016). In this study, we stained for B cells to answer the question of whether this difference could also be seen in cSCC in addition to gaining experience in skin histology which would ultimately aid in microdissection for proteomics. However, this current study found no difference in CD20+ cell numbers between P-M and P-NM in cSCCs (**Figure 3:3**), although, higher numbers of CD20+ cells was associated with a longer time to metastasis (**Figure 3:5**). Staining for CD1a, a Langerhans cell marker, revealed there were significantly more Langerhans cells in P-NM than P-M tumours (**Figure 3:4**). Furthermore, a significant relationship was found between intratumoral CD1a cells and time to metastasis (**Figure 3:6**). It is unclear whether this is a causal association, i.e. whether the tumour has gone on to metastasise because there are less Langerhans cells in the primary cancer, or whether the reduced numbers of Langerhans cells are an epiphenomenon of tumours that metastasise. Langerhans cells are a specialised type of dendritic cell residing in the epidermis (Merad et al., 2002). Much of the published research in the literature on Langerhans cells and cSCC has investigated Langerhans cells contribution to cSCC development from normal skin (Lewis et al., 2015, Schwarz et al., 2010), rather than on Langerhans cells contribution to metastasis.

Langerhans cells can promote T cell activation (Fujita et al., 2012) and it has been reported that there are more CD8+ cytotoxic T cells in P-NMs than P-Ms (Lai et al., 2015, Lai et al., 2016). It is possible that the increase in Langerhans cells in P-NMs may promote T cell activity, hindering carcinogenesis progression and therefore preventing it from metastasising, but it is not clear whether this would occur via antigen presentation to CD4+ T cells by MHC class II or via cross-presentation to CD8+ T cells. However, it has been reported that murine epidermis lacking Langerhans cells developed carcinogenesis less

readily than those with Langerhans cells intact, suggesting a pro-oncogenic role for Langerhans cells (Lewis et al., 2015). One could argue that the latter observation is not relevant to UVR-induced skin cancer because that study used chemical carcinogenesis and it is known that Langerhans cells metabolise 7,12-dimethylbenz[a]anthracene (DMBA) to the carcinogen DMBA-trans-3,4-diol (Modi et al., 2012). In the current study, the fewer Langerhans cells in P-M may be due to the malignant keratinocytes causing a reduction in the number of Langerhans cells in the tumour, or possibly the tumour not being seen as "threatening" by the immune system, or due to an increased migration of Langerhans cells to lymph nodes in an attempt to activate an immune response to a more aggressive cSCC. Related to this, it is known that Langerhans cells can induce Tregs in some circumstances (Seneschal et al., 2012, Gomez de Aguero et al., 2012); for example UVR damaged Langerhans cells have been reported to migrate to lymph nodes and activate T regulatory cells, a mechanism which may promote tumorigenesis (Schwarz et al., 2010). In P-M cSCCs, it is possible that migration of Langerhans cells to the lymph nodes might result in the increase in T regulatory cells that has been reported in P-Ms previously (Lai et al., 2015, Lai et al., 2016) thus resulting in a dampened immune response and allowing the tumour to grow and metastasise. Furthermore, ingenuity pathway analysis revealed upstream regulator, TGBF1, a protein which is in part, responsible for Treg cell's suppressive function (Wu et al., 2016), as significantly activated in the P-M group (**Figure 3:25**).

Extracting proteins from FFPE samples is a complex technique, with no optimal protocol established to date (**Appendix 1**). The current study found that using a combination of 2D online fractionation and an extraction method using RapiGest surfactant gave a high protein yield, consistently producing better yields than many results in the current literature (**Appendix 1**). Furthermore, analysis of this method revealed that almost half of the proteins identified were found in all three repeats and that the protein abundancies were very similar between samples, suggesting good reproducibility. 1D fractionation is an established method in our laboratory for other types of proteomics and therefore it was decided to undertake both 1D and 2D fractionation as independent methodologies, producing two separate sets of results for cSCCs. Over 4,000 unique proteins were identified in this study, which is one of the highest proteome coverages achieved from proteomics studies using FFPE samples (**Appendix 1**). Furthermore, this study has achieved the highest number of protein IDs compared to any other proteomic studies on FFPE cSCCs

in the published literature, including the most recent studies by Azimi *et al* (Azimi et al., 2016, Azimi et al., 2019)and Foll *et al (Foll et al., 2017)* which achieved 1,310, 3574, and 2,102 protein IDs, respectively.

To establish which statistical analysis to carry out to compare the P-M and P-NM groups, the proteomics data had to be assessed to see whether it had a normal (Gaussian) distribution. As previously stated, proteomics data suffers from the inability to detect proteins below certain concentrations and therefore it can suffer from the "floor effect". The floor effect (also referred to as the basement effect) occurs when data cannot be recorded below a certain level and thus the distribution of data is skewed because the *true* lowest values may not be present (Karp and Lilley, 2007). Therefore, a conservative non-parametric approach was used for statistical analysis of the proteomics data in this current study. Similarly to other 'omics' studies, significance testing large data sets has a probability of producing false positives. In smaller data sets, family-wise error rates through tests such as Bonferroni, are usually applied. However, doing this in large data sets can cause false negatives as criteria become more stringent with every variable measured (Noble, 2009). Statistical advice (from Research Design and Methodology, University of Southampton) suggested that plotting all p-values, obtained through statistical analysis, into a histogram which would clarify whether the data was enriched in true significant results (higher in the $P<0.05$ region) or whether it represented a set of significant p-values which were due to chance alone (where the histogram of p-values would be expected to look flat). Different percentages of missing values were considered during the analysis, but a missing value of 50% was chosen in order to include the highest numbers of samples per protein yet still have confidence in the results. The option to impute missing values was considered, but it is known that not imputing data results in higher statistical power and confidence (Bantscheff et al., 2012, Webb-Robertson et al., 2015), therefore no imputation was carried out.

Proteomic quantification data produced through 1D and 2D fractionation was explored in a variety of ways. Although a number of driver genetic mutations are known to exist in cSCC, many of the proteins encoded by these genes (including NOTCH1, NOTCH2, p53, HRAS, NRAS, BRAF, PI3K (South et al., 2014, Pickering et al., 2014, Li et al., 2015) did not appear in the data in the current study. Some proteins such as CDKN2A, Kras and others

encoded by mutated genes reported in the South *et al* (2014), Pickering *et al* (2014) and Li *et al* (2015) publications appeared but in very few samples. However, the effects of the mutations are more likely to alter function than abundance of the mutated protein. Admittedly, IHC studies on cSCC have identified elevated levels of nuclear p53 in cSCCs (Missero and Antonini, 2014) but the use of FFPE samples in the proteomics is likely to have resulted in preferential enrichment of cytosolic proteins rather than those in the nucleus and/or membranes. Support for this comes from the use of the software programme, Panther, in the present study which indicated that approximately 10% - 15% of the proteins identified by mass spectrometry following 1D and 2D separation were attributed to a nuclear or membrane location.

In order to identify proteins relevant to cSCC metastasis, multiple analyses were performed; these included classical statistics (Mann Whitney U test), String analysis/KEGG pathway enrichment, Gene ontology analysis, IPA, WGCNA and TDA. These analyses aimed to explore the data in different perspectives in order to provide a comprehensive view of the differences between P-M and P-NM cSCCs and to ultimately identify important pathways and proteins involved in the development of metastases from this cancer. In view of the vast numbers of differentially expressed proteins in P-Ms and P-NMs, it is not possible to discuss each individual protein in detail, but some of these proteins reoccur in several different types of analyses, suggesting that they play a role in driving metastasis. For example, Tenascin C (TNC) was identified in almost all of the analyses, including its involvement in multiple KEGG pathways, having a statistically significant difference in its levels between P-Ms and P-NMs, and being a driver between P-M and P-NM groups in the TDA. Tenascins are large extracellular glycoproteins involved in various cell functions including adhesion, signalling, proliferation and migration (Pas et al., 2006). TNC is elevated in several cancers including brain (McLendon et al., 2000), breast (Ioachim et al., 2002), cervical (Buyukbayram and Arslan, 2002), gastro-intestinal (Gazzaniga et al., 2005), head and neck (Atula et al., 2003), and melanoma (Ilmonen et al., 2004) and there is evidence that it may be involved in metastases of cancers from breast (Oskarsson et al., 2011) and colorectal origins (Gulubova and Vlaykova, 2006). Another protein found in most of the above analyses is glucosidase-α neutral AB (GANAB), an enzyme involved in cleaving glycoproteins. Interestingly, GANAB has been reported to show reduced expression in head

and neck cancers (Chiu et al., 2014), but there has been little research on this protein in other cancers.

Some proteins identified, such as transketolase (TKT), Rab GDP dissociation inhibitor 2 (GDI2), lumican (LUM), periostin (POSTN) and fibrinogen beta chain (FGB) were found to be significantly differentially expressed in multiple analyses but were not seen in the KEGG pathway analysis. Nonetheless, many of these proteins have been recognised as involved in cancer development. For example, TKT is a cellular enzyme involved in metabolism and has been reported to counteract oxidative stress and promote liver cancer development (Xu et al., 2016). GDI2 has been reported as a metastasis suppressor in bladder cancer by inhibiting Rho GTPases in the cytoplasm (Moissoglu et al., 2009). In this current study, however, we found an increase in both GDI2 and RAC1 (a Rho GTPase) in P-Ms and, moreover, it has been reported previously that GDI2 promoted epithelial-mesenchymal transition (which is relevant for metastasis (Kalluri and Weinberg, 2009)) through RAC1 mediated NF-κB activation (Cho et al., 2014). LUM, an extracellular matrix protein has been reported as a potential biomarker in cisplatin-resistant head and neck cancer (Yamano et al., 2010) and POSTN is an extracellular protein recognised as a promoter of epithelial-mesenchymal transition and has been reported in several cancers, promoting metastasis (Morra and Moch, 2011).

Many ribosomal proteins were identified as being differentially expressed, particularly ribosomal protein 13 (RPS13) which was identified in almost all analyses. RPS13 has been found in other cancers, for instance in gastric cancer (Guo et al., 2011) and in addition to RPS13, other ribosomal proteins such as RPS7, RPS20, RPS2, RPS28, RPL4, RPL6 and RPS10 were also found to be significantly different between P-Ms and P-NMs in many analyses performed in the present study. Cancerous cells proliferate rapidly and as a result require increased protein synthesis (White-Gilbertson et al., 2009), but it is not clear whether the increase in ribosomal proteins in P-Ms is causally contributing to development of metastasis or whether this increase is simply a result of a more aggressive type of cancer.

STRING/KEGG pathway analysis and gene ontology analysis identifies “areas of interest” based on significantly differentially expressed proteins. STRING analysis in the current study revealed clusters of interacting proteins, and mapping KEGG pathways on top of this structure revealed several pathways significantly enriched between P-Ms and P-NMs. One

such “pathway”, involving multiple proteins, was extracellular matrix receptor interactions (**Figure 3:22**). In addition, gene ontology analysis identified enrichment in extracellular matrix and extracellular matrix organisation (**Figure 3:23**, **Figure 3:24**). These findings suggest a strong involvement of extracellular matrix interactions in development of cSCC metastasis. The extracellular matrix is a complex network of proteins secreted by cells of a specific tissue (Venning et al., 2015). For metastasis to occur, a cancerous cell must separate from the primary tumour and invade adjacent tissue (Brodland and Zitelli, 1992). This is dependent on the composition of the extracellular matrix (ECM) and the ability of the cell to invade and migrate within this. Several proteins have been identified in this study as promoters of cancer cell migration in the ECM, including POSTN (Siriwardena et al., 2006, Michaylira et al., 2010), TNC (Venning et al., 2015) and ANXA5 (Peng et al., 2016, Ding et al., 2017). Furthermore, promotion of epithelial-mesenchymal transition of cells through interactions with ECM proteins has been reported, including several identified in this current study; examples are POSTN (Morra and Moch, 2011), GDI2 (Cho et al., 2014) and TNC (Nagaharu et al., 2011).

Additional to these findings, this current study identified an enrichment in focal adhesion in gene ontology (**Figure 3:23**, **Figure 3:24**) and KEGG pathway analysis (**Figure 3:20**, **Figure 3:22**), involving multiple highly significantly differentially expressed proteins (e.g. TNC, FN1, RAC1, VCL and multiple collagens). Focal adhesion is the formation of large assemblies between the proteins in the extracellular matrix and the integrins on the cell surface, and the resulting cell-matrix adhesions can have an effect on cell adhesion, migration and intracellular signalling (Nagano et al., 2012). IPA revealed significant activation of integrin signalling in P-Ms compared to P-NM (**Figure 3:25**). A number of integrins have been identified, as have their ligands, in the published literature, many of which were seen to differ in amount between P-Ms and P-NMs in this current study (e.g. TNC, FN1, POSTN, VCL and various collagens). Integrins are transmembrane proteins which can alter intracellular signalling, and have been reported to be involved in cSCC tumorigenesis and metastasis (and have been reviewed in a number of papers, e.g. (Eke and Cordes, 2015, Duperret and Ridky, 2013, Janes and Watt, 2006). A study investigating pancreatic cancer identified periostin (POSTN) as a ligand for α6 β4 integrin complex, resulting in an activation of PI3K-AKt signalling (Baril et al., 2006). Furthermore, TGFβ has also been reported to induce

expression and intracellular localisation of EGFR (increased expression seen in **Figure 3:25**) which in turn activates the PI3K-Akt signalling pathway (Wendt et al., 2010).

In this current study, PI3K-AKt signalling was found to be significantly enriched through KEGG pathway analysis (**Figure 3:20**, **Figure 3:22**) and IPA (**Figure 3:25**, **Figure 3:26**). PI3K-AKt signalling plays a major role in cellular function, regulating proliferation, growth and survival, and when dysregulated is well known to play an important role in cancer (Osaki et al., 2004, Danielsen et al., 2015) and development of metastasis (Yao et al., 2017, Li et al., 2017). The data in the current study also suggests an upregulation of ILK signalling in P-Ms (**Figure 3:25**), a pathway which has been reported to act with PI3K-Akt signalling to promote epithelial-mesenchymal transition (EMT), an important process in tumorigenesis and metastasis (Li et al., 2014). Many of the proteins identified in PI3K-AKt signalling in this current study were also involved in focal adhesion and/or extracellular matrix receptor interactions. These data collectively suggest an interaction between ECM, focal adhesion and PI3K-AKt signalling.

Another significantly enriched area involving the ECM comprised exosomes, exocytosis and extracellular signalling seen in GO analysis (**Figure 3:23**, **Figure 3:24**) and IPA (**Figure 3:25**). Exocytosis is the process of ejecting molecules from the cell via vesicle fusion with the plasma membrane. A role for ANXA5 has been reported in exocytosis and membrane repair (Bouter et al., 2011), as has the calcium ion binding protein CALML5 (albeit in neurones) (Burgoyne and Clague, 2003), with both of these proteins found in the present study to be significantly differentially expressed between P-Ms and P-NMs. Furthermore, it has been suggested that primary tumours can secrete factors (through exocytosis) to transform distant sites into pre-metastatic niches (Kaplan et al., 2005), through effect on extracellular signalling. Exosomes secreted from primary tumours have received a lot of recent attention due to their ability to prime these pre-metastatic niches for metastatic spread (Costa-Silva et al., 2015, Hoshino et al., 2015, Peinado et al., 2011).

WGCNA is a tool used to identify groups of heavily correlated genes in a dataset and relate them to clinical characteristics or identify pathway enrichment within them. Six modules of heavily correlated proteins were identified in this study, and these were arbitrarily labelled with a colour. The only module which was correlated (either positively or negatively) with metastasis was the blue one. This blue module expressed enrichment in many pathways

including various ribosomal activities and translation. Cancerous cells undergo rapid proliferation and therefore require timely protein synthesis and, generally, more aggressive tumours proliferate faster than less aggressive tumours. The blue module and its enrichment in protein processing and folding, seen in this current study, are therefore of no surprise. Furthermore, proliferation requires nucleotide binding and therefore the enrichment seen in these areas is also to be expected. The yellow module displayed enrichment in keratinisation and is understandably positively correlated to differentiation. The turquoise module showed significant pathway enrichment and was positively correlated to Clarks and CD1a peritumoural staining.

Another analytical tool for the exploration of large data sets being used in 'omics' studies is topological data analysis (Bigler et al., 2016). Using 3 current major prognostic markers of cSCC (differentiation, depth and diameter), a TDA structure was generated in the current study that largely distinguished outcome between most samples into P-Ms and P-NMs. This in itself suggests that TDA is a useful method to distinguish between P-M and P-NM tumours using certain histological characteristics. The proteomic data in the current study generated a TDA structure that largely distinguished between P-Ms and P-NMs. This, taken in conjunction with the earlier analysis of the proteomic data, suggests that TDA might be useful as an approach to identify a useful prognostic marker to distinguish between primary cSCC lesions which will and those which won't go on to metastasise. It is unclear at the present time whether a whole proteomic TDA or a targeted assay comprising of several proteins would produce a better prognostic marker, however, each approach will require further investigation and validation on a larger sample size.

To date, this is one of the largest mass spectrometry based proteomics studies on cSCC. Furthermore, it is one of the largest mass spectrometry based proteomic studies carried out on FFPE samples. More importantly, this study identified a number of potential biomarkers for metastasis in cSCC in addition to identifying several key pathways involved.

Chapter 3

Chapter 4: Verification and validation of cutaneous squamous cell carcinoma protein biomarkers using targeted mass spectrometry and machine learning

4.1 Introduction

Multiple reaction monitoring (MRM) is a targeted proteomic approach which utilises isotopically labelled peptides by the incorporation of heavy isotopes, to accurately determine the concentration of the native counterpart in a sample. Depending on the preciousness of the tissue samples and the number of samples, it is common to create a reference calibration curve of increasing amounts of heavy peptides to avoid the need to create a calibration curve within each sample. Typically, a sample is 'spiked' with a known concentration of heavy labelled peptide which is then ionised, usually by ESI. Ions are then separated in an initial mass analyser (often a quadrupole) to only allow the heavy and native peptides of interest to pass through. Ions are then fragmented using collision-induced dissociation and the resulting products, known as transitions, are detected with a final mass analyser (Picotti and Aebersold, 2012). Using the known amount of heavy peptide and the ratio of heavy to light peptide, it is possible to calculate the concentration of the light (i.e. native) peptide (**Figure 4.1**).

The development of an MRM experiment requires prior information of the target to be measured, usually inferred from previous data acquired. Peptides specific to proteins are usually identified from a discovery experiment or can be selected based upon previously acquired proteomic data from the Proteomics IDEntification (PRIDE) database (Vizcaino et al., 2016). However, peptides identified using PRIDE must first be validated in one's tissue samples to ensure that the peptides are indeed present (and in sufficient quantities to be able to be measured). Alternatively, it is possible to create an archive of spectra from previous experiments, where peptides identified have corresponding spectra in a library known as a "spectral library". This library can then be used to identify unique proteins, peptides and fragment ions (transition ions) within the samples measured. Skyline is a software designed to aid in the development, implementation and analysis of MRM experiments. Skyline is capable of creating a spectral library from previous data and can

identify unique peptides to specific proteins (MacLean et al., 2010, Liebler and Zimmerman, 2013). In addition to choosing unique peptides, it is important to pick unique fragments (transition ions) for each peptide. This is because MRM, as the name implies, looks at multiple targets, therefore it is imperative that two or more fragments with the same mass are not selected, otherwise the measurement of these ions will give inaccurate results.

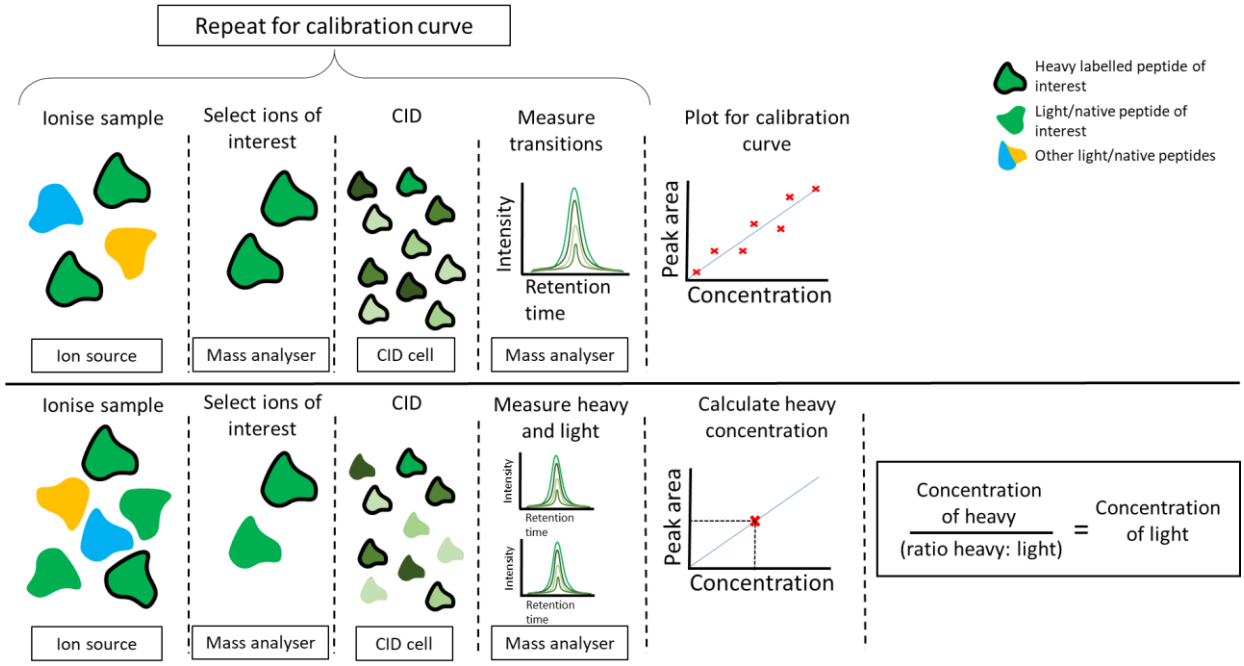


Figure 4:1: Flow diagram of multiple reaction monitoring.

Top; A calibration curve of heavy labelled peptide is created: Samples are ionised and heavy labelled peptides are then isolated in a mass analyser. Peptides are fragmented in a CID cell, then precursors and transitions are measured. The peak area and known concentration are plotted to give a calibration curve. **Bottom;** Samples are spiked with heavy labelled peptides of interest and then ionised: Peptides of interest are isolated and these peptides are fragmented in the CID cell. A mass analyser measures the amount of heavy and light ions of interest. The peak area of heavy peptide is plotted on a calibration curve to find it's true concentration. The concentration of the heavy peptide is divided by the ratio of the heavy to light peptides to calculate the concentration of the light peptide. CID, collisional induced dissociation.

MRM has been used in a wide array of disciplines (Liebler and Zimmerman, 2013), including animal husbandry (Kusebauch et al., 2018) and meat authentication (Watson et al., 2015) but is most notably used in systems biology (Elschenbroich and Kislinger, 2011) and biomarker discovery/validation (Picotti and Aebersold, 2012). In terms of systems biology,

examples where MRM has been used include elucidation of the beta-catenin signalling pathway (Chen et al., 2010) in addition to measuring a network of growth proteins from S.*cerevisiae* (Picotti et al., 2009). Biomarker discovery and validation using MRM has led to candidate biomarkers for ovarian cancer (Wang et al., 2017), colorectal cancer (Kume et al., 2014), oral cancer (Chen et al., 2017b) and prostate cancer (Yocum et al., 2010) amongst others.

As MRM produces an absolute quantification measurement, it is possible to do classical statistical analysis on the results, in addition to more elaborate interrogation. An increasingly attractive area in medicine and biomarker discovery is predictive modelling. Predictive modelling is the technique of using machine learning on known data to predict unknown data. Machine learning has been used in a variety of fields, including, but not limited to, diabetes (Kavakiotis et al., 2017), depression (Dipnall et al., 2016) ageing (Fabris et al., 2017) and cancer (Hornbrook et al., 2017, Lynch et al., 2017, Yu et al., 2016, Gupta et al., 2014).

4.2 Methods

A total of 101 samples were used in this chapter, consisting of 22 P-M samples (of the 24 P-M samples used in discovery) and 22 P-NM samples (of the 24 P-NM samples used in discovery) for verification and 28 P-M and 29 P-NM samples for validation.

4.2.1 Selecting suitable proteins for verification and validation

Our laboratory had already carried out some previous work on transketolase (TKT) and therefore had some isotopically labelled heavy peptides of this protein. As TKT was identified as significantly differentially expressed in both 1D data and 2D data, it was decided to take forward for verification and validation. A list of all combinations of proteins that were significantly differentially expressed in both 1D data and 2D data, where one protein was always TKT, was created. For example, the first row of the list would be TKT, protein 2, protein 3, followed by the second row which would be TKT, protein 2, protein 4, and so on.

1D data and 2D data were both, separately, split into a training data set (67% of all data) whereby a glm model (using each combination of proteins as predictors) was trained through 5-fold cross validation repeated 3 times and allowing an automatic tune length of 5. The remaining 33% of data was used as a test set to test the final models produced from the training stage. The 1D and 2D AUCs of each model from the different combinations of proteins were summed to rank models and identify which proteins could best predict metastasis in the discovery cSCC proteomic data.

4.2.2 Using targeted proteomics to verify and validate original findings

A spectral library of the discovery proteomics data was created as outlined in chapter 2.9.3.1. Targeted proteomics was carried out as described in chapter 2.9.3.2. Briefly, isotopically heavy labelled peptides were initially analysed using a Synapt G2-Si high resolution mass spectrometer, to assess their suitability as MRM targets. A serial halving dilution of each peptide was then analysed in a background cSCC peptide matrix to determine calibration data.

100fmol of each isotopically heavy labelled peptide was spiked into 22 P-M samples (of the 24 P-M discovery samples) and 22 P-NM samples (of the 24 P-NM discovery samples) and analysed on a Synapt G2-si mass spectrometer. Calibration data achieved in the initial analysis of heavy labelled peptides was used to calculate the true amount of each heavy labelled peptide in each sample. Using the light: heavy ratio, it was then possible to calculate the amount of native peptide in each sample.

4.2.3 Time to metastasis plot

R and the packages survminer and survival were used to create “time to metastasis” plots. Time “0” was determined as the date of excision and counted in days until metastasis occurred or in occasions where metastasis was present at excision, time of metastasis was also defined at 0. In instances where metastasis was not present (i.e. P-NMs), time to metastasis was set at 5 years (1,825 days), which was the cut off for the P-NM criteria. High and low expression was defined as either above or below the median, respectively. P values were obtained by log-rank test.

4.2.4 Predictive modelling on verification and validation data

Verification and validation MRM data was pooled together (to total 101 samples) and split into training (67%) and test sets (33%). Various algorithms were used to create models on the training set, whereby the predictors were the MRM peptide data. Models were created using 10 fold cross validation repeated 3 times. Models were compared to assess conformity and correlation to select best models for stacking.

The final model was tested on the test data to produce a ROC curve. Current guidelines were used on the data to acquire a sensitivity and specificity which was then plotted on the ROC curve. Sensitivities and specificities were also taken from the Roscher *et al* (Roscher et al., 2018) paper.

4.3 Results

4.3.1 Selecting suitable proteins for Multiple Reaction Monitoring (MRM) analysis

Machine learning on the discovery proteomic data was performed to identify which proteins had the best power to predict metastasis. As previously mentioned, our laboratory had access to transketolase (TKT) heavy labelled peptides and, as a result, all models were created to incorporate TKT (uniprot ID: P29401), the other 2 markers being one of every combination of proteins that were significantly differentially expressed between P-M and P-NM in both the 1D and 2D data. The top model consisted of P29401 (TKT), P39656 (Dolichyl-diphosphooligosaccharide--protein glycosyltransferase 48 kDa subunit, DDOST) and P08758 (Annexin A5, ANXA5) (**Table 4:1**). An AUC of the ROC curve of 0.9 and 0.938 was produced on 1D and 2D data, respectively.

Table 4.1: Top 10 protein combination models produced to classify samples as P-M or P-NM.

Marker1	*Marker2*	*Marker3*	*1D ROC AUC*	*1D ROC Sensitivity*	*1D ROC Specificity*	*2D ROC AUC*	*2D ROC Sensitivity*	*2D ROC Specificity*	*SUM 1D & 2D ROC AUCs*
P29401	P39656	P08758	0.9	1	0.667	0.938	1	0.75	1.838
P29401	P51884	Q9NZT1	0.889	1	0.667	0.767	1	0.6	1.656
P29401	Q15582	P08758	0.875	0.75	1	0.778	1	0.5	1.653
P29401	P51884	P08758	0.813	0.75	1	0.833	1	0.6	1.646
P29401	P50990	P08758	0.933	0.833	1	0.7	0.5	1	1.633
P29401	P39656	P51884	0.88	0.8	1	0.75	0.5	1	1.63
P29401	P62937	P08758	0.939	1	0.714	0.667	0.429	1	1.605
P29401	P07237	P51884	0.905	0.857	1	0.7	0.833	0.6	1.605
P29401	Q15063	P62277	0.85	0.6	1	0.75	0.857	0.75	1.6
P29401	P08758	Q9NZT1	0.857	0.833	0.857	0.738	0.429	1	1.595

ROC, receiver operating characteristic. AUC, area under curve.
Identifiers are Uniprot IDs

4.3.2 Selecting suitable peptides for Multiple Reaction Monitoring (MRM) proteins

Once the biomarker candidates of interest had been selected, they were matched against the Skyline spectral library (created from proteomics data produced in the discovery phase) to assess their suitability to be used for MRM analysis. Skyline was also used to identify unique peptides for DDOST and ANXA5 which had multiple, high intensity transition ions (**Figure 4.2**). The three heavy labelled TKT peptides we had in our laboratory were matched against the spectral library where it was highlighted that only 1 of the 3 were detected in the discovery data and so could not possibly be correctly referenced (as the ms/ms spectra is unknown). Nonetheless, the remaining 7 peptides were present in the discovery data with high spectral counts (number of spectra per peptide) and at high intensities, indicating they would be good MRM candidates**. Table 4.2** is a table of the unique peptides selected for each protein with their m/z and the most suitable transition ions. Therefore the 6 peptides for DDOST and ANXA5 were synthesised with heavy isotopes incorporated into their structure.

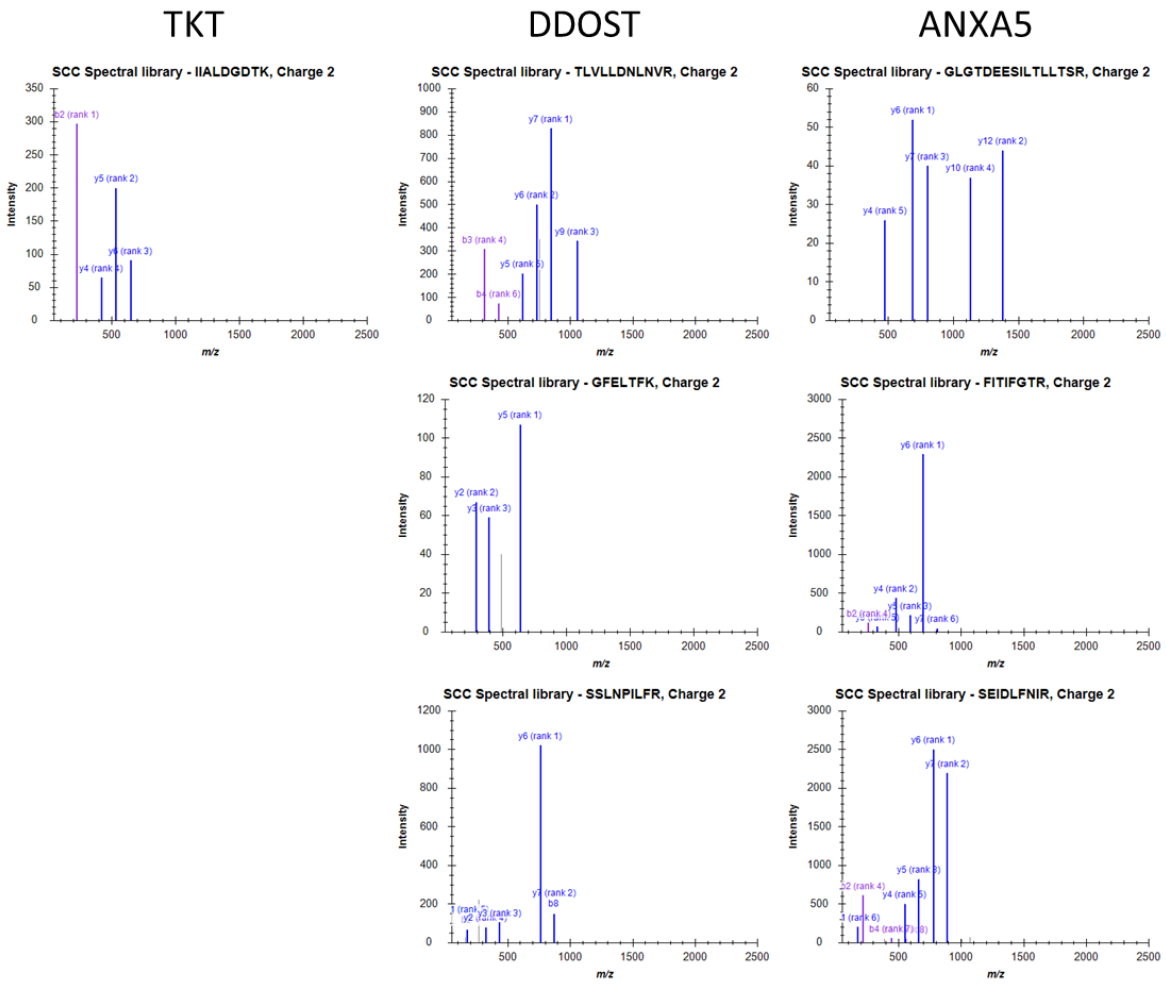


Figure 4:2: Unique peptides of candidate biomarkers showed several transition ions identified in spectral library.

Three Unique peptides for both DDOST and ANXA5 were identified and compared to the spectral library created from the discovery proteomic data. Each peptide displayed a number of transition ions found in the spectral library and at high intensities. As work on TKT had been carried out previously in our lab, we used the heavy labelled peptides already purchased. 1 of these 3 peptides was identified in the cSCC spectral library with good transition ions, the remaining 2 were not.

Table 4.2. The unique peptides selected for each protein of interest with their m/z and transition ions

PROTEIN	PEPTIDE	M/Z AT CHARGE 2	TRANSITION IONS	MODIFIED PEPTIDE SEQUENCE	M/Z AT CHARGE 2	TRANSITION IONS
TKT	IIALDGDTK	473.2662	648.3199, 535.2358, 420.2089	IIALDGDTK[$^{13}C_6$, $^{15}N_2$]	477.2733	656.3341, 543.25, 428.2231
DDOST	TLVLLDNLNVR	635.3799	1055.6208, 843.4683, 615.3573	TLVLLDNLNVR[$^{13}C_6$, $^{15}N_4$]	640.384	1065.629, 853.4766, 625.3656
	GFELTFK	421.2264	637.3556, 395.2289, 294.1812	GFELTFK[$^{13}C_6$, $^{15}N_2$]	425.2335	645.3698, 403.2431, 302.1954
	SSLNPILFR	523.8033	872.5352, 759.4512, 435.2714, 175.1190	SSLNPILFR[$^{13}C_6$, $^{15}N_2$]	528.8074	882.5435, 769.4595, 445.2797, 185.1272
ANXA5	GLGTDEESILTLLTSR	852.9543	1376.7268, 1132.6572, 803.4985, 690.4145, 476.2827	GLGTDEESILTLLTSR[$^{13}C_6$, $^{15}N_4$]	857.9585	1386.7350, 1142.6655, 813.5068, 700.4227, 486.2910
	FITIFGTR	477.774	807.4723, 964.3883, 593.3406, 480.2565, 333.1881	FITIFGTR[$^{13}C_6$, $^{15}N_4$]	482.7781	817.4806, 704.3965, 603.3488, 490.2648, 343.1964
	SEIDLFNIR	553.7957	890.5094, 777.4254, 662.3984, 549.3144, 175.1190	SEIDLFNIR[$^{13}C_6$, $^{15}N_4$]	558.7998	900.5177, 787.4336, 672.4067, 559.3226, 185.1272

TKT, Transketolase. DDOST, Dolichyl-diphosphooligosaccharide--protein glycosyltransferase 48 kDa subunit. ANXA5, Annexin A5. m/z, mass to charge ratio

4.3.3 Predictive power of DDOST and ANXA5

As DDOST and ANXA5 were selected by creating a predictive model in conjunction with TKT and it was unclear whether there would be sufficient data from TKT to contribute to such a model, DDOST and ANXA5 were assessed for their predictive power alone (**Figure 4.3**). Using discovery proteomic data of DDOST and ANXA5, a number of different algorithms were used to build predictive models, including glm (Nelder and Wedderburn, 1972), knn (Cover and Hart, 1967), svm (Vapnik and Chervonenkis, 1974), nb (Duba and Hart, 1973), c5.0 (Breiman et al., 1984), rf (Breiman, 1999), bagging (Breiman, 1996) and cart (Steinberg, 2009). A full list of the algorithms used in this thesis can be found in **Appendix 4.** Models were trained on DDOST and ANXA5 discovery proteomic data using 5 fold cross validation repeated 3 times.

The glm algorithm produced a model with the highest ROC score and therefore this was explored further. 1D and 2D proteomic data were trained and tuned using the glm algorithm (**Figure 4.4**) and an average AUC (area under the curve) of 0.82 was obtained from and AUC of 0.84 on 2D proteomic data and 0.80 on 1D proteomic data. These models suggest that DDOST and ANXA5 would have potential predictive power without incorporating TKT into the prediction.

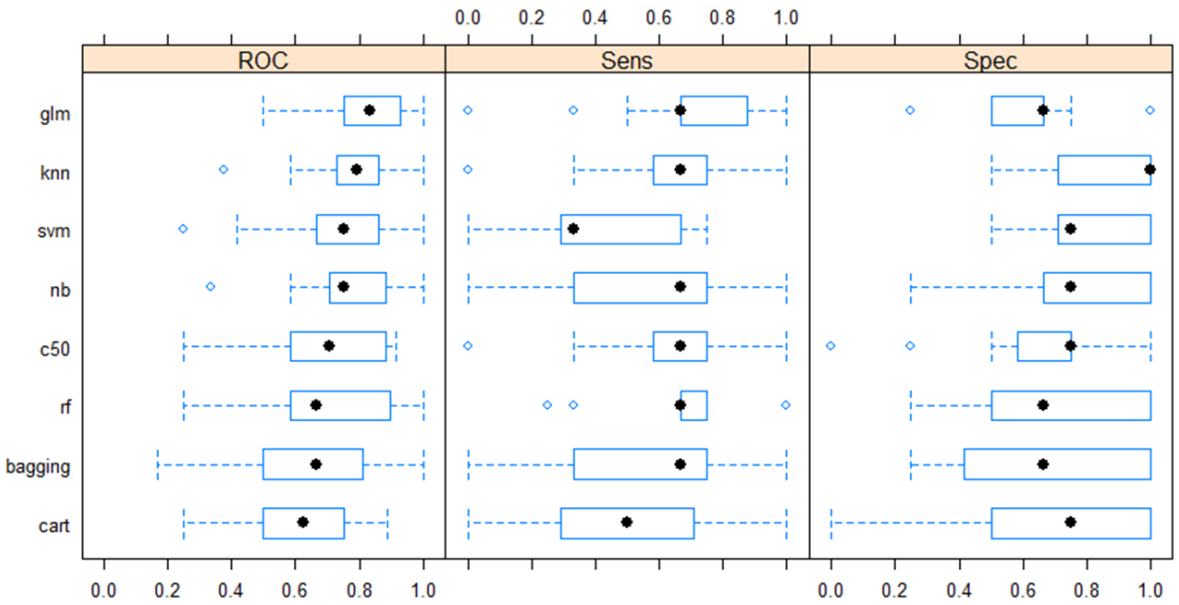


Figure 4:3: Demonstrating the predictive ability of DDOST and ANXA without TKT.

To establish whether a predictive model could be produced if the MRM on TKT didn't produce useful data, a number of different, simple, machine learning algorithms were trained using 5 fold cross-validation repeated 3 times on DDOST and ANXA5 discovery proteomic data. It was evident that, in the absence of TKT data, DDOST and ANXA5 still have significant potential to predict outcome. Box and whiskers depict resampling performance range. Dots report median, boxes give interquartile range and whiskers give range. Hollow points depict outliers. Glm, generalised linear model. Knn, K's nearest neighbour. SVM, support vector machine. Nb, Naïve Bayes. Rf, Random forest. Cart, classification and regression trees.

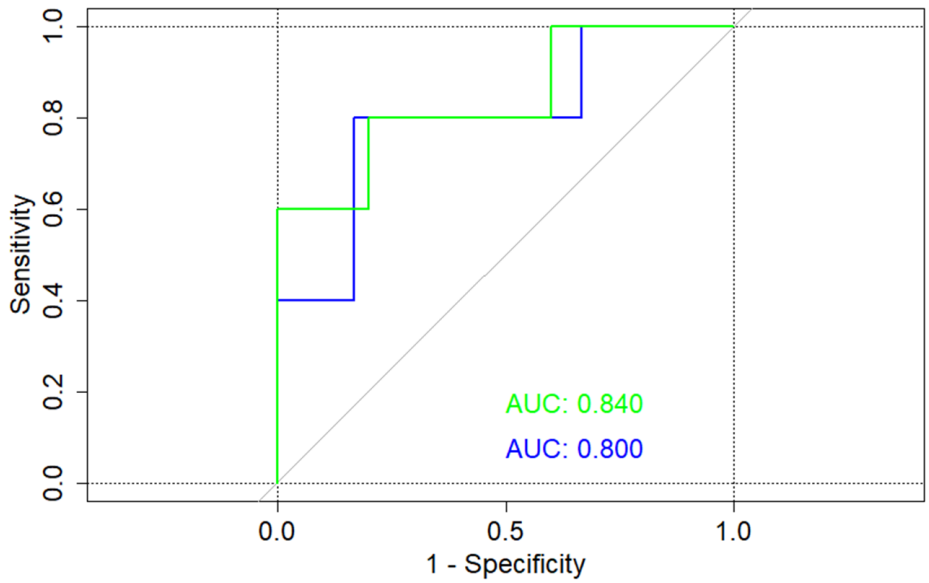


Figure 4:4: DDOST and ANXA5 have an average AUC of 0.82 when trained using 5 fold cross validation repeated 3 times on DDOST and ANXA5 discovery proteomic data.
glm was identified as a potentially suitable algorithm so it was trained and tuned on 1D (blue) and 2D (green) DDOST and ANXA5 discovery proteomic data to reveal a model capable of categorising P-M and P-NM samples. AUC, area under the curve

4.3.4 MRM peptide calibration curves

Heavy labelled peptides were initially analysed on a synapt G2-Si high resolution mass spectrometer in targeted acquisition mode at a concentration of 100fmol to assess the chromatography of each peptide. Chromatograms and accompanying mass spectra were imported into Skyline for analysis (**Figure 4.5**). Chromatography demonstrated good intensities and good peak widths/shapes of precursor ions and associated transitions ions. The chromatograms of all of the peptides were within the predicted retention time window apart from Peptide 1, ANXA5, which was earlier than predicted. Nonetheless, it was the only peak of significant intensity where the precursors and transition ions matched and was therefore selected as the peptide peak.

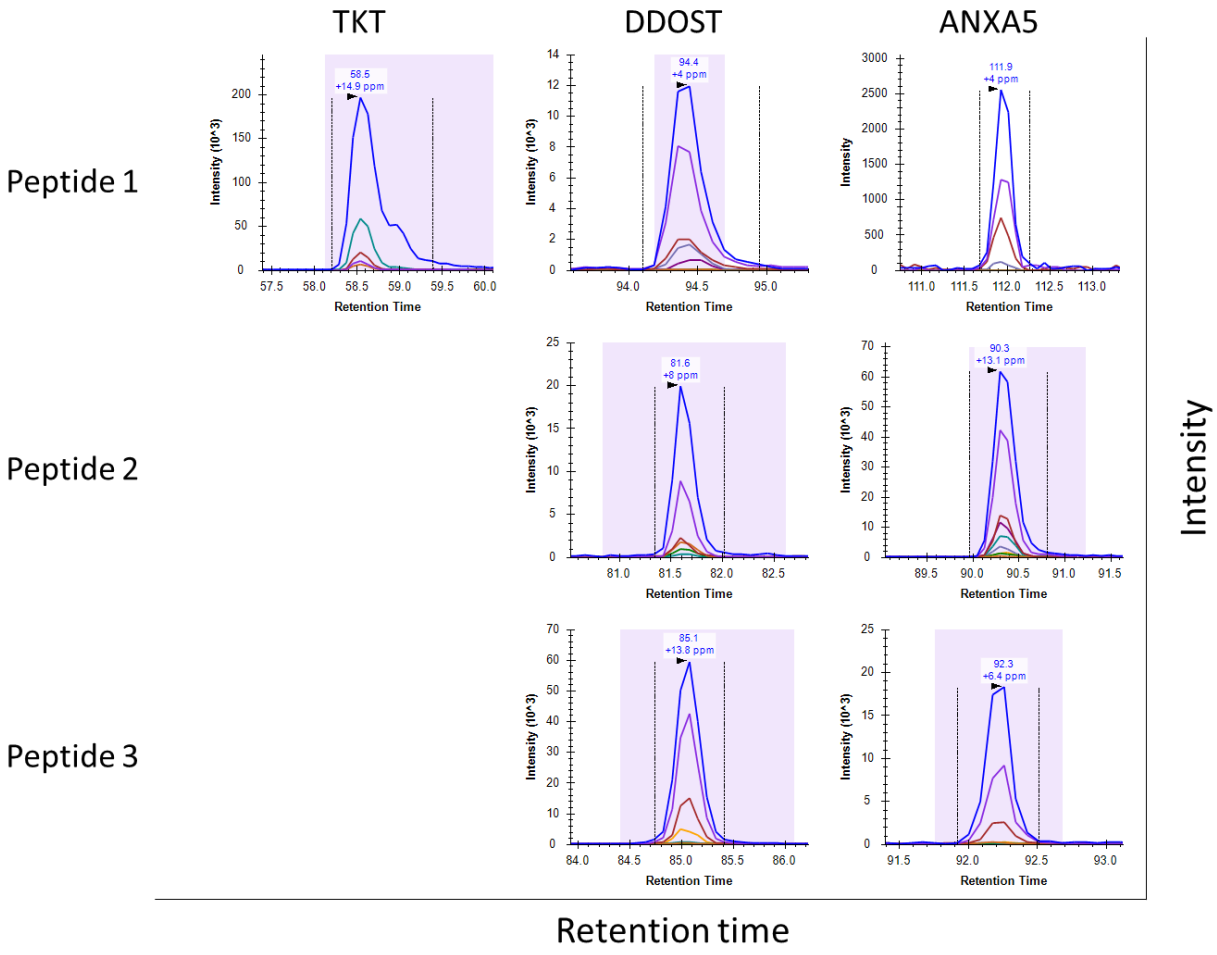


Figure 4:5: MRM chromatography of the selected heavy peptides.

Each of the seven peptides produced clear chromatographic peaks, with matching peaks corresponding to MRM transitions. Intensities were sufficient for reliable measurements to allow quantification of the heavy peptide and to correlate it with the original amount loaded heavy peptid. Peptide 1 for ANXA5 was not observed in the predicted retention time window (blue shading) but was confirmed via its corresponding transition ions.

Once quantification of heavy peptides had been established by assessing chromatography, calibration curves for each heavy labelled peptide were produced to enable later determination of the native peptide concentration in unknown samples. A two fold dilution series from 200fmol down to 0.78125fmol of each heavy peptide was used for the MRM experiments, with 1µg of cSCC peptide background matrix added to each sample as an internal standard. Peak areas of TKT followed the expected linear trend of the dilution series and the standard cSCC background matrix appeared consistent (**Figure 4:6**).

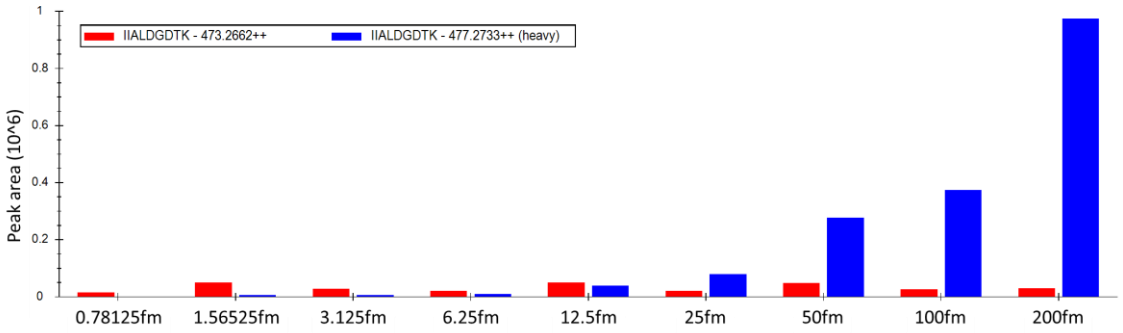


Figure 4:6: Calibration of TKT heavy peptide 1 for MRM

Calibration data of the heavy labelled TKT peptide was created using MRM on a twofold dilution of heavy peptide amounts from 200fmol to 0.78125fmol (blue). 1µg of cSCC peptide background matrix was used as internal standard (red). Peptide sequence and m/z shown for both light (red) and heavy, (blue) peptides.

A similar dilution calibration curve was produced for DDOST (**Figure 4.7**) and ANXA5 (**Figure 4.8**) peptides. Each peptide had a linear trend, as was to be expected of a dilution series. The internal standard was also fairly consistent between peptides.

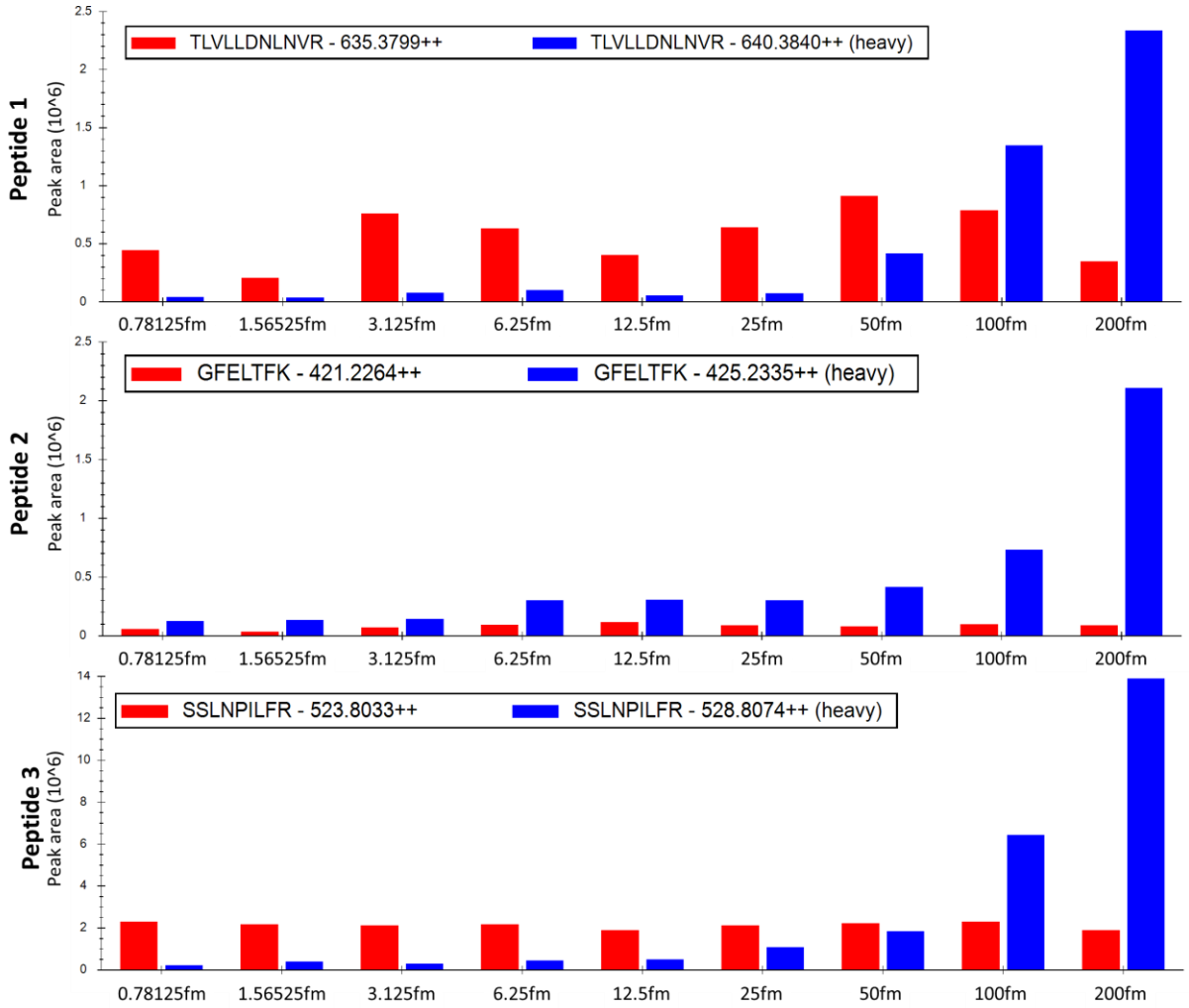


Figure 4:7: Calibration of DDOST heavy peptides 1,2 and 3 for MRM.

Calibration data of the heavy labelled DDOST peptide was created using MRM on a twofold dilution of heavy peptide amounts from 200fmol to 0.78125fmol (blue). 1µg of cSCC peptide background matrix was used as internal standard (red). Peptide sequence and m/z shown for both light (red) and heavy, (blue) peptides.

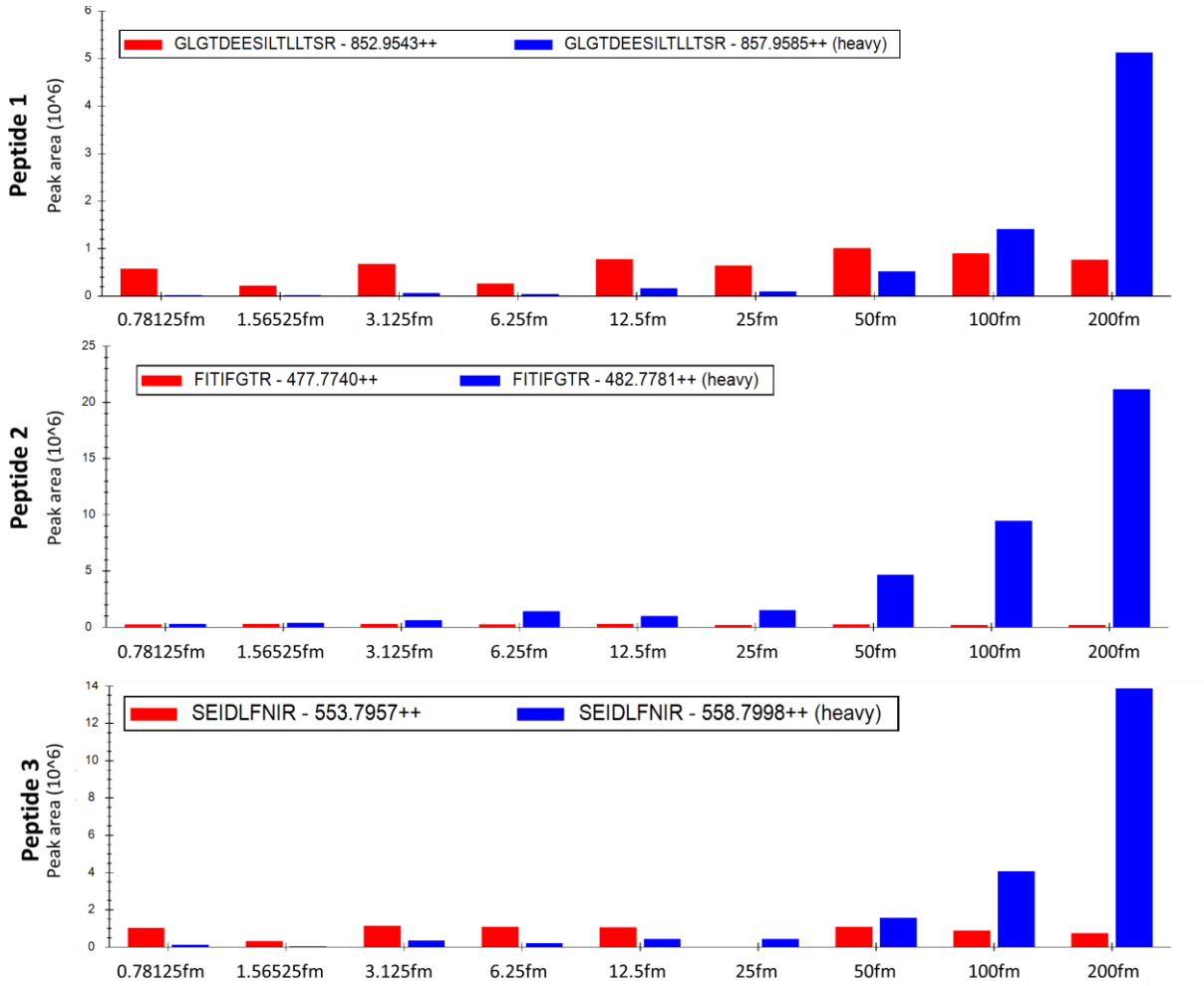


Figure 4:8: ANXA5 peptide calibration data from MRM of ANXA5 heavy peptides 1, 2 and 3.

Calibration data of the heavy labelled ANXA5 peptide was created using MRM on a two fold dilution of heavy peptide amounts from 200fmol to 0.78125fmol (blue). 1µg of cSCC peptide background matrix was used as internal standard (red). Peptide sequence and m/z shown for both light (red) and heavy, (blue) peptides.

A linear regression of the MRM results of each heavy peptide compared with the sample amount of each heavy peptide was performed to obtain a slope and an R^2 value. The R^2 value is also known as the coefficient of determination and is an indication of how predictive one variable is on the other, in this case, how much the peak area can predict the analyte concentration of the heavy peptide. The slope can later be used to calculate the analyte concentration in a sample, when the peak area of the light and heavy peptides of that sample has been obtained using MRM. All of the selected heavy peptides had an R^2 value above 0.9, indicating a good ability to quantify analyte concentration using MRM. TKT had an R^2 of 0.9827 (**Figure 4.9**) and each of the three peptides for DDOST had R^2 values of

0.9773, 0.9288, and 0.9814 (**Figure 4.10**) and for ANXA5 had R^2 values of 0.9245, 0.9916, and 0.9355, (**Figure 4.11**).

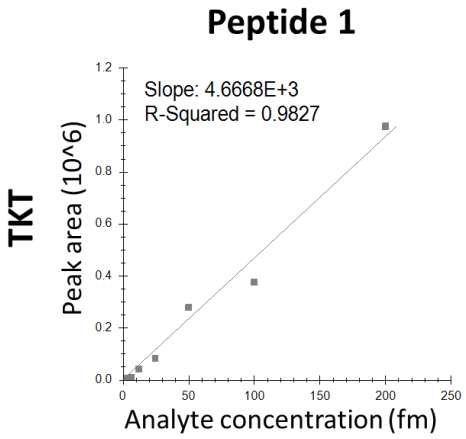


Figure 4:9: Linear regression of TKT peak area of MRM peptide to inputted analyte concentration.

MRM heavy peptide data was imported into Skyline to acquire peak areas for each concentration and a linear regression model fitted between MRM peak area and analyte concentration.

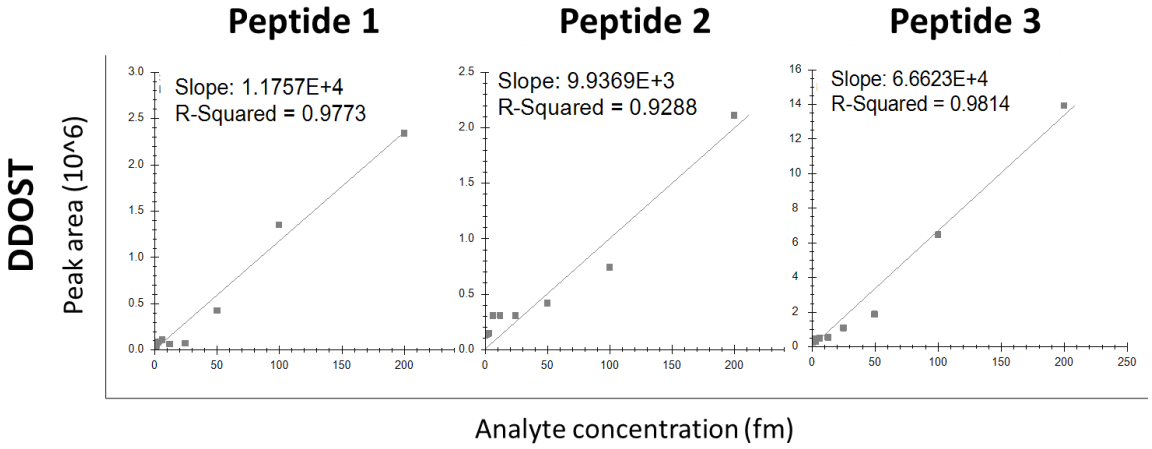


Figure 4:10: Linear regression of DDOST peak area of MRM peptides to inputted analyte concentrations.

Dilution calibration curve was imported into Skyline to acquire peak areas for each concentration and a linear regression model fitted between MRM peak area and analyte concentration.

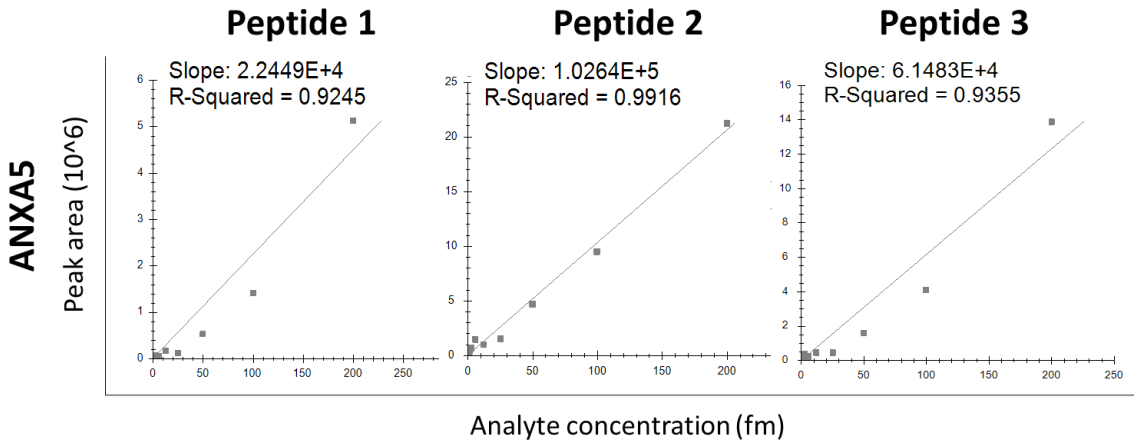


Figure 4:11: Linear regression of ANXA5 peak area of MRM peptides to inputted analyte concentration.

Dilution calibration curve was imported into Skyline to acquire peak areas for each concentration and a linear regression model fitted between MRM peak area and analyte concentration.

4.3.5 Verification of protein biomarkers from discovery proteomics

After the production of suitable calibration data for each MRM peptide, each discovery proteomic sample was investigated using MRM, with the addition of 100fmol heavy peptide. Results were imported into Skyline and corresponding peaks identified and selected. The peak area of the heavy peptide was used with the linear regression from the calibration data to calculate the "corrected" amount of heavy labelled peptide in each sample. Using the corrected amount of heavy peptide and dividing it by the ratio of light (native) to heavy, the amount of native peptide in the sample could be accurately quantified. There was no significant difference of the TKT peptide measured between P-M and P-NM (**Figure 4.12**). However, the MRM results indicated that there was significantly more DDOST in the P-M than P-NM samples, consistent with what had been previously identified in the discovery proteomics (**Figure 4.13**). Furthermore, there was also significantly more ANXA5 in the P-M than P-NM cSCCs, which was also similar to that observed in the discovery proteomics (**Figure 4.14**). An average for each of the proteins, DDOST and ANXA5, were calculated used the mean of their three respective peptides. Both of these means were significantly different between P-M and P-NM.

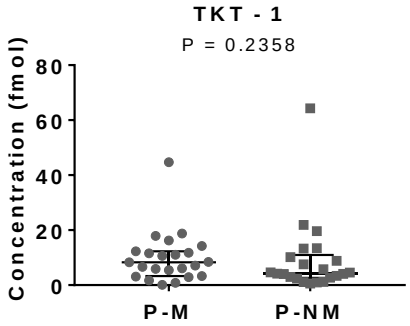


Figure 4:12: MRM verification of TKT.

Concentration of the TKT native peptide was determined by dividing the corrected concentration of the heavy peptide by the ratio of heavy:light peptide peaks (as per figure 4.1). Mann-Whitney U test for significance. Error bars are the median with interquartile range

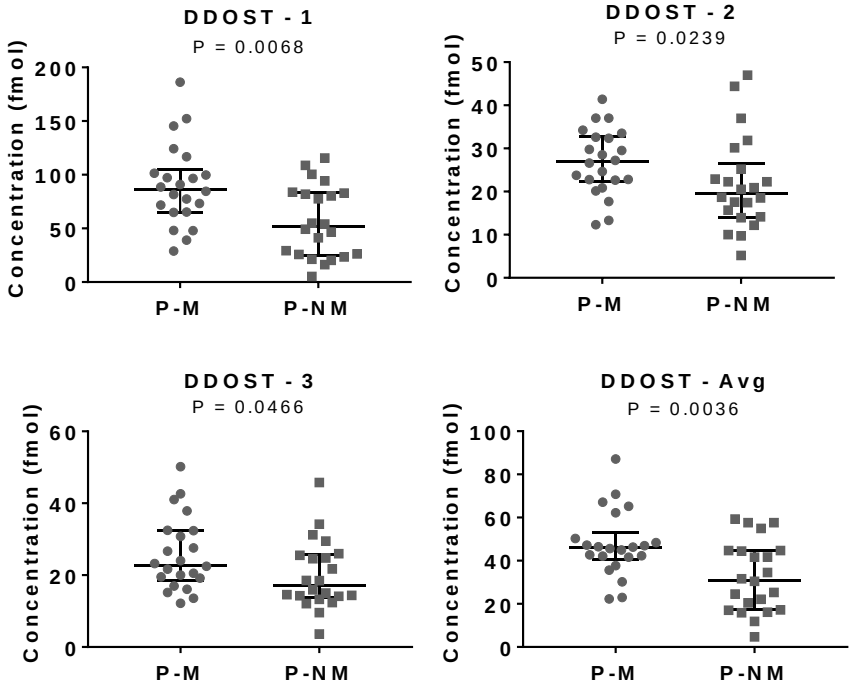


Figure 4:13: MRM verification of DDOST peptides and overall protein.

Concentration of native peptide was determined by dividing the calculated concentration of the heavy peptide by the ratio of heavy:light peptide peaks. An average of all 3 peptides was calculated using the mean. Mann-Whitney U test for significance. Error bars are median with interquartile range

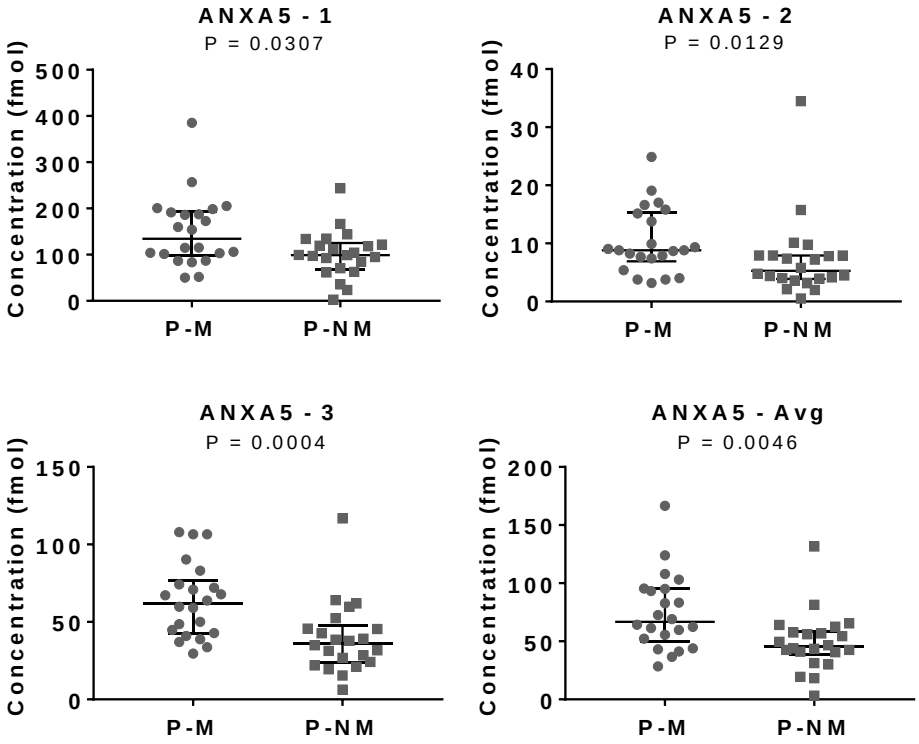


Figure 4:14: MRM verification of ANXA5 peptides and overall protein.

Concentration of native peptide was determined by dividing the calculated concentration of the heavy peptide by the ratio of heavy:light peptide peaks. An average of all 3 peptides was calculated using the mean. Mann-Whitney U test for significance. Error bars are median with interquartile range

4.3.5.1 Histological verification of L-Plastin

Although MRM successfully verified the discovery findings, we chose to investigate whether these findings could also be validated via a third, independent method, immunohistochemistry. L-plastin is a protein found in 68% of carcinomas and 53% of solid tumours (Lin et al., 1993) and is also known to have an important role in the activation of T-cells (Wabnitz et al., 2007). For these reasons, it was hypothesised there would be more L-plastin+ cells in the P-M group than the P-NM group. To investigate this, immunohisochemical staining for L-Plastin was carried out on the discovery proteomic samples (**Figure 4.15**). L-Plastin+ cells were identified predominantly in the stroma of the cSCCs with few L-Plastin+ cells in the tumour islands. Consistent with the result of the discovery proteomics, there was significantly more L-Plastin+ cells in P-M samples compared to P-NM cSCCs (P=0.0136).

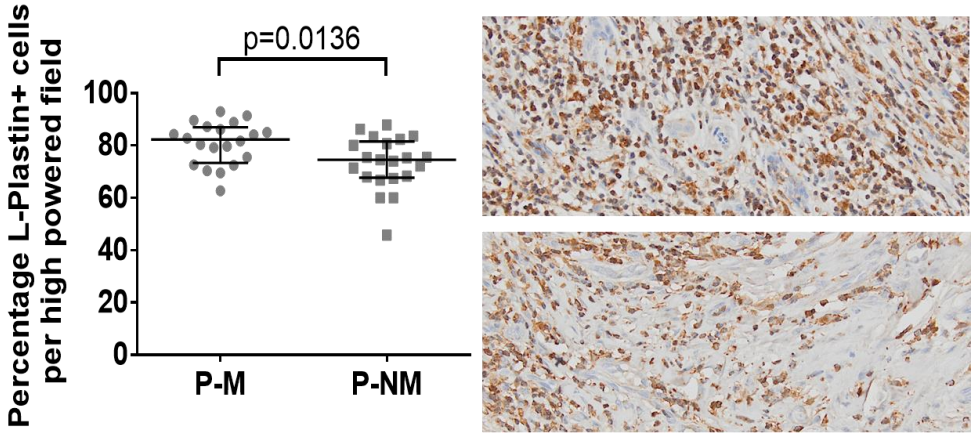


Figure 4:15: Immunohistochemical staining of L-Plastin in P-M and P-NM samples. Slides were subjected to standard IHC protocols and stained using a rabbit monoclonal antibody to L-Plastin. Representative images of L-Plastin stain are shown on the right side of the figure, where top panel is P-M and bottom panel is P-NM. P values obtained through Mann Whitney U test for significance.

4.3.6 Machine learning on Multiple Reaction Monitoring (MRM) verification data

Using the data from the MRM performed on the cSCCs above (henceforth referred to as the "MRM peptide verification data"), different machine learning algorithms were trained using 5 fold cross validation repeated 3 times. The peptide MRM data (**Figure 4.16**) and "protein" MRM data (mean of peptides) (**Figure 4.17**) was trained on 13 different machine learning algorithms. The results from these algorithms suggested that although the MRM data for DDOST and ANXA5 have some predictive power, there wasn't an obvious model that outperforms the others; in fact most of the models appear to be relatively weak learners, scoring <0.8 AUC. To further validate the above findings, it was decided upon to carry out DDOST and ANXA5 MRM analysis on a previously unseen separate sample cohort of cSCCs.

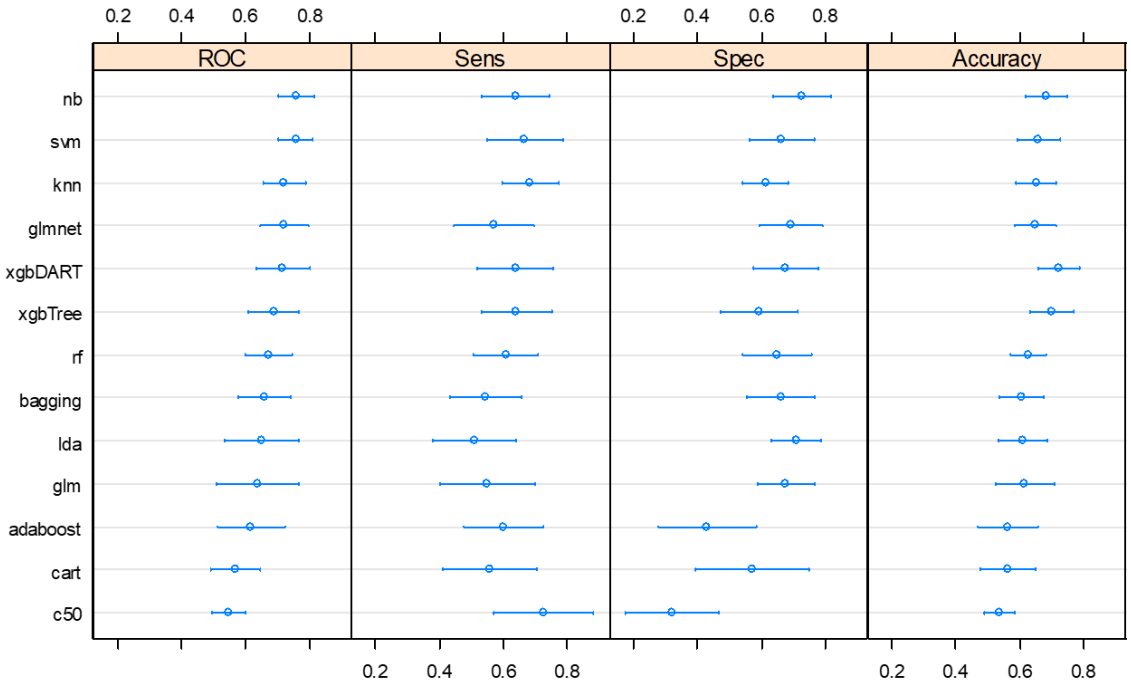


Figure 4:16: MRM peptide verification data for DDOST and ANXA5 were subjected to several different machine learning algorithms to assess the predictive power of the data. 5 fold cross validation repeated 3 times was carried out on all MRM peptide verification data. Error bars are confidence intervals. Nb, Naïve Bayes. SVM, support vector machine. Knn, K's nearest neighbour. Glm, generalised linear model. Xgb, extreme gradient boosting. Rf, random forest. Lda, linear discriminant analysis. Cart, classification and regression trees.

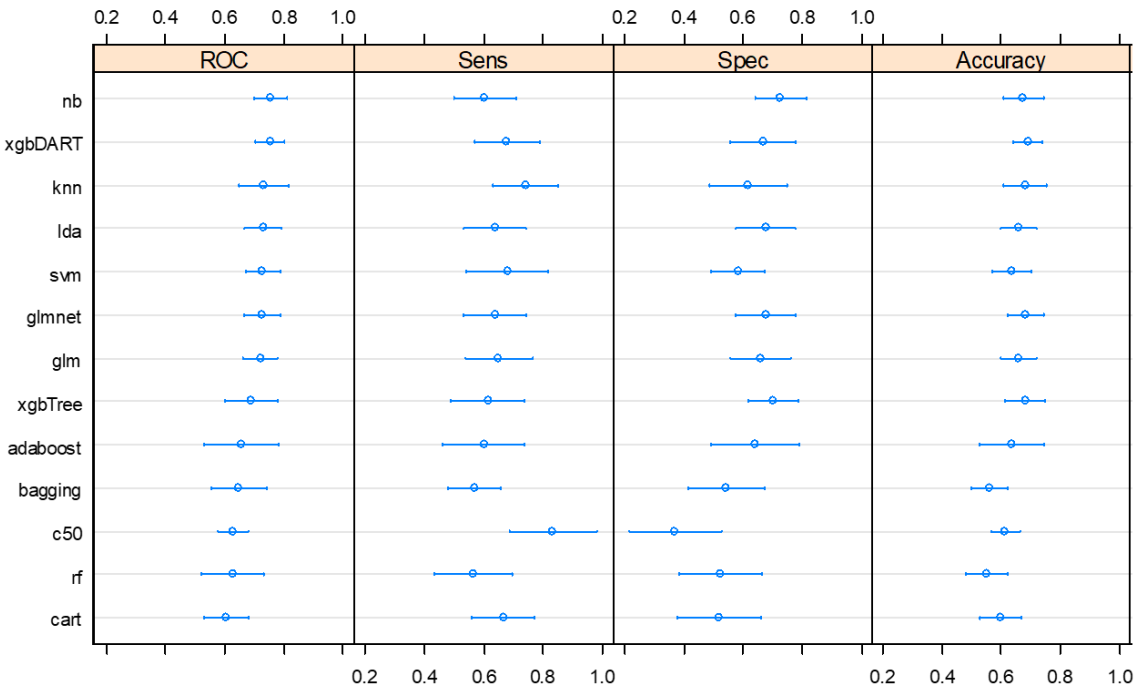


Figure 4:17: MRM protein verification data for DDOST and ANXA5 were subjected to several different machine learning algorithms to assess the predictive power of the data. 5 fold cross validation repeated 3 times was carried out on all MRM peptide averages (the "protein" data). Error bars are confidence intervals. Nb, Naïve Bayes. SVM, support vector machine. Knn, K's nearest neighbour. Glm, generalised linear model. Xgb, extreme gradient boosting. Rf, random forest. Lda linear discriminant analysis. Cart, classification and regression trees.

4.3.7 Validation of DDOST and ANXA5 results on new set of cutaneous Squamous Cell Carcinoma (cSCC) samples

A new set of cSCCs were selected and processed for proteomic analysis according to the same criteria as the discovery samples. A table of clinical characterisations of the samples used for validation can be found in **Table 4:2**. Briefly, there was 28 P-M samples and 29 P-NM samples, each with a similar male to female ratio. There were more, well differentiated tumours in the P-NM group and more, poorly differentiated tumours in the P-M group, as was to be expected. P-M samples were also typically larger in diameter and depth, compared to P-NM samples.

Table 4.3: Clinical and histological details of cSCC samples used for validation MRM analysis.

	P-M	P-NM
Number of Samples	28	29
Male	21 (75.00%)	20 (68.96%)
Female	7 (25.00%)	9 (31.03%)
Well differentiated	0 (0.00%)	11 (37.93%)
Moderately differentiated	8 (28.57%)	12 (41.38%)
Poorly differentiated	20 (71.43%)	6 (20.69%)
Perivascular invasion	5 (17.86%)	1 (3.45%)
Perineural invasion	6 (21.43%)	1 (3.45%)
Immunosuppressed	4 (14.29%)	5 (17.24%)
Mean Tumour depth (mm)	8.54 ± 7.19	4.45 ± 2.93
Mean Tumour diameter (mm)	32.91 ± 38.85*	13.33 ± 8.19

P-M, Primary metastatic. P-NM, Primary non-metastatic.
**outlier with 210mm diameter included*

MRM was carried out in the same manner as for the earlier verification MRM analysis. Calibration data that were created for the verification MRM experiments were also used in the validation cohort of samples to calculate the "corrected" amount of heavy peptide in the cSCC samples and subsequently the amount of the native peptide of interest. Similar to the results of the discovery proteomics and the verification MRM experiments, there was significantly more DDOST (**Figure 4.18**) and ANXA5 (**Figure 4.19**) in the P-M than in the P-NM samples.

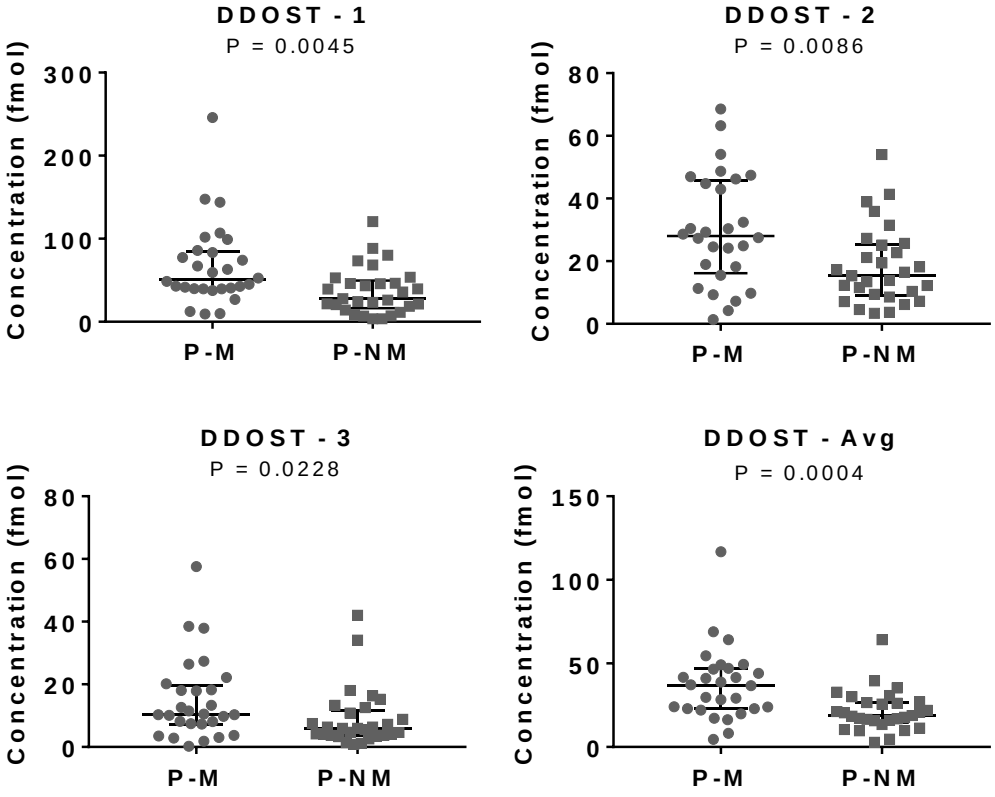


Figure 4:18: MRM Validation of DDOST peptides and overall protein.

Concentration of native peptide was determined by dividing the calculated concentration of the heavy peptide by the ratio of heavy:light peptide peaks on MRM. An average of all 3 peptides was calculated using the mean. Mann-Whitney U test for significance. Error bars are median with interquartile range

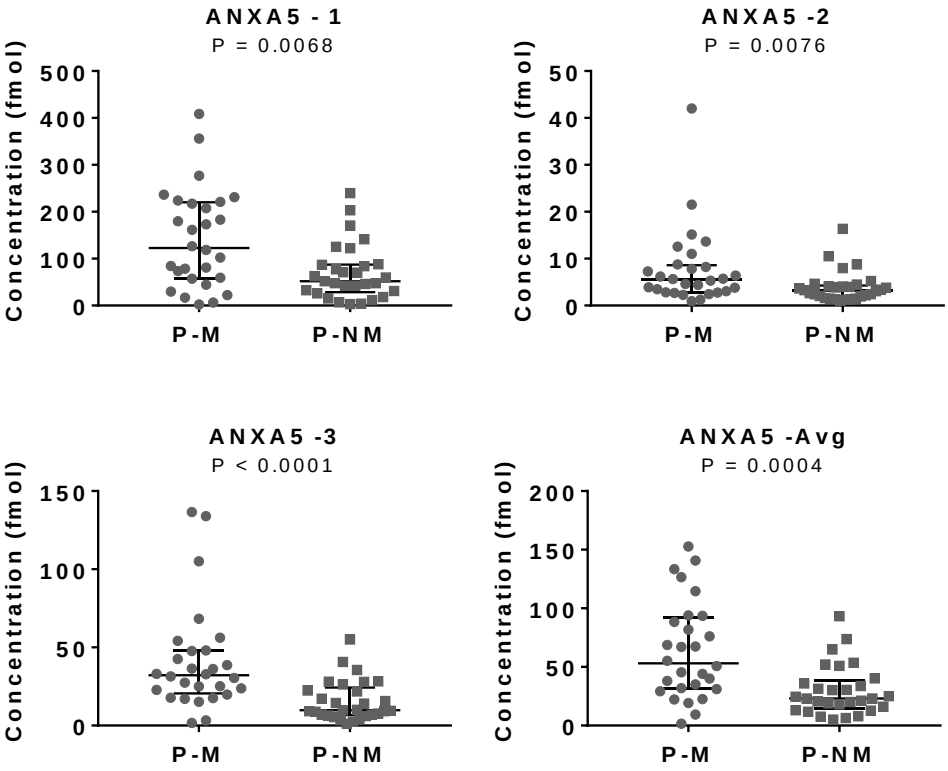


Figure 4:19: MRM validation of ANXA5 peptides and overall protein.

Concentration of native peptide was determined by dividing the calculated concentration of the heavy peptide by the ratio of heavy:light peptide peaks on MRM. An average of all 3 peptides was calculated using the mean. Mann-Whitney U test for significance. Error bars are median with interquartile range

4.3.8 DDOST and ANXA5's effect on time to metastasis

The MRM data from all the cSCC samples (discovery/verification groups and validation groups) along with the number of days until metastasis occurred, from the initial presentation in the dermatology clinic, were used to create a Kaplin-Meier survival plot. Where patients had metastasis at the time of their presentation to the dermatology clinic, time to metastasis was recorded as 0 days. Expression of DDOST and ANXA5 was categorised as either high or low depending on whether it was above or below the median, respectively. There was a significant positive association between high DDOST expression and a quicker time to metastasis (**Figure 4.20**). There was also a significant positive association between high ANXA5 expression and a quicker time to metastasis (**Figure 4.21**).

Furthermore, high expression of both DDOST and ANXA5, combined, was significantly associated with a shorter time to metastasis (**Figure 4.22**).

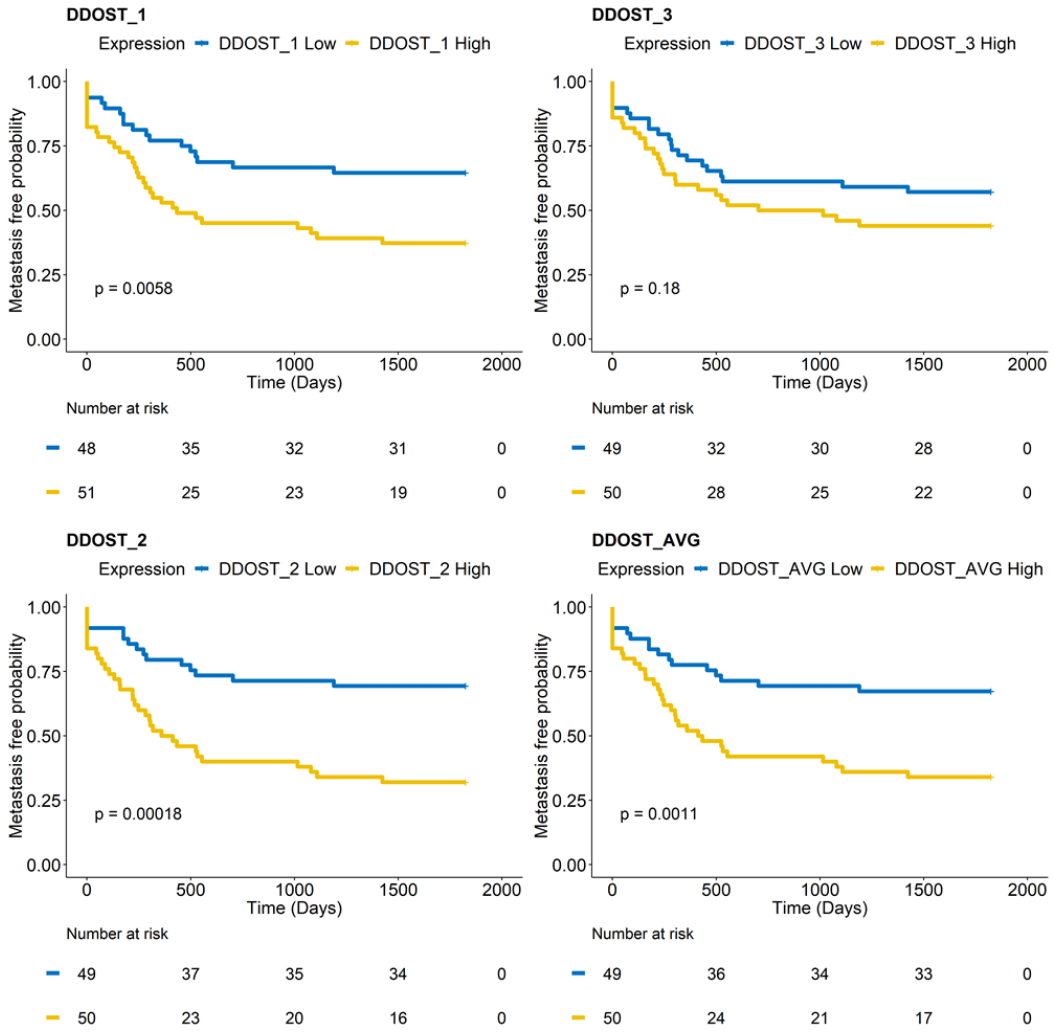


Figure 4:20: The effect MRM DDOST data has on time to metastasis.

Time to metastasis was deduced from the number of days between the initial dermatology clinic attendance and the identification of metastasis. High and low expression was defined as above or below the median as per **Figure 4:18**. P value obtained through Log Rank test.

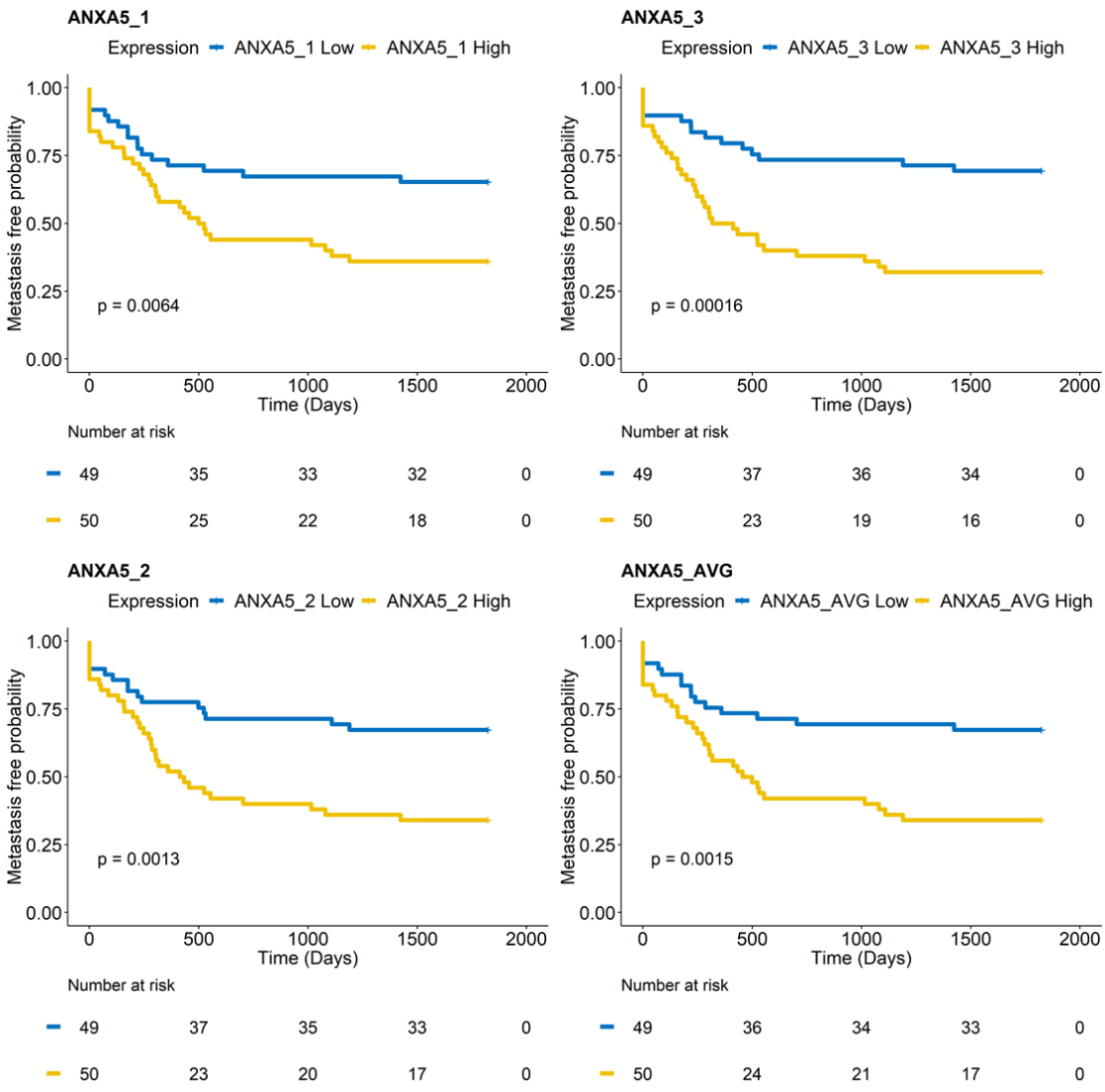


Figure 4:21: The effect MRM ANXA5 data has on time to metastasis.

Time to metastasis was determined from the number of days between the initial dermatology clinic attendance and the identification of metastasis. High and low expression was defined as above or below the median from **Figure 4:19**. P value obtained through Log Rank test.

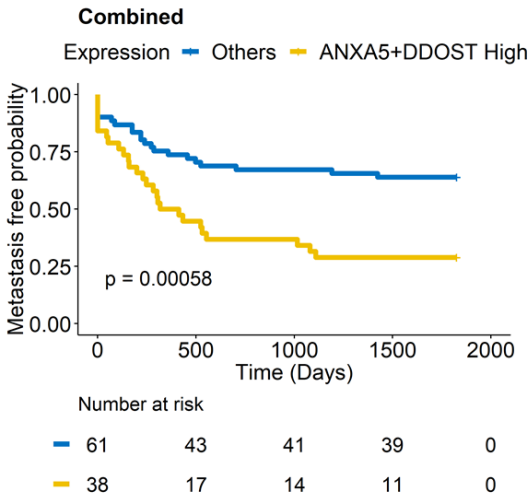


Figure 4:22: The effect combined high expression of MRM DDOST and ANXA5 data has on time to metastasis.

High expression denotes samples that had high DDOST and high ANXA5 expression (defined as above the median value for each of these proteins). Time to metastasis was deduced from number of days between the initial dermatology clinic attendance and the identification of metastasis. P value calculated using Log Rank test.

4.3.9 Machine learning on all Multiple Reaction Monitoring (MRM) data

It was decided that the peptide MRM data would be more suitable for modelling than the protein MRM data as this is a simplification of the original (peptide) data. For instance, the proteins MRM data is derived from the peptide data, and so by using the "raw" peptide data, hidden trends should be maintained, whereas if the averaged protein data was used, these might become less prominent. All MRM samples (from the verification and validation data) were pooled together, totalling 50 P-M samples and 51 P-NM samples. For machine learning and to build a predictive model which could later be tested, these samples were randomly split into training (67%) and testing (33%) cohorts. Models were trained using 10 fold cross validation repeated 3 times. 13 widely varying machine learning algorithms were trained on the training cohort (**Figure 4.23**). None of the tested models appeared significantly better than the others. Furthermore, it appeared that similar to the verification MRM, all of the tested models were relatively weak learners, highlighted by their relatively low AUC scores.

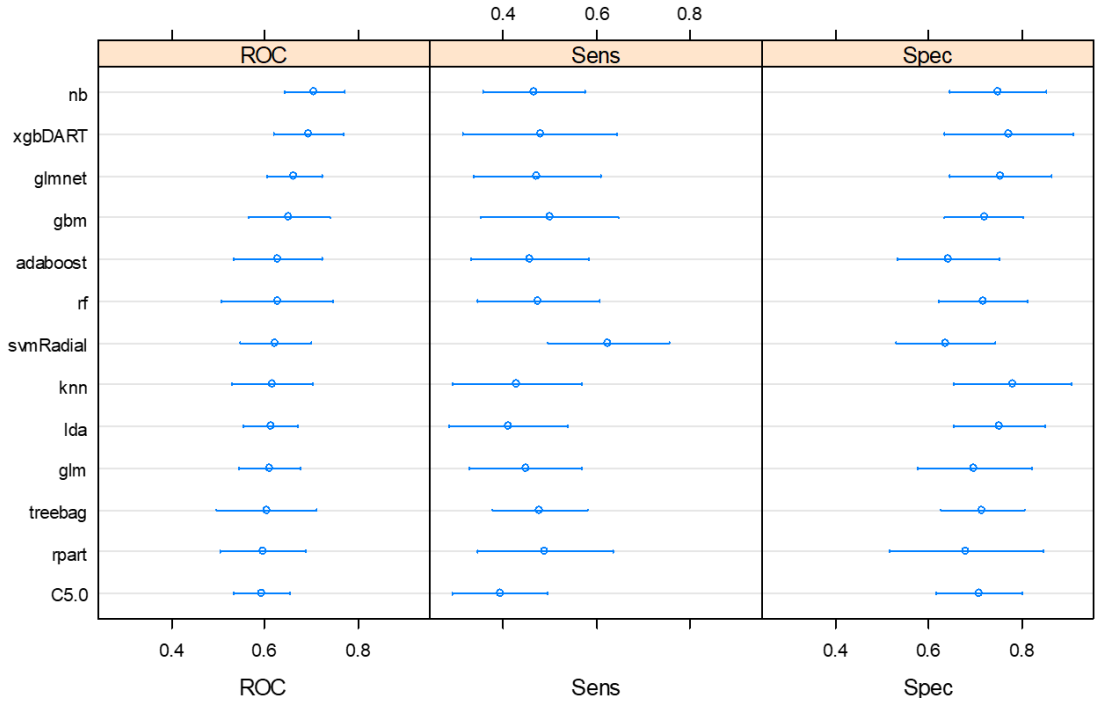


Figure 4:23: Applying different machine learning algorithms to the MRM data from the combined verification and validation samples.

All peptide MRM data was split into training (67%) and testing (33%) cohorts. Different algorithms were trained on the training set using 10 fold cross validation repeated 3 times. Error bars are confidence intervals. Nb, Naïve Bayes. Xgb, extreme gradient boosting. Gbm, gradient boosting model. Rf, random forest. Svm, Support vector machine. Knn, K's nearest neighbour. Lda, linear discriminant analsysis, glm, generalised linear model.

"Ensemble modelling" is the approach of taking several weak learners and using them together to create a strong, top level learner. Stacked ensemble modelling is a type of ensemble modelling and is the process of creating several weak learners which attempt to solve a problem and using a meta-learner on these models' predictions to solve the same problem better. For a stacked model to work, there needs to be little conformity between model predictions as if there is conformity amongst models, the correlated predictions will inherently be weighted more by the meta-learner. With this is mind, a correlation matrix of predictions produced by the models tested was generated (**Figure 4.24**). A high correlation is hard to define, but typically any pairwise correlation over 0.75 is probably too correlated for a stacked model. Using the correlation matrix, it was identified that glmnet, knn, adaboost, xgbDART and gbm did not seem overly correlated but still had good AUCs suggesting each of the models are correctly classifying different samples.

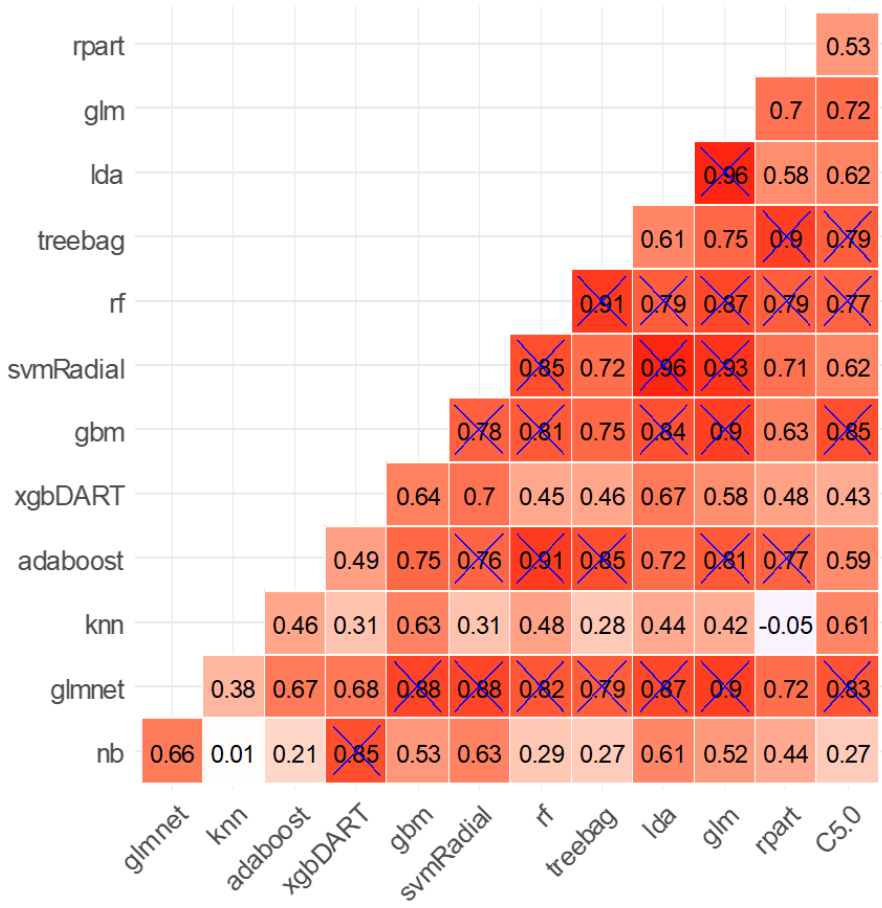


Figure 4:24: Testing for correlation of the mathematical models applied to the MRM data.

Predictions from models were correlated and visualised in a correlation matrix to assess their conformity. A group of mathematical models of the MRM data which exhibit low correlation is desirable for generation of a stacked ensemble model, because a meta-level learner applied to multiple uncorrelated models can learn from each model and correctly classify samples where models disagree. Nb, glmnet, knn, adaboost, xgbDART and gbm show minimal correlation and thus these models displayed suitability for stacking. Nb, Naïve Bayes. Xgb, extreme gradient boosting. Gbm, gradient boosting model. Rf, random forest. Svm, Support vector machine. Knn, K's nearest neighbour. Lda, linear discriminant analysis, glm, generalised linear model. Blue crosses indicate correlation above 0.75.

A typical meta-learner is either a decision tree or neural network as it is able to identify complex patterns in simple data, i.e. predictions from weaker learners. For this reason, an extreme gradient boosted tree algorithm was used. This in itself is a type of ensembling called boosting whereby weak learners are sequentially applied to a dataset and learn after each iteration with the weight of each learner depending on its accuracy. An overview of the stacked ensemble model produced can be seen in **Figure 4.25**.

Chapter 4

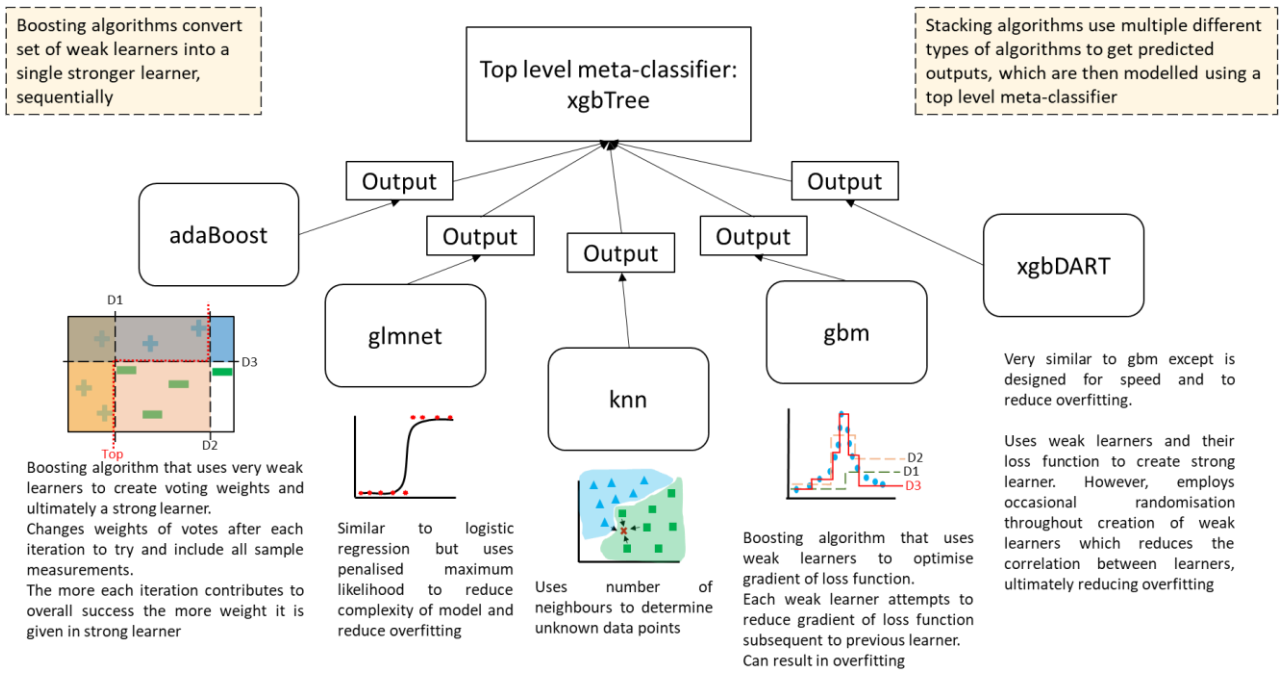


Figure 4:25: Overview of stacked model used to predict metastasis in cSCC from DDOST and ANXA5 MRM data.

Selected models, based on their High ROC score and low correlation to other models, were trained on the DDOST and ANXA5 MRM data from the training group of cSCCs (67%) using 10 fold cross validation repeated 3 times. Predictions from each model were then submitted to a top level meta-classifier which was also trained using 10 fold cross validation repeated 3 times. The final top level model was then used to predict the likelihood of metastases in the testing data set of cSCCs (33%).

The stacked ensemble model produced a ROC curve with an AUC of 0.929 (confidence interval 0.8277 – 1) (**Figure 4.26**). This suggested that the model is better at predicting likelihood of metastases from primary cSCC than any other clinical scoring systems currently in use, including the British Association of Dermatologists (BAD) and American Joint Committee on Cancer (AJCC) scoring systems. An optimal threshold in the ROC curve generated by the model would give rise to a sensitivity of 88.24% and a specificity of 94.12%, or if sensitivity were to be favoured, a sensitivity of 94.12% and a specificity of 82.36%.

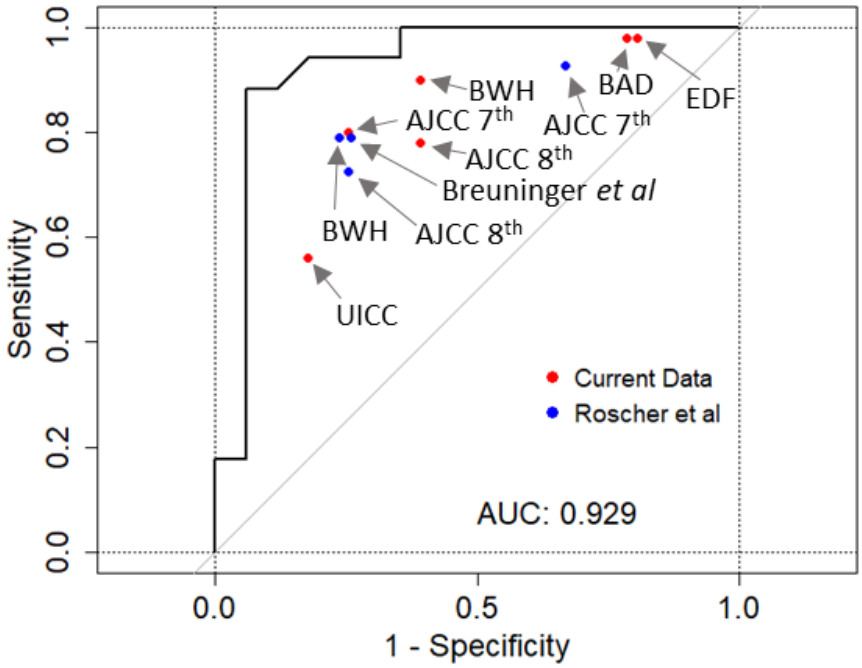


Figure 4:26: The final predictive model using DDOST and ANXA5 MRM data.

The stacked ensemble model was tested on the test set of data and a corresponding ROC curve produced. Evaluation of current clinical scoring systems in use that determine high and low risk of metastasis were applied to the cSCC samples used for MRM (red points). A study by Roscher *et al* assessed current clinical scoring systems and reported their sensitivities and specificities (blue points). AUC confidence interval = 0.8277 - 1. BAD, British Association of Dermatologists. EDF, European Dermatology Forum. BWH, Brigham Women's Hospital. AJCC, American Joint Committee on Cancer. UICC, Union for International Cancer Control. AUC, area under the curve.

4.4 Discussion

The aim of this chapter was to identify potentially important proteins from the discovery proteomic data generated in chapter 3 and validate those findings as well as assess their potential as prognostic biomarkers. From the discovery proteomics, there was a total of 133 potential biomarkers, however when reduced to the number of significantly differentially expressed proteins in both the 1D and 2D data, this number reduced to 33. Nonetheless, it is still challenging to explore this many proteins in depth (e.g. using MRM) and assess their prognostic value.

MRM was chosen as the most practical and viable option to verify the results of the discovery proteomics. Although other assays can be performed on peptides such as peptide ELISAs and western blots, these could be very expensive because antibodies would need to be raised against specific sequences of peptides identified in the initial proteomics. Furthermore, there is a risk of variable specificity because certain domains of peptide sequences might be homologous to other similar proteins. MRM is a specific and sensitive targeted MS approach capable of accurately quantifying peptides in a sample. As our laboratory had previously done MRM work on TKT (a protein identified as significantly differentially expressed in both 1D and 2D discovery proteomics), it was decided to use this protein, with 2 others for the MRM analysis.

Logistic regression was carried out on each combination of proteins, where the number in each combination was 3, one of which was always TKT. This machine learning gave rise to hundreds of models, of which the one with the highest ROC AUC in both the 1D and 2D data was TKT, DDOST and ANXA5. Unfortunately, it was later discovered that of the three peptides our laboratory currently had for TKT, only one was present in our cSCC discovery proteomics spectral library, meaning that the others had not been identified in the discovery proteomics. This resulted in just one peptide for TKT, which was likely to be of value, and which subsequently revealed no significant difference in TKT between P-M and P-NM. Typically, it is suggested to have 3 or more transitions (fragmentations) per peptide and have 3 or more peptides per protein to generate an accurate representation of the true abundance of a protein. Therefore, the results from a single TKE peptide meant that it could not be concluded with any confidence whether there was or was not more TKT in P-M than P-NM cSCCs, and thus TKT was omitted from future analysis. Nonetheless, all 6 peptides for DDOST and ANXA5 (3 for each) could be detected in the spectral library and what's more with a high spectral count and high intensity.

In light of this, DDOST and ANXA5 were employed for MRM verification/validation and modelling, without TKT. ANXA5 is a protein commonly used in flow cytometry to identify apoptotic cells as it gets localised to the outer membrane during this process. It is also known to have anticoagulative properties and may indeed be partly responsible for several placenta-mediated pregnancy complications (Aranda et al., 2018, Rogenhofer et al., 2018). moreover, it has been suggested that ANXA5 may be a suitable prognostic marker for other

types of cancers, including colorectal cancer (Xue et al., 2009, Sun et al., 2017) , renal cell carcinoma (Tang et al., 2017) and liver cancer (Peng et al., 2016). The mode of action of ANXA5 in promoting development of metastases is not fully understood, but it has been proposed that the increase in ANXA5 associated with poorer prognosis could be due to activation of the PI3K/Akt/mTOR signalling pathway (Tang et al., 2017), or could be partially attributed to ANXA5's effect on integrin signalling (Sun et al., 2018). It has been found that knockdown of ANXA5 leads to suppressed expression of various molecules in the integrin signalling pathway, which can have an anti-progressive effect on tumours (Sun et al., 2018, Janes and Watt, 2006).

There is limited research surrounding DDOST and any role it may play in cancer. Its typical cellular function is to catalyse the transfer of high mannose oligosaccharides to asparagine residues on newly formed polypeptides (Roboti and High, 2012). Nonetheless, one study found that there was higher average expression of DDOST in positive metastatic lymph nodes (1.29) compared to negative lymph nodes (0.4) in gastric cancer (Hasegawa et al., 2002). The human protein atlas also recognises DDOST as an unfavourable marker in renal cancer, liver cancer, and head and neck cancer but a favourable marker in endometrial cancer (Human Protein Atlas, 2018). The mechanisms by which DDOST could enable metastasis is unknown but may involve glycosylation and the impact this has on cancer progression (Pinho and Reis, 2015). DDOST has several aliases, including OST-48 (oligosaccharide transferase-4) and AGE-R1 (advanced glycosylation end products - receptor 1). AGE's are proteins or lipids which have been glycated, potentially altering their function (Baraka-Vidot et al., 2015), and were first identified in their role in degenerative diseases such as diabetes (Yamamoto and Sugimoto, 2016), chronic kidney disease (Clarke et al., 2016) and Alzheimer's disease (Drenth et al., 2017). AGE's are a ligand for DDOST and upon binding have been associated with a pro-inflammatory effect (Byun et al., 2017). Another significantly different protein identified in this study which has been found to have a pro-inflammatory effect is L-plastin.

As previously stated, L-plastin has been reported to be expressed in 68% of carcinomas and 53% of other solid tumours of nonepithelial origin (Lin et al., 1993). Furthermore, it has been identified to be important in the activation of T-cells (Wabnitz et al., 2007). Given that previous studies have identified higher numbers of T-cells in P-M samples compared to P-

NM samples and specifically higher T-Reg cells in P-M compared to P-NM (Lai et al., 2016, Lai et al., 2015), it was decided to immunohistochemically stain for L-plastin to identify if there is also more L-plastin in P-M samples compared to P-NM samples. This study found this to be significantly true (p=0.0136) and therefore it is possible that L-plastin is activating T-cells and potentially T-Reg cells, in a pro-oncogenic fashion. Simultaneously, in Chapter 3 of this thesis, PI3K-Akt signalling was identified as a potential key pathway involved in the metastasis of cSCC. In light of this, it has been identified that PI3K-Akt signalling promotes prostate cancer metastasis via upregulating L-plastin (Chen et al., 2017a).

Using modelling, DDOST and ANXA5 showed predictive prognostic value in the discovery proteomic data, and verification/validation MRM found more DDOST and ANXA5 in P-M than P-NM cSCCs. A previous study by Azimi et al (Azimi et al., 2016), used FFPE cSCC samples to identify biomarkers of cSCC by comparing normal skin to cSCC, but they did not identify ANXA5 or DDOST in their list of significantly differentially expressed proteins. Moreover a very recent study by the same group, investigating proteomic differences between Bowen's disease, actinic keratosis and cSCC FFPE samples did not identify ANXA5 or DDOST (Azimi et al., 2019).

There are currently no widely used protein prognostic biomarkers for metastasis from primary cSCC used in the clinic today. This current study identified significantly more DDOST and ANXA5 in P-M samples than P-NM samples in three separate phases; discovery, verification on the same discovery samples using a targeted approach and finally validation in a previously unseen separate sample cohort. Using verification and validation MRM data, there was a total of 101 samples, each with 6 accurate peptide measurements, i.e. 3 for DDOST and 3 for ANXA5. With the size of this sample set and the accurate nature of the measurements of DDOST and ANXA5, it was decided to create a machine learning model of the data to identify to what extent these proteins could predict metastasis in cSCC. Each of the models produced were relatively weak learners and so a stacked ensemble approach was taken.

There are three main types of ensemble modelling; bagging, boosting and stacking. Typically, bagging is used to decrease variance, boosting used to reduce biasing and stacking employed to generally improve predictions. The resulting stacked model produced an AUC of 0.929 (CI = 0.8277 - 1), with an optimal accuracy of 91.18% (sensitivity = 88.24%,

specificity = 94.12%). Moreover, the model out performed every current clinical scoring system at every threshold (every edge of the ROC curve) in both sensitivity and specificity. In addition, high DDOST and ANXA5 each showed a positive association with a quicker time to metastasis from the primary cSCC and their combined high expression has a positive, highly significant association with a quicker time to metastasis. This highlights that not only are these two proteins good indicators of future metastasis but that they seem to be associated with how long it takes for cSCC to metastasise. In conclusion, this study verified and validated findings outlined in chapter 3 in addition to identifying the prognostic predictive potential of two key proteins; DDOST and ANXA5.

Chapter 4

Chapter 5: Proteomic characterisation of Melanoma skin tumours

5.1 Introduction

Melanoma arises when melanocytes acquire genetic mutations and become cancerous. Melanocytes are pigment producing cells, typically residing in the skin but can be found in other areas of the body such as the uvea (Shain et al., 2019). Cutaneous melanoma rates in the UK have increased by 175% in males between 1993-1995 and 2013-2015 and 95% in females during the same period (Cancer Research UK, 2018). Melanoma accounts for 5% of all cancer in the UK population and is the 5th most common type of cancer, excluding NMSCs (Cancer Research UK). Melanoma is the 5th leading cancer in males and the 7th leading cancer in females in the USA (not including NMSCs) (Siegel et al., 2013). It has been reported that on average, an individual suffering from melanoma could lose 20.4 years of potential life, almost 4 years more than that of all other malignant cancers where the potential loss of life has been estimated 16.6 years (Ekwueme et al., 2011).

Melanoma diagnosis is typically carried out through visual examination by a healthcare professional, usually a dermatologist or a doctor with a special interest in dermatology. In addition to histological confirmation on the excised lesion, immunohistochemical staining is used to support the diagnosis. Several histological stains have been identified as markers which help to identify melanoma, these include S-100, HMB45, melan-A/MARTI and MITF (Kashani-Sabet, 2014). More recently however, a type of artificial intelligence known as a deep convolutional neural network (CNN) was employed to diagnose malignant melanoma clinically from a library of almost 130,000 images. The resulting model was able to predict malignant melanoma with as much accuracy as human experts (Esteva et al., 2017).

Although the use of CNN allows for early diagnosis and subsequently faster treatment, it is still inevitable that some of these primary tumours may go on to metastasise. It is believed that one third of melanoma patients will experience recurrence (Soong et al., 1998), whether it be local, nodal or distant. Despite the capability of melanoma to spread to any organ, the most frequent sites of distant metastasis are the liver, bone and brain and the 5

year survival rate for metastatic melanoma is <15% (Tas, 2012). Furthermore, it has been identified that the site of recurrence has a significant effect on the mortality rate, with those experiencing metastasis to visceral sites being at the greatest risk (Soong et al., 1998).

Due to the high mortality rate associated with metastatic melanoma, it is crucial to identify prognostic markers for melanoma and indeed markers for metastasis as well. LDH has been described as “an independent and highly significant predictor of survival outcome among patients with stage IV [melanoma]” (Balch et al., 2009). Despite its use in the clinic, studies have criticised its ability to predict prognosis due to its low sensitivity and specificity (Bougnoux and Solassol, 2013). Several other biomarkers have been proposed as potential prognostic markers; these have included, but are not limited to, CXCR1, CXCR2, CXCR3, CXCR4, CCR5, CCR7 and CCR10. However, studies have suggested that out of this list, only CXCR4 has enough prognostic data to be appropriately used as a prognostic marker for melanoma (Scala et al., 2005, Gould Rothberg et al., 2009). Furthermore, a study investigating the efficacy of several suggested prognostics marker for melanoma (BRAF, MMP2, P27, Dicer, Fbw7 and Tip60) found that although BRAF and MMP2 are strong prognostic markers for stage 1 and stage 2 melanoma respectively, there are very few prognostic markers useful for late stage AJCC melanomas and metastatic melanoma (Cheng et al., 2015).

A proteomics study which utilised raw mass spectra of 205 serum samples from 101 AJCC stage 1 melanomas and 104 AJCC stage 4 melanoma was able to correctly predict the stage over 80% of the time (Mian et al., 2005). More recently however, proteomic studies have begun to look at FFPE tissue due to the abundance of available samples and the amount of complimentary clinical data which comes with them. A study on melanoma FFPE identified 171 proteins which varied between benign nevi, primary melanoma and metastatic melanomas. Despite this being the largest proteomic FFPE melanoma study to date, it focuses on the differences between different lesions (i.e. benign nevi and melanoma) and melanomas at notably different stages (primary and metastatic) (Byrum et al., 2013).

In this chapter, primary melanomas which subsequently metastasised and primary melanomas which did not metastasise were subjected to proteomic analysis to identify protein biomarkers of metastasis. It is vital to identify markers of metastasis in patients who have not gone on to metastasise yet in an attempt to identify factors which might

allow clinicians to prevent the development of metastases. Furthermore, it is crucial to identify whether a tumour is likely to metastasise after excision because melanoma is frequently excised with no evidence of metastasis at that stage yet metastases present at a later date.

5.2 Methods

A total of 48 samples were used in this chapter, consisting of 24 Pmel-M samples and 24 Pmel-NM samples. These samples were stratified for Breslow depth to ensure there was no significant bias in the sample cohort. The optimised method developed in chapter 3 was used for the discovery phase in this chapter.

5.2.1 Proteomic analysis of melanoma samples

Full materials and methods of the discovery proteomics methods can be found in chapter 2.6. Samples were quantified using a Direct Detect infrared spectrometer outlined in chapter 2.7 and cleaned up using a C18 reverse phase technique (full material and methods can be found in chapter 2.8). Samples were then analysed using a Waters Synapt G2-Si high resolution mass spectrometer using the methods described in chapter 2.9.

5.2.2 Bioinformatics and data analysis

Chapter 2.10.2 describes the way in which protein concentrations were normalised. Statistical analysis was performed on the results by comparing Pmel-M data to Pmel-NM data. Whole proteome analysis was carried out through the use of volcano plots as described in chapter 2.11.1 and topological data analysis as outlined in chapter 2.11.5. Significantly differentially expressed proteins were further analysed using STRING, gene ontology and WGCNA as outlined in chapters; 2.11.2, 2.11.3, 2.11.4, respectively.

5.2.3 Targeted mass spectrometry of melanoma

Only one protein, Keratin 9, was identified as significantly differentially expressed between Pmel-M and Pmel-NM and so was selected as one of the three proteins to progress to targeted mass spectrometry. Our laboratory had already performed a targeted analysis of GSN in another experiment and as such we had isotopically heavy labelled peptides for this

protein, therefore it too was selected for MRM analysis. The final protein selected was based off of the discovery data and biological relevance.

A spectral library of the discovery proteomics data was created as outlined in chapter 2.9.3.1. Targeted proteomics was carried out as described in chapter 2.9.3.2. Briefly, isotopically heavy labelled peptides were initially analysed using a Synapt G2-Si high resolution mass spectrometer, to assess their suitability as MRM targets. A serial halving dilution of each peptide was then analysed in a background melanoma peptide matrix to determine calibration data.

100fmol of each isotopically heavy labelled peptide was spiked into 24 Pmel-M samples and 24 Pmel-NM samples and analysed on a Synapt G2-si mass spectrometer. Calibration data achieved in initial analysis of heavy labelled peptides was used to calculate the true amount of each heavy labelled peptide in each sample. Using the light: heavy ratio, it was then possible to calculate the amount of native peptide in each sample.

5.3 Results

The proteomic discovery method used in this chapter has been described in chapter 2 and was also carried out in chapter 3. To ensure the same methodology was suitable for the melanoma portion of this project, bioreplicate experiments of melanoma were undertaken and the reproducibility assessed (**Figure 5.1**). A coefficient or correlation was determined by correlating every protein abundancy in one sample to each other sample, which gives an r value where the closer the number to 1, the better the correlation. An acceptable r value is dependent on many things, but an r value of > 0.8 is generally acceptable; in comparison of the bioreplicate experiments, the r values obtained were 0.7933, 0.7938 and 0.8224. Furthermore, 47.8% of the proteins identified were identified in at least 2 of the 3 bioreplicates.

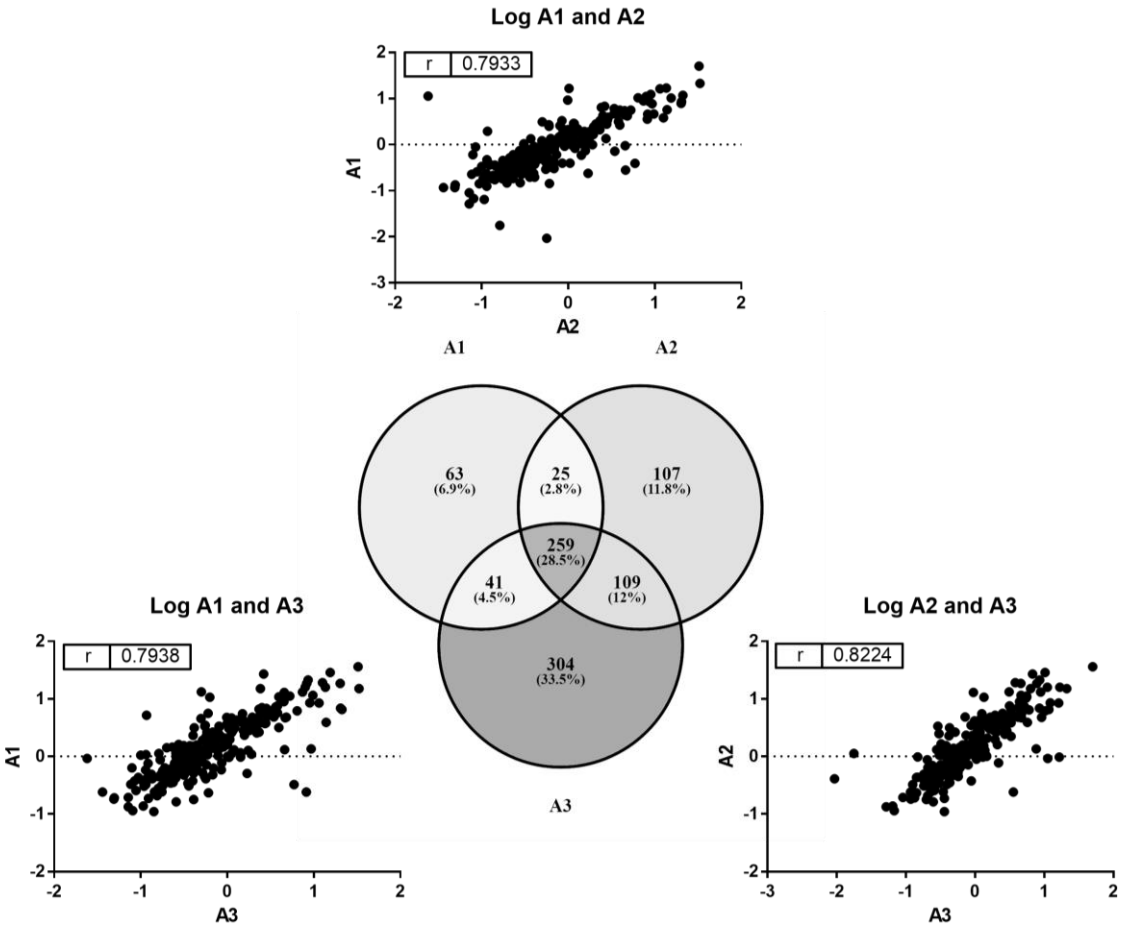


Figure 5:1: investigating the technical reproducibility of the RapiGest method in melanoma.

Proteins from melanoma sample "A" was extracted and quantified with mass spectrometry in triplicate. 28.5% of all the unique proteins in three experiments were identified in all experiments. 47.8% of the unique proteins were identified in two or more experiments. There was a high positive correlation of all proteins between experiments.

5.3.1 Clinical characteristics of melanoma samples

Similar to chapter 3, 24 Pmel-M and 24 Pmel-NM were used in the discovery proteomics (**Table 5:1**). Briefly, there were slightly more males than females in the Pmel-M group and slightly more females in the Pmel-NM group. The majority of melanomas were classified as superficial spreading in both groups. The majority of samples had a reported Clark's levels of IV. As Breslow thickness is a known indicator of prognosis and risk of metastasis, we decided to stratify for Breslow thickness, and as such, there was no significant difference in this parameter between the Pmel-M and Pmel-NM groups. Similar to chapter 3, information on geographic ancestry was not collected and therefore was not available.

Table 5.1: A Table of clinical characteristics of melanoma samples used for discovery proteomics

	Pmel-M	*Pmel-NM*
Number of Samples	24	24
Male	14 (58.33%)	9 (37.5%)
Female	10 (41.67%)	15 (62.5%)
Superficial spreading	17 (70.83%)	20 (83.33%)
Nodular	6 (25%)	4 (16.67%)
Desmoplastic	1 (4.17%)	0 (0%)
Pigmentation – High	6 (25%)	4 (16.67%)
Pigmentation – Moderate	7 (29.17%)	7 (29.17%)
Pigmentation -Low	4 (16.67%)	7 (29.17%)
Clark's Level V	2 (8.33%)	0 (0%)
Clark's Level IV	16 (66.67%)	15 (62.5%)
Clark's Level ≤ III	6 (25%)	9 (37.5%)
Breslow thickness (mm)	2.76 ± 1.63	2.08 ± 1.46

Pmel-M, Primary tumours which metastasised. Pmel-NM, Primary tumours which did not metastasise. pigmentation could not be acquired for all as several FFPE blocks were returns to histopathology either by request or because too little tissue was left for research

5.3.2 Protein quantitation and identification

Following extraction and digestion of proteins from the melanomas, peptide quantification was carried out to ensure later loading onto the mass spectrometer was standardised (**Table 5:2**). There was a wide variety in the total yield of peptides from each sample, ranging from 18.3µg to 217.5µg. The median total peptide concentration was 76.6µg in the Pmel-M group and 55.9µg in the Pmel-NM group.

Table 5.2: Quantification of total peptide concentration using DirectDetect

Metastatic			Non-Metastatic		
Sample	**mg/ml protein**	**Total peptide (µg)**	**Sample**	**mg/ml protein**	**Total peptide (µg)**
Pmel-M1	0.814	81.4	**Pmel-NM1**	0.475	47.5
Pmel-M2	0.183	18.3	**Pmel-NM2**	0.405	40.5
Pmel-M3	0.991	99.1	**Pmel-NM4**	0.519	51.9
Pmel-M4	1.298	129.8	**Pmel-NM7**	0.645	64.5
Pmel-M5	1.365	136.5	**Pmel-NM11**	0.358	35.8
Pmel-M6	0.23	23	**Pmel-NM13**	0.338	33.8
Pmel-M7	0.661	66.1	**Pmel-NM14**	1.269	126.9
Pmel-M8	0.361	36.1	**Pmel-NM16**	0.506	50.6
Pmel-M9	0.312	31.2	**Pmel-NM17**	0.394	39.4
Pmel-M10	0.683	68.3	**Pmel-NM20**	0.529	52.9
Pmel-M11	1.238	123.8	**Pmel-N21**	0.682	68.2
Pmel-M15	0.277	27.7	**Pmel-NM22**	0.597	59.7
Pmel-M16	0.76	76	**Pmel-NM23**	0.589	58.9
Pmel-M17	0.75	75	**Pmel-NM24**	0.631	63.1
Pmel-M18	0.47	47	**Pmel-NM25**	0.404	40.4
Pmel-M20	0.905	90.5	**Pmel-NM26**	0.66	66
Pmel-M21	2.175	217.5	**Pmel-NM30**	0.358	35.8
Pmel-M22	0.903	90.3	**Pmel-NM31**	0.413	41.3
Pmel-M24	1.615	161.5	**Pmel-NM32**	0.613	61.3
Pmel-M26	0.807	80.7	**Pmel-NM33**	0.428	42.8
Pmel-M27	1.439	143.9	**Pmel-NM34**	1.535	153.5
Pmel-M28	0.47	47	**Pmel-NM35**	1.23	123
Pmel-M29	0.67	67	**Pmel-NM36**	1.723	172.3
Pmel-M30	0.772	77.2	**Pmel-NM37**	0.845	84.5

Pmel-M, Primary tumours which metastasised. Pmel-NM, Primary tumours which did not metastasise.

The total number of protein IDs per sample varied, both between Pmel-M and Pmel-NM and between the 1D and 2D LC separation experiments (**Figure 5.2**). Less proteins were identified in samples separated by 2D compared to those separate by 1D, and fewer proteins were identified in Pmel-NM than in Pmel-M cases. The number of IDs ranged from 114 to 1,111 and the mean number of IDs was 435 with a standard deviation of 188.

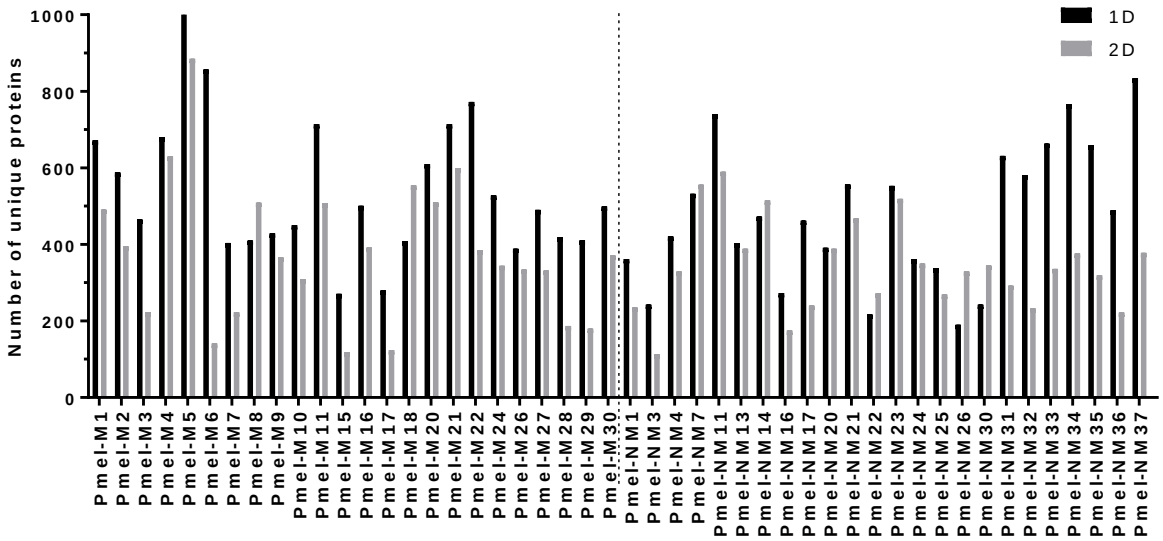


Figure 5:2: Numbers of unique proteins identified using 1D, and separately 2D, fractionation prior to MS.

3.75µg of protein per melanoma sample was analysed using mass spectrometry and the results were processed in the Protein Lynx Global Server to identify individual proteins.

A total of 3,447 unique proteins were identified from all 48 melanoma samples, which consisted of 2,750 IDs in samples separate by 1D and 2,259 IDs in samples separate by 2D. 45.32% of proteins were identified in the results from both 1D and 2D which equated to a total of 1,562 IDs.

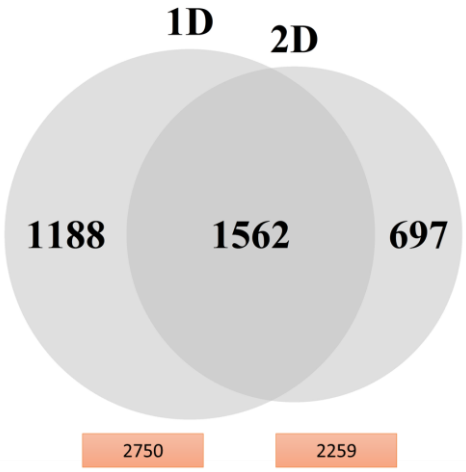


Figure 5:3: Number of unique proteins identified from 1D and 2D fractionation

The total number of proteins identified following 1D fractionation of the 48 melanoma samples (which included 24 Pmel-M and 24 Pmel-NM) was 2750, and following 2D separation was 2259. 45.32% of proteins were identified in both 1D and 2D fractionation methods

As described in chapter 3, proteomic data often suffers from the “floor effect” as the limitations of instruments causes a bottom level below which proteins cannot be detected. For this reason, it was suggested non-parametric tests were used to statistically analyse the data. Alternatively, log transforming the data can sometimes result in a Gaussian distribution and thus create a suitable dataset for parametric tests. Log transforming this data from the melanoma samples resulted in a mixture of normally and non-normally distributed samples (**Figure 5.4**, **Appendix 3**). For this reason, a conservative approach was taken and non-parametric tests were used to analyse the data (unless otherwise stated).

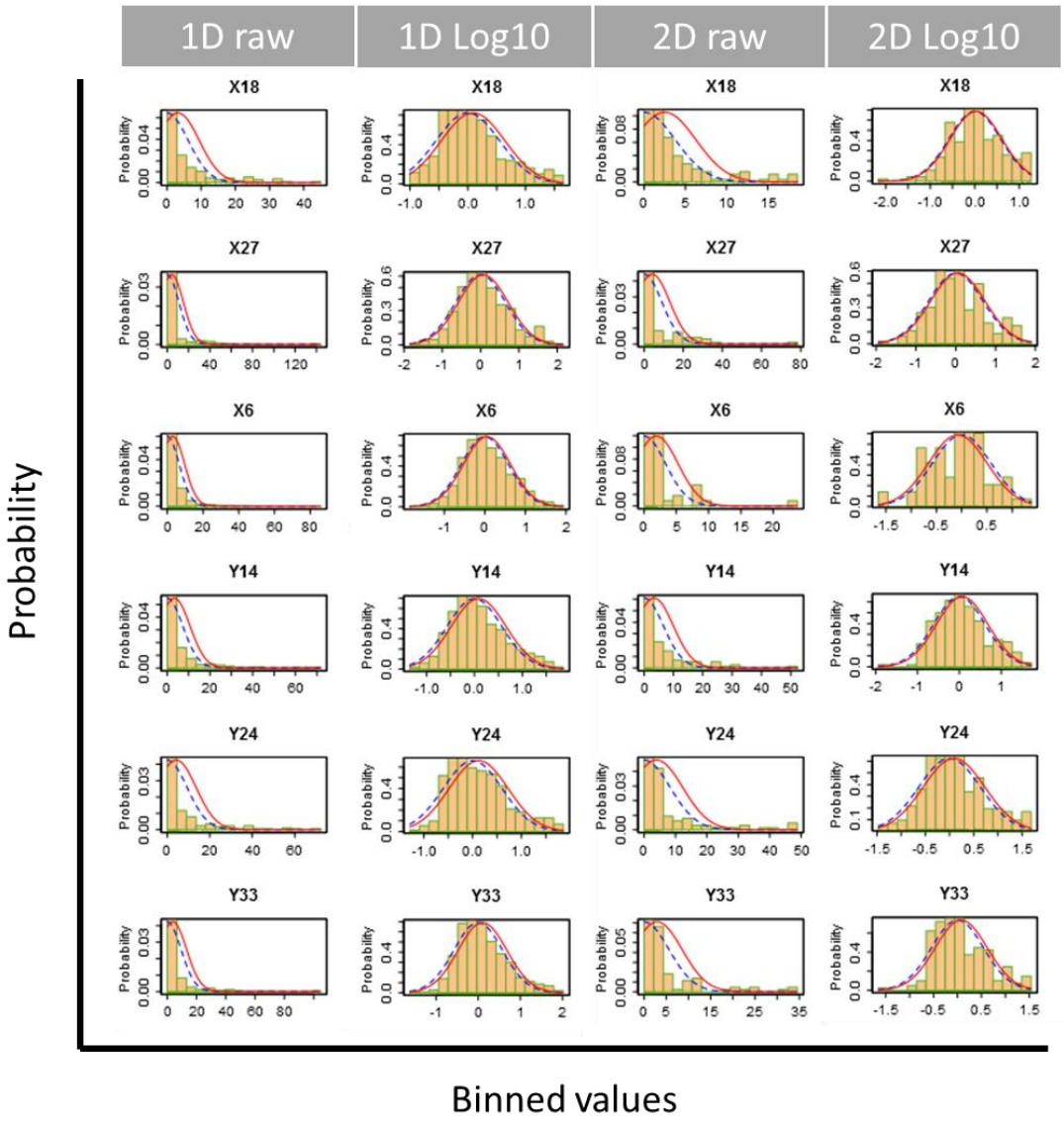


Figure 5:4: Histograms of proteomics data from the melanoma samples.

Histograms of proteins abundancies were created using Inferno. As the sensitivity of the MS instrument is finite, a "floor effect" can be seen on the raw data and therefore the data was log10 transformed to assess the distribution. A mixture of normal and non-normal distributions can be seen in the log10 transformed data; for this reason, a conservative approach was taken and subsequent analysis used non-parametric tests. Histograms of all melanoma samples can be found in **Appendix 3**.

5.3.3 Significantly differentially expressed proteins

A combined total of 31 significantly differentially expressed proteins (P<0.05) were identified between Pmel-M and Pmel-NM melanomas in the 1D and 2D data, (**Figure 5.5**). 16 significantly differentially expressed proteins were identified between the Pmel-M and Pmel-NM tumours in the 1D data and 16 significantly differentially expressed proteins identified between Pmel-M and Pmel-NM tumours in the 2D data. One protein was identified as significantly differentially expressed between Pmel-M and Pmel-NM melanomas in both the 1D and 2D data; this protein was keratin 9 (KRT9).

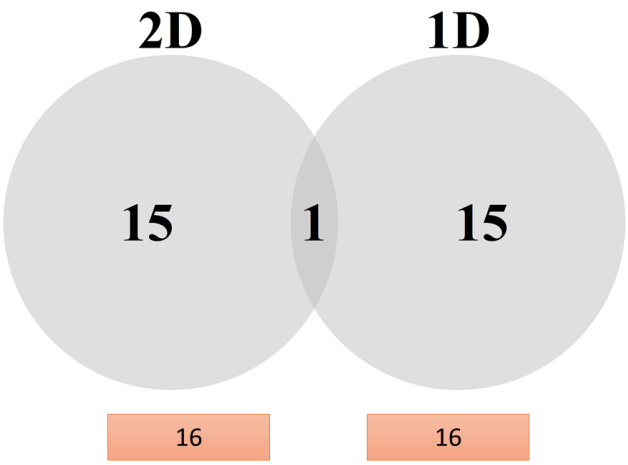


Figure 5:5: Numbers of proteins that were differentially expressed between Pmel-M and Pmel-NM melanomas in the 1D and the 2D data

A total of 16 proteins were identified as significantly differentially expressed between Pmel-M and Pmel-NM in the 1D data and also in the 2D data, however, only one protein was identified as significantly differentially expressed between Pmel-M and Pmel-NM in both the 1D and the 2D data. P<0.05, Man Whitney U test for significance.

A volcano plot of each data set (1D and 2D) of the melanoma samples was created. The volcano plot for the 1D data highlighted proteins P01860 (IGHG3), Q9H6N6 (MYH16), Q92817 (EVPL) and Q7KZF4 (SND1) by their low P value and high fold changes. Volcano plot for the 2D data highlighted proteins Q8N1N4 (KRT78), Q14240 (WIF4A2), P19012 (KRT15), P35527 (KRT9), P40926 (MDH2) and Q8NBS9 (TXNDC5) with a low P value and high fold changes.

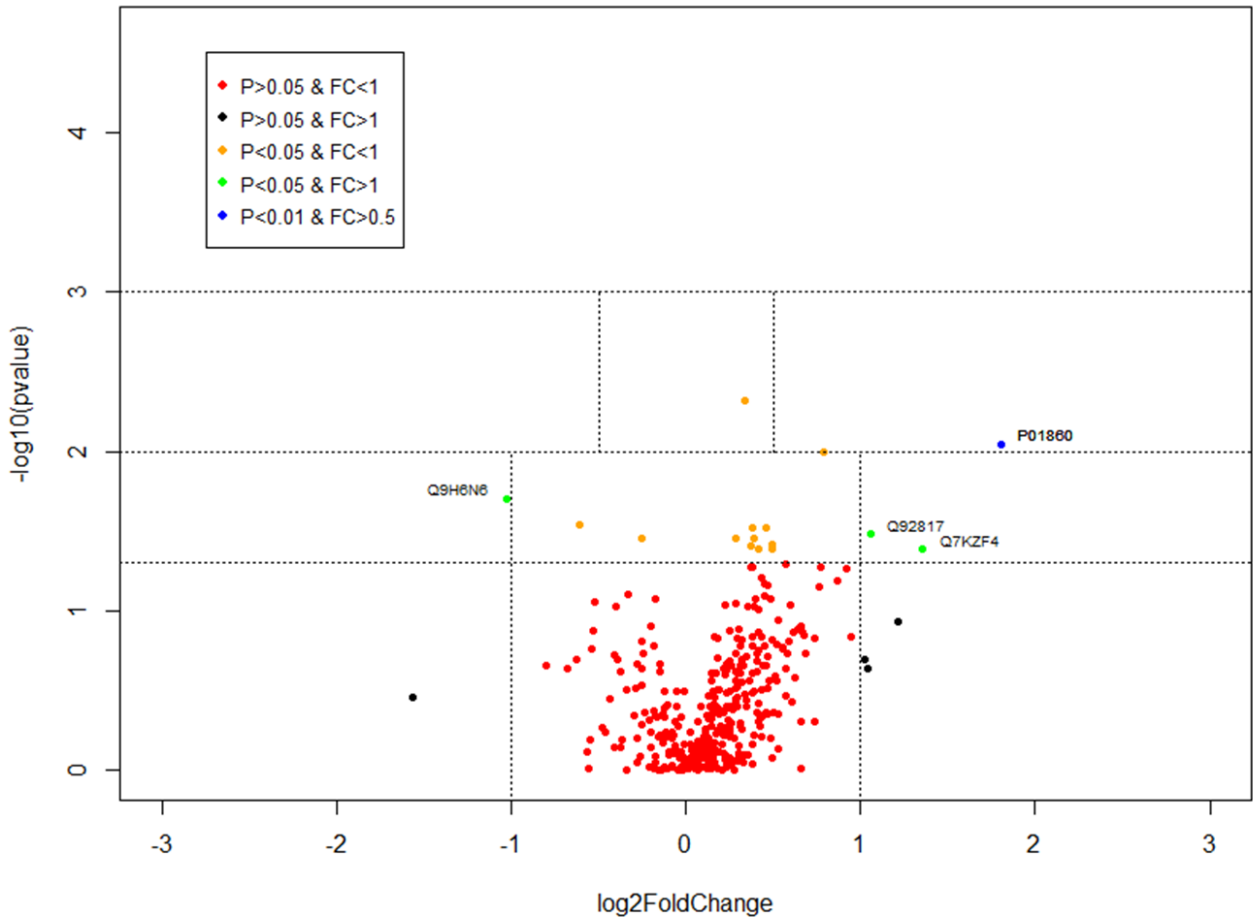


Figure 5:6: Volcano plot of 1D data highlights proteins of interest.

P-values obtained from Pmel-M vs Pmel-NM using Mann Whitney U test and subsequently log10 transformed for visualisation in volcano plots. Fold changes (FC) were acquired by subtracting mean quantification of Pmel-M from Pmel-NM and were then log2 transformed to generate the graph. Red points indicate non-significant p-value (p>0.05) and fold change (<1 log2). Black points indicate non-significant p-value (p>0.05) but significant fold change (>1 log2). Orange points indicate significant p-value (p<0.05) but not significant fold change (<1 log2). Green points indicate significant fold change (>1 log2) and significant p-value (p<0.05). Labels are Uniprot protein accession numbers.

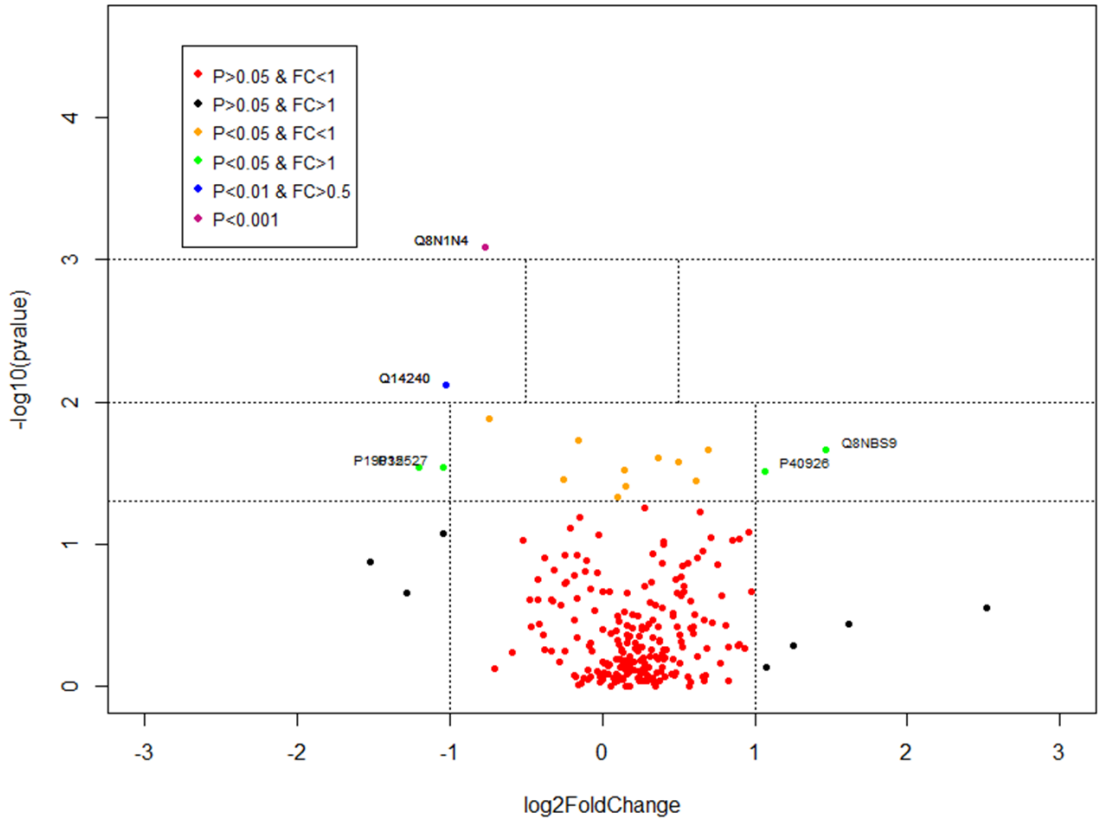


Figure 5:7: Volcano plot of 2D data highlights proteins of interest.

P-values obtained from Pmel-M vs Pmel-NM using Mann Whitney U test and subsequently log10 transformed for visualisation in volcano plots. Fold changes (FC) were acquired by subtracting mean quantification of Pmel-M from Pmel-NM and were then log2 transformed to generate the graph. Red points indicate non-significant p-value (p>0.05) and fold change (<1 log2). Black points indicate non-significant p-value (p>0.05) but significant fold change (>1 log2). Orange points indicate significant p-value (p<0.05) but not significant fold change (<1 log2). Green points indicate significant fold change (>1 log2) and significant p-value (p<0.05). Labels are Uniprot protein accession numbers.

There were two proteins with P values below 0.001 for differential expression between Pmel-M and Pmel-NM melanomas; these were actin gamma 1 (ACTG1) and keratin type II cytoskeletal 78 (KRT78) in the 1D and 2D data, respectively. There was a range in the fold changes for the same comparison of the proteins in the 1D data with the lowest being -0.252 and the highest being +1.802 (**Table 5:3**). Similarly, there was a wide range of fold changes between Pmel-M and Pmel-NM melanomas in the 2D data with the lowest being +0.097 and the highest being +1.461 (**Table 5:4**). The one protein identified as significantly

differential expressed between the Pmel-M and Pmel-NM groups in both 1D and 2D, was KRT9, which is indicated by the green shading in Tables 5.3 and 5.4.

Table 5.3: List of significantly differentially expressed proteins between Pmel-M and Pmel-NM in the 1D data and their respective fold changes and P values

Uniprot ID	*Gene ID*	*Protein Name*	*log2FoldChange*	*P value*
P63261	ACTG1	Actin Gamma 1	0.338013	4.82E-03
P01860	IGHG3	Immunoglobulin Heavy Constant Gamma 3	1.802058	0.009022
P25786	PSMA1	Proteasome Subunit Alpha 1	0.788837	0.010045
Q9H6N6	MYH16	myosin heavy chain 16	-1.02722	0.019726
P35527	KRT9	Keratin 9	-0.61257	0.02913
P62491	RAB11a	Ras-related protein Rab-11A	0.382207	0.030177
P13639	EEF2	Eukaryotic elongation factor 2	0.457982	0.030267
Q92817	EVPL	Envoplakin	1.057036	0.03326
P30086	PEB1	Periplasmic amino acid-binding protein	-0.25176	0.035189
P68104	EEF1A1	Eukaryotic translation elongation factor 1 α1	0.386652	0.035508
P62805	HISTH4	Histone H4	0.284845	0.035508
P46776	RPL27A	Ribosomal Protein L27a	0.496393	0.037992
P35579	MYH9	Myosin-9	0.375744	0.039408
Q15063	POSTN	Periostin	0.493123	0.041138
Q7KZF4	SND1	Staphylococcal nuclease domain containing 1	1.356476	0.041184
Q7L7L0	HIST3H2A	Histone H2A type 3	0.414143	0.041316

p-values were obtained through Mann Whitney U test for significance between Pmell-Ms and Pmel-NMs. Fold change was calculated from the mean of protein abundancies between each group. Green shading indicates single significantly differentially expressed protein in both the 1D and 2D data

Table 5.4: List of significantly differentially expressed proteins between Pmel-M and Pmel-NM in the 2D data and their respective fold changes and P values

Uniprot ID	*Gene ID*	*Protein Name*	*log2FoldChange*	*P value*
Q8N1N4	KRT78	Keratin 78	-0.76589	8.25E-04
Q14240	EIF4A2	Eukaryotic Translation Initiation Factor 4A2	-1.02262	0.007681
P01859	IGHG2	Immunoglobulin Heavy Constant Gamma 2	-0.74507	0.013166
Q15019	SEPT2	Septin 2	-0.15355	0.01868
Q8NBS9	TXNDC5	hioredoxin Domain Containing 5	1.461222	0.021754
P20700	LMNB1	Lamin B1	0.695982	0.021874
P62249	RPS16	Ribosomal Protein S16	0.370449	0.024652
P05023	ATP1A1	ATPase Na+/K+ Transporting Subunit Alpha 1	0.497553	0.026246
P19012	KRT15	Keratin 15	-1.206	0.028958
P35527	KRT9	Keratin 9	-1.04325	0.029021
P16070	CD44	Cluster of Differentiation 44	0.140834	0.030051
P40926	MDH2	Malate Dehydrogenase 2	1.063803	0.030735
Q16555	DPYSL2	Dihydropyrimidinase-related protein 2	-0.25623	0.034954
O75083	WDR1	WD Repeat Domain 1	0.613523	0.035621
P06396	GSN	Gelsolin	0.155568	0.039034
P23528	CFL1	Cofilin 1	0.096728	0.046927

p-values were obtained through Mann Whitney U test for significance between Pmel-Ms and Pmel-NMs. Fold change calculated from mean of protein abundancies between each group. Green shading indicates significantly differentially expressed protein in both the 1D and 2D data.

Examples of significantly differentially expressed proteins between Pmel-M and Pmel-NM groups from the 1D data can be seen in **Figure 5.8**. The median abundancy of proteins varied a lot with the lowest being 0.37ng of PSMA1 in the Pmel-NM group to the highest being 20.72ng of ACTG1in the Pmel-M group.

Examples of significantly differentially expressed proteins from the 2D data can be seen in **Figure 5:9**. The lowest median abundancy of a protein was 0.109ng of KRT78 in the Pmel-M group and the highest was 2.086ng of LMNB1in the Pmel-M group.

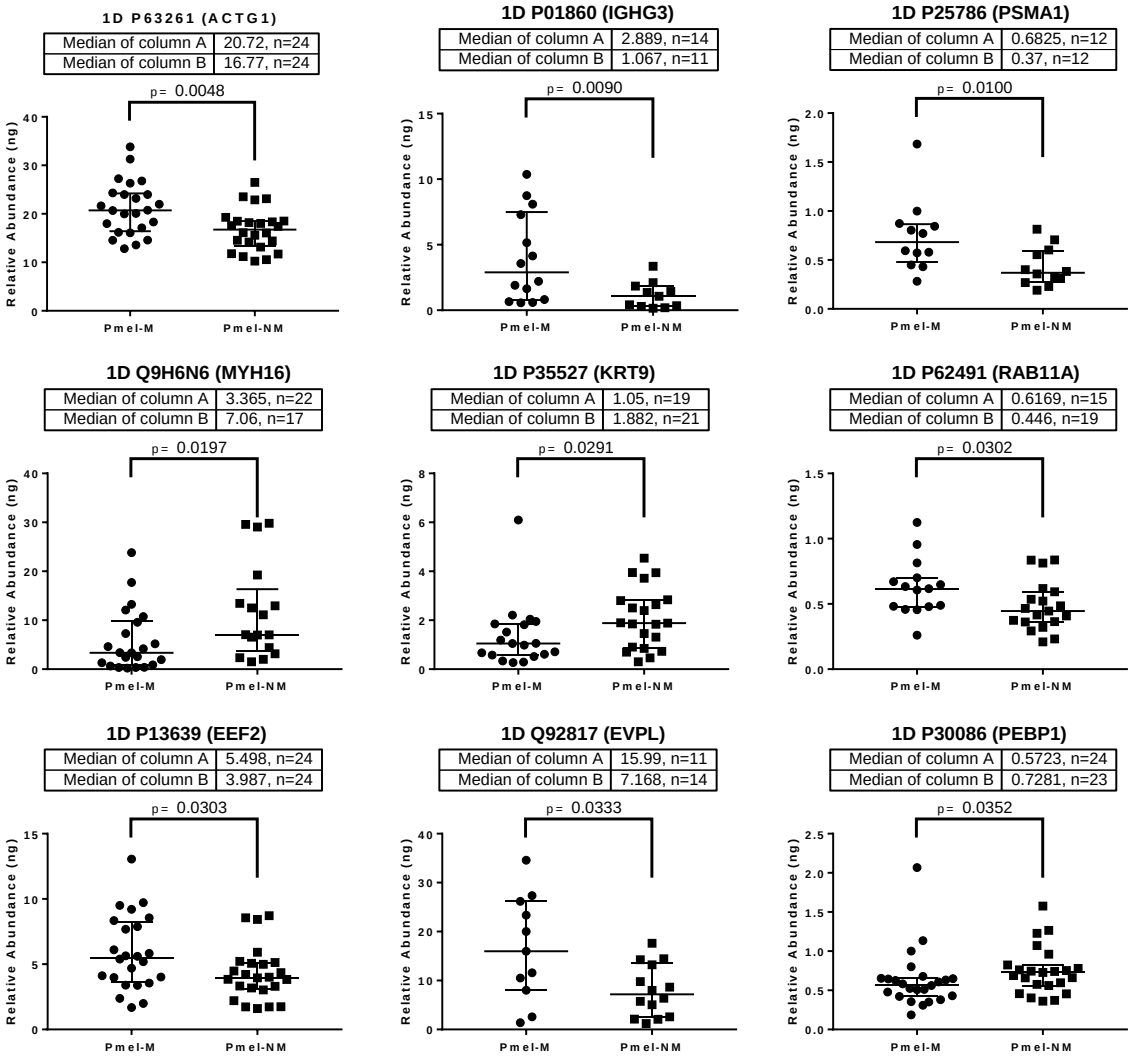


Figure 5:8: Examples of significantly differentially expressed proteins between Pmel-M and Pmel-NM melanomas in 1D proteomic data

P values obtained through Mann Whitney U test for significance. Median with interquartile range shown.

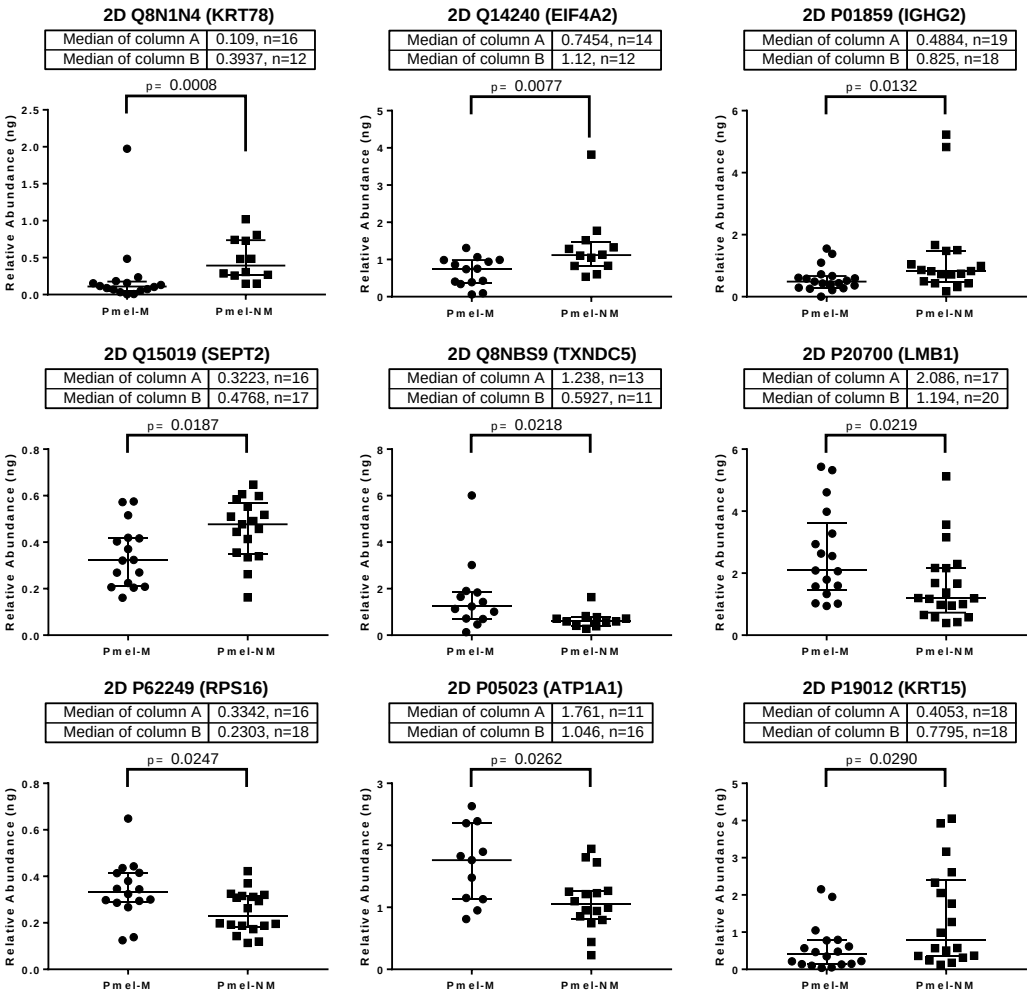
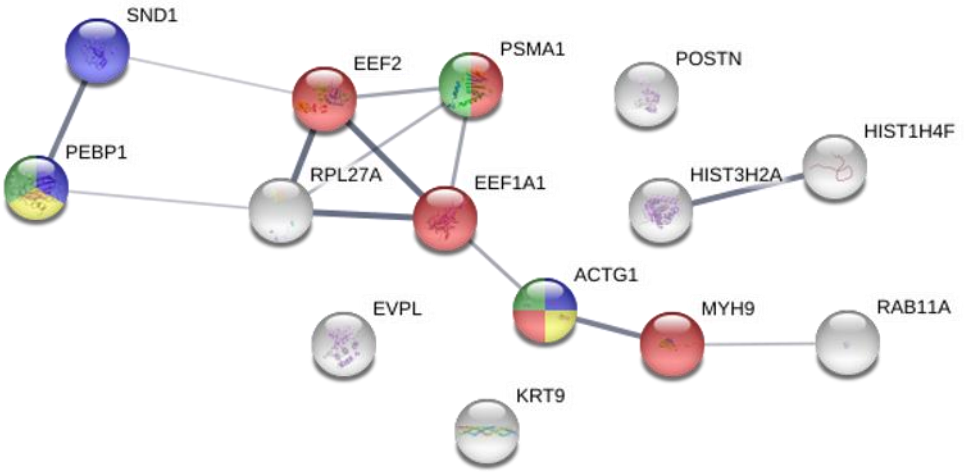


Figure 5:9: Examples of significantly differentially expressed proteins between Pmel-M and Pmel-NM melanomas in 2D proteomic data

P values obtained through Mann Whitney U test for significance. Median with interquartile range shown.

5.3.4 Search tool for the retrieval of interacting genes/proteins (STRING) analysis

STRING analysis was used to create a network from the significantly differentially expressed proteins between Pmel-M and Pmel-NM identified in the 1D data (**Figure 5:10**) and 2D data (**Figure 5:11**). Significantly differentially expressed proteins from the 1D data produced a structure with 13 edges. A network with this many proteins would be expected to have 3 edges by chance alone and thus the network is significantly enriched in interactions (P=0.000004) and suggests that these proteins are likely to be connected in determining biological aspects of melanoma metastasis. Reactome pathway overlay highlighted several proteins involved in the MAPK pathway, including BRAF and RAF. Furthermore, Reactome enrichment displayed a strong coverage of an innate immune response.

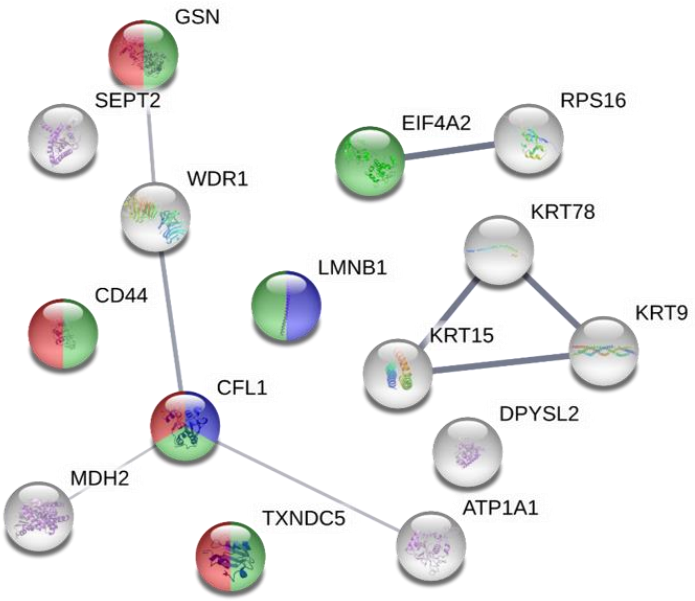


	Reactome Pathway	FDR
	Signalling by BRAF and RAF fusions	0.0026
	paradoxical activation of RAF signalling by kinase inactive BRAF	0.0084
	Innate immune response	0.0084
	MAPK family signalling cascades	0.0111

Figure 5:10: STRING analysis of significantly differentiated proteins between Pmel-M and Pmel-NM from the 1D data

Significantly differentiated proteins between Pmel-M and Pmel-NM from 1D discovery proteomics were analysed using STRING. A medium confidence score of 0.4 was allowed in the structure creation as recommended by the software manufacturers. Thickness of lines (edges) indicates confidence in association between two proteins. Total number of nodes is 14. Total number of edges is 13. The Reactome pathway enrichment has been overlaid onto the STRING structure. FDR, False discovery rate.

Significantly differentially expressed proteins between Pmel-M and Pmel-NM from the 2D data yielded a structure with 8 edges. From a similar set of proteins and a network of similar size, only one edge would be expected by chance and the network is therefore enriched in interactions (P= 0.000106), again suggesting that the proteins in the network are at least in part biologically connected to melanoma metastases. Amongst other pathways, Reactome enrichment was identified in the immune system. There was also enrichment in JAK-STAT signalling after IL12 stimulation, a cytokine secreted by antigen presenting cells (Dorman and Holland, 2000).



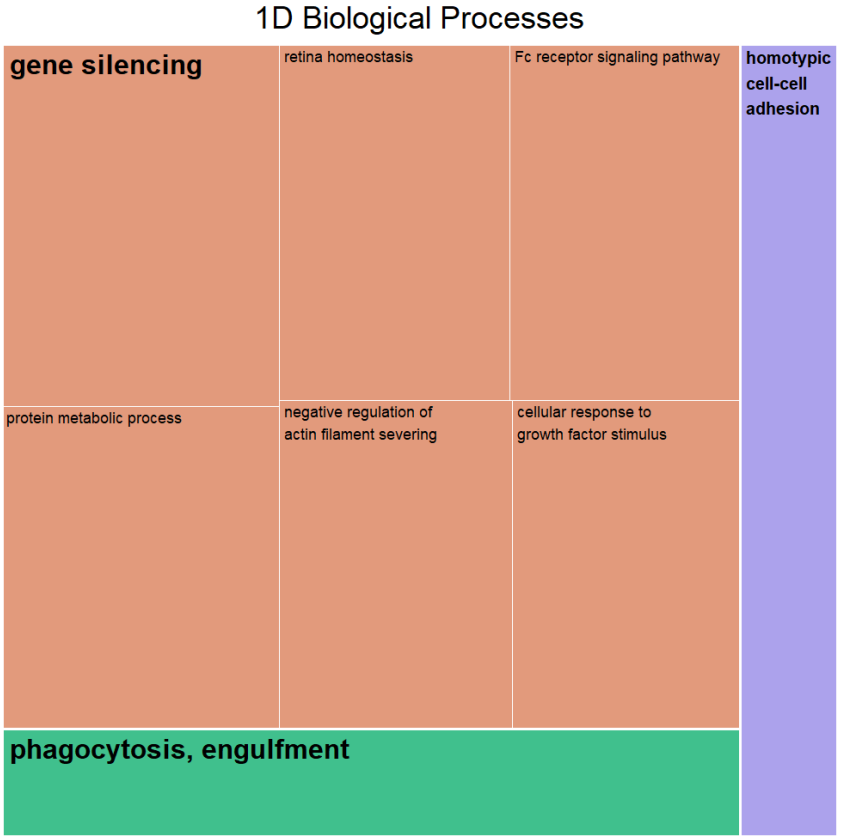
	Reactome Pathway	FDR
	gene and protein expression by JAK-STAT signalling after IL-12 stimulation	0.0166
	immune system	0.0186
	innate immune response	0.0353

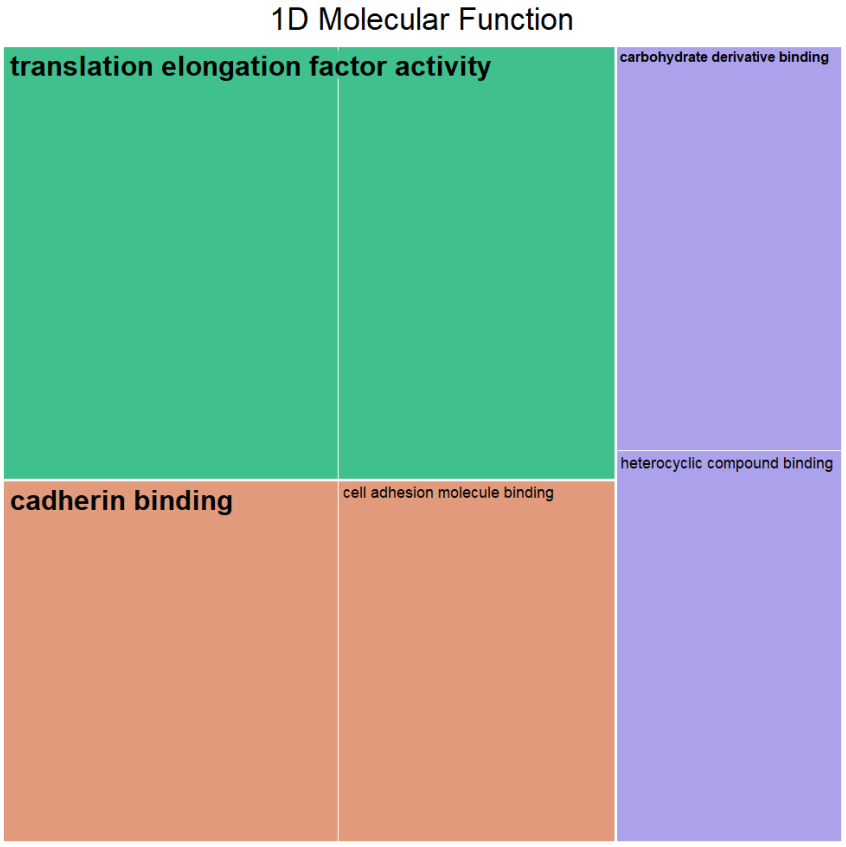
Figure 5:11: STRING analysis of significantly differentiated proteins between Pmel-M and Pmel-NM from the 2D data

Significantly differentiated proteins between Pmel-M and Pmel-NM from 2D discovery proteomics were inputted into STRING. A medium confidence score of 0.4 was allowed in the structure creation. Thickness of lines (edges) indicates confidence in association between two proteins. Total number of nodes is 15. Total number of edges is 8. The Reactome pathway enrichment has been overlaid onto the STRING structure. FDR, False discovery rate.

5.3.5 Gene ontology analysis

Gene ontology analysis of significantly differentially expressed proteins between Pmel-M and Pmel-NM from 1D proteomic data were reduced and visualised using ReViGO and this highlighted several enriched gene ontology terms (**Figure 5:12**). This included “cadherin binding”, “cell adhesion molecule binding” and “cell-cell adhesion”. Further enrichment in “protein metabolic processes”, “cellular response to growth factor stimulus” and “phagocytosis” was observed, in addition to “vesicle” and “extracellular region part”.


1D Biological Processes
gene silencing
retina homeostasis
Fc receptor signaling pathway
homotypic cell-cell adhesion
protein metabolic process
negative regulation of actin filament severing
cellular response to growth factor stimulus
phagocytosis, engulfment


1D Molecular Function
translation elongation factor activity
carbohydrate derivative binding
heterocyclic compound binding
cadherin binding
cell adhesion molecule binding

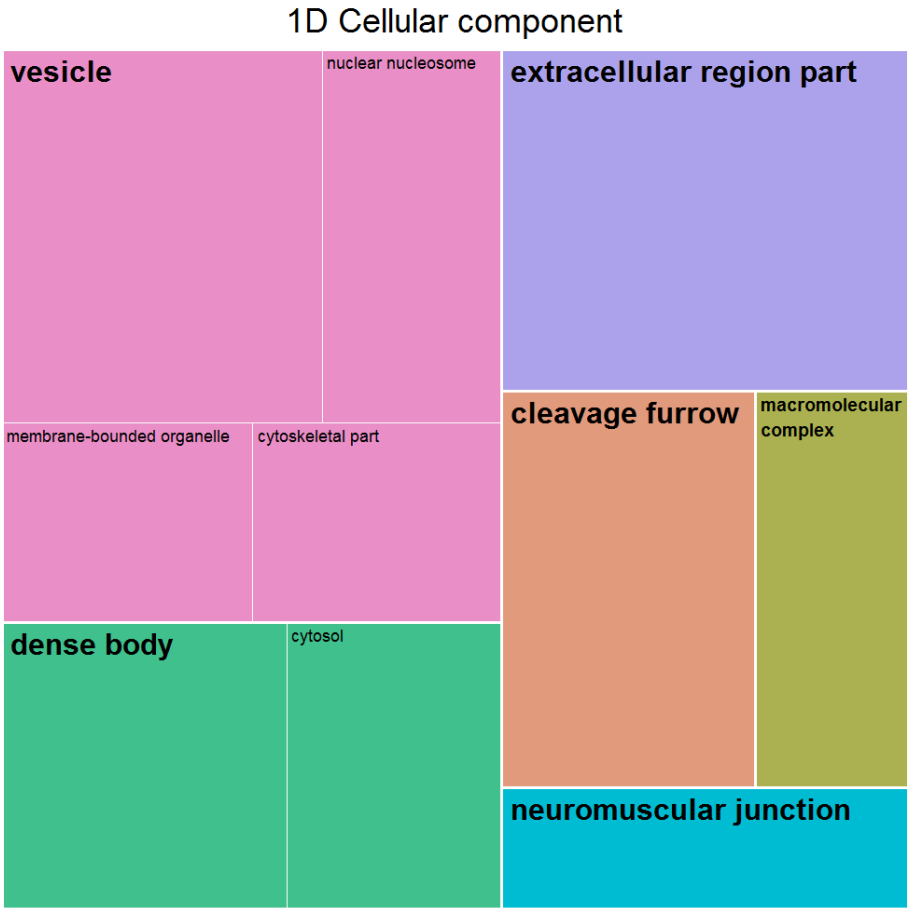
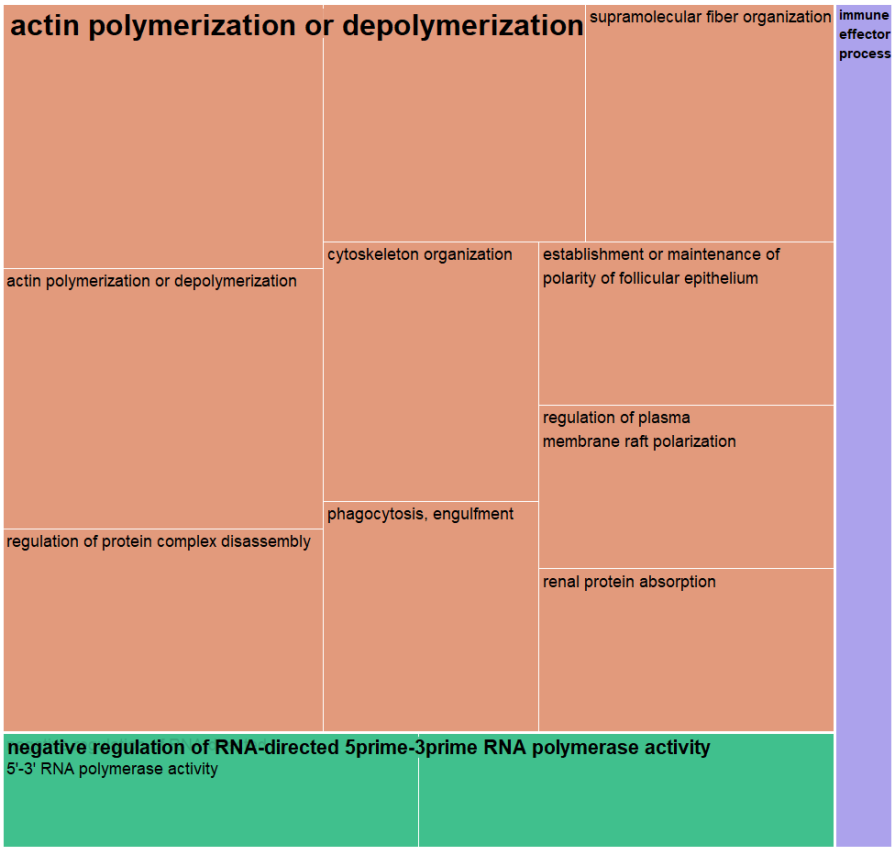


Figure 5:12: Gene ontology analysis of significantly expressed proteins between Pmel-M and Pmel-NM in 1D data.

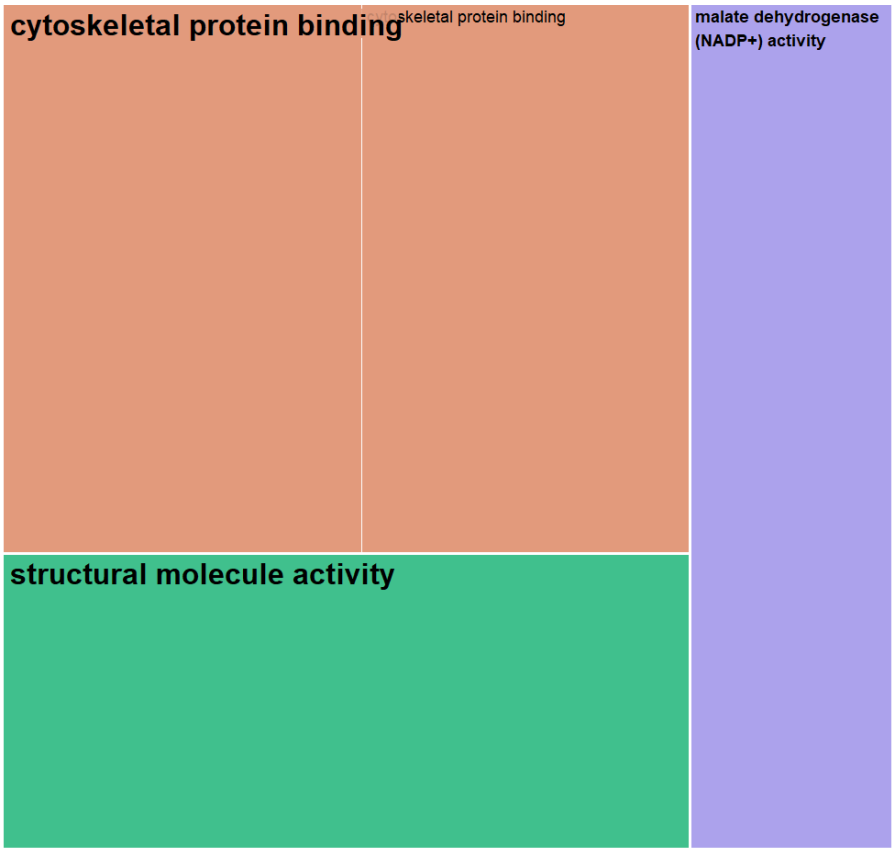
Significantly differentially expressed proteins were inputted into GoGorilla and reduced using ReViGO. The area of each of the rectangles or boxes is representative of the amount of enrichment of that gene ontology term.

Gene ontology analysis of significantly differentially expressed proteins between Pmel-M and Pmel-NM in the 2D data also revealed several area of enrichment (**Figure 5:13**). Similar to the 1D data, there appeared to be enrichment in binding including, "cytoskeletal binding" and "skeletal protein binding". Indeed, several areas related to cytoskeletal polymerisation and organisation. There was also enrichment in "extracellular exosome", "vesicle" and "focal adhesion".

2D Biological Processes



2D Molecular Function



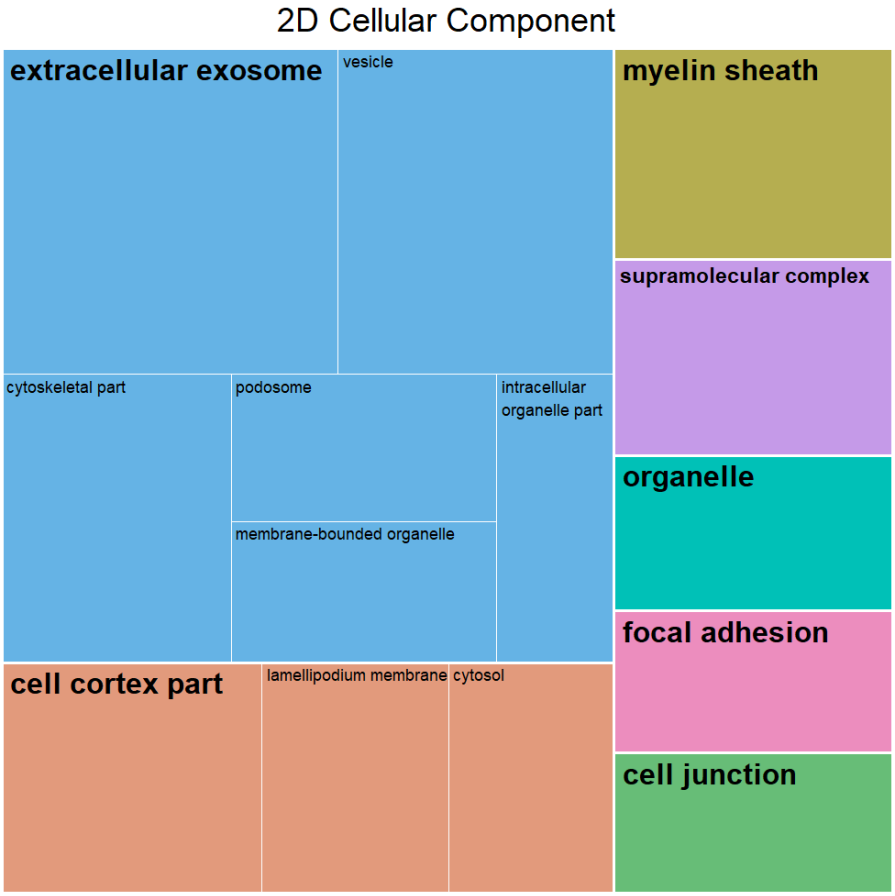


Figure 5:13: Gene ontology analysis of significantly differentially expressed proteins between Pmel-M and Pmel-NM in the 2D data.

Significantly differentially expressed proteins were analysed for GO enrichment using GoGorilla and reduced using REViGO. The area of each of the rectangles or boxes is representative of the amount of the enrichment of that gene ontology term.

5.3.6 Ingenuity pathway analysis

Ingenuity pathway analysis of significantly differentially expressed proteins in both the 1D and 2D melanoma proteomic data revealed no strong enrichment of pathways, therefore only whole 1D and 2D melanoma proteomic data was used. This 1D and 2D proteomic data revealed a number of significantly enriched pathways (**Figure 5:14**). As was the case in chapter 3, a combined p value cut off of <0.00001 (>5 log10 p value) was employed. EIF2 signalling was the most enriched pathway in both the 1D and 2D data, and was predicted to be activated in both cases. Multiple Rho signalling pathways were enriched in these data, however some were predicted to be activated, such as “RhoA signalling” and “signalling by Rho family GTPases”, whereas some were predicted to be inhibited, such as “RhoGDI

signalling" and "regulation of actin-based motility by Rho". "Integrin signalling" and "integrin-linked kinase (ILK) signalling" were both predicted to be activated. In addition, there was enrichment in several immune related pathways, including "acute phase response signalling", "leukocyte extravasation signalling", "granzyme B signalling" and "Fc receptor mediated phagocytosis in macrophages and monocytes". IPA is also able to predict upstream regulators based on the proteomic data provided, therefore this was also explored (**Figure 5:15**). The most significant upstream regulator identified in the 1D melanoma proteomic data was PCGEM1 and in the 2D proteomic data was IL15. The most inhibited upstream regulator was miR-122-5p and most activated was either IL15 or HIF1A. Several of the upstream regulators were only detected in one dataset (i.e. the 1D or 2D melanoma proteomic dataset), including PCGEM1, HSP90B1, CUL4B, SYVN1, SPDEF, EOMES, ERK1/2 and EGLN.

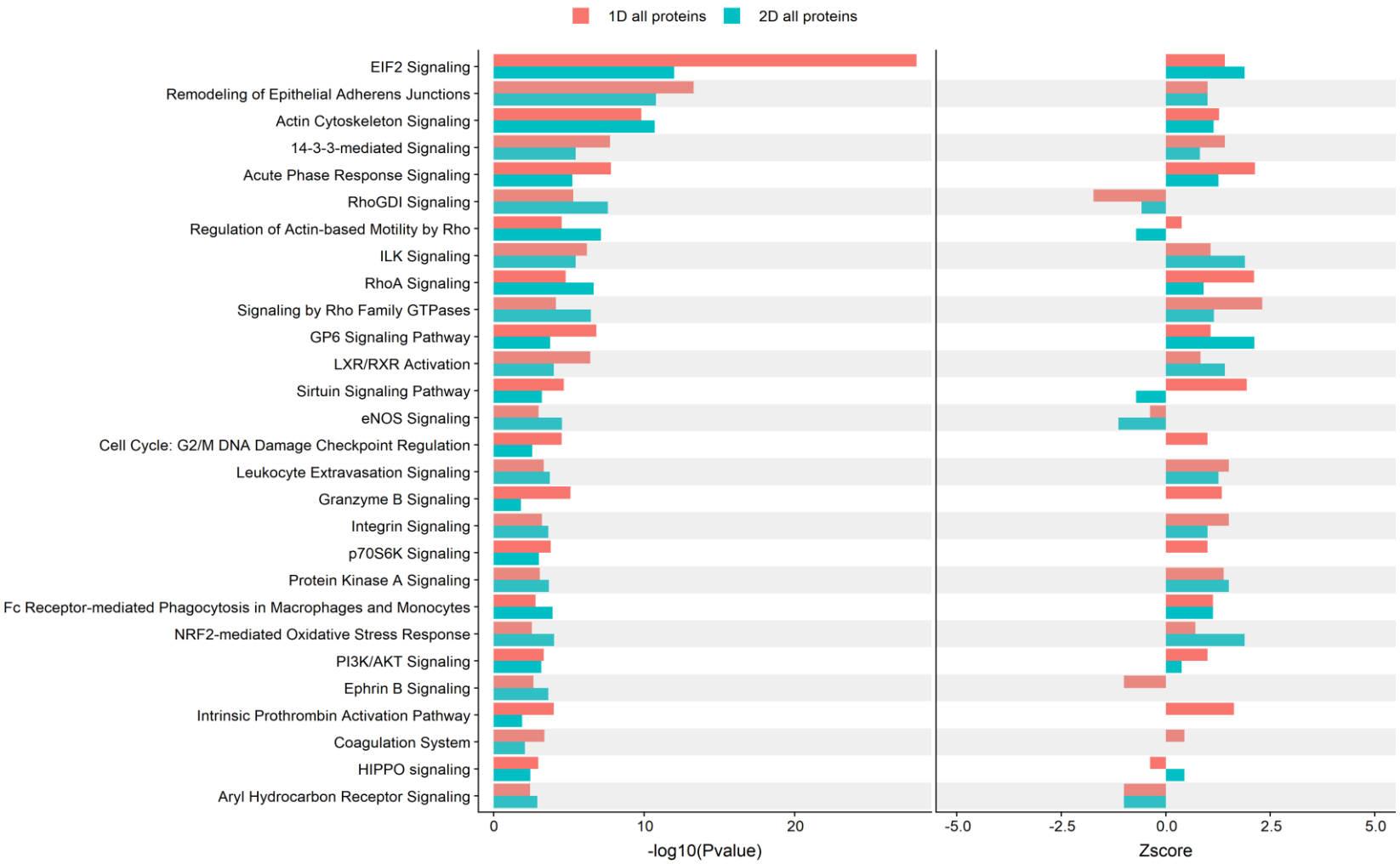


Figure 5:1 Ingenuity pathway analysis of melanoma proteomic data

Whole 1D and 2D melanoma proteomic data were analysed using IPA to identify pathway enrichment. Results were stacked and ranked by the sum of the log10 p values. Z score is the value given for the activation state where positive numbers represent an activation and negative numbers represent an inhibition. Only those pathways with an enrichment p value <0.00001 were included.

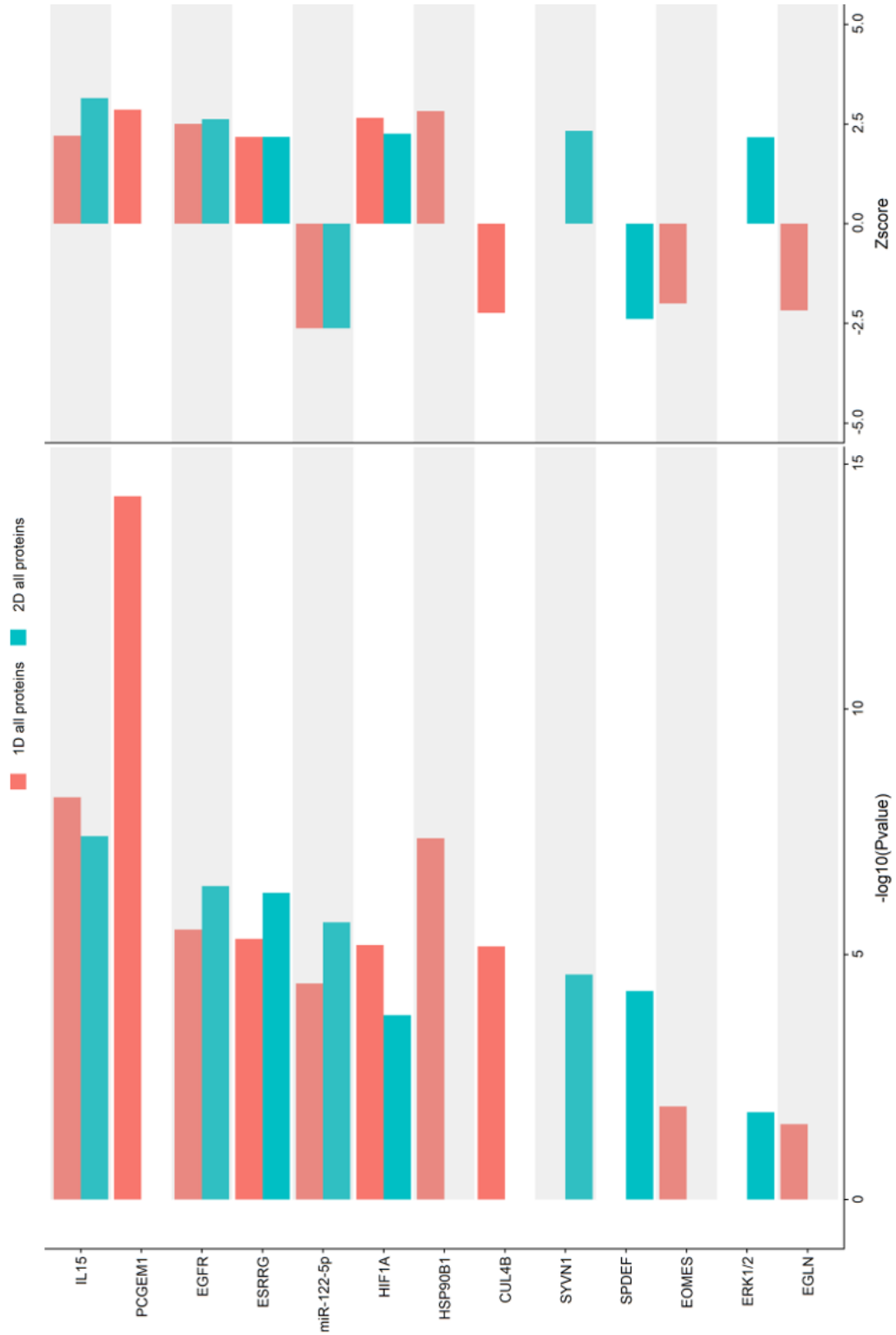


Figure 5:2 Upstream analysis of melanoma proteomic data using ingenuity pathway analysis

1D and 2D melanoma proteomic data sets were analysed for potential upstream regulators using IPA. Regulators were stacked and ranked by their sum log10 p value. Z score indicated activation or inhibition where positive numbers indicate activation and negative numbers indicate inhibition.

5.3.7 Weighted gene co-expression network analysis

The number of clinical and histological characteristics our laboratory had available for the melanoma samples were much less than that for cSCC in Chapter 3, and therefore this limited the number of results provided in the weighted gene co-expression network analysis. After pre-processing and network creation, there were 4 modules of proteins identified (**Figure 5:16**). No module correlated significantly with development of metastasis. The only characteristics which significantly correlated with protein modules were Clark's level and Breslow thickness. The brown, blue and turquoise modules correlated positively with Clark's level, of which the turquoise module gave the highest and most significant correlation. The brown, blue and turquoise modules also correlated positively with Breslow thickness. Blue and turquoise modules both positively correlated significantly with Breslow thickness at a high level of significance.

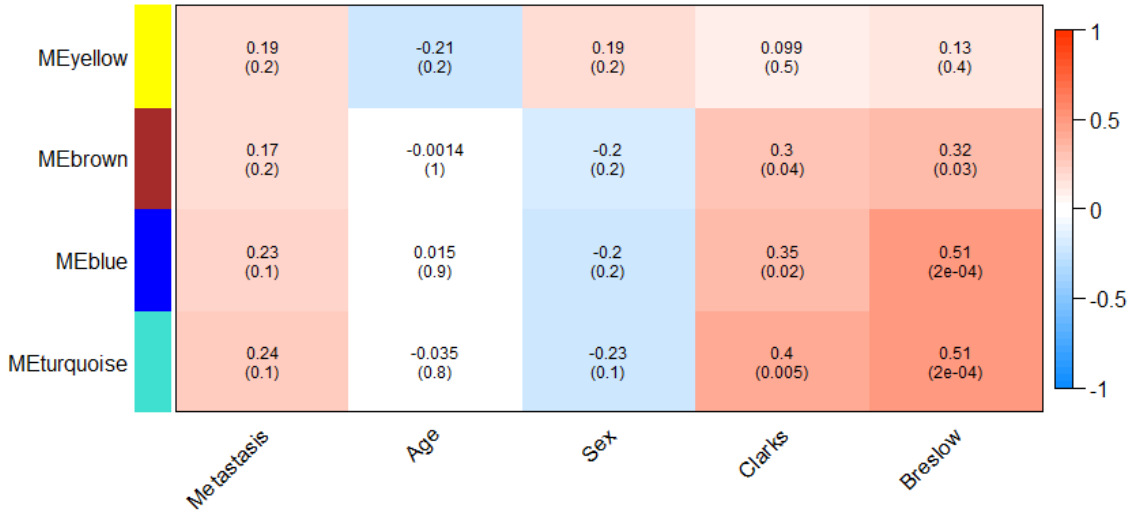


Figure 5:16: Module-trait correlation analysis, using WGCNA.

Proteomic data from the melanomas was used to create a network and identify modules of correlated clusters. Correlation values are from Pearson's correlation. In each box, the upper values are correlations and lower values in parentheses are p values. ME, module eigengene.

The modules identified from WGCNA can be input into various pathway analysis tools to better understand why proteins are correlated and what effect it might have on biological systems. The blue module displayed an enrichment in various protein production related pathways including ribosomal, translation and elongation pathways. Interestingly, the blue module was also enriched for the Reactome pathway "regulation of expression of SLITs and

ROBOs". The majority of enrichment within the turquoise module related to MAPK signalling, specifically BRAF. These included enrichment in "signalling by high-kinase activity BRAF mutants", "paradoxical activation of RAF signalling by kinase inactive BRAF2, "signalling by moderate kinase activity BRAF mutants", "MAP2K and MAPK activation", "p130cas linkage to MAPK signalling for integrins", and "signalling by BRAF and RAF fusions". Indeed, the turquoise module also expressed enrichment in various integrin related pathways, including those already mentioned relating to MAPK activation in addition to "focal adhesion" and "ECM-receptor interaction". Although there was some enrichment in the brown and yellow modules, there was little of great significance and coverage.

source	term name Protein databases (CORUM protein complexes)	term ID	n. of term genes	corrected p-value	TURQUOISE	YELLOW	BLUE	BROWN
cor	Ribosome, cytoplasmic	CORUM:306	80	3.95e-24				27
cor	40S ribosomal subunit, cytoplasmic	CORUM:338	31	4.15e-17				16
cor	Nop56p-associated pre-rRNA complex	CORUM:3055	104	9.73e-17				24
cor	40S ribosomal subunit, cytoplasmic	CORUM:305	33	1.56e-16				16
cor	60S ribosomal subunit, cytoplasmic	CORUM:308	47	5.95e-07				11
source	term name Biological pathways (KEGG)	term ID	n. of term genes	corrected p-value	TURQUOISE	YELLOW	BLUE	BROWN
keg	Focal adhesion	KEGG:04510	199	2.68e-05	15			
keg	Protein digestion and absorption	KEGG:04974	90	7.40e-04	9			
keg	Citrate cycle (TCA cycle)	KEGG:00020	30	3.11e-04				6
keg	Estrogen signaling pathway	KEGG:04915	136	5.91e-06		5		
keg	ECM-receptor interaction	KEGG:04512	82	3.55e-05	10			
keg	Ribosome	KEGG:03010	136	4.02e-22				27
keg	Carbon metabolism	KEGG:01200	117	5.68e-04				10
keg	Pathogenic Escherichia coli infection	KEGG:05130	55	1.04e-05	9			
source	term name Biological pathways (Reactome)	term ID	n. of term genes	corrected p-value	TURQUOISE	YELLOW	BLUE	BROWN
rea	Formation of tubulin folding intermediates by CCT/TriC	R-HSA-389960	26	1.59e-04				6
rea	Signaling by high-kinase activity BRAF mutants	R-HSA-6802948	34	9.22e-05	7			
rea	Post-translational protein phosphorylation	R-HSA-8957275	108	5.03e-08	14			
rea	Paradoxical activation of RAF signaling by kinase inactive BRAF	R-HSA-6802955	38	2.07e-04	7			
rea	Selenocysteine synthesis	R-HSA-2408557	94	1.95e-28				27
rea	Signaling by moderate kinase activity BRAF mutants	R-HSA-6802946	38	2.07e-04	7			
rea	MAP2K and MAPK activation	R-HSA-5674135	38	2.07e-04	7			
rea	Apoptotic cleavage of cell adhesion proteins	R-HSA-351906	11	1.23e-04		3		
rea	Extracellular matrix organization	R-HSA-1474244	296	2.34e-09	23			
rea	Activation of BAD and translocation to mitochondria	R-HSA-111447	15	3.80e-04	5			
rea	Eukaryotic Translation Elongation	R-HSA-156842	94	1.81e-33				30
rea	Protein methylation	R-HSA-8876725	17	4.12e-04				5
rea	Viral mRNA Translation	R-HSA-192823	90	5.08e-29				27
rea	Platelet degranulation	R-HSA-114608	127	2.01e-11	18			5
rea	Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex	R-HSA-75035	13	1.67e-04	5			
rea	Regulation of expression of SLITs and ROBOs	R-HSA-9010553	172	8.46e-25				30
rea	Innate Immune System	R-HSA-168249	1094	4.00e-08	42	27	11	
rea	p130Cas linkage to MAPK signaling for integrins	R-HSA-372708	15	3.80e-04	5			
rea	Signaling by BRAF and RAF fusions	R-HSA-6802952	58	3.21e-04	8			
rea	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	R-HSA-975956	96	8.56e-30				28
rea	GRB2:SOS provides linkage to MAPK signaling for Integrins	R-HSA-354194	15	3.80e-04	5			
rea	Smooth Muscle Contraction	R-HSA-445355	35	5.10e-06	8			

Figure 5:17: WGCNA module pathway analysis of the melanoma proteomic data. Modules identified in WGCNA were subjected to pathway analysis and CORUM database alignment. Strong hierarchical filtering was employed to reduce the number of terms and optimise interpretation, therefore only results with P<0.001 are shown.

5.3.8 Topological Data analysis

Similar to WGCNA, TDA is a type of unsupervised machine learning which uses all the data available, in this case proteomic data, to establish trends and correlate sample proteomes. Whole sample proteomes from the 1D and 2D melanoma proteomic data were inputted into Ayasdi workbench to create representative TDA structures (**Figure 5:18**). Both the Pmel-M and the Pmel-NM groups appeared to have high homology between nodes as there was relatively high interconnectivity between them. Nonetheless, the interconnectivity between nodes of the Pmel-M group was very high and therefore suggested that the Pmel-M samples were, by and large, very similar. Although proteomic data was able to effectively separate Pmel-M from Pmel-NM samples using TDA, there appeared to be different subgroups within the Pmel-NM group. This potential for a subgroup within the Pmel-M group was seen more clearly in the 2D data compared to the 1D data but was visible in both sets of data. Due to this, it was decided to recreate the structures but with the intention of driving these groups to separate to allow comparison between them (**Figure 5:19**). Driving the separation of these groups revealed several proteins which, if these subgroups of Pmel-M and Pmel-NM are correct (i.e. exist in biological terms), might be “driving proteins” (i.e. proteins which are likely causative in generating the biological subgroups). Splitting the groups in the 1D data produced four obvious groups, two for Pmel-M and two for Pmel-NM. There were 3 and 15 driving proteins within the Pmel-NM tumours and within the Pmel-M tumours respectively. Several driving proteins were identified as different between M1 and NM1/NM2, which is to be expected as these are essentially subgroups of Pmel-M vs Pmel-NM. Surprisingly, however, M2 vs NM1/NM2 only highlighted a few driving proteins that differed between them. Conversely, the 2D melanoma proteomic data produced only one Pmel-M group and one Pmel-NM group along with one mixed group which couldn’t be defined as completely Pmel-M or Pmel-NM. An interesting driver factor identified between the mixed group and the NM group was age. A table of all driving factors can be seen in **Table 5:5**.

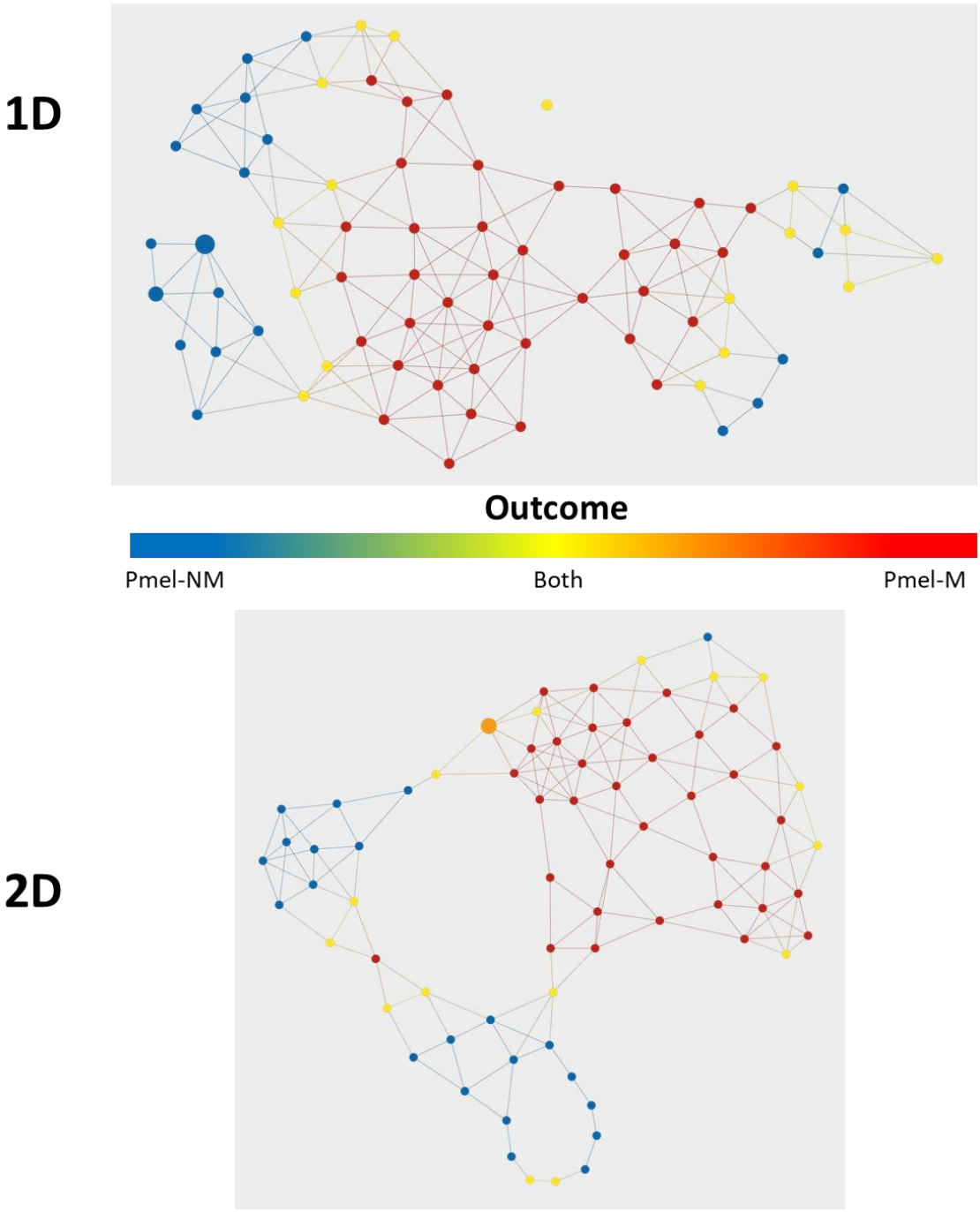


Figure 5:18: Topological structures created from 1D and 2D melanoma proteomic data. A hamming metric and two neighbourhood lenses were used in Ayasdi to create topological structures. Outcome was colour mapped on top of these structures to illustrate how the structure relates to outcome. Blue nodes represent Pmel-NM, red represents Pmel-M and yellow indicate nodes containing both Pmel-NM and Pmel-M.

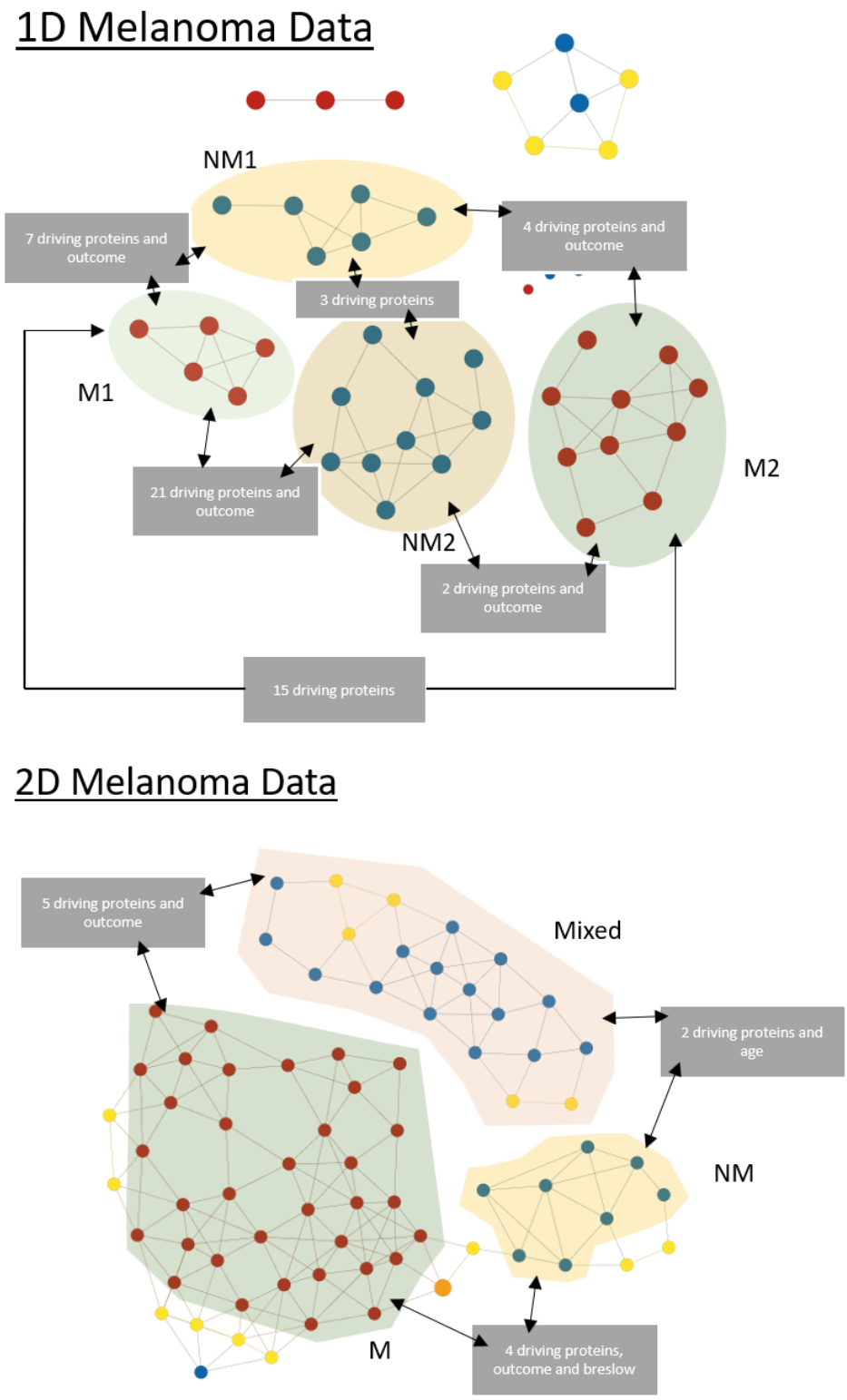


Figure 5:19: Driving separation of TDA identified melanoma sub-groups.

Groups were driven apart by modifying resolution and gain of the neighbourhood lenses until separation was achieved. Kolmogorov Smirnov test of fitness was used to determine potential driver proteins ($P<0.05$). Upper topological structure generated from1D data, lower topological structure from 2D data.

Table 5.5: Proteins identified as possible driver proteins in TDA subgroups of melanoma

1D

M1 vs NM1		M1 vs NM2		M2 vs NM2		M2 vs NM1		M1 vs M2		NM1 vs NM2	
Factor	P-value	Factor	P-value	Factor	P-value	Factor	P-value	Factor	P-value	Factor	P-value
Outcome	0.008	Outcome	0.004	Outcome	4.12E-04	Outcome	0.001	P02790	0.004	P42224	0.035
P62917	0.035	P26373	0.012	P37837	0.035	P02647	0.022	P15153	0.012	P25786	0.047
P62269	0.035	Q15365	0.022	P21333	0.044	P42224	0.030	O00299	0.030	Q7KZF4	0.047
Q14764	0.035	Q9Y490	0.022			P61313	0.047	P22314	0.032		
O43175	0.035	P17987	0.025			P31948	0.047	P00338	0.032		
Q15063	0.044	P37802	0.035			P23381	0.065	P16403	0.032		
P42224	0.047	P62269	0.035					P02768	0.032		
P09382	0.047	P11940	0.035					P26373	0.033		
		P68104	0.038					P37837	0.033		
		Q8NBS9	0.038					P02647	0.035		
		P62805	0.038					P17987	0.035		
		P63261	0.038					P12956	0.038		
		P13639	0.038					P04899	0.040		
		P02790	0.044					Q07065	0.044		
		P25786	0.047					Q9BVC6	0.047		
		P49189	0.047								
		O00299	0.047								
		P46776	0.047								
		Q7L7L0	0.047								
		P17931	0.047								
		P26641	0.047								
		P23246	0.047								

2D

M vs Mixed		Mixed vs NM		M vs NM	
Factor	P-value	Factor	P-value	Factor	P-value
Outcome	8.24E-06	P50995	0.013	Outcome	5.08E-05
Q8N1N4	0.008	O00299	0.030	O43390	0.014
Q13813	0.010	Age	0.042	O00299	0.017
P17931	0.021	P19012	0.049	P50995	0.019
P19012	0.027			P61978	0.021
P41219	0.033			Breslow	0.035

5.3.9 MRM analysis

The number of proteins that were identified as differentially expressed between Pmel-M and Pmel-NM was much lower than that seen as differentially expressed between P-M cSCCs and P-NM cSCCs, however, similar to the cSCC portion of this project, MRM was employed to see whether it would validate the discovery proteomic results of the melanoma data. As KRT9 was the only significantly differentially expressed protein to appear in both the 1D and 2D discovery proteomics of the melanomas, it seemed suitable to try to verify this by MRM. As our laboratory already had GSN heavy labelled peptides, GSN was also chosen for MRM verification. The prospect of doing machine learning to identify a third target differentially expressed protein in the melanomas was considered,

but decided against because the main reason for doing it on the cSCC data was to reduce the number of options from which to choose proteins for MRM verification / validation in that project. In the case of melanoma, there were fewer options, and so LMNB1 was chosen as the third candidate due to its high fold change, low p value, novelty and credibility in terms of likely biological influence.

5.3.9.1 Selecting suitable peptides from proteins of interest for Multiple Reaction Monitoring (MRM) analysis

To identify suitable unique peptides for the proteins of interest, the data were imported into skyline and matched to a melanoma spectral library created from the melanoma proteomic discovery data (**Figure 5:20**). All 3 of the GSN peptides held in our laboratory were identified in the melanoma spectral library. Peptides for KRT9 and LMNB1 were selected based on being unique to their respective proteins and having a high spectral count with transitions (fragment ions) at a relatively high intensity. Examples of the spectral library matches for these peptides, including their transitions, are shown in figure 5.20. The final peptides selected, along with their m/z and suitable transitions can be seen in **Table 5.6**.

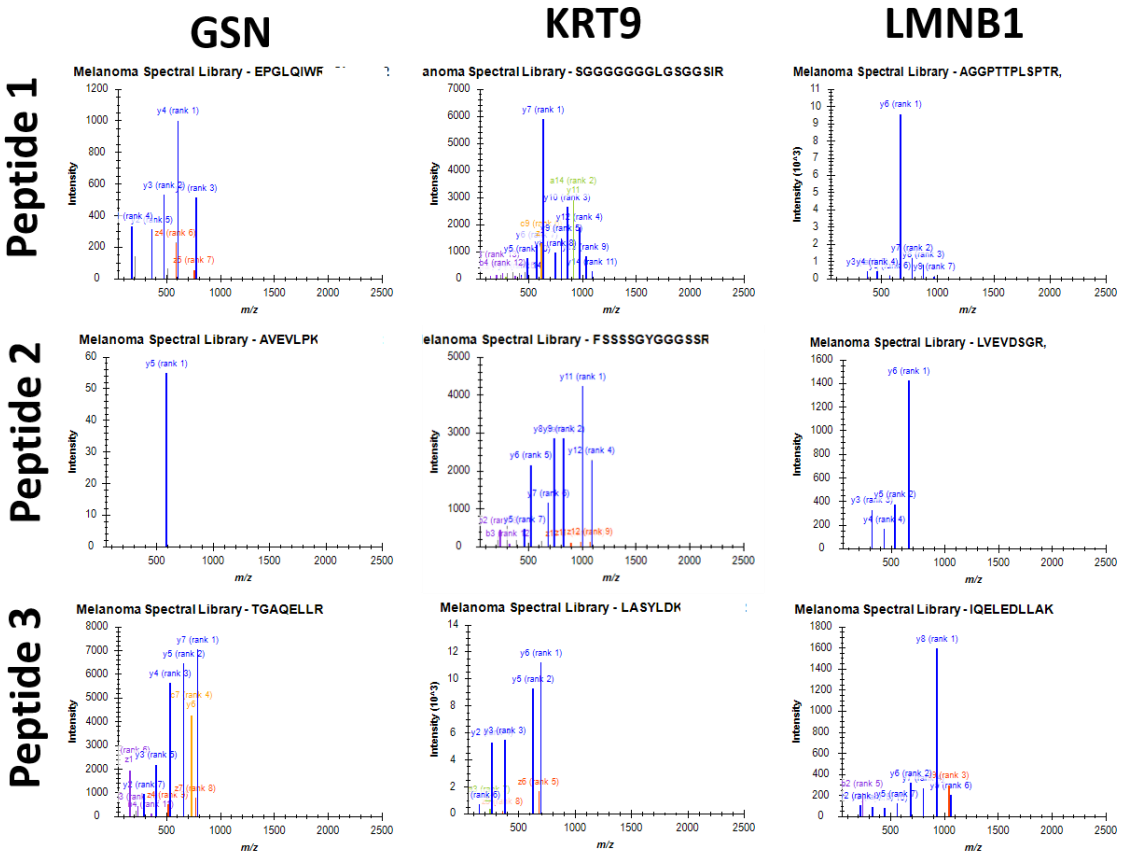


Figure 5:20: Spectral library matching of GSN, KRT9 and LMNB1 in the melanoma samples

The three GSN peptides already present in our laboratory appeared in the melanoma spectral library, however, peptide two displayed only one viable transition ion and thus highlighted a potential weakness of this peptide. 3 peptides for each KRT9 and LMNB1 were selected based on their spectral count (i.e. the number of the discovery melanoma samples they appeared in) and the intensities of the transition ions. Peptide sequence given at the top of each spectra.

Table 5.6: The unique peptides selected for each protein of interest with their m/z and transition ions

PROTEIN	PEPTIDE	M/Z AT CHARGE 2	TRANSITION IONS	MODIFIED PEPTIDE SEQUENCE	M/Z AT CHARGE 2	TRANSITION IONS
GSN	EPGLQIWR	499.7754	772.4464, 602.3409, 474.2823, 175.1190	EPGLQIWR[$^{13}C_6$, $^{15}N_4$]	504.7787	782.4547, 612.3492, 484.2906, 185.1272
	AVEVLPK	378.2367	585.3606	AVEVLPK[$^{13}C_6$, $^{15}N_2$]	382.2438	593.3748
	TGAQELLR	444.2509	786.4468, 729.4254, 658.3882, 530.3297	TGAQELLR[$^{13}C_6$, $^{15}N_4$]	449.255	796.4551, 739.4336, 668.3965, 540.3379
KRT9	SGGGGGGGLGSGGSIR	616.8025	974.5014, 917.4799, 860.4585, 633.3315	SGGGGGGGLGSGGSIR[$^{13}C_6$, $^{15}N_4$]	621.8067	984.5097, 927.4882, 870.4667, 643.3397
	FSSSSGYGGGSSR	618.268	1088.4603, 1001.4283, 827.3642, 740.3322	FSSSSGYGGGSSR[$^{13}C_6$, $^{15}N_4$]	623.2721	1098.4686, 1011.4365, 837.3725, 750.3405
	LASYLDK	405.2238	696.3563, 625.3192, 375.2238, 262.1397	LASYLDK[$^{13}C_6$, $^{15}N_2$]	409.2309	704.3705, 633.3334, 383.2380, 270.1539
LMNB1	AGGPTTPLSPTR	577.8118	872.4863, 771.4359, 670.3883, 460.2514	AGGPTTPLSPTR[$^{13}C_6$, $^{15}N_4$]	582.816	882.4919, 781.4442, 680.3965, 470.2597
	LVEVDSGR	437.7351	662.3104, 533.2678, 434.1994, 319.1724	LVEVDSGR[$^{13}C_6$, $^{15}N_4$]	442.7392	672.3187, 543.2761, 444.2077, 329.1807
	IQELEDLLAK	586.3321	930.5142, 801.4716, 688.3876, 242.1499	IQELEDLLAK[$^{13}C_6$, $^{15}N_2$]	590.3392	938.5284, 809.4858, 696.4018, 242.1499

5.3.9.2 MRM peptide calibration curves

Peptides were synthesised and isotopically heavy labelled by Cambridge Research Biochemicals. Each peptide was subsequently tested on a Synapt G2-Si mass spectrometer at a concentration of 100fmol to establish their suitability (**Figure 5:21**). Peptide 3 of KRT9 proved inadequate as a corresponding peak could not be reliably selected. Peptides 1 and 2 for proteins KRT9 and LMNB1 were found outside of their predicted retention time window (blue shading), however, these were the only logical peaks in the chromatogram and thus these peptides were included in the subsequent MRM experiments. Nonetheless, peptide 1 of KRT9 and peptide 2 of GSN had slightly un-uniformed peaks (tailing on either side) but were included (with some caution) in the subsequent MRM investigations.

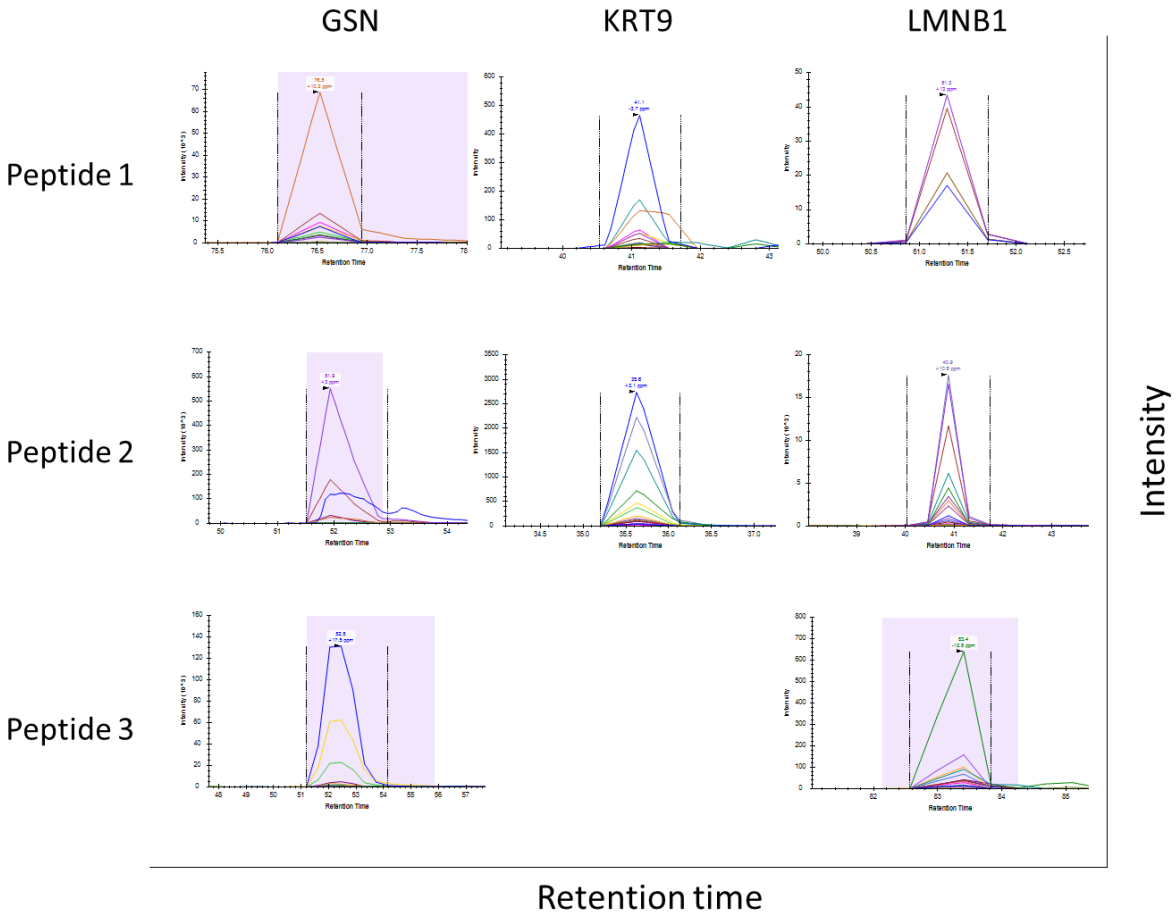


Figure 5:21: Chromatography of MRM peptides

100fmol of each peptide was investigated on a Synapt G2-Si mass spectrometer in targeted acquisition mode and results imported into skyline. Peptide 3 of KRT9 produced unreliable peaks which would have led to future confusion over the correct peak in tumour samples and was therefore omitted from future experiments. Peaks from peptides 1 and 2 from KRT9 and LMNB1 were found outside of the predicted retention time window (depicted as shaded area in above graphs) but these appeared to be the only logical peaks and these peptides were therefore included in subsequent experiments on melanoma. The peak shape of peptide 1, KRT9 appeared suboptimal (i.e. non-gaussian) but this peptide was used in subsequent MRM experiments on melanoma samples in order to have more than one peptide for KRT9 in those investigations. Peptide 2, GSN peak was also imperfect as it showed tailing with some transitions not having distinct peaks, but was used with caution in subsequent melanoma MRM experiments.

Once suitability of the relevant peptides had been established as above, calibration curves of each peptide were created to enable calculation of the amount of heavy and subsequently the corresponding light peptides, in each melanoma sample. Each calibration

curve for GSN peptides produced an R^2 value above 0.9 and suggested a good linear trend for comparison of added amount of heavy peptide in the sample with the MS estimated amount of heavy peptide in the sample (**Figure 5:22**). However, the light peptide (background melanoma matrix), used as an internal melanoma standard, was detected at a very low intensity.

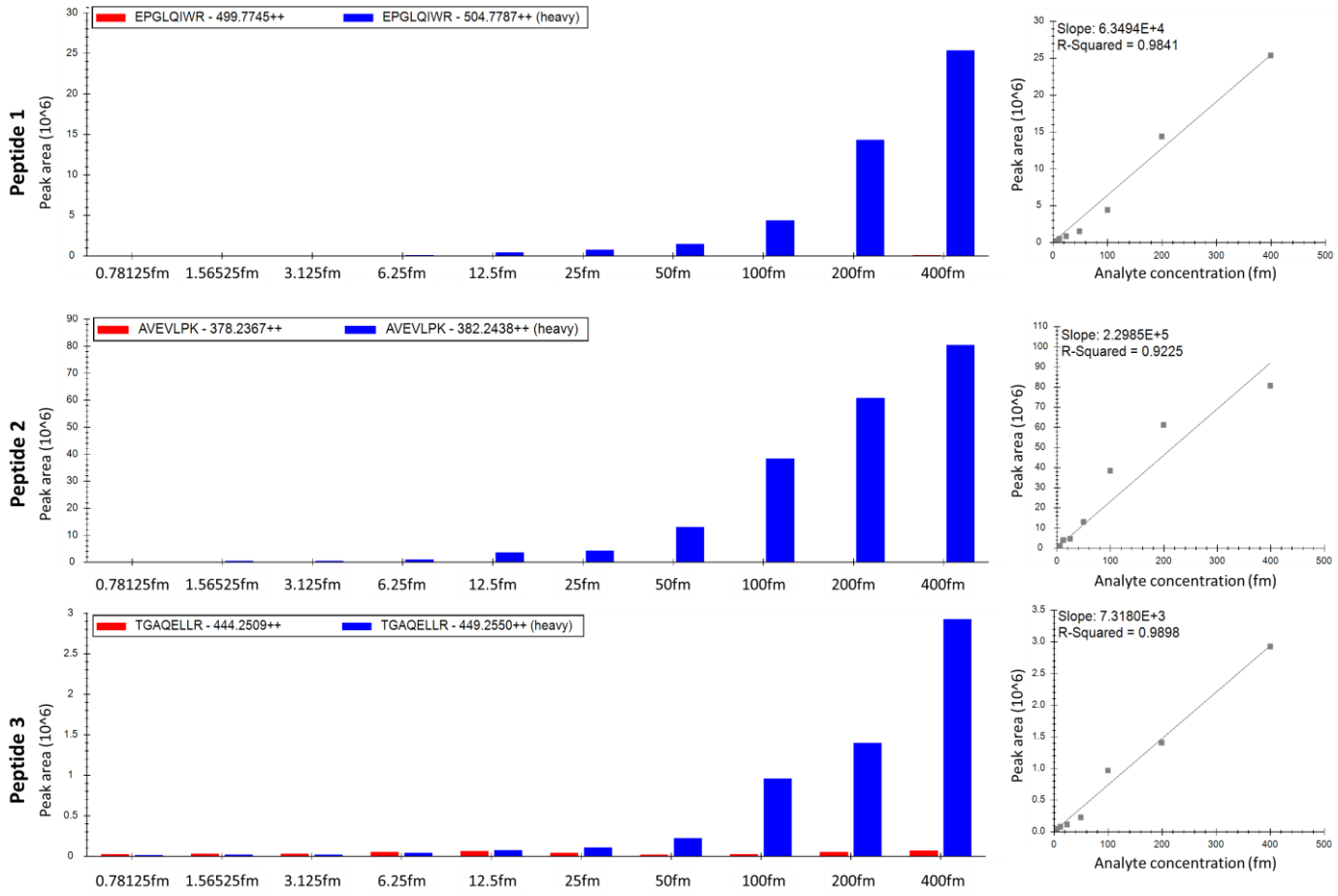


Figure 5:22: MRM calibration curves for GSN.

Calibration data of the heavy labelled GSN peptides was created using MRM on a twofold dilution series of the amount of relevant heavy peptide. 1µg of melanoma peptide sample was used as a background matrix (red).

Calibration curves for peptides 1 and 2 of KRT9 also produced R^2 values above 0.9 with an acceptable linear trend for MS estimated amount versus added amount of heavy peptide (**Figure 5:23**). Although the internal melanoma standard was present at a constant concentration, it appeared as though it might be increasingly detected as the amount of heavy labelled peptide increased.

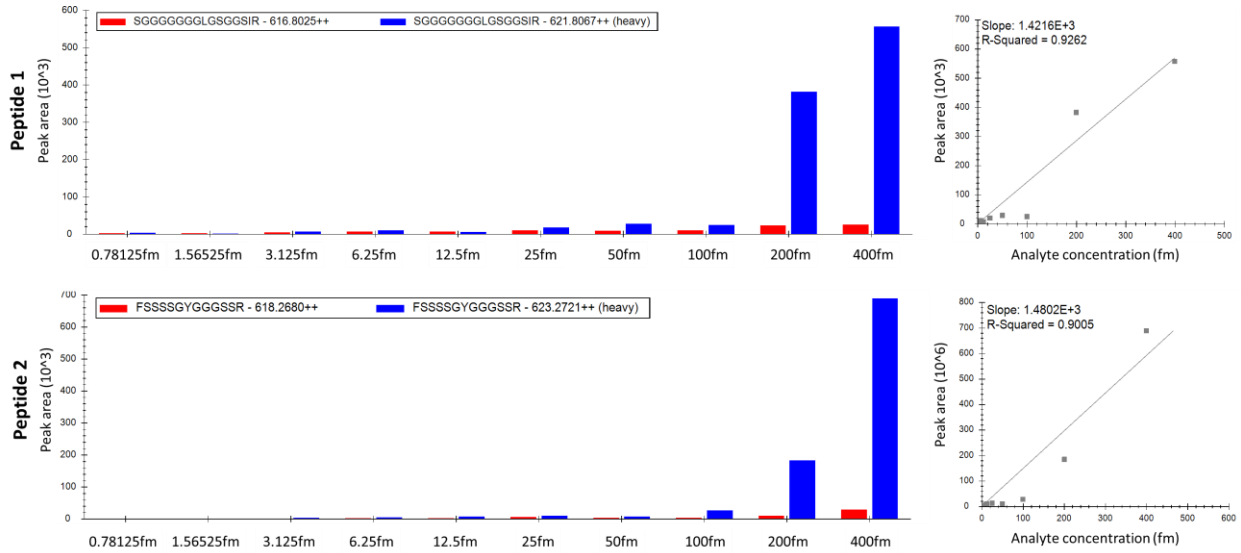


Figure 5:23: Calibration curves for KRT9

Calibration data of heavy labelled peptides for KRT was created using MRM on a twofold dilution series of the amount of relevant KRT9 heavy peptide. 1µg of melanoma peptide sample was used as internal standard (red).

Calibration curves for peptides 1 and 3 of LMNB1 produced R^2 values above 0.9, however peptide 2 did not and therefore could not accurately determine the true amount of heavy peptide in a sample and, subsequently, the level of the native peptide within a sample (**Figure 5:24**).

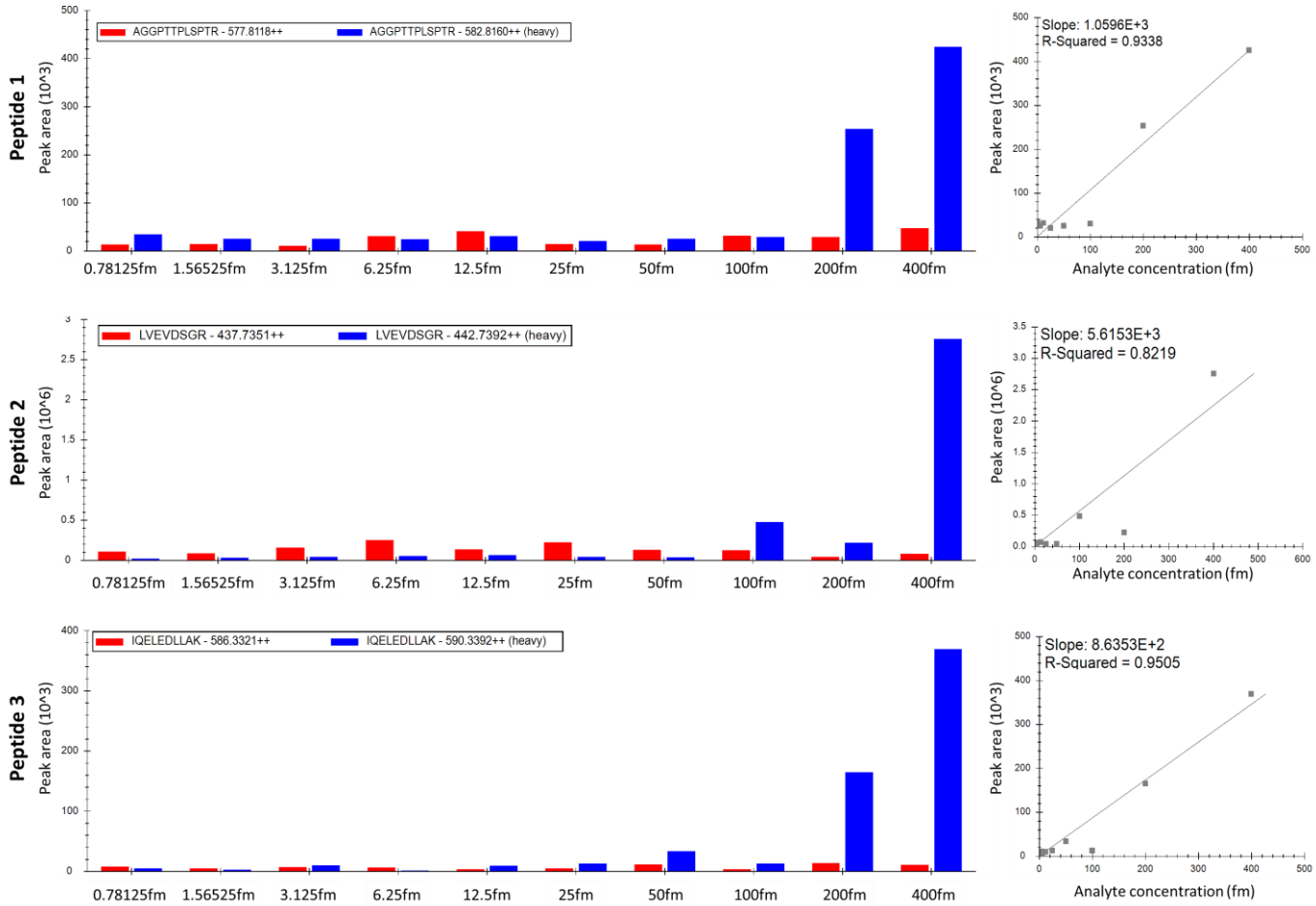


Figure 5:24: Calibration curves for LMNB1

Calibration data of heavy labelled peptides for LMNB1 was created using MRM on a twofold dilution series of the amount of relevant heavy peptide. 1µg of melanoma peptide sample was used as internal standard (red).

5.3.9.3 Verification

Following the production of the heavy peptide calibration curves as above, each melanoma sample was investigated with MRM with the addition of 100fmol of each heavy labelled peptide. The calibration curve was then used to determine the corrected concentration of heavy peptide in a sample which was subsequently used, with the heavy to light ratio, to find out the amount of light peptide in each sample. However, MRM analysis of the Pmel-M and Pmel-NM samples demonstrated no significant difference in the amount of any of these peptides between the Pmel-M and Pmel-NM groups.

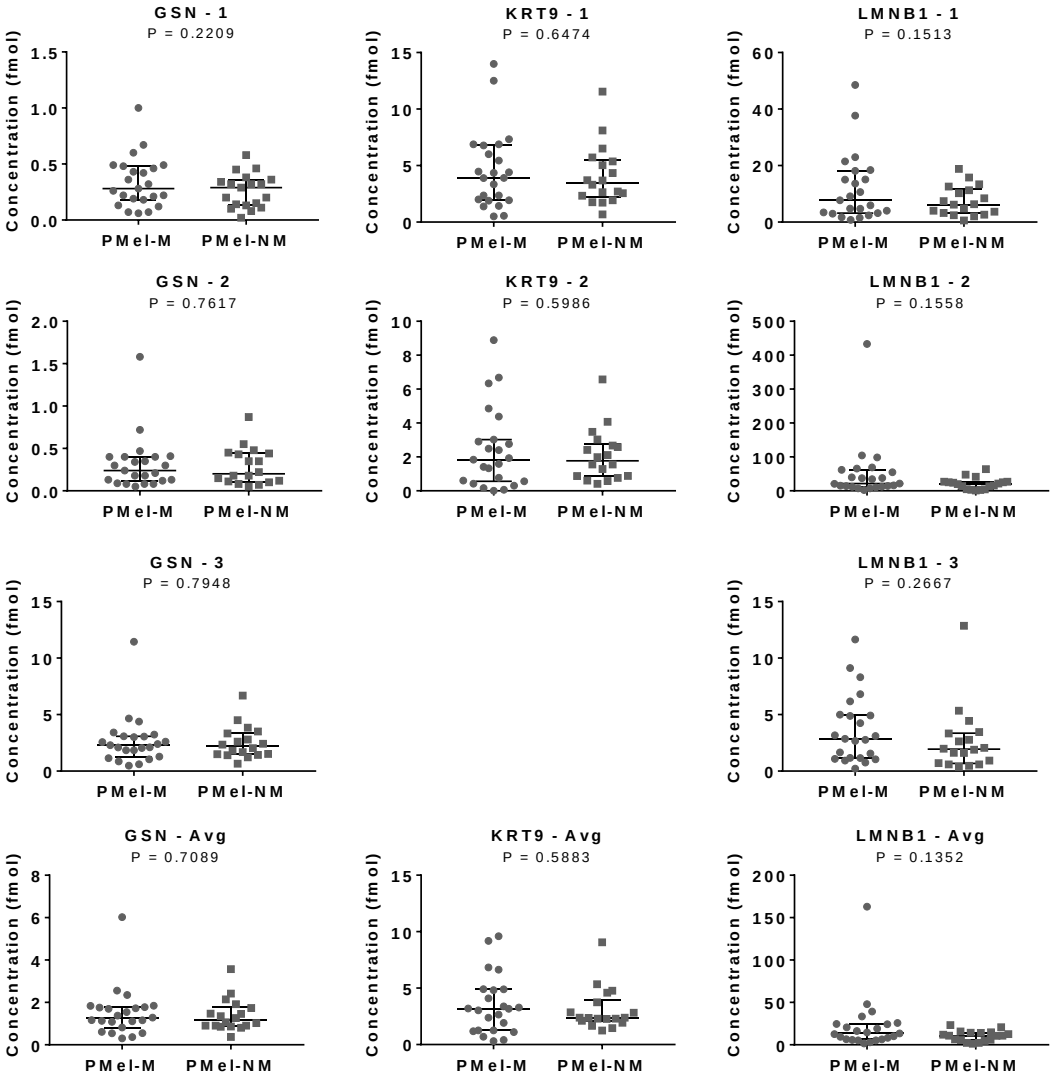


Figure 5:25: MRM verification data between Pmel-M and Pmel-NM.

Each melanoma sample was investigated with the addition of 100fmol of the relevant heavy labelled peptide. The "true" heavy peptide concentration was calculated using the previously determined calibration curves for the heavy peptides. The true heavy concentration was then used with the heavy:light ratio to calculate the concentration and amount of the light peptide in the sample. The averages (Avg) of the peptides for the different proteins were calculated using the mean. Error bars are interquartile range with median plotted on. Mann Whitney U test for significance.

5.3.10 cSCC-melanoma proteome comparison

As cSCC and melanoma arise from the same tissue, albeit from different cells within skin, comparing the discovery proteomic data from the experiments in this thesis might provide some insight into the similarities and differences between these two malignancies in relation to the development of metastases from this tissue. One simple method of doing this is to compare the proteomes of the primary samples which subsequently metastasised (P-M cSCCs and Pmel-Ms), primary samples which did not metastasise (P-NM cSCCs and Pmel-NMs) and the significantly differentially expressed proteins that differed between the P-M cSCCs and P-NM cSCCs and between the Pmel-Ms and Pmel-NMs. There were fewer melanoma proteins in every comparison against cSCC. For example, there were 265 P-M cSCC specific proteins, 370 shared proteins between P-M and Pmel-Ms, and 55 Pmel-M specific proteins. Likewise, 190 cSCC specific proteins were found in P-NM cSCC samples, 316 shared between P-NM cSCCs and Pmel-NMs, and 73 Pmel-NM specific proteins. Interestingly, the large majority of proteins were identified in both the primary tumours (cSCC and melanoma) which subsequently metastasised and the primary tumours (cSCC and melanoma) which did not metastasise. However, only 136 proteins were specific to tumours which metastasised and 25 proteins specific to tumours which did not metastasise. Comparison of the entire SCC proteomic data compared to the entire melanoma proteomic data revealed a number of unique IDs found only in one set of samples (cSCC or melanoma). Furthermore, in cases where the analysis was restricted to proteins present in at least 50% of the samples, 210 of these proteins were found only in SCC and 57 were found exclusively in melanoma. There were 8 proteins identified as being significantly differentially expressed between primary tumours which metastasised and primary tumours which had not metastasised in both melanoma and cSCC; these included EEF2, SEPT2, POSTN, HISTH4, EEF1A1, GSN, RSP16 and LMNB1.

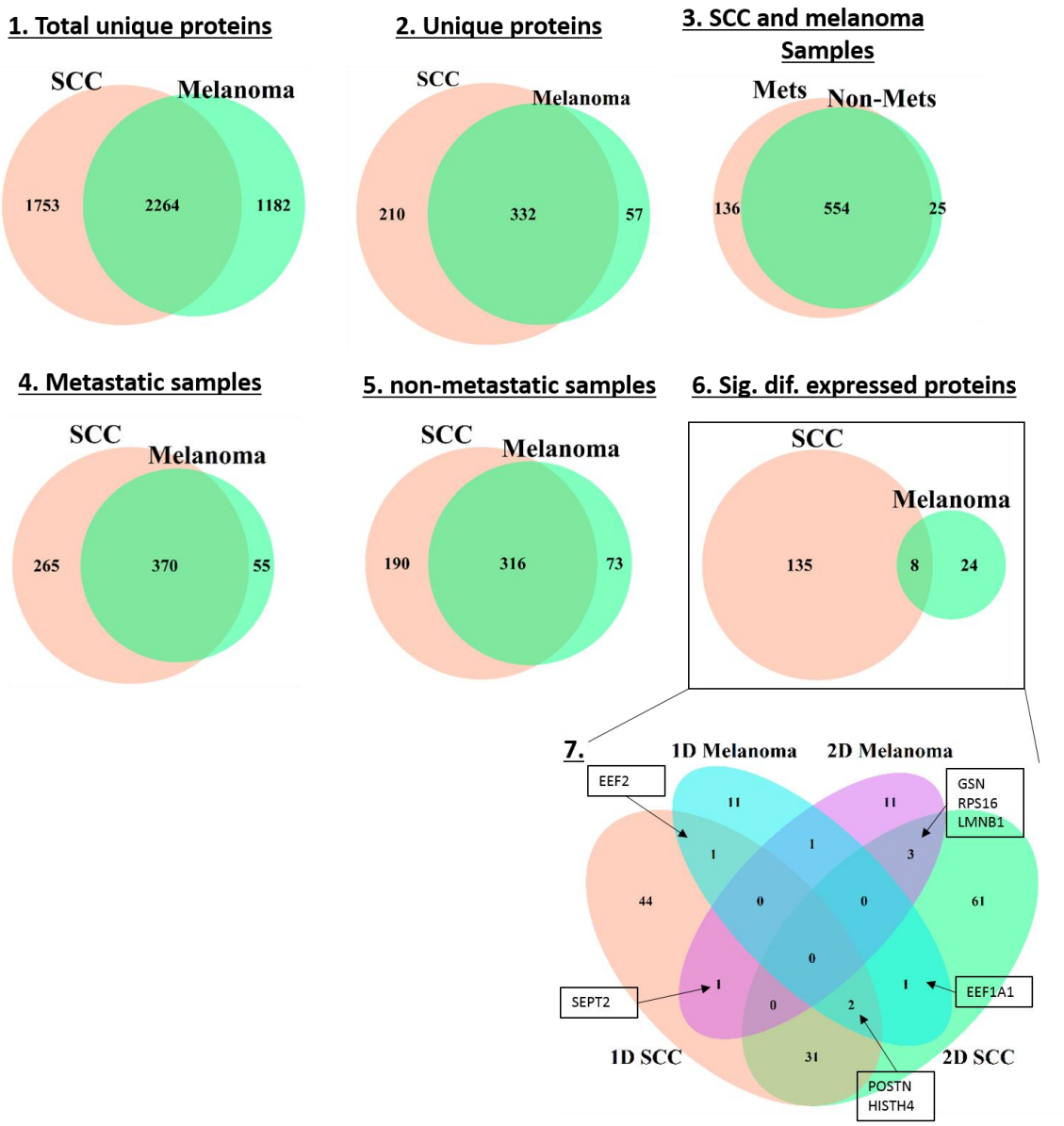


Figure 5:26: Comparison of the cSCC and melanoma discovery proteomics data.

cSCC and melanoma discovery proteomic data, generated by LC-MS, were compared to identify similarities and differences between these two tumours. Venn diagrams were created in R using package "Venndiagram". **1.** Total number of unique proteins identified in SCC and melanoma were compared using no missing value filter (as is the case for all other analyses which utilised a 50% missing value filter, where each protein was identified in at least 50% of samples). **2.** Unique protein IDs (from proteins which appeared in at least 50% of samples) from SCC and melanoma were compared. **3.** IDs from both SCC and melanoma primary tumours which had subsequently metastasised were compared to IDs from both SCC and melanoma primary tumours which had not metastasised. **4.** IDs from SCC primary tumours which subsequently metastasised were compared to IDs from melanoma primary tumours which subsequently metastasised. **5.** IDs from SCC primary

tumours which had not metastasised were compared to IDs from melanoma primary tumours which had not metastasised. **6.** Significantly differentially expressed proteins between P-M and P-NM in SCC were compared to significantly differentially express proteins between Pmel-M and Pmel-NM in melanoma. **7.** Breakdown of **6**, comparing significantly differentially expressed proteins identified in 1D and 2D data from SCC and melanoma.

5.4 Discussion

The aim of this chapter was to identify potential prognostic markers for subsequent development metastasis in primary melanoma. There are currently very few prognostic biomarkers available that can complement staging systems of melanoma or are helpful when used on their own (Weiss et al., 2015). Many other studies have aspired to identify prognostic biomarkers in melanoma, including some which used FFPE melanoma samples (Byrum et al., 2011, Byrum et al., 2013), however, a lot of these studies focussed on differences between primary melanomas and metastatic deposits from this cancer. In the current study, we set out to identify prognostic markers from primary tumours which had metastasised and primary tumours which had not metastasised. Furthermore, as Breslow depth is a known prognostic marker, samples were stratified for this feature in order to ensure there was no significant difference between the Breslow thicknesses of the Pmel-M and Pmel-NM groups.

Although identical methods were used for the proteomic investigations of cSCCs and melanomas, **Figure 5:26** highlights that there were fewer proteins identified in melanoma samples compared to cSCC samples. Indeed, although melanoma and cSCC are skin cancers, throughout this project the proteomic analysis of melanoma samples proved more challenging than that for cSCCs. Fewer proteins were identified in melanoma, and a lower number of these were significantly differentially expressed between primary tumours which metastasised and primary tumours which had not metastasised. Moreover, MRM analysis failed to validate the initial discovery findings in melanoma, in part due to difficulties faced in the MRM method development and seemingly low native peptide concentration. It is possible that the difficulties encountered with the melanoma samples could have arisen from the melanin present within these melanomas. Melanins are polymers produced in a process known as melanogenesis, where the amino acid tyrosine

is oxidised and undergoes a series of changes to become predominantly either eumelanin or pheomelanin (Prota, 2000). Furthermore, melanin is renowned for its "sticky" properties and ability to bind to other substances, including proteins (Mani et al., 2001). In fact, it has been reported that melanoproteins (melanin bound to proteins) are better scavengers of UV-induced free radicals (Pascutti and Ito, 1992) and thus melanin may have a biological need to bind to proteins. Furthermore, the pH environment surrounding melanin has been found to have an important effect on its binding capabilities and, indeed, more acidic environments induce more melanin polymerisation in addition to a stronger bond between melanin and other proteins/peptides and the formation of potential melanin bridges between proteins (Mani et al., 2001). Much of the protein extraction protocol used for LC-MS is carried out in an acidic environment and moreover, most of the LC separation is carried out in acidic conditions. It is therefore possible that the protein extraction from the melanoma samples released high concentrations of melanin (e.g. as a result of lysis of melanosomes), which subsequently bound to free proteins (or peptides after digestion) and modified their mass. Furthermore, melanin is made from the amino acid tyrosine, which has an amine group, and dopaquinone in the melanin synthetic pathway also contains tyrosine's amine group. This may have encouraged binding of melanin precursors to proteins and/or peptides during the extraction of proteins from the melanoma samples because protein/protein cross-linkages can be formed between nucleophilic groups of amino acids, and it is likely that similar cross linkages can form between proteins/peptides and melanin (and its precursors), thus modifying the mass of the protein in question (Hoffman et al., 2015).

In addition to acidic environments producing more polymerisation of melanin and melanoproteins, alkaline environments can allow auto-oxidation which can also induce polymerisation (Mani et al., 2001). It is therefore possible that the higher pH used in the 2D LC separation resulted in more polymerisation of melanin and possibly promoted additional melanin polymer/protein binding. This extra alkaline step in the 2D fractionation process could account for the fewer protein IDs seen in the 2D melanoma data than in the 1D melanoma data, as seen in **Figure 5:3**. Melanins have also been reported to bind to chromatographic columns, thus degrading LC/MS performance, resulting in studies attempting to remove melanin from samples containing melanosomes (Chi et al., 2006).

Another possibility for the difficulties encountered with attempting to identify proteins associated with subsequent development of metastases in the melanoma proteomics could be due to the presence of subgroups within Pmel-M and Pmel-NM tumours, as suggested in the TDA analysis. It is possible that within the sample cohort, these subgroups might have differed in their mechanisms of developing metastasis and therefore the power to identify significant differences was low because more samples would have been needed in each of the subgroups. Whereas the melanoma samples used in this project have not been investigated genetically, these different subgroups might possess different pathogenic genotypes such as that of a BRAF mutant (Haluska et al., 2006), CDKN2A mutant (Harland et al., 2014) or PTEN mutant (Davies et al., 2008).

The modules identified in the WGCNA could also be explained by different subgroups and could also potentially clarify why there was no correlation with propensity to metastasise. For example, it is possible that modules of correlated proteins, specifically blue and turquoise, represent different genotypes which could explain the BRAF/MAPK enrichment seen in the turquoise module.

Despite the possibility of subgroups within the melanoma groups, 31 proteins were identified as significantly differentially expressed between Pmel-M and Pmel-NM. STRING analysis and subsequent pathway analysis of these proteins revealed an enrichment in immune response. It is well known that immunosuppressed individuals have an increased risk of developing melanoma (Euvrard et al., 2003) and indeed an increased risk of metastasis (Martinez et al., 2003). The 1D proteomic data also highlighted enrichment in several MAPK pathways in the melanomas. It has been reported that BRAF mutations appear in ~60-70% of melanomas (Haluska et al., 2006) and BRAF mutations are associated with a higher risk of metastasis (Adler et al., 2017). These proteomic data are consistent with this, which in turn supports the biological relevance of the results presented in the discovery proteomics part of this project.

Cytoskeletal remodelling and motility appeared to be highlighted in several pathway enrichment analysis tools including GO and IPA. Significant activation of remodelling of adherens junctions, actin cytoskeleton signalling and enrichment in actin polymerisation and cytoskeletal organisation was noted. These are likely to be relevant to the biology of the melanomas in this project, because as cancers progress, they often lose polarity

resulting in growth and invasion of the cancer into surrounding tissue as a result of alterations in actin polymerisation, cytoskeletal reorganisation and adherens junctions (Gandalovicova et al., 2016). An activation of many Rho related pathways including regulation of actin-based motility, RhoA signalling and signalling by Rho GTPases was observed. A dysregulation of Rho signalling is known to have an effect on cell polarity and thus, in this instance could be promoting progression by increasing invasiveness (Ellenbroek and Collard, 2007).

IPA upstream analysis also predicted a number of proteins which it inferred were either activated or inhibited. IPA indicated a strong activation of IL15 in the 1D and 2D melanoma proteomic data, but as highlighted in the results, there was no strong enrichment of pathways in the significantly differentially expressed proteins in Pmel-M compared to Pmel-NM. Higher levels of IL-15 have been associated with less metastasis than IL-15 knockouts in breast and melanoma cell lines (Gillgrass et al., 2014) and with a better immune response in melanoma. There were several other noteworthy predicted upstream regulators, including EGFR which is known to have an important role in many cancers, but is also believed to play a possible role in vemurafenib (BRAF inhibitor) resistance (Gross et al., 2015). Furthermore, the most inhibited regulator was miR-122, which conversely in other studies have been found to be increased in metastatic cancers (Fan et al., 2018).

The MRM experiments failed to verify the results which suggested that KRT9, GSN and LMNB1 were associated with development of melanoma metastases in the discovery proteomics. The decision to investigate KRT9 as an MRM verification target was a simple choice because KRT9 was the only significantly different protein present in both the 1D and 2D melanoma data. Likewise, the decision to investigate GSN further was based on the fact that our laboratory already had the heavy labelled GSN peptides. The criteria for selecting the final target was more challenging. The graph for LMNB1 in the discovery proteomics (**Figure 5:9**) looked promising as there appeared to be a good split between Pmel-M and Pmel-NM in addition to having a relatively high fold change. Furthermore, LMNB1 appeared to be relatively high in abundance, which reduces the risk of instrument sensitivity becoming a limiting factor. Another factor involved in choosing LMNB1 was that there was little research on this area and thus there was an element of novelty to investigating this further.

As previously mentioned, MRM analysis throughout the melanoma verification process proved more difficult than compared to cSCCs. The reason for this is unclear but, as mentioned above, it is possible that the high concentrations of melanin might have affected the results of the heavy and native peptides alike, resulting in unreliable results. Indeed, several of the chromatograms for the heavy labelled peptides were inconsistent, in addition to being present at low intensities. It would have been possible to increase the concentration of the heavy peptides in order to increase the intensity of the peaks, but the fact that 100fmol of peptide were of low intensities suggests that the problem was associated with the peptides themselves, or the cells / tissue that the peptides resided in. Furthermore, the intensities of the native peptides in the melanoma sample seen in the calibration experiments were extremely low, suggesting again that something in the tissue or preparation of the mixture was responsible for the reduced number of ions detected for those specific masses of peptides / proteins. All samples were kept at -20°C until their use in the discovery proteomics and MRM experiments, whereupon they were kept at 4°C in order to avoid repeated freeze/thaw cycles. It is possible that during this 4°C period some degradation of proteins / peptides occurred, or that there was sufficient time for melanoproteins to form, resulting in differences in mass, thus lowering the intensity of the peptides on the MRM chromatograms.

Nonetheless, no significant difference between Pmel-M and Pmel-NM for any of the targeted peptides was detected by MRM. Although the proteins could not be validated using MRM, other proteins from the list of differentially expressed proteins in the melanoma discovery proteomics could be investigated, and/or the MRM for the three investigated proteins (GSN, KRT9, LMNB1) could be optimised to a greater extent and repeated. Alternatively, a different methodology could be used to verify and subsequently validate some of the discovery proteomic data. Another possibility might be to cleave any melanoproteins in the samples and filter out melanin (Chi et al., 2006) to optimise the MRM experiments. Despite the MRM data being inconclusive, it should be noted that the proteomic discovery data yielded many results that would have been expected (as highlighted in the results and discussion of this chapter) in addition to novel areas of interest such as potential subgroups with the Pmel-M and Pmel-NM groups of melanoma.

Chapter 6: Modelling Clinical Characteristics of cutaneous Squamous Cell Carcinoma (cSCC)

6.1 Introduction

Modelling of proteomics data provided interesting results in the earlier chapters in this thesis, and in view of having gained experience in machine learning and the lack of an optimal clinical staging system for cSCC, the work in this chapter aimed to determine whether it was possible to use modelling of clinical and/or histological parameters of cSCC (as this data is routinely available in clinical practice) to better predict prognosis in cSCC. Current clinical staging systems for cSCC vary greatly in their criteria for high/low risk or aggressive/non-aggressive cSCCs (Stratigos et al., 2015, Farasat et al., 2011, Lydiatt et al., 2017, Motley et al., 2003). A systematic review and meta-analysis of risk factors for cSCC has highlighted many factors which contribute to the development and progression of cSCC (Thompson et al., 2016). These factors which relate to the primary tumour include a Breslow thickness of >2mm, Clarks level 5, perineural invasion, diameter >20mm, site (e.g. temple, ear, lip) and poor differentiation status. Furthermore, many of these risk factors were also noted when the authors reported risk factors for development of cSCC metastasis (Thompson et al., 2016).

The British Association of Dermatologists (BAD) guidelines for evaluation of cSCCs have limitations of being very sensitivity but rather unspecific. The BAD defines a low risk SCC as a tumour which appears on a sun-exposed site (excluding ear or lip), has a diameter of less than 20mm, is less than 4mm in depth, has a Clarks level below 5 and is well differentiated (Motley et al., 2003). The European Dermatology Forum (EDF) guidelines also suggest that any tumour with moderate or poor differentiation is of high risk, but state that depth of less than 6mm is low risk (compared to 4mm in the BAD guidelines) (Stratigos et al., 2015). The American Joint Committee on Cancer (AJCC) 7th edition use similar characteristics except that they classify poorly differentiated (rather than including moderately differentiated) tumours as high risk; furthermore, they state a tumour of 2mm thickness or Clarks level ≥4 is high risk (Karia et al., 2014). Guidelines are constantly adapting in an

attempt to improve their efficacy and, as such, the new AJCC (8th) edition was modified so that any tumour >6mm or Clarks level 5 are classified as high risk (Sood et al., 2019).

Other staging systems for cSCC have also been proposed; these include the Brigham and Women's Hospital (BWH) (Karia et al., 2014), Brueninger *et al* system (Breuninger et al., 2012) and the Union for International Cancer Control (UICC) (Fang, 2017), each with their own criteria. Although different institutions use different staging systems for cSCC, it is clear that a global and improved cSCC staging system is desperately needed, especially one that has a good sensitivity *and* specificity. Currently, the AJCC 8th version has arguably the best system, producing a sensitivity and specificity of 72.5% and 74.6% respectively (Roscher et al., 2018). Based on the limitations of the current staging systems, it was envisioned that the generation of a staging system which employed clinical and/or histological parameters, through machine learning and modelling of relevant clinical and histological parameters, would be extremely helpful because it could be employed in practise and utilised in a clinical setting with ease and, moreover, relatively quickly.

6.2 Methods

Clinical features of the cSCC samples used in this thesis were subjected to predictive modelling to assess their power to successfully predict which samples are metastatic.

6.2.1 Predictive modelling

As described in chapter 2, predictive modelling was carried out using the statistical programming language R with machine learning packages; caret, caret ensemble, pROC and doParallel for multithread, parallel processing. Clinical and histological data from cSCC samples used for the earlier proteomics studies were combined into one large dataset, totalling 101 samples. This dataset was then randomly (but consistently for each subsequent modelling approach) split into training (67%) and test (33%) sets. Machine learning was then carried out using 10-fold cross validation repeated 3 times for each model. When creating stacked models, base learners were assessed for their correlation to identify a suitable set of models. Stacked models were also trained using 10 fold cross validation repeated 3 times. A full list of the algorithms used in this thesis can be found in **Appendix 4.**

6.3 Results

6.3.1 Initial modelling

Selecting which algorithm to use, or which algorithms to use in the case of a stacked ensemble model, is not an obvious choice because there is not one specific algorithm which has been identified as the optimal one when applying modelling to a clinical problem. Much of the selection process is based on a process of trial and error, and gaining experience in identifying which models perform best with the relevant data. For binomial classification problems, which this is (has metastatised *or* has not metastatised), a first point of call is usually logistic regression due to its simplicity, speed and understandability. Therefore, a generalised linear model (glm) approach was taken as an initial starting point to generate a model which could potentially predict likelihood of development of metastases in patients with primary cSCC.

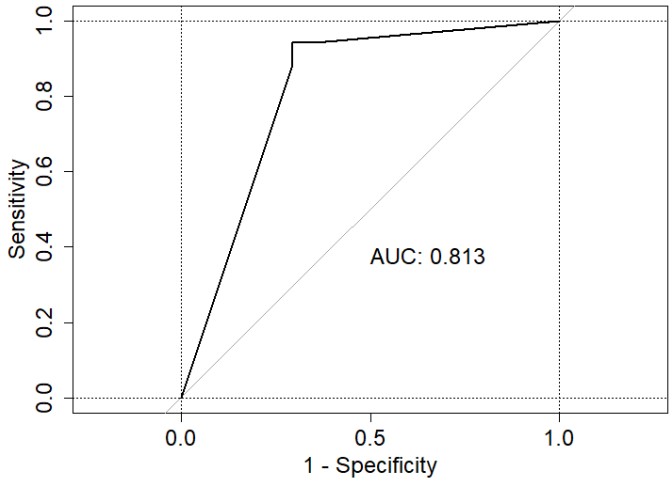


Figure 6:1: A glm model with clinical and histological characteristics of cSCC as predictor variables produced a model with an AUC of 0.813.

A glm model, using cSCC histological parameters (differentiation, diameter, Clarks level, depth, perivascular invasion, perineural invasion and site of tumour) and clinical parameters (age, sex of the patient) as predictive variables. The model was trained using 10-fold cross validation repeated 3 times on a training set (67%) and tested on a test set (33%).

6.3.2 Feature selection

A glm model was created for cSCC, using histological parameters (differentiation, diameter, depth, Clarks level, perivascular invasion, perineural invasion and site of the primary tumour) and clinical parameters (age and sex of the patient) to predict which samples were P-M and P-NM (**figure 6.1**). The ROC curve of this glm model produced an AUC of 0.813 with a 95% confidence interval of 0.680-0.946 (DeLong) and an optimal sensitivity and specificity of 94.1% and 70.6%, respectively. Understanding the characteristics used as predictors, and thus identifying those which add most value to the model, requires evaluation to assess the usefulness (i.e. the predictive power) of each one of the characteristics. Therefore, this was undertaken for the parameters that had been used in the initial glm model.

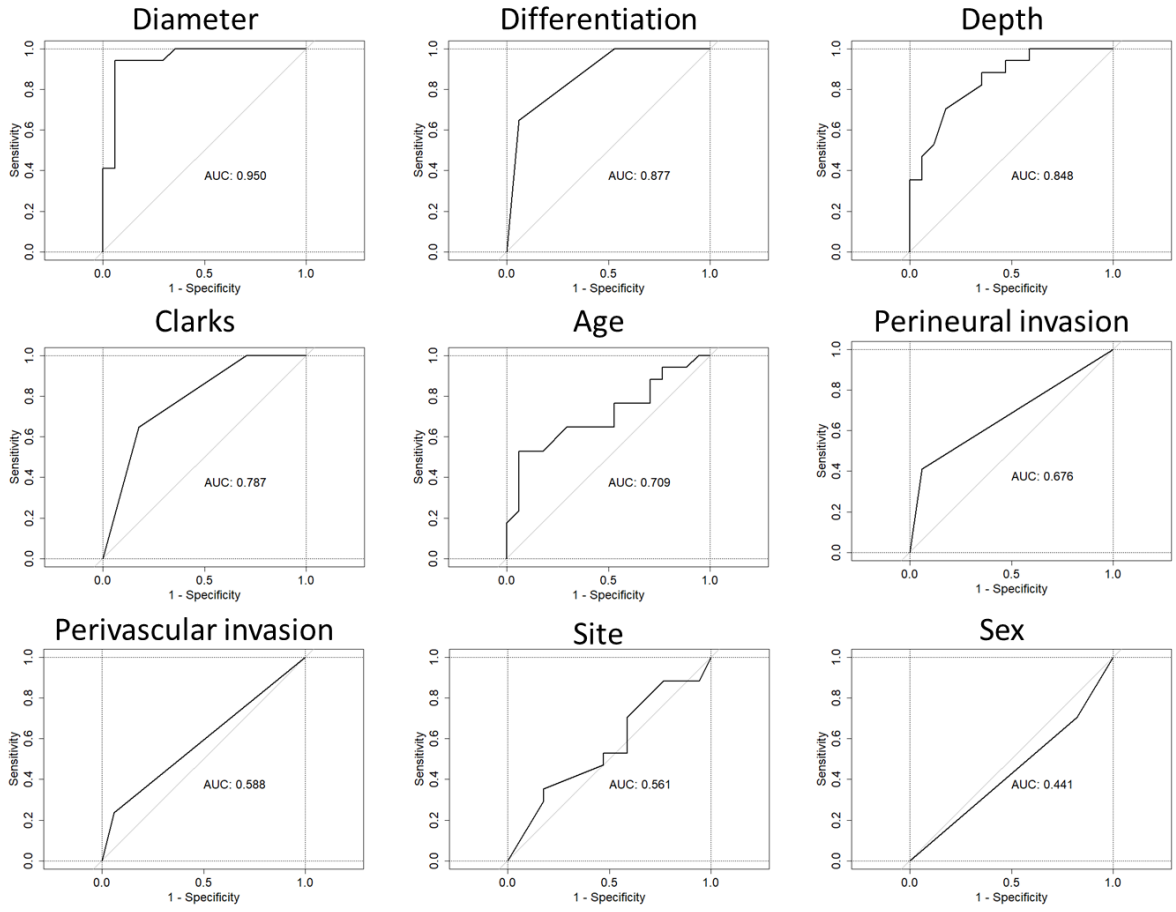


Figure 6:2: The individual predictive power of each variable in identifying the likelihood of development of metastases from primary cSCCs.

Using 10-fold cross validation repeated 3 times, a glm model was created for each clinical characteristic to help identify useful predictors. Models were trained on 67% of the data and tested on 33% of the data.

To determine the individual predictive power of each characteristic in the initial model, a glm model of each one was created (**Figure 6:2**). Diameter was the best individual predictor with an AUC of 0.95 and differentiation and depth were second and third respectively, with AUCs of 0.877 and 0.848. Perivascular invasion and site of tumour were only slightly better than that observed by chance alone, whereas sex (gender) of the patient was actually worse than simply using chance alone. It is important to note that this highlights the *individual* predictive power of the characteristic whereas, in truth, many of these characteristics will be more powerful when used in combination due to the true biological connections between them, i.e. the fact that more aggressive tumours are likely to exhibit several features that associate with being aggressive.

Furthermore, the diameter, differentiation and depth (henceforth dubbed the 3 D's) of an SCC have been recognised for many years to influence clinical outcome in cSCC as well as in other cancers such as melanoma, and therefore these factors represent a good set of variables to use as predictors. A glm model was therefore created, using diameter, depth and differentiation as the predictors (**Figure 6:3**).

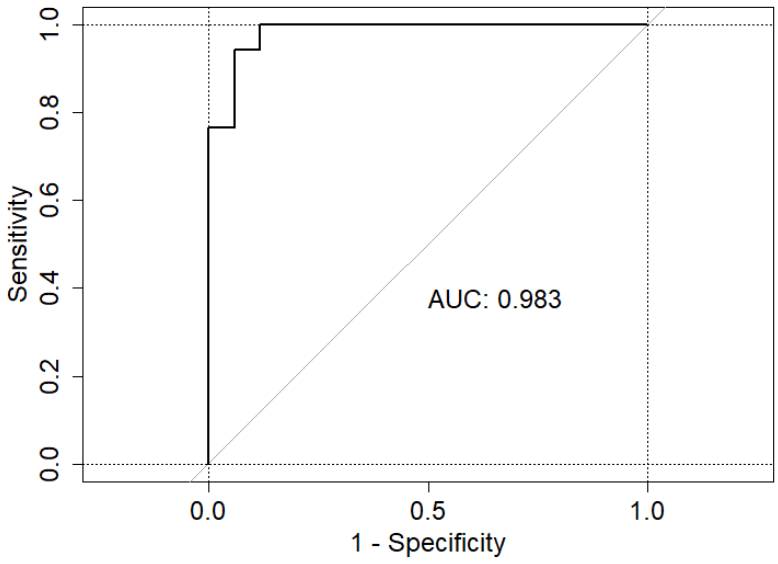


Figure 6:3: A glm prediction model using diameter, differentiation and depth produces a model with an AUC of 0.983.

10-fold cross validation repeated 3 times was employed when creating the glm model using differentiation, diameter and depth as predictive variables. The model was trained on 67% of the cSCC samples and tested on 33% of the samples.

A glm model using the 3 D's produced an AUC of 0.983 with a 95% confidence interval of 0.9506-1 (DeLong). The optimal threshold produced a sensitivity of 94.1% and a specificity of 94.1%. This alone is a good model and warrants the application of this model to a larger sample cohort to validate it further in a larger set of patients. However, this model still produced several misclassifications in the 101 cSCCs in this study, which if extrapolated to the general population could result in many patients being misclassified and therefore incorrectly treated worldwide if the model was employed in clinical practice.

Although the model was trained using repeated cross validation, there was still a risk of overfitting which may not have been detected here due to an inherent bias secondary to the greater number of poorly differentiated tumours in the P-M group compared to the P-NM group. For this reason, it was decided to try different algorithms and, indeed, ensemble algorithms in an attempt to reduce the likelihood of overfitting because many algorithms have built-in defence strategies to avoid this. For this reason, glmnet was henceforth used in place of glm because glmnet uses regularisation to reduce the chance of overfitting. Regularisation is a function in mathematical models which constrains certain components in the data and reduces the impact attributable to these parameters (i.e. by reducing coefficients of features towards zero) thus decreasing the potential for overfitting. Furthermore, although depth performed better as an individual predictor of metastasis (**Figure 6:2**), it was found that with differentiation and diameter, Clark's level and depth performed equally well. In addition, the use of Clark's level attempts to account for different thicknesses of the dermis at different skin sites. Therefore, Clark's level was selected in place of depth in subsequent models. These models were generalised linear model with convex penalties (glmnet), linear discriminant analysis (lda), extreme gradient boositing: dropouts meet multiple additive regression trees (xgbDART), neural network (nnet), naïve bayes (nb), C5.0, gradient boosted machines (gbm), support vector model radial (svmRadial), regularised random forest (RRF), adaptive boosting (adaboost), treebag, K's nearest neighbour (knn) and recursive partitioning and regression trees (rpart).

6.3.3 Algorithm selection

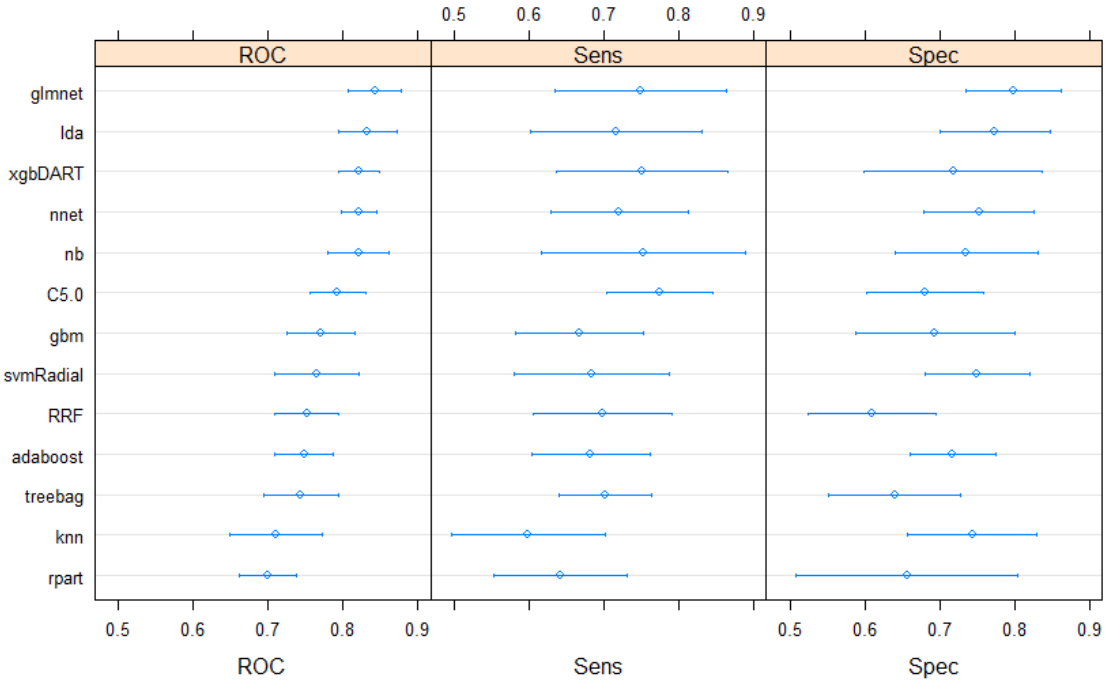


Figure 6:4: Testing multiple machine learning algorithms

Multiple different machine learning algorithms were 10fold cross validated on a training set of data (67% of total 101 cSCCs) and then tested on the remaining 33% of the data. Predictive variables; diameter, differentiation and Clark's level were used. Error bars indicate the variance amongst models within cross validation.

Many of the models tested, produced high ROC scores (**Figure 6:4**). However, as expected, no single model was perfect and each model misclassified different samples in terms of predicting outcome. Stacked ensemble modelling, as previously used in chapter 4 of this thesis, is an excellent way to get good coverage of correct classifications and often increases the possibility of "catching" hard to classify samples so that the resulting model is more accurate.

It is reasonable, but incorrect, to assume that the more mathematical models used, the better the final prediction model will be. In fact, the importance and benefits of combining models lies within the use of those models which corroborate appropriately each other's prediction, specifically in those cases which are classified as positive results (in this case, metastasis). Nonetheless, a stacked model of all the above tested algorithms was created (**Figure 6:5**). This stacked model proved worse than the use of glm by itself, supporting the notion that too many algorithms can be detrimental.

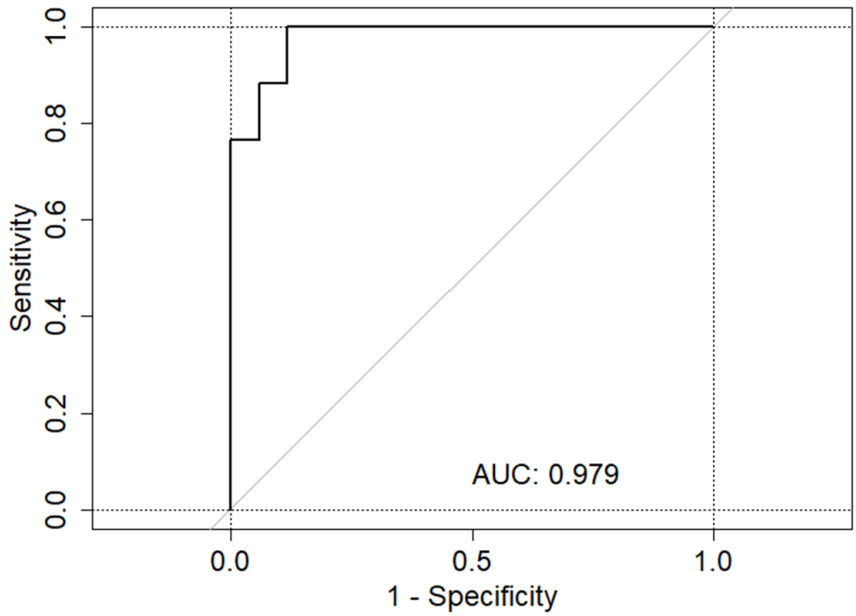


Figure 6:5: Using multiple machine learning algorithms to predict development of metastases from cSCC.

Using the 13 different algorithms outlined in **Figure 6:4**, a model was trained using 10fold cross validation on 67% of the cSCC data and tested on the remaining 33% of the data. The predictors which were used were differentiation, diameter and Clark's level.

6.3.4 Stacked ensemble modelling

To successfully combine models for a stacked ensemble model, it is important to make sure that there isn't too high a correlation between the models, because this could result in excessive weighting for incorrectly classified samples. Similar to the model produced for the MRM data in chapter 4, a correlation matrix of all the tested models was created to help select suitable algorithms (**Figure 6:6**).

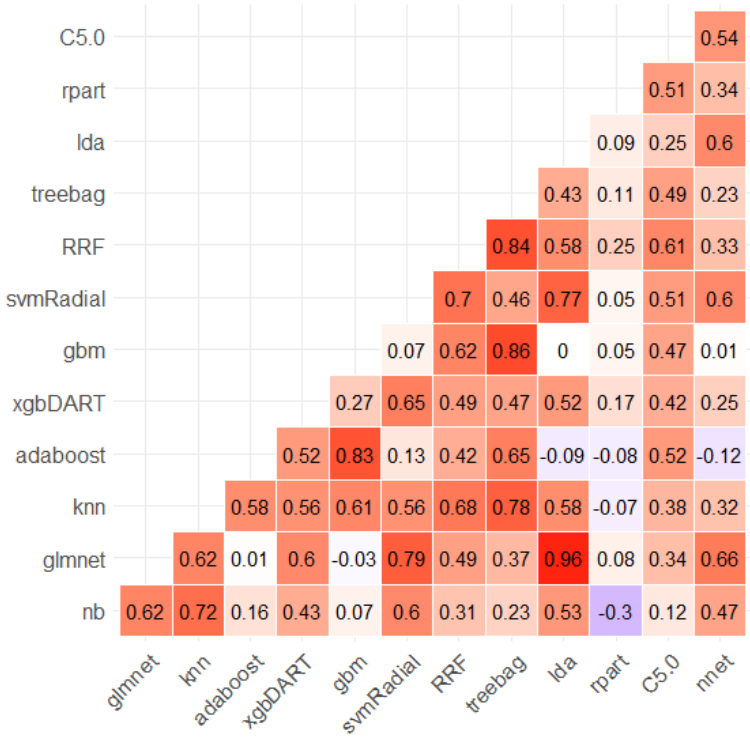


Figure 6:6: Correlation matrix between different model classifications.

The predictions of the 13 models shown in **Figure 6:4** were correlated to each other. In general, pairwise correlations >0.75 are considered too correlated for use in a stacked model.

Although the correlation matrix provides useful information when deciding which models to try in conjunction with each other, there is no ideal way of choosing the most complimentary models and the best method for picking which models to use remains trial and error. Therefore, through trial and error, glmnet, xgbDART, nnet and RRF were found to be a good combination of algorithms. Each of these individual algorithms produced high ROC scores and a range of sensitivities and specificities (**Figure 6:7** and **Figure 6:8**). Furthermore, each of these models expressed little correlation (positive or negative) to one another, suggesting their suitability for stacking (**Figure 6:9**).

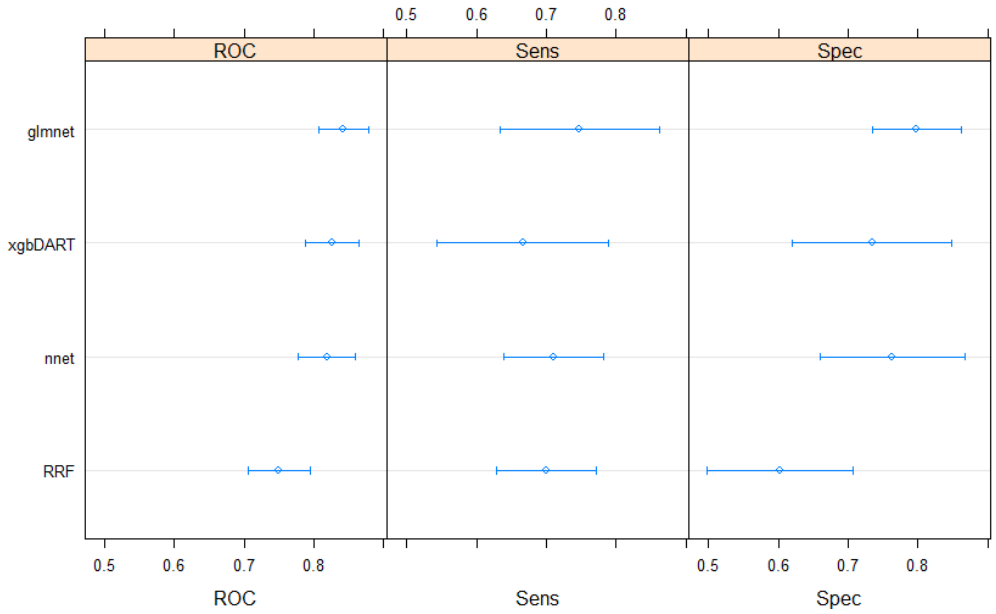


Figure 6:7: investigating the predictive power of glmnet, xgbDART, nnet and RRF algorithms previously identified.

Each of the models were trained on 67% of the cSCC samples using 10fold cross validation repeated 3 times, and tested on 33% of the cSCC samples. Diameter, differentiation and Clark's level used as predictors. Error bars indicate the variance amongst models within the cross validation.

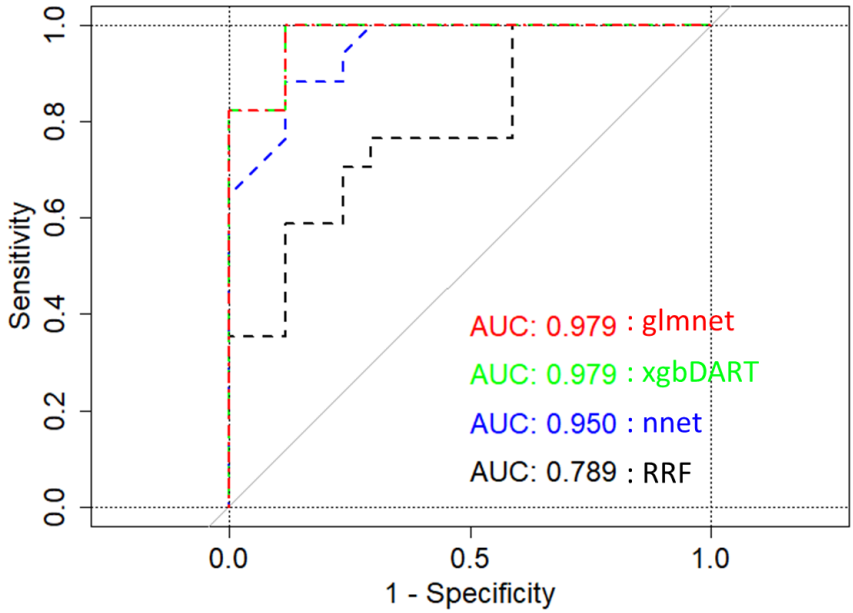


Figure 6:8: ROC curves of glment, xgbDART, nnet and RRF as individual models to predict development of metastases from primary cSCCs.

Diameter, differentiation and Clark's level were used as predictors when training each model on 67% of 101 cSCC samples using 10fold cross validation repeated 3 times. Models were then tested on the remaining 33% of the cSCC samples.

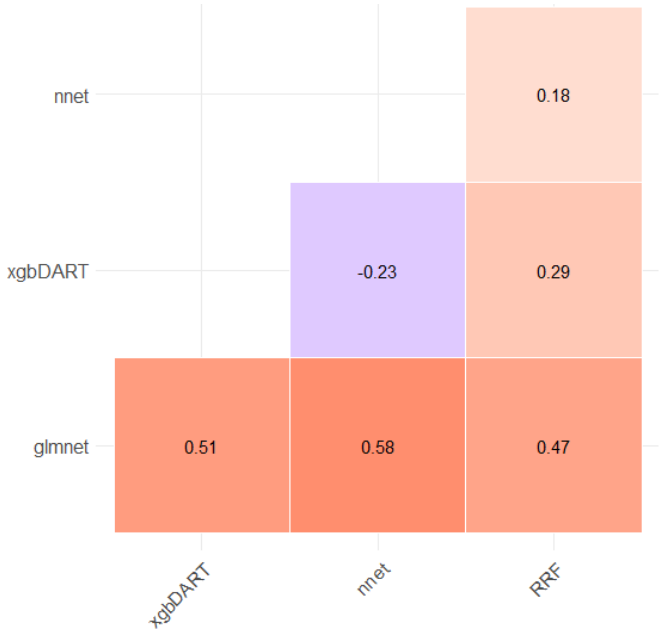


Figure 6:9: There is relatively low correlation between the individual models (also called base level learners) nnet, xgbDART, glmnet and RRF in the prediction of metastases from primary cSCCs.

Predictions of each model, trained on 67% of data using 10 fold cross validation repeated 3 times, were correlated to each other. As each model performed quite well, some correlation was to be expected.

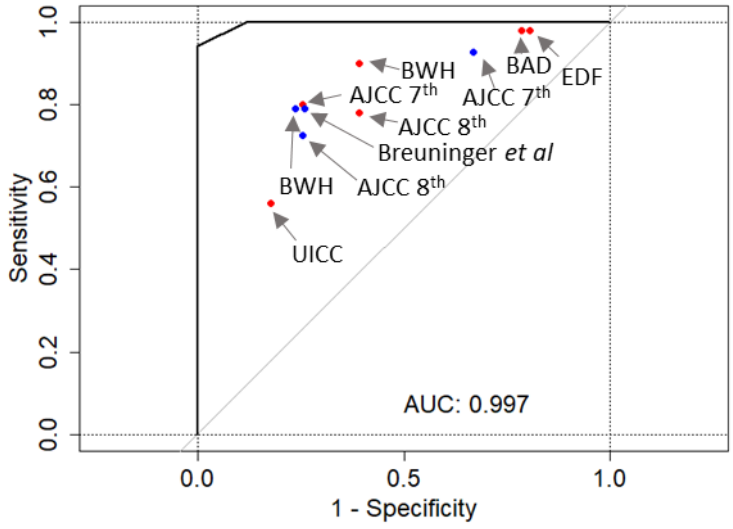


Figure 6:10: ROC curve analysis of the stacked ensemble model produced an AUC of 0.997.

A stacked ensemble model using glmnet, xgbDART, nnet and RRF as base learners was trained on 67% of the data using10 fold cross validation repeated 3 times. Predictions were then submitted to a top level Meta learner, xgbTree, to stack the model and test on the remaining 33% of data.

The nnet, xgbDART, glmnet and RRF prediction models were therefore combined to generate a stacked ensemble model. This stacked ensemble model produced an AUC of 0.997 (**Figure 6:10**) with a 95% confidence interval of 0.9883-1 (DeLong). This resulted in an optimal threshold providing a sensitivity of 94.1% and specificity of 100%. If sensitivity were favoured over the optimal threshold, one could obtain a sensitivity of 100% and a specificity of 88.2%. This model was compared to the prediction of metastases using clinical scoring systems on the same data, and compared to the data published by Roscher et al (2018) as previously depicted in **Figure 4:26**. The result suggested that this stacked ensemble model was better at predicting development of metastases from primary cSCCs than any current cSCC scoring system in use today.

6.4 Discussion

The original model produced (**Figure 6:1**) when using cSCC histological parameters (differentiation, diameter, Clarks level, depth, perivascular invasion, perineural invasion and site of tumour) and clinical parameters (age, sex of the patient) as predictors of development of metastases from cSCC achieved a relatively good AUC value and, as such, highlighted the well-recognised potential of each of these variables to predict clinical outcome in cSCC. Nonetheless, using features which are loosely related to the classification state (for example in this case, sex of patient) in conjunction with strong features which actually hold predictive potential, will result in a weaker model because those loose features decrease the effect of the important ones (Bastanlar and Ozuysal, 2014). It is for this reason that understanding the data and carefully selecting biologically relevant features (or at least those which have strong predictive power in a given input data) is critical. Furthermore, features (predictive variables) which are strongly correlated with other features can lead to unnecessary bias and so should be avoided. For these reasons, the number of features in subsequent models was limited to three from the initial nine that were used originally. Diameter and differentiation are two histological parameters that are always recorded when excising a cSCC and are known to be important risk factors of metastasis. This coupled with the fact that, individually, they were the best performers out of all the other parameters when used in separate models, supported their use as features in the later models. Many of the guidelines in use today use differentiation state as a factor when establishing high/low risk of metastases and, furthermore, often classify cSCCs with

poor differentiation or poor and moderate differentiation as high risk (Motley et al., 2003, Stratigos et al., 2015). The issue with this, however, is that there are some moderate differentiated cSCCs with no other high risk features thus, for example, they are lower risk than those tumours which are moderately differentiated and show evidence of perineural invasion, and so using differentiation alone could readily misclassify cSCCs inappropriately.

Although depth was the second best performing individual feature for predicting metastases, it was reasoned that Clark's level is essentially a simplification of depth in relation to skin site because different skin sites have different thicknesses of the dermis between the epidermis and the subcutaneous tissue (Losquadro, 2017). Of course, site also plays an important role in metastasis because different areas of the body have more or less vascularity and lymphatic channels (amongst other things) (Nedelec et al., 2016). However, in order to include site in the models, the training data set would have needed to be extremely large so as to ensure that overfitting did not occur. Furthermore, the WGCNA analysis of cSCC in chapter 3 also revealed a module (turquoise) which not only expressed a positive correlation with Clark's level but also displayed enrichment in multiple pathways involved in metastasis. Therefore, it was considered that Clark's level as a feature would give the algorithm a combined appreciation of depth and site as well as the more complex biological involvement between these (because they are correlated). Indeed, the combination of diameter, differentiation and Clark's level in a glm model produced an excellent model with a high ROC AUC. Nonetheless, this was the result of using just one (relatively simple) model and it was considered that perhaps an ensemble model (or indeed another more complex model) might generate a ROC curve with a higher AUC.

There is no defined rule for which model or algorithm best suits which data; this largely still relies on experience and on trial and error (Bastanlar and Ozuysal, 2014). Several different algorithms were tested individually in this chapter but then a process of trial and error was used to select the combination of models used for stacking. Each of the individual algorithms employed has its strengths, for instance, glm is a common algorithm used for binomial classification problems and its universality makes it a preferred first choice. xgbDART is a powerful algorithm as, being a boosting algorithm, it can learn from all the data provided in a sequential manner and specifically xgbDART uses randomisation to reduce the chance of overfitting. Boosting algorithms are good at identifying and ignoring

noise in the data, which if this current study were to be performed on an even bigger cohort in the future, could significantly reduce overfitting and increase the overall accuracy of the model. Nnet is a powerful machine learning algorithm which is very good at identifying patterns within data. Given that diameter, differentiation and Clark's level are all likely to be biologically connected, and so probably influence each other in certain ways, nnet may be able to identify patterns in this data that would be hard to identify without machine learning. RRF is a random forest algorithm which utilises regularisation to reduce the chance of overfitting. The use of regularisation here is important because random forest models can have a tendency to overfit especially when they have a huge amount of trees, each with multiple branches and nodes. Regularisation works by shrinking the coefficient estimates towards zero which, in turn, produces a simpler model, less capable of overfitting. Although xgbDART and RRF are both decision tree based, they provide different coverage (**Figure 6:9**), as xgbDART uses mostly very small decision trees to learn sequentially from the last, whereas RRF creates one big "forest" of trees which results in one model that has learnt from all the data as a whole. The resulting stacking model, which used glmnet, xgbDART, nnet and RRF in combination, gave a ROC curve with a very high AUC and outperformed all current guidelines/clinical scoring systems in use currently. In addition, this ensemble model gave a ROC curve with a better AUC than that of the model derived from MRM data outlined in chapter 4. Admittedly, the mass spectrometry-based proteomics in chapters 3 and 4 gave better insight into the biology of cSCC, but the work in the current chapter shows, perhaps surprisingly, that a fresh look (i.e. mathematical modelling) at old systems (i.e. simple histological parameters) can sometimes have the potential to predict clinical outcome to a better extent than novel observations with "omics" approaches.

Chapter 7: General Discussion

The title and aim of this thesis were to identify factors within skin cancer that contribute to the development of metastasis. This was accomplished using a mass spectrometry-based proteomic approach on cSCC and melanoma primary tumours which had metastasised and primary tumours which had not metastasised within a minimum of 5 years since excision of the primary cancer. This alone highlights a way in which this study is fairly unique because most other studies investigate for factors within primary tumours and their subsequent metastases or for differences between the primary and metastatic tumours (Corbo et al., 2017). The use of primary tumours which metastasised and primary tumours which did not metastasise means that these two groups are likely to be more similar than when comparing primary with metastatic tumours. Therefore, less differences in the proteins identified were expected between the groups but it was likely that any such differences would be of interest as a potential driving influence of the development of metastasis.

Although both melanoma and cSCC arise in the skin, they obviously arise from different cell types and, as such, the proteomic profiles and possible contributing factors to development of metastasis were expected to differ between them. The total number of proteins identified in the cSCC samples was 4,018 (**Figure 3:11**) whereas the total number of proteins identified in the melanomas was 3,447 (**Figure 5:3**). We believe this to be the highest number of unique proteins (and indeed proteome coverage) from any cSCC or melanoma FFPE proteomic study, for example the number of proteins identified in a previous melanoma study was 1,528 (Byrum et al., 2013) and in cSCC was 2,120 (Foll et al., 2017). More recently, a proteomics study of actinic keratoses, Bowen's disease and cSCCs, published in the Journal of Investigative Dermatology, identified 3574 proteins across 93 samples of actinic keratosis, Bowen's disease and cSCC (Azimi et al., 2019). The number of IDs in this thesis are also on par with cell line proteomics investigations on cSCC and melanoma (Konstantakou et al., 2017, Paulitschke et al., 2015). The reason for the lower number of proteins identified in the melanoma samples in this thesis was hypothesised to be a result of melanin binding to proteins. For example, it was likely that the higher amount of melanin found in melanomas could bind to proteins in the tumour either during formalin fixation or during the subsequent protein extraction process prior to the LC-MS (Hoffman

et al., 2015). Despite attempts to isolate pure proteins without melanin during the protein extraction, this may have been impossible because melanin is frequently bound to protein *in vivo* (Mani et al., 2001) and may in fact have a biological disposition to do so (Pascutti and Ito, 1992); (Sharma et al., 2002).

A total of 144 proteins and 31 proteins were identified as being significantly differentially expressed between P-M and P-NM cSCCs (**Figure 3:14**) and between Pmel-M and Pmel-NM melanomas (**Figure 5:5**), respectively. The 144 proteins identified in cSCC gave insight into the possible factors contributing to the development of metastasis in cSCC. It is well established that the immune system plays a key role in cSCC development and progression (Rangwala and Tsai, 2011). This is predominantly shown though the effect of immunosuppression, where the risk of developing skin cancer can increase 50-250 fold in immunosuppressed individuals (Alter et al., 2014, Euvrard et al., 2003) and the risk of metastasis is also raised in this patient group (Martinez et al., 2003). Ingenuity pathway analysis revealed significant activation of the TCR in P-M compared to P-NM (**Figure 3:26**) suggesting that there was significantly more activation of T cells in P-M than in P-NM tumours. Although more T cell activation in P-M samples may seem counter-intuitive, it has been found that P-M samples have increased numbers of T-reg lymphocytes (Lai et al., 2016, Lai et al., 2015), so it is possible that part of the observed TCR activation might be due to activation of higher number of Tregs, which in turn would suppress the immune system and allow the cancer to metastasise. Furthermore, IPA also highlighted activation of TGFB1 in P-M samples, which might support the hypothesis that the increase in TCR activation is related to Treg immunosuppression, because TGFB1 is, at least in part, responsible for Treg suppressive function (Wu et al., 2016).

Immunosuppression from Treg cells would facilitate an environment suitable for cancer progression but does not necessarily explain how the cancer itself progresses and metastasises. The bioinformatic analysis of the mass spectrometry results in this thesis identified an enrichment in several connected biological pathways which eluded to possible systems involved in the development of cSCC metastasis. Simultaneously, gene ontology analysis (**Figure 3:23**, **Figure 3:24**), KEGG enrichment analysis (**Figure 3:20**, **Figure 3:22**) and IPA (**Figure 3:25**) revealed significantly more extracellular matrix/focal adhesion and integrin signalling in P-M compared to P-NM. It is known that extracellular stimuli can

promote PI3K-Akt signalling (Thorpe et al., 2015), which in turn is known to promote cancer progression (Yao et al., 2017, Li et al., 2017) and indeed, there was evidence of enrichment and activation of the PI3K-Akt signalling in P-M compared to P-NM (**Figure 3:20**, **Figure 3:22**, **Figure 3:25**, **Figure 3:26**). Furthermore, activation of PI3K-Akt signalling and ERK (which was also seen in the IPA in this chapter 4) has been found to promote metastasis in oropharyngeal SCC as it induces resistance to anoikis (Zeng et al., 2002), a kind of programmed cell death specific to epithelial cells which lose polarity and separate from their normal environment. The IPA (**Figure 3:25**) and gene ontology results also suggested significantly more exocytosis in P-M samples compared to P-NM samples. One possible explanation for this is that P-M samples are priming distant sites into pre-metastatic niches, ready for metastasis (Costa-Silva et al., 2015, Hoshino et al., 2015, Peinado et al., 2011).

The 31 significantly differentially expressed proteins identified in the melanoma samples also indicated an immune system involvement because significant enrichment in this was identified in KEGG analysis (**Figure 5:10**, **Figure 5:11**) and IPA (**Figure 5:14**), albeit less so than that seen in the cSCC analysis. KEGG pathway analysis also revealed a significant enrichment in "signalling by BRAF and RAF fusions" (**Figure 5:10**), which supports findings that over-activation of BRAF (usually through mutations) can contribute to melanoma metastasis (Adler et al., 2017). IPA and gene ontology analysis also indicated enrichment of cytoskeletal remodelling through remodelling of adheren junction, actin cytoskeleton and actin polymerisation/cytoskeletal organisation and activated Rho signalling pathways in Pmel-M compared to Pmel-NM. As many cancer progress, they often lose their polarity which can promote metastasis (Rejon et al., 2016, Halaoui and McCaffrey, 2015, Gandalovicova et al., 2016), with evidence that reduced polarity could be caused by dysregulation of Rho signalling (Ellenbroek and Collard, 2007) which leads to cytoskeletal rearrangements, increasing invasive potential (van de Merbel et al., 2018).

Although 31 differentially expressed proteins were identified between the Pmel-M and Pmel_NM groups, TDA revealed that there could be molecular subgroups within the Pmel-M and Pmel-NM groups. It is possible that these subgroups represented clusters of samples with known mutations in genes such as BRAF, CDKN2A or PTEN. It is also possible that these subgroups represent, until now, unidentified clusters of melanoma samples based on alterations within tumour proteomes independently of mutations in the

aforementioned genes. To determine which of these is correct, future studies could benefit from increasing the sample numbers, undertaking proteomic investigations as were performed in this thesis, and conducting targeted sequencing analysis to look for known driver gene mutations in each sample. Furthermore, future studies could benefit from stratifying melanoma cohorts into their known subgroups, i.e. superficial spreading melanoma, lentigo maligna melanoma, nodular melanoma and acral melanoma. Stratifying for these known subgroups could enable investigation into the molecular biology underpinning each but perhaps moreover, could identify if the molecular phenotypes identified here, correspond to these sub-groups or if indeed they are independent molecular phenotypes.

Interestingly, only one protein was identified in both the 1D and 2D data from the melanoma proteomics; this was Keratin 9. Keratin 9 is almost exclusively found in the suprabasal layers of palmoplantar epidermis (Fuchs-Telem et al., 2013) and therefore identifying this protein in melanoma samples from various different body sites was interesting. Unfortunately, this difference in expression could not be verified using MRM, and so future studies could investigate this further and, if confirmed, could investigate how this keratin might influence melanomas to promote metastasis. A brief comparison of melanoma and cSCC proteomics was performed (**Figure 5:26**), however, the results were only discussed in minor detail as careful consideration was given to the fact that the melanoma proteomic data was not verified. Nonetheless, if future studies could verify the validity of the melanoma data in this thesis, interesting analyses could proceed from the differences in the metastases-related proteomes between cSCCs and melanoma.

Although MRM was unable to verify the findings of the preceding mass spectrometry-based proteomic in this tumour, MRM proved successful in verifying the results of the discovery proteomics in cSCCs (**Figure 4:13**, **Figure 4:14**) as well as validating this on a completely new sample cohort (**Figure 4:18**, **Figure 4:19**). Furthermore, additional verification was performed through IHC staining of L-plastin (**Figure 4:15**). Selecting ANXA5 and DDOST for MRM verification was done by machine learning and modelling of the discovery proteomic cSCC data. Machine learning involves the use of an algorithm to identify trends in data and then apply them to unknown cases (Bastanlar and Ozuysal, 2014). Machine learning and artificial intelligence have been used in biological sciences and medicine to create

diagnostic and prognostic tests (Lao et al., 2017) as well as in predictive analysis to calculate the chance of drug; resistance, suitability and even identify targets (Korotcov et al., 2017). It has also been used in image analysis (Erickson et al., 2017) and has proved to be invaluable when combined with medical professional inputs (Wang et al., 2016) and has even proved more accurate in some cases (Esteva et al., 2017).

Nonetheless, employing machine learning and modelling on the MRM cSCC data produced a model capable of predicting metastasis with an optimal accuracy of 91.18% (sensitivity = 88.24%, specificity = 94.12%). Moreover, the ROC curve of that model produced an AUC of 0.929 and performed better at predicting development of cSCC metastases than clinical scoring systems in current use. Although this doesn't necessarily mean that ANXA5 or DDOST are contributing to metastasis directly, it does suggest that these proteins are either influencers or are influenced by the metastatic process. ANXA5 has been found to promote metastasis in several types of cancer (Xue et al., 2009, Sun et al., 2017, Tang et al., 2017), however, little is known about DDOSTs role in cancer progression. It could be beneficial for future studies to identify, whether ANXA5 and DDOST are causally involved in the development of metastases from cSCCs. This could be done by assessing cellular location of these proteins as well as ablating the expression of their respective genes in the relevant cell types to determine the effect this has on the metastatic potential of cSCC samples. The identification of ANXA5 and DDOST as biomarkers of cSCC metastasis could also lead to potential new treatment options for patients with cSCCs that are likely to metastasise in order to prevent the development of future metastases.

Despite the predictive power of the ANXA5 and DDOST proteins, the final chapter in this thesis showed that using diameter, differentiation and Clark's level of invasion of the primary cSCC in a machine learning approach produced a prediction model with a ROC AUC of 0.997. At an optimal threshold (that is one that produces the highest sensitivity and specificity summed), a sensitivity of 94.1% and specificity of 100% could be achieved giving a summed sensitivity and specificity of 194.1% (out of 200%). Currently, according to a study by Roscher *et al* (2018), the systems outlined by Breuninger *et al (2012)* and the Brigham's and Women's hospital (Karia et al., 2014) have the best predictive ability with a summed sensitivity and specificity of 153.2% and 155.3 (each out of 200%), respectively. Admittedly, Roscher et al investigated 184 cSCC samples and we only studied 101 cSCC

samples; nonetheless, the model presented in chapter 6 is better than the current guidelines/scoring systems used to predict the development of metastasis from cSCCs.

This latter model is potentially very easy to test further and, if confirmed, to subsequently employ in a clinical setting because, theoretically, a software with the built in coefficients of each feature could be created. A pathologist could then input the differentiation, diameter and Clark's level of a cSCC section and obtain a more accurate high/low risk of metastasis for the individual patient. Such a system could result in optimal patient care and ensure that cSCC patients are followed up appropriately. Furthermore, employing such a system could save the NHS time and money through avoiding unnecessary follow up appointments as well as potentially identifying metastatic spread at an earlier stage in those requiring follow-up following excision of the primary cSCC.

In conclusion, this thesis has identified multiple factors within skin cancer that contribute to metastasis. Proteomic and subsequent bioinformatics analysis of cSCC and melanoma samples highlighted several key pathways likely to be involved in the metastatic process. Verification and validation of cSCC proteomics using MRM confirmed these results and moreover revealed that two proteins, ANXA5 and DDOST could accurately predict metastasis in cSCC, therefore, making them biomarkers for metastasis in cSCC. It was also identified that a model, produced using only differentiation, diameter and Clark's level is more accurate at predicting development of metastasis from primary cSCCs than that used in any current clinical scoring system, and that, following further confirmation, this latter model could be integrated into current clinical practice with relative ease.

References

THPA DDOST [Online]. The Human Protein Atlas. Available: https://www.proteinatlas.org/ENSG00000244038-DDOST/pathology [Accessed 31/12/18 2018].

AARNISALO, A. A., GREEN, K. M., O'MALLEY, J., MAKARY, C., ADAMS, J., MERCHANT, S. N. & EVANS, J. E. 2010. A method for MS(E) differential proteomic analysis of archival formalin-fixed celloidin-embedded human inner ear tissue. *Hear Res,* 270**,** 15-20.

ABEL, E. L., ANGEL, J. M., KIGUCHI, K. & DIGIOVANNI, J. 2009. Multi-stage chemical carcinogenesis in mouse skin: fundamentals and applications. *Nat Protoc,* 4**,** 1350-62.

ABERER, W., SCHULER, G., STINGL, G., HONIGSMANN, H. & WOLFF, K. 1981. Ultraviolet light depletes surface markers of Langerhans cells. *J Invest Dermatol,* 76**,** 202-10.

ADDIS, M. F., TANCA, A., PAGNOZZI, D., CROBU, S., FANCIULLI, G., COSSU-ROCCA, P. & UZZAU, S. 2009. Generation of high-quality protein extracts from formalin-fixed, paraffin-embedded tissues. *Proteomics,* 9**,** 3815-23.

ADLER, N. R., WOLFE, R., KELLY, J. W., HAYDON, A., MCARTHUR, G. A., MCLEAN, C. A. & MAR, V. J. 2017. Tumour mutation status and sites of metastasis in patients with cutaneous melanoma. *Br J Cancer,* 117**,** 1026-1035.

AEBERSOLD, R. & MANN, M. 2003. Mass spectrometry-based proteomics. *Nature,* 422**,** 198-207.

ALAM, M. & RATNER, D. 2001. Cutaneous squamous-cell carcinoma. *N Engl J Med,* 344**,** 975-83.

ALBERT, M. R. & WEINSTOCK, M. A. 2003. Keratinocyte carcinoma. *CA Cancer J Clin,* 53**,** 292-302.

ALKHAS, A., HOOD, B. L., OLIVER, K., TENG, P. N., OLIVER, J., MITCHELL, D., HAMILTON, C. A., MAXWELL, G. L. & CONRADS, T. P. 2011. Standardization of a sample preparation and analytical workflow for proteomics of archival endometrial cancer tissue. *J Proteome Res,* 10**,** 5264-71.

ALTER, M., SATZGER, I., SCHREM, H., KALTENBORN, A., KAPP, A. & GUTZMER, R. 2014. Non-melanoma skin cancer is reduced after switch of immunosuppression to mTOR-inhibitors in organ transplant recipients. *J Dtsch Dermatol Ges,* 12**,** 480-8.

ARANDA, F., UDRY, S., PERES WINGEYER, S., AMSHOFF, L. C., BOGDANOVA, N., WIEACKER, P., LATINO, J. O., MARKOFF, A. & DE LARRANAGA, G. 2018. Maternal carriers of the ANXA5 M2 haplotype are exposed to a greater risk for placenta-mediated pregnancy complications. *J Assist Reprod Genet,* 35**,** 921-928.

ARMSTRONG, B. K. & KRICKER, A. 2001. The epidemiology of UV induced skin cancer. *J Photochem Photobiol B,* 63**,** 8-18.

ASCO. *Melanoma: Statistics* [Online]. ASCO. Available: https://www.cancer.net/cancer-types/melanoma/statistics [Accessed 24/09/19].

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

References

ATHAR, M. 2002. Oxidative stress and experimental carcinogenesis. *Indian J Exp Biol,* 40**,** 656-67.

ATULA, T., HEDSTROM, J., FINNE, P., LEIVO, I., MARKKANEN-LEPPANEN, M. & HAGLUND, C. 2003. Tenascin-C expression and its prognostic significance in oral and pharyngeal squamous cell carcinoma. *Anticancer Res,* 23**,** 3051-6.

AZIMI, A., KAUFMAN, K. L., ALI, M., KOSSARD, S. & FERNANDEZ-PENAS, P. 2016. In Silico Analysis Validates Proteomic Findings of Formalin-fixed Paraffin Embedded Cutaneous Squamous Cell Carcinoma Tissue. *Cancer Genomics Proteomics,* 13**,** 453-465.

AZIMI, A., YANG, P., ALI, M., HOWARD, V., MANN, G. J., KAUFMAN, K. L. & FERNANDEZ-PENAS, P. 2019. Data independent acquisition proteomic analysis can discriminate between actinic keratosis, Bowen's disease and cutaneous squamous cell carcinoma. *J Invest Dermatol*.

AZIMZADEH, O., BARJAKTAROVIC, Z., AUBELE, M., CALZADA-WACK, J., SARIOGLU, H., ATKINSON, M. J. & TAPIO, S. 2010. Formalin-fixed paraffin-embedded (FFPE) proteome analysis using gel-free and gel-based proteomics. *J Proteome Res,* 9**,** 4710-20.

AZIMZADEH, O., SCHERTHAN, H., YENTRAPALLI, R., BARJAKTAROVIC, Z., UEFFING, M., CONRAD, M., NEFF, F., CALZADA-WACK, J., AUBELE, M., BUSKE, C., ATKINSON, M. J., HAUCK, S. M. & TAPIO, S. 2012. Label-free protein profiling of formalin-fixed paraffin-embedded (FFPE) heart tissue reveals immediate mitochondrial impairment after ionising radiation. *J Proteomics,* 75**,** 2384-95.

BALAJEE, A. S., MAY, A. & BOHR, V. A. 1999. DNA repair of pyrimidine dimers and 6-4 photoproducts in the ribosomal DNA. *Nucleic Acids Res,* 27**,** 2511-20.

BALCH, C. M. 1992. Cutaneous melanoma: prognosis and treatment results worldwide. *Semin Surg Oncol,* 8**,** 400-14.

BALCH, C. M., GERSHENWALD, J. E., SOONG, S. J., THOMPSON, J. F., ATKINS, M. B., BYRD, D. R., BUZAID, A. C., COCHRAN, A. J., COIT, D. G., DING, S., EGGERMONT, A. M., FLAHERTY, K. T., GIMOTTY, P. A., KIRKWOOD, J. M., MCMASTERS, K. M., MIHM, M. C., JR., MORTON, D. L., ROSS, M. I., SOBER, A. J. & SONDAK, V. K. 2009. Final version of 2009 AJCC melanoma staging and classification. *J Clin Oncol,* 27**,** 6199-206.

BALDWIN, M. A. 2004. Protein identification by mass spectrometry - Issues to be considered. *Molecular & Cellular Proteomics,* 3**,** 1-9.

BANTSCHEFF, M., LEMEER, S., SAVITSKI, M. M. & KUSTER, B. 2012. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem,* 404**,** 939-65.

BARAKA-VIDOT, J., PLANESSE, C., MEILHAC, O., MILITELLO, V., VAN DEN ELSEN, J., BOURDON, E. & RONDEAU, P. 2015. Glycation alters ligand binding, enzymatic, and pharmacological properties of human albumin. *Biochemistry,* 54**,** 3051-62.

BARIL, P., GANGESWARAN, R., MAHON, P. C., CAULEE, K., KOCHER, H. M., HARADA, T., ZHU, M., KALTHOFF, H., CRNOGORAC-JURCEVIC, T. & LEMOINE, N. R. 2006. Periostin promotes invasiveness and resistance of pancreatic cancer cells to hypoxia-induced cell death: role of the [beta]4 integrin and the PI3k pathway. *Oncogene,* 26**,** 2082-2094.

BARTH, A., WANEK, L. A. & MORTON, D. L. 1995. Prognostic factors in 1,521 melanoma patients with distant metastases. *J Am Coll Surg,* 181**,** 193-201.

References

BASTANLAR, Y. & OZUYSAL, M. 2014. Introduction to machine learning. *Methods Mol Biol,* 1107**,** 105-28.

BATAILLE, V., BISHOP, J. A., SASIENI, P., SWERDLOW, A. J., PINNEY, E., GRIFFITHS, K. & CUZICK, J. 1996. Risk of cutaneous melanoma in relation to the numbers, types and sites of naevi: a case-control study. *British Journal of Cancer,* 73**,** 1605-1611.

BATEMAN, N. W., SUN, M., BHARGAVA, R., HOOD, B. L., DARFLER, M. M., KOVATICH, A. J., HOOKE, J. A., KRIZMAN, D. B. & CONRADS, T. P. 2011. Differential proteomic analysis of late-stage and recurrent breast cancer from formalin-fixed paraffin-embedded tissues. *J Proteome Res,* 10**,** 1323-32.

BAYR, H. 2005. Reactive oxygen species. *Critical care medicine,* 33**,** S498-S501.

BEAUMONT, K. A., MOHANA-KUMARAN, N. & HAASS, N. K. 2013. Modeling Melanoma In Vitro and In Vivo. *Healthcare (Basel),* 2**,** 27-46.

BECKER, J. C., HOUBEN, R., SCHRAMA, D., VOIGT, H., UGUREL, S. & REISFELD, R. A. 2010. Mouse models for melanoma: a personal perspective. *Exp Dermatol,* 19**,** 157-64.

BEISSERT, S., HOSOI, J., GRABBE, S., ASAHINA, A. & GRANSTEIN, R. D. 1995. IL-10 inhibits tumor antigen presentation by epidermal antigen-presenting cells. *J Immunol,* 154**,** 1280-6.

BELL, L. N., SAXENA, R., MATTAR, S. G., YOU, J., WANG, M. & CHALASANI, N. 2011. Utility of formalin-fixed, paraffin-embedded liver biopsy specimens for global proteomic analysis in nonalcoholic steatohepatitis. *Proteomics Clin Appl,* 5**,** 397-404.

BENJAMIN, C. L. & ANANTHASWAMY, H. N. 2007. P53 and the pathogenesis of skin cancer. *Toxicology and Applied Pharmacology,* 224**,** 241-248.

BENNIKE, T. B., KASTANIEGAARD, K., PADURARIU, S., GAIHEDE, M., BIRKELUND, S., ANDERSEN, V. & STENSBALLE, A. 2016. Comparing the proteome of snap frozen, RNAlater preserved, and formalin-fixed paraffin-embedded human tissue samples. *EuPA Open Proteomics,* 10**,** 9-18.

BICKERS, D. R. & ATHAR, M. 2000. Novel approaches to chemoprevention of skin cancer. *J Dermatol,* 27**,** 691-5.

BICKERS, D. R. & ATHAR, M. 2006. Oxidative stress in the pathogenesis of skin disease. *J Invest Dermatol,* 126**,** 2565-75.

BICKERS, D. R., LIM, H. W., MARGOLIS, D., WEINSTOCK, M. A., GOODMAN, C., FAULKNER, E., GOULD, C., GEMMEN, E., DALL, T., AMERICAN ACADEMY OF DERMATOLOGY, A. & SOCIETY FOR INVESTIGATIVE, D. 2006. The burden of skin diseases: 2004 a joint project of the American Academy of Dermatology Association and the Society for Investigative Dermatology. *J Am Acad Dermatol,* 55**,** 490-500.

BIGLER, J., BOEDIGHEIMER, M., SCHOFIELD, J. P. R., SKIPP, P. J., CORFIELD, J., ROWE, A., SOUSA, A. R., TIMOUR, M., TWEHUES, L., HU, X., ROBERTS, G., WELCHER, A. A., YU, W., LEFAUDEUX, D., DE MEULDER, B., AUFFRAY, C., CHUNG, K. F., ADCOCK, I. M., STERK, P. J., DJUKANOVIC, R., PLATFORM, U. B. S. G. W. I. F. T. U.-B. P. I., PATIENT REPRESENTATIVES FROM THE ETHICS, B. & SAFETY MANAGEMENT, B. 2016. A Severe Asthma Disease Signature from Gene Expression Profiling of Peripheral Blood from U-BIOPRED Cohorts. *Am J Respir Crit Care Med*.

BOUGNOUX, A. C. & SOLASSOL, J. 2013. The contribution of proteomics to the identification of biomarkers for cutaneous malignant melanoma. *Clin Biochem,* 46**,** 518-23.

References

BOUTER, A., GOUNOU, C., BERAT, R., TAN, S., GALLOIS, B., GRANIER, T., D'ESTAINTOT, B. L., POSCHL, E., BRACHVOGEL, B. & BRISSON, A. R. 2011. Annexin-A5 assembled into two-dimensional arrays promotes cell membrane repair. *Nat Commun,* 2**,** 270.

BRADFORD, P. T., GOLDSTEIN, A. M., TAMURA, D., KHAN, S. G., UEDA, T., BOYLE, J., OH, K. S., IMOTO, K., INUI, H., MORIWAKI, S., EMMERT, S., PIKE, K. M., RAZIUDDIN, A., PLONA, T. M., DIGIOVANNA, J. J., TUCKER, M. A. & KRAEMER, K. H. 2011. Cancer and neurologic degeneration in xeroderma pigmentosum: long term follow-up characterises the role of DNA repair. *J Med Genet,* 48**,** 168-76.

BRANTSCH, K. D., MEISNER, C., SCHONFISCH, B., TRILLING, B., WEHNER-CAROLI, J., ROCKEN, M. & BREUNINGER, H. 2008. Analysis of risk factors determining prognosis of cutaneous squamous-cell carcinoma: a prospective study. *Lancet Oncol,* 9**,** 713-20.

BRASH, D. E., RUDOLPH, J. A., SIMON, J. A., LIN, A., MCKENNA, G. J., BADEN, H. P., HALPERIN, A. J. & PONTEN, J. 1991. A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. *Proc Natl Acad Sci U S A,* 88**,** 10124-8.

BREIMAN, L. 1996. Bagging predictors. *Machine Learning,* 24**,** 123-140.

BREIMAN, L. 1997. Arcing the edge. Technical Report 486, Statistics Department, University of California at ….

BREIMAN, L. 1999. 1 RANDOM FORESTS--RANDOM FEATURES.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. & STONE, C. J. 1984. Classification and regression trees. Belmont, CA: Wadsworth. *International Group,* 432**,** 151-166.

BRENNER, M. & HEARING, V. J. 2008. The protective role of melanin against UV damage in human skin. *Photochem Photobiol,* 84**,** 539-49.

BRESLOW, A. 1970. Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma. *Ann Surg,* 172**,** 902-8.

BRESLOW, A. 1979. Prognostic Factors in the Treatment of Cutaneous Melanoma. *Journal of Cutaneous Pathology,* 6**,** 208-212.

BREUNINGER, H., BLACK, B. & RASSNER, G. 1990. Microstaging of squamous cell carcinomas. *Am J Clin Pathol,* 94**,** 624-7.

BREUNINGER, H., BRANTSCH, K., EIGENTLER, T. & HAFNER, H. M. 2012. Comparison and evaluation of the current staging of cutaneous carcinomas. *J Dtsch Dermatol Ges,* 10**,** 579-86.

BRODERS, A. C. 1921. Squamous-Cell Epithelioma of the Skin: A Study of 256 Cases. *Ann Surg,* 73**,** 141-60.

BRODLAND, D. G. & ZITELLI, J. A. 1992. Mechanisms of metastasis. *J Am Acad Dermatol,* 27**,** 1-8.

BROECKX, V., BOONEN, K., PRINGELS, L., SAGAERT, X., PRENEN, H., LANDUYT, B., SCHOOFS, L. & MAES, E. 2016. Comparison of multiple protein extraction buffers for GeLC-MS/MS proteomic analysis of liver and colon formalin-fixed, paraffin-embedded tissues. *Mol Biosyst,* 12**,** 553-65.

BRONSERT, P., WEISSER, J., BINIOSSEK, M. L., KUEHS, M., MAYER, B., DRENDEL, V., TIMME, S., SHAHINIAN, H., KUSTERS, S., WELLNER, U. F., LASSMANN, S., WERNER, M. & SCHILLING, O. 2014. Impact of routinely employed procedures for tissue processing on the proteomic analysis of formalin-fixed paraffin-embedded tissue. *Proteomics Clin Appl,* 8**,** 796-804.

References

BROŻYNA, A. A., JÓŹWICKI, W., ROSZKOWSKI, K., FILIPIAK, J. & SLOMINSKI, A. T. 2016. Melanin content in melanoma metastases affects the outcome of radiotherapy. *Oncotarget,* 7**,** 17844-17853.

BURGOYNE, R. D. & CLAGUE, M. J. 2003. Calcium and calmodulin in membrane fusion. *Biochim Biophys Acta,* 1641**,** 137-43.

BUYUKBAYRAM, H. & ARSLAN, A. 2002. Value of tenascin-C content and association with clinicopathological parameters in uterine cervical lesions. *Int J Cancer,* 100**,** 719-22.

BYRUM, S., AVARITT, N. L., MACKINTOSH, S. G., MUNKBERG, J. M., BADGWELL, B. D., CHEUNG, W. L. & TACKETT, A. J. 2011. A quantitative proteomic analysis of FFPE melanoma. *J Cutan Pathol,* 38**,** 933-6.

BYRUM, S. D., LARSON, S. K., AVARITT, N. L., MORELAND, L. E., MACKINTOSH, S. G., CHEUNG, W. L. & TACKETT, A. J. 2013. Quantitative Proteomics Identifies Activation of Hallmark Pathways of Cancer in Patient Melanoma. *J Proteomics Bioinform,* 6**,** 43-50.

BYSTROM, S., FREDOLINI, C., EDQVIST, P. H., NYAIESH, E. N., DROBIN, K., UHLEN, M., BERGQVIST, M., PONTEN, F. & SCHWENK, J. M. 2017. Affinity Proteomics Exploration of Melanoma Identifies Proteins in Serum with Associations to T-Stage and Recurrence. *Transl Oncol,* 10**,** 385-395.

BYUN, K., YOO, Y., SON, M., LEE, J., JEONG, G. B., PARK, Y. M., SALEKDEH, G. H. & LEE, B. 2017. Advanced glycation end-products produced systemically and by macrophages: A common contributor to inflammation and degenerative diseases. *Pharmacol Ther,* 177**,** 44-55.

C.R.UK. 2015. *Skin Cancer Incidence Statistics* [Online]. Available: http://www.who.int/uv/faq/skincancer/en/index1.html [Accessed 09/03/2016 2015].

CANCER RESEARCH UK, M. I. S.

CHEN, C., CAI, Q., HE, W., LAM, T. B., LIN, J., ZHAO, Y., CHEN, X., GU, P., HUANG, H., XUE, M., LIU, H., SU, F., HUANG, J., ZHENG, J. & LIN, T. 2017a. AP4 modulated by the PI3K/AKT pathway promotes prostate cancer proliferation and metastasis of prostate cancer via upregulating L-plastin. *Cell Death Dis,* 8**,** e3060.

CHEN, T. & GUESTRIN, C. 2016. Xgboost: A scalable tree boosting system. 785-794.

CHEN, Y., GRUIDL, M., REMILY-WOOD, E., LIU, R. Z., ESCHRICH, S., LLOYD, M., NASIR, A., BUI, M. M., HUANG, E., SHIBATA, D., YEATMAN, T. & KOOMEN, J. M. 2010. Quantification of beta-catenin signaling components in colon cancer cell lines, tissue sections, and microdissected tumor cells using reaction monitoring mass spectrometry. *J Proteome Res,* 9**,** 4215-27.

CHEN, Y. T., CHEN, H. W., WU, C. F., CHU, L. J., CHIANG, W. F., WU, C. C., YU, J. S., TSAI, C. H., LIANG, K. H., CHANG, Y. S., WU, M. & OU YANG, W. T. 2017b. Development of a Multiplexed Liquid Chromatography Multiple-Reaction-Monitoring Mass Spectrometry (LC-MRM/MS) Method for Evaluation of Salivary Proteins as Oral Cancer Biomarkers. *Mol Cell Proteomics,* 16**,** 799-811.

CHEN, Y. T., DUBROW, R., ZHENG, T. Z., BARNHILL, R. L., FINE, J. & BERWICK, M. 1998. Sunlamp use and the risk of cutaneous malignant melanoma: a population-based case-control study in Connecticut, USA. *International Journal of Epidemiology,* 27**,** 758-765.

References

CHENG, K. C., CAHILL, D. S., KASAI, H., NISHIMURA, S. & LOEB, L. A. 1992. 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G----T and A----C substitutions. *J Biol Chem,* 267**,** 166-72.

CHENG, Y., LU, J., CHEN, G., ARDEKANI, G. S., ROTTE, A., MARTINKA, M., XU, X., MCELWEE, K. J., ZHANG, G. & ZHOU, Y. 2015. Stage-specific prognostic biomarkers in melanoma. *Oncotarget,* 6**,** 4180-9.

CHERPELIS, B. S., MARCUSEN, C. & LANG, P. G. 2002. Prognostic factors for metastasis in squamous cell carcinoma of the skin. *Dermatol Surg,* 28**,** 268-73.

CHI, A., VALENCIA, J. C., HU, Z. Z., WATABE, H., YAMAGUCHI, H., MANGINI, N. J., HUANG, H., CANFIELD, V. A., CHENG, K. C., YANG, F., ABE, R., YAMAGISHI, S., SHABANOWITZ, J., HEARING, V. J., WU, C., APPELLA, E. & HUNT, D. F. 2006. Proteomic and bioinformatic characterization of the biogenesis and function of melanosomes. *J Proteome Res,* 5**,** 3135-44.

CHITSAZZADEH, V., COARFA, C., DRUMMOND, J. A., NGUYEN, T., JOSEPH, A., CHILUKURI, S., CHARPIOT, E., ADELMANN, C. H., CHING, G., NGUYEN, T. N., NICHOLAS, C., THOMAS, V. D., MIGDEN, M., MACFARLANE, D., THOMPSON, E., SHEN, J., TAKATA, Y., MCNIECE, K., POLANSKY, M. A., ABBAS, H. A., RAJAPAKSHE, K., GOWER, A., SPIRA, A., COVINGTON, K. R., XIAO, W., GUNARATNE, P., PICKERING, C., FREDERICK, M., MYERS, J. N., SHEN, L., YAO, H., SU, X., RAPINI, R. P., WHEELER, D. A., HAWK, E. T., FLORES, E. R. & TSAI, K. Y. 2016. Cross-species identification of genomic drivers of squamous cell carcinoma development across preneoplastic intermediates. *Nat Commun,* 7**,** 12601.

CHIU, C.-C., LI, H.-F., CHEN, Y.-J. & CHEN, A.-J. 2014. Abstract 3067: Differential proteomic profilingidentifies HNSCCinvasion genes: GANAB is negative regulator in HNSCC invasion. *Cancer Research,* 70**,** 3067.

CHO, H. J., PARK, S. M., KIM, I. K., NAM, I. K., BAEK, K. E., IM, M. J., YOO, J. M., PARK, S. H., RYU, K. J., HAN, H. T., KIM, H. J., HONG, S. C., KIM, K. D., PAK, Y., KIM, J. W., LEE, C. W. & YOO, J. 2014. RhoGDI2 promotes epithelial-mesenchymal transition via induction of Snail in gastric cancer cells. *Oncotarget,* 5**,** 1554-64.

CHOI, H. & NESVIZHSKII, A. I. 2008. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J Proteome Res,* 7**,** 47-50.

CLARK, W. H., FROM, L., BERNARDINO, E. A. & MIHM, M. C. 1969. The Histogenesis and Biologic Behavior of Primary Human Malignant Melanomas of the Skin. *Cancer Research,* 29**,** 705.

CLARKE, R. E., DORDEVIC, A. L., TAN, S. M., RYAN, L. & COUGHLAN, M. T. 2016. Dietary Advanced Glycation End Products and Risk Factors for Chronic Disease: A Systematic Review of Randomised Controlled Trials. *Nutrients,* 8**,** 125.

COHEN, P., WEST, S. G. & AIKEN, L. S. 2014. *Applied multiple regression/correlation analysis for the behavioral sciences*, Psychology Press.

COLE, R. B. 2011. *Electrospray and MALDI mass spectrometry: fundamentals, instrumentation, practicalities, and biological applications*, John Wiley & Sons.

COMMANDEUR, S., DE GRUIJL, F. R., WILLEMZE, R., TENSEN, C. P. & EL GHALBZOURI, A. 2009. An in vitro three-dimensional model of primary human cutaneous squamous cell carcinoma. *Exp Dermatol,* 18**,** 849-56.

References

CORBO, C., CEVENINI, A. & SALVATORE, F. 2017. Biomarker discovery by proteomics-based approaches for early detection and personalized medicine in colorectal cancer. *Proteomics Clin Appl,* 11**,** 1600072.

CORNISH, D., HOLTERHUES, C., VAN DE POLL-FRANSE, L. V., COEBERGH, J. W. & NIJSTEN, T. 2009. A systematic review of health-related quality of life in cutaneous melanoma. *Annals of Oncology,* 20**,** vi51-vi58.

COSTA-SILVA, B., AIELLO, N. M., OCEAN, A. J., SINGH, S., ZHANG, H., THAKUR, B. K., BECKER, A., HOSHINO, A., MARK, M. T., MOLINA, H., XIANG, J., ZHANG, T., THEILEN, T. M., GARCIA-SANTOS, G., WILLIAMS, C., ARARSO, Y., HUANG, Y., RODRIGUES, G., SHEN, T. L., LABORI, K. J., LOTHE, I. M., KURE, E. H., HERNANDEZ, J., DOUSSOT, A., EBBESEN, S. H., GRANDGENETT, P. M., HOLLINGSWORTH, M. A., JAIN, M., MALLYA, K., BATRA, S. K., JARNAGIN, W. R., SCHWARTZ, R. E., MATEI, I., PEINADO, H., STANGER, B. Z., BROMBERG, J. & LYDEN, D. 2015. Pancreatic cancer exosomes initiate pre-metastatic niche formation in the liver. *Nat Cell Biol,* 17**,** 816-26.

COVER, T. M. & HART, P. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory,* 13**,** 21-27.

CRAVEN, R. A., CAIRNS, D. A., ZOUGMAN, A., HARNDEN, P., SELBY, P. J. & BANKS, R. E. 2013. Proteomic analysis of formalin-fixed paraffin-embedded renal tissue samples by label-free MS: assessment of overall technical variability and the impact of block age. *Proteomics Clin Appl,* 7**,** 273-82.

D'SOUZA, B., MIYAMOTO, A. & WEINMASTER, G. 2008. The many facets of Notch ligands. *Oncogene,* 27**,** 5148-67.

DANG, C., GOTTSCHLING, M., MANNING, K., O'CURRAIN, E., SCHNEIDER, S., STERRY, W., STOCKFLETH, E. & NINDL, I. 2006. Identification of dysregulated genes in cutaneous squamous cell carcinoma. *Oncol Rep,* 16**,** 513-9.

DANIELSEN, S. A., EIDE, P. W., NESBAKKEN, A., GUREN, T., LEITHE, E. & LOTHE, R. A. 2015. Portrait of the PI3K/AKT pathway in colorectal cancer. *Biochim Biophys Acta,* 1855**,** 104-21.

DAVIES, M. A., STEMKE-HALE, K., TELLEZ, C., CALDERONE, T. L., DENG, W., PRIETO, V. G., LAZAR, A. J., GERSHENWALD, J. E. & MILLS, G. B. 2008. A novel AKT3 mutation in melanoma tumours and cell lines. *Br J Cancer,* 99**,** 1265-8.

DE GRUIJL, F. R. & FORBES, P. D. 1995. UV-induced skin cancer in a hairless mouse model. *Bioessays,* 17**,** 651-60.

DE HERTOG, S. A., WENSVEEN, C. A., BASTIAENS, M. T., KIELICH, C. J., BERKHOUT, M. J., WESTENDORP, R. G., VERMEER, B. J., BOUWES BAVINCK, J. N. & LEIDEN SKIN CANCER, S. 2001. Relation between smoking and skin cancer. *J Clin Oncol,* 19**,** 231-8.

DIFFEY, B. L. 2004. The future incidence of cutaneous melanoma within the UK. *Br J Dermatol,* 151**,** 868-72.

DINEHART, S. M. & POLLACK, S. V. 1989. Metastases from squamous cell carcinoma of the skin and lip. An analysis of twenty-seven cases. *J Am Acad Dermatol,* 21**,** 241-8.

DING, X. M., LI, J. X., WANG, K., WU, Z. S., YAO, A. H., JIAO, C. Y., QIAN, J. J., BAI, D. S. & LI, X. C. 2017. Effects of silencing annexin A5 on proliferation and invasion of human cholangiocarcinoma cell line. *Eur Rev Med Pharmacol Sci,* 21**,** 1477-1488.

References

DIPNALL, J. F., PASCO, J. A., BERK, M., WILLIAMS, L. J., DODD, S., JACKA, F. N. & MEYER, D. 2016. Fusing Data Mining, Machine Learning and Traditional Statistics to Detect Biomarkers Associated with Depression. *PLoS One,* 11**,** e0148195.

DOBSON, C. M. 2003. Protein folding and misfolding. *Nature,* 426**,** 884-90.

DOBSON, C. M., ŠALI, A. & KARPLUS, M. 1998. Protein Folding: A Perspective from Theory and Experiment. *Angewandte Chemie International Edition,* 37**,** 868-893.

DOERR, A. 2015. DIA mass spectrometry. *Nature Methods,* 12**,** 35-35.

DOMON, B. & AEBERSOLD, R. 2006. Mass spectrometry and protein analysis. *Science,* 312**,** 212-7.

DONADIO, E., GIUSTI, L., CETANI, F., DA VALLE, Y., CIREGIA, F., GIANNACCINI, G., PARDI, E., SAPONARO, F., TORREGROSSA, L., BASOLO, F., MARCOCCI, C. & LUCACCHINI, A. 2011. Evaluation of formalin-fixed paraffin-embedded tissues in the proteomic analysis of parathyroid glands. *Proteome Sci,* 9**,** 29.

DONEANU, C. E., XENOPOULOS, A., FADGEN, K., MURPHY, J., SKILTON, S. J., PRENTICE, H., STAPELS, M. & CHEN, W. 2012. Analysis of host-cell proteins in biotherapeutic proteins by comprehensive online two-dimensional liquid chromatography/mass spectrometry. *MAbs,* 4**,** 24-44.

DORMAN, S. E. & HOLLAND, S. M. 2000. Interferon-gamma and interleukin-12 pathway defects and human disease. *Cytokine Growth Factor Rev,* 11**,** 321-33.

DOWNWARD, J. 2003. Targeting RAS signalling pathways in cancer therapy. *Nat Rev Cancer,* 3**,** 11-22.

DRENTH, H., ZUIDEMA, S. U., KRIJNEN, W. P., BAUTMANS, I., VAN DER SCHANS, C. & HOBBELEN, H. 2017. Advanced Glycation End-Products Are Associated With the Presence and Severity of Paratonia in Early Stage Alzheimer Disease. *J Am Med Dir Assoc,* 18**,** 636 e7-636 e12.

DRUMMOND, E. S., NAYAK, S., UEBERHEIDE, B. & WISNIEWSKI, T. 2015. Proteomic analysis of neurons microdissected from formalin-fixed, paraffin-embedded Alzheimer's disease brain tissue. *Sci Rep,* 5**,** 15456.

DUBA, R. O. & HART, P. E. 1973. Pattern Classification and Scene Analysis Wiley. *New York*.

DUPERRET, E. K. & RIDKY, T. W. 2013. Focal adhesion complex proteins in epidermis and squamous cell carcinoma. *Cell Cycle,* 12**,** 3272-3285.

DURINCK, S., HO, C., WANG, N. J., LIAO, W., JAKKULA, L. R., COLLISSON, E. A., PONS, J., CHAN, S. W., LAM, E. T., CHU, C., PARK, K., HONG, S. W., HUR, J. S., HUH, N., NEUHAUS, I. M., YU, S. S., GREKIN, R. C., MAURO, T. M., CLEAVER, J. E., KWOK, P. Y., LEBOIT, P. E., GETZ, G., CIBULSKIS, K., ASTER, J. C., HUANG, H., PURDOM, E., LI, J., BOLUND, L., ARRON, S. T., GRAY, J. W., SPELLMAN, P. T. & CHO, R. J. 2011. Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov,* 1**,** 137-43.

EDEN, E., NAVON, R., STEINFELD, I., LIPSON, D. & YAKHINI, Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics,* 10**,** 48.

EGERTSON, J. D., KUEHN, A., MERRIHEW, G. E., BATEMAN, N. W., MACLEAN, B. X., TING, Y. S., CANTERBURY, J. D., MARSH, D. M., KELLMANN, M., ZABROUSKOV, V., WU, C. C. & MACCOSS, M. J. 2013. Multiplexed MS/MS for improved data-independent acquisition. *Nat Methods,* 10**,** 744-6.

References

EGGERMONT, A. M., SUCIU, S., SANTINAMI, M., TESTORI, A., KRUIT, W. H., MARSDEN, J., PUNT, C. J., SALES, F., GORE, M., MACKIE, R., KUSIC, Z., DUMMER, R., HAUSCHILD, A., MUSAT, E., SPATZ, A., KEILHOLZ, U. & GROUP, E. M. 2008. Adjuvant therapy with pegylated interferon alfa-2b versus observation alone in resected stage III melanoma: final results of EORTC 18991, a randomised phase III trial. *Lancet,* 372**,** 117-26.

EGGERMONT, A. M. M., CHIARION-SILENI, V., GROB, J.-J., DUMMER, R., WOLCHOK, J. D., SCHMIDT, H., HAMID, O., ROBERT, C., ASCIERTO, P. A., RICHARDS, J. M., LEBBÉ, C., FERRARESI, V., SMYLIE, M., WEBER, J. S., MAIO, M., KONTO, C., HOOS, A., DE PRIL, V., GURUNATH, R. K., DE SCHAETZEN, G., SUCIU, S. & TESTORI, A. 2015. Adjuvant ipilimumab versus placebo after complete resection of high-risk stage III melanoma (EORTC 18071): a randomised, double-blind, phase 3 trial. *The Lancet Oncology,* 16**,** 522-530.

EKE, I. & CORDES, N. 2015. Focal adhesion signaling and therapy resistance in cancer. *Semin Cancer Biol,* 31**,** 65-75.

EKWUEME, D. U., GUY, G. P., JR., LI, C., RIM, S. H., PARELKAR, P. & CHEN, S. C. 2011. The health burden and economic costs of cutaneous melanoma mortality by race/ethnicity-United States, 2000 to 2006. *J Am Acad Dermatol,* 65**,** S133-43.

EL GHISSASSI, F., BAAN, R., STRAIF, K., GROSSE, Y., SECRETAN, B., BOUVARD, V., BENBRAHIM-TALLAA, L., GUHA, N., FREEMAN, C., GALICHET, L., COGLIANO, V. & GROUP, W. H. O. I. A. F. R. O. C. M. W. 2009. A review of human carcinogens--part D: radiation. *Lancet Oncol,* 10**,** 751-2.

ELLENBROEK, S. I. & COLLARD, J. G. 2007. Rho GTPases: functions and association with cancer. *Clin Exp Metastasis,* 24**,** 657-72.

ELSCHENBROICH, S. & KISLINGER, T. 2011. Targeted proteomics by selected reaction monitoring mass spectrometry: applications to systems biology and biomarker discovery. *Mol Biosyst,* 7**,** 292-303.

ERICKSON, B. J., KORFIATIS, P., AKKUS, Z. & KLINE, T. L. 2017. Machine Learning for Medical Imaging. *Radiographics,* 37**,** 505-515.

ESTEVA, A., KUPREL, B., NOVOA, R. A., KO, J., SWETTER, S. M., BLAU, H. M. & THRUN, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature,* 542**,** 115-118.

EUVRARD, S., KANITAKIS, J. & CLAUDY, A. 2003. Skin cancers after organ transplantation. *N Engl J Med,* 348**,** 1681-91.

FABRIS, F., MAGALHAES, J. P. & FREITAS, A. A. 2017. A review of supervised machine learning applied to ageing research. *Biogerontology,* 18**,** 171-188.

FAN, Y., MA, X., LI, H., GAO, Y., HUANG, Q., ZHANG, Y., BAO, X., DU, Q., LUO, G., LIU, K., MENG, Q., ZHAO, C. & ZHANG, X. 2018. miR-122 promotes metastasis of clear-cell renal cell carcinoma by downregulating Dicer. *Int J Cancer,* 142**,** 547-560.

FANG, W. 2017. Interpretation of 2017 National Comprehensive Cancer Network (NCCN) guidelines for the diagnosis and treatment of esophageal squamous cell carcinoma through the new TNM staging of esophageal carcinoma (eighth edition) by the Union for International Cancer Control (UICC) and the American Cancer Commission (AJCC). *Zhonghua Wei Chang Wai Ke Za Zhi,* 20**,** 1122-1126.

References

FARAGE, M. A., MILLER, K. W. & MAIBACH, H. I. 2009. *Textbook of aging skin*, Springer Science & Business Media.

FARASAT, S., YU, S. S., NEEL, V. A., NEHAL, K. S., LARDARO, T., MIHM, M. C., BYRD, D. R., BALCH, C. M., CALIFANO, J. A., CHUANG, A. Y., SHARFMAN, W. H., SHAH, J. P., NGHIEM, P., OTLEY, C. C., TUFARO, A. P., JOHNSON, T. M., SOBER, A. J. & LIEGEOIS, N. J. 2011. A new American Joint Committee on Cancer staging system for cutaneous squamous cell carcinoma: creation and rationale for inclusion of tumor (T) characteristics. *J Am Acad Dermatol,* 64**,** 1051-9.

FENN, J. B., MANN, M., MENG, C. K., WONG, S. F. & WHITEHOUSE, C. M. 1989. Electrospray ionization for mass spectrometry of large biomolecules. *Science,* 246**,** 64-71.

FERLAY, J., SHIN, H. R., BRAY, F., FORMAN, D., MATHERS, C. & PARKIN, D. M. 2010. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer,* 127**,** 2893-917.

FIELDS, S. & JOHNSTON, M. 2005. Cell biology. Whither model organism research? *Science,* 307**,** 1885-6.

FLANAGAN, N., HEALY, E., RAY, A., PHILIPS, S., TODD, C., JACKSON, I. J., BIRCH-MACHIN, M. A. & REES, J. L. 2000. Pleiotropic effects of the melanocortin 1 receptor (MC1R) gene on human pigmentation. *Hum Mol Genet,* 9**,** 2531-7.

FOLL, M. C., FAHRNER, M., GRETZMEIER, C., THOMA, K., BINIOSSEK, M. L., KIRITSI, D., MEISS, F., SCHILLING, O., NYSTROM, A. & KERN, J. S. 2017. Identification of tissue damage, extracellular matrix remodeling and bacterial challenge as common mechanisms associated with high-risk cutaneous squamous cell carcinomas. *Matrix Biol*.

FORLONI, M., DOGRA, S. K., DONG, Y., CONTE, D., JR., OU, J., ZHU, L. J., DENG, A., MAHALINGAM, M., GREEN, M. R. & WAJAPEYEE, N. 2014. miR-146a promotes the initiation and progression of melanoma by activating Notch signaling. *Elife,* 3**,** e01460.

FOWLER, C. B., WAYBRIGHT, T. J., VEENSTRA, T. D., O'LEARY, T. J. & MASON, J. T. 2012. Pressure-assisted protein extraction: a novel method for recovering proteins from archival tissue for proteomic analysis. *J Proteome Res,* 11**,** 2602-8.

FRANCESCHI, P., GIORDAN, M. & WEHRENS, R. 2013. Multiple comparisons in mass-spectrometry-based -omics technologies. *Trac-Trends in Analytical Chemistry,* 50**,** 11-21.

FRANCESCHINI, A., SZKLARCZYK, D., FRANKILD, S., KUHN, M., SIMONOVIC, M., ROTH, A., LIN, J., MINGUEZ, P., BORK, P., VON MERING, C. & JENSEN, L. J. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res,* 41**,** D808-15.

FRANSEN, M., KARAHALIOS, A., SHARMA, N., ENGLISH, D. R., GILES, G. G. & SINCLAIR, R. D. 2012. Non-melanoma skin cancer in Australia. *Med J Aust,* 197**,** 565-8.

FRIERSON, H. F., JR. & COOPER, P. H. 1986. Prognostic factors in squamous cell carcinoma of the lower lip. *Hum Pathol,* 17**,** 346-54.

FU, Z., YAN, K., ROSENBERG, A., JIN, Z., CRAIN, B., ATHAS, G., HEIDE, R. S., HOWARD, T., EVERETT, A. D., HERRINGTON, D. & VAN EYK, J. E. 2013. Improved protein extraction and protein identification from archival formalin-fixed paraffin-embedded human aortas. *Proteomics Clin Appl,* 7**,** 217-24.

References

FUCHS-TELEM, D., PADALON-BRAUCH, G., SARIG, O. & SPRECHER, E. 2013. Epidermolytic palmoplantar keratoderma caused by activation of a cryptic splice site in KRT9. *Clin Exp Dermatol,* 38**,** 189-92: quiz 192.

FUJITA, H., SUAREZ-FARINAS, M., MITSUI, H., GONZALEZ, J., BLUTH, M. J., ZHANG, S., FELSEN, D., KRUEGER, J. G. & CARUCCI, J. A. 2012. Langerhans cells from human cutaneous squamous cell carcinoma induce strong type 1 immunity. *J Invest Dermatol,* 132**,** 1645-55.

GALLAGHER, R. P., HILL, G. B., BAJDIK, C. D., COLDMAN, A. J., FINCHAM, S., MCLEAN, D. I. & THRELFALL, W. J. 1995. Sunlight exposure, pigmentation factors, and risk of nonmelanocytic skin cancer. II. Squamous cell carcinoma. *Arch Dermatol,* 131**,** 164-9.

GAMEZ-POZO, A., SANCHEZ-NAVARRO, I., CALVO, E., DIAZ, E., MIGUEL-MARTIN, M., LOPEZ, R., AGULLO, T., CAMAFEITA, E., ESPINOSA, E., LOPEZ, J. A., NISTAL, M. & VARA, J. A. 2011. Protein phosphorylation analysis in archival clinical cancer samples by shotgun and targeted proteomics approaches. *Mol Biosyst,* 7**,** 2368-74.

GANDALOVICOVA, A., VOMASTEK, T., ROSEL, D. & BRABEK, J. 2016. Cell polarity signaling in the plasticity of cancer cell invasiveness. *Oncotarget,* 7**,** 25022-49.

GARBE, C., EIGENTLER, T. K., KEILHOLZ, U., HAUSCHILD, A. & KIRKWOOD, J. M. 2011. Systematic review of medical treatment in melanoma: current status and future prospects. *Oncologist,* 16**,** 5-24.

GARBE, C. & LEITER, U. 2009. Melanoma epidemiology and trends. *Clin Dermatol,* 27**,** 3-9.

GARBE, C., PERIS, K., HAUSCHILD, A., SAIAG, P., MIDDLETON, M., BASTHOLT, L., GROB, J. J., MALVEHY, J., NEWTON-BISHOP, J., STRATIGOS, A. J., PEHAMBERGER, H., EGGERMONT, A. M., EUROPEAN DERMATOLOGY, F., EUROPEAN ASSOCIATION OF, D.-O., EUROPEAN ORGANISATION FOR, R. & TREATMENT OF, C. 2016. Diagnosis and treatment of melanoma. European consensus-based interdisciplinary guideline - Update 2016. *Eur J Cancer,* 63**,** 201-17.

GARBE, C., TERHEYDEN, P., KEILHOLZ, U., KÖLBL, O. & HAUSCHILD, A. 2008. Treatment of melanoma. *Dtsch Arztebl Int,* 105**,** 845-851.

GARG, K., MAURER, M., GRISS, J., BRUGGEN, M. C., WOLF, I. H., WAGNER, C., WILLI, N., MERTZ, K. D. & WAGNER, S. N. 2016. Tumor-associated B cells in cutaneous primary melanoma and improved clinical outcome. *Hum Pathol,* 54**,** 157-64.

GAST, M. C., SCHELLENS, J. H. & BEIJNEN, J. H. 2009. Clinical proteomics in breast cancer: a review. *Breast Cancer Res Treat,* 116**,** 17-29.

GAZZANIGA, P., NOFRONI, I., GANDINI, O., SILVESTRI, I., FRATI, L., AGLIANO, A. M. & GRADILONE, A. 2005. Tenascin C and epidermal growth factor receptor as markers of circulating tumoral cells in bladder and colon cancer. *Oncol Rep,* 14**,** 1199-202.

GERSTENBLITH, M. R., GOLDSTEIN, A. M., FARGNOLI, M. C., PERIS, K. & LANDI, M. T. 2007. Comprehensive evaluation of allele frequency differences of MC1R variants across populations. *Hum Mutat,* 28**,** 495-505.

GIBLIN, A. V. & THOMAS, J. M. 2007. Incidence, mortality and survival in cutaneous melanoma. *J Plast Reconstr Aesthet Surg,* 60**,** 32-40.

References

GIBNEY, G. T., MESSINA, J. L., FEDORENKO, I. V., SONDAK, V. K. & SMALLEY, K. S. 2013. Paradoxical oncogenesis--the long-term effects of BRAF inhibition in melanoma. *Nat Rev Clin Oncol,* 10**,** 390-9.

GILLGRASS, A., GILL, N., BABIAN, A. & ASHKAR, A. A. 2014. The absence or overexpression of IL-15 drastically alters breast cancer metastasis via effects on NK cells, CD4 T cells, and macrophages. *J Immunol,* 193**,** 6184-91.

GLOBOCAN. 2017. *GLOBOCAN* [Online]. Available: http://globocan.iarc.fr/Default.aspx [Accessed].

GLOSTER, H. M., JR. & NEAL, K. 2006. Skin cancer in skin of color. *J Am Acad Dermatol,* 55**,** 741-60; quiz 761-4.

GODDEN, D., BRENNAN, P. A. & MILNE, J. 2010. Update on melanoma: the present position. *Br J Oral Maxillofac Surg,* 48**,** 575-8.

GOEL, V. K., LAZAR, A. J., WARNEKE, C. L., REDSTON, M. S. & HALUSKA, F. G. 2006. Examination of mutations in BRAF, NRAS, and PTEN in primary cutaneous melanoma. *J Invest Dermatol,* 126**,** 154-60.

GOEPFERT, H., DICHTEL, W. J., MEDINA, J. E., LINDBERG, R. D. & LUNA, M. D. 1984. Perineural invasion in squamous cell skin carcinoma of the head and neck. *Am J Surg,* 148**,** 542-7.

GOGAS, H., EGGERMONT, A. M., HAUSCHILD, A., HERSEY, P., MOHR, P., SCHADENDORF, D., SPATZ, A. & DUMMER, R. 2009. Biomarkers in melanoma. *Ann Oncol,* 20 Suppl 6**,** vi8-13.

GOMEZ DE AGUERO, M., VOCANSON, M., HACINI-RACHINEL, F., TAILLARDET, M., SPARWASSER, T., KISSENPFENNIG, A., MALISSEN, B., KAISERLIAN, D. & DUBOIS, B. 2012. Langerhans cells protect from allergic contact dermatitis in mice by tolerizing CD8(+) T cells and activating Foxp3(+) regulatory T cells. *J Clin Invest,* 122**,** 1700-11.

GOON, P. K., GREENBERG, D. C., IGALI, L. & LEVELL, N. J. 2016. Squamous Cell Carcinoma of the Skin has More Than Doubled Over the Last Decade in the UK. *Acta Derm Venereol,* 96**,** 820-1.

GORHAM, E. D., MOHR, S. B., GARLAND, C. F., CHAPLIN, G. & GARLAND, F. C. 2007. Do sunscreens increase risk of melanoma in populations residing at higher latitudes? *Ann Epidemiol,* 17**,** 956-63.

GOULD ROTHBERG, B. E., BRACKEN, M. B. & RIMM, D. L. 2009. Tissue biomarkers for prognosis in cutaneous melanoma: a systematic review and meta-analysis. *J Natl Cancer Inst,* 101**,** 452-74.

GREAVES, M. & MALEY, C. C. 2012. Clonal evolution in cancer. *Nature,* 481**,** 306-13.

GREEN, A., NEALE, R., KELLY, R., SMITH, I., ABLETT, E., MEYERS, B. & PARSONS, P. 1996. An animal model for human melanoma. *Photochem Photobiol,* 64**,** 577-80.

GREENMAN, C., STEPHENS, P., SMITH, R., DALGLIESH, G. L., HUNTER, C., BIGNELL, G., DAVIES, H., TEAGUE, J., BUTLER, A., STEVENS, C., EDKINS, S., O'MEARA, S., VASTRIK, I., SCHMIDT, E. E., AVIS, T., BARTHORPE, S., BHAMRA, G., BUCK, G., CHOUDHURY, B., CLEMENTS, J., COLE, J., DICKS, E., FORBES, S., GRAY, K., HALLIDAY, K., HARRISON, R., HILLS, K., HINTON, J., JENKINSON, A., JONES, D., MENZIES, A., MIRONENKO, T., PERRY, J., RAINE, K., RICHARDSON, D., SHEPHERD, R., SMALL, A., TOFTS, C., VARIAN, J., WEBB, T., WEST, S., WIDAA, S., YATES, A., CAHILL, D. P., LOUIS, D. N., GOLDSTRAW, P., NICHOLSON, A. G., BRASSEUR, F., LOOIJENGA, L., WEBER, B. L., CHIEW, Y. E., DEFAZIO, A., GREAVES, M. F., GREEN, A. R., CAMPBELL, P., BIRNEY, E., EASTON, D. F., CHENEVIX-TRENCH, G., TAN, M. H., KHOO, S. K.,

TEH, B. T., YUEN, S. T., LEUNG, S. Y., WOOSTER, R., FUTREAL, P. A. & STRATTON, M. R. 2007. Patterns of somatic mutation in human cancer genomes. *Nature,* 446**,** 153-8.

GRIEWANK, K. G. 2016. Biomarkers in melanoma. *Scand J Clin Lab Invest Suppl,* 245**,** S104-12.

GROSS, A., NIEMETZ-RAHN, A., NONNENMACHER, A., TUCHOLSKI, J., KEILHOLZ, U. & FUSI, A. 2015. Expression and activity of EGFR in human cutaneous melanoma cell lines and influence of vemurafenib on the EGFR pathway. *Target Oncol,* 10**,** 77-84.

GUAN, M., CHEN, X., MA, Y., TANG, L., GUAN, L., REN, X., YU, B., ZHANG, W. & SU, B. 2015. MDA-9 and GRP78 as potential diagnostic biomarkers for early detection of melanoma metastasis. *Tumour Biol,* 36**,** 2973-82.

GUILHAUS, M., MLYNSKI, V. & SELBY, D. 1997. Perfect timing: Time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry,* 11**,** 951-962.

GULUBOVA, M. & VLAYKOVA, T. 2006. Immunohistochemical assessment of fibronectin and tenascin and their integrin receptors alpha5beta1 and alpha9beta1 in gastric and colorectal cancers with lymph node and liver metastases. *Acta Histochem,* 108**,** 25-35.

GUO, T., WANG, W., RUDNICK, P. A., SONG, T., LI, J., ZHUANG, Z., WEIL, R. J., DEVOE, D. L., LEE, C. S. & BALGLEY, B. M. 2007. Proteome analysis of microdissected formalin-fixed and paraffin-embedded tissue specimens. *J Histochem Cytochem,* 55**,** 763-72.

GUO, X., SHI, Y., GOU, Y., LI, J., HAN, S., ZHANG, Y., HUO, J., NING, X., SUN, L., CHEN, Y., SUN, S. & FAN, D. 2011. Human ribosomal protein S13 promotes gastric cancer growth through down-regulating p27(Kip1). *J Cell Mol Med,* 15**,** 296-306.

GUPTA, S., TRAN, T., LUO, W., PHUNG, D., KENNEDY, R. L., BROAD, A., CAMPBELL, D., KIPP, D., SINGH, M., KHASRAW, M., MATHESON, L., ASHLEY, D. M. & VENKATESH, S. 2014. Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ Open,* 4**,** e004007.

HAASS, N. K., SMALLEY, K. S., LI, L. & HERLYN, M. 2005. Adhesion, migration and communication in melanocytes and melanoma. *Pigment Cell Res,* 18**,** 150-9.

HADDADEEN, C., LAI, C., CHO, S. Y. & HEALY, E. 2015. Variants of the melanocortin-1 receptor: do they matter clinically? *Exp Dermatol,* 24**,** 5-9.

HAIGH, P. I., DIFRONZO, L. A. & MCCREADY, D. R. 2003. Optimal excision margins for primary cutaneous melanoma: a systematic review and meta-analysis. *Can J Surg,* 46**,** 419-26.

HALAOUI, R. & MCCAFFREY, L. 2015. Rewiring cell polarity signaling in cancer. *Oncogene,* 34**,** 939-50.

HALUSKA, F. G., TSAO, H., WU, H., HALUSKA, F. S., LAZAR, A. & GOEL, V. 2006. Genetic alterations in signaling pathways in melanoma. *Clin Cancer Res,* 12**,** 2301s-2307s.

HAMES, B. D. 1998. *Gel electrophoresis of proteins: a practical approach*, OUP Oxford.

HAMMER, E., ERNST, F. D., THIELE, A., KARANAM, N. K., KUJATH, C., EVERT, M., VOLKER, U. & BARTHLEN, W. 2014. Kidney protein profiling of Wilms' tumor patients by analysis of formalin-fixed paraffin-embedded tissue samples. *Clin Chim Acta,* 433**,** 235-41.

HANAHAN, D. & WEINBERG, R. A. 2000. The Hallmarks of Cancer. *Cell,* 100**,** 57-70.

HANAHAN, D. & WEINBERG, R. A. 2011. Hallmarks of cancer: the next generation. *Cell,* 144**,** 646-74.

References

HARLAND, M., CUST, A. E., BADENAS, C., CHANG, Y. M., HOLLAND, E. A., AGUILERA, P., AITKEN, J. F., ARMSTRONG, B. K., BARRETT, J. H., CARRERA, C., CHAN, M., GASCOYNE, J., GILES, G. G., AGHA-HAMILTON, C., HOPPER, J. L., JENKINS, M. A., KANETSKY, P. A., KEFFORD, R. F., KOLM, I., LOWERY, J., MALVEHY, J., OGBAH, Z., PUIG-BUTILLE, J. A., ORIHUELA-SEGALES, J., RANDERSON-MOOR, J. A., SCHMID, H., TAYLOR, C. F., WHITAKER, L., BISHOP, D. T., MANN, G. J., NEWTON-BISHOP, J. A. & PUIG, S. 2014. Prevalence and predictors of germline CDKN2A mutations for melanoma cases from Australia, Spain and the United Kingdom. *Hered Cancer Clin Pract,* 12**,** 20.

HASEGAWA, S., FURUKAWA, Y., LI, M., SATOH, S., KATO, T., WATANABE, T., KATAGIRI, T., TSUNODA, T., YAMAOKA, Y. & NAKAMURA, Y. 2002. Genome-wide analysis of gene expression in intestinal-type gastric cancers using a complementary DNA microarray representing 23,040 genes. *Cancer Res,* 62**,** 7012-7.

HEALY, E., FLANNAGAN, N., RAY, A., TODD, C., JACKSON, I. J., MATTHEWS, J. N., BIRCH-MACHIN, M. A. & REES, J. L. 2000. Melanocortin-1-receptor gene and sun sensitivity in individuals without red hair. *Lancet,* 355**,** 1072-3.

HECK, D. E., VETRANO, A. M., MARIANO, T. M. & LASKIN, J. D. 2003. UVB light stimulates production of reactive oxygen species: unexpected role for catalase. *J Biol Chem,* 278**,** 22432-6.

HO, C. S., LAM, C. W., CHAN, M. H., CHEUNG, R. C., LAW, L. K., LIT, L. C., NG, K. F., SUEN, M. W. & TAI, H. L. 2003. Electrospray ionisation mass spectrometry: principles and clinical applications. *Clin Biochem Rev,* 24**,** 3-12.

HODIS, E., WATSON, I. R., KRYUKOV, G. V., AROLD, S. T., IMIELINSKI, M., THEURILLAT, J. P., NICKERSON, E., AUCLAIR, D., LI, L., PLACE, C., DICARA, D., RAMOS, A. H., LAWRENCE, M. S., CIBULSKIS, K., SIVACHENKO, A., VOET, D., SAKSENA, G., STRANSKY, N., ONOFRIO, R. C., WINCKLER, W., ARDLIE, K., WAGLE, N., WARGO, J., CHONG, K., MORTON, D. L., STEMKE-HALE, K., CHEN, G., NOBLE, M., MEYERSON, M., LADBURY, J. E., DAVIES, M. A., GERSHENWALD, J. E., WAGNER, S. N., HOON, D. S., SCHADENDORF, D., LANDER, E. S., GABRIEL, S. B., GETZ, G., GARRAWAY, L. A. & CHIN, L. 2012. A landscape of driver mutations in melanoma. *Cell,* 150**,** 251-63.

HOFFMAN, E. A., FREY, B. L., SMITH, L. M. & AUBLE, D. T. 2015. Formaldehyde crosslinking: a tool for the study of chromatin complexes. *J Biol Chem,* 290**,** 26404-11.

HORNBROOK, M. C., GOSHEN, R., CHOMAN, E., O'KEEFFE-ROSETTI, M., KINAR, Y., LILES, E. G. & RUST, K. C. 2017. Early Colorectal Cancer Detected by Machine Learning Model Using Gender, Age, and Complete Blood Count Data. *Dig Dis Sci,* 62**,** 2719-2727.

HOSHINO, A., COSTA-SILVA, B., SHEN, T. L., RODRIGUES, G., HASHIMOTO, A., TESIC MARK, M., MOLINA, H., KOHSAKA, S., DI GIANNATALE, A., CEDER, S., SINGH, S., WILLIAMS, C., SOPLOP, N., URYU, K., PHARMER, L., KING, T., BOJMAR, L., DAVIES, A. E., ARARSO, Y., ZHANG, T., ZHANG, H., HERNANDEZ, J., WEISS, J. M., DUMONT-COLE, V. D., KRAMER, K., WEXLER, L. H., NARENDRAN, A., SCHWARTZ, G. K., HEALEY, J. H., SANDSTROM, P., LABORI, K. J., KURE, E. H., GRANDGENETT, P. M., HOLLINGSWORTH, M. A., DE SOUSA, M., KAUR, S., JAIN, M., MALLYA, K., BATRA, S. K., JARNAGIN, W. R., BRADY, M. S., FODSTAD, O., MULLER, V., PANTEL, K., MINN, A. J., BISSELL, M. J., GARCIA, B. A., KANG, Y., RAJASEKHAR, V. K., GHAJAR, C. M., MATEI, I., PEINADO, H., BROMBERG, J. & LYDEN, D. 2015. Tumour exosome integrins determine organotropic metastasis. *Nature,* 527**,** 329-35.

IIZUKA, H. 1995. Epidermal Architecture That Depends on Turnover Time. *Journal of Dermatological Science,* 10**,** 220-223.

IKEHATA, H., KAWAI K FAU - KOMURA, J.-I., KOMURA J FAU - SAKATSUME, K., SAKATSUME K FAU - WANG, L., WANG L FAU - IMAI, M., IMAI M FAU - HIGASHI, S., HIGASHI S FAU - NIKAIDO, O., NIKAIDO O FAU - YAMAMOTO, K., YAMAMOTO K FAU - HIEDA, K., HIEDA K FAU - WATANABE, M., WATANABE M FAU - KASAI, H., KASAI H FAU - ONO, T. & ONO, T. 2008. UVA1 genotoxicity is mediated not by oxidative damage but by cyclobutane pyrimidine dimers in normal mouse skin.

ILMONEN, S., JAHKOLA, T., TURUNEN, J. P., MUHONEN, T. & ASKO-SELJAVAARA, S. 2004. Tenascin-C in primary malignant melanoma of the skin. *Histopathology,* 45**,** 405-11.

IOACHIM, E., CHARCHANTI, A., BRIASOULIS, E., KARAVASILIS, V., TSANOU, H., ARVANITIS, D. L., AGNANTIS, N. J. & PAVLIDIS, N. 2002. Immunohistochemical expression of extracellular matrix components tenascin, fibronectin, collagen type IV and laminin in breast cancer: their prognostic value and role in tumour invasion and progression. *Eur J Cancer,* 38**,** 2362-70.

ITO, S. & WAKAMATSU, K. 2003. Quantitative analysis of eumelanin and pheomelanin in humans, mice, and other animals: a comparative review. *Pigment Cell Res,* 16**,** 523-31.

JABLONSKI, N. G. & CHAPLIN, G. 2010. Human skin pigmentation as an adaptation to UV radiation. *Proceedings of the National Academy of Sciences of the United States of America,* 107**,** 8962-8968.

JACOBS, J. M., ADKINS, J. N., QIAN, W. J., LIU, T., SHEN, Y., CAMP, D. G., 2ND & SMITH, R. D. 2005. Utilizing human blood plasma for proteomic biomarker discovery. *J Proteome Res,* 4**,** 1073-85.

JAIN, M. R., LI, Q., LIU, T., RINAGGIO, J., KETKAR, A., TOURNIER, V., MADURA, K., ELKABES, S. & LI, H. 2012. Proteomic identification of immunoproteasome accumulation in formalin-fixed rodent spinal cords with experimental autoimmune encephalomyelitis. *J Proteome Res,* 11**,** 1791-803.

JAIN, M. R., LIU, T., HU, J., DARFLER, M., FITZHUGH, V., RINAGGIO, J. & LI, H. 2008. Quantitative Proteomic Analysis of Formalin Fixed Paraffin Embedded Oral HPV Lesions from HIV Patients. *Open Proteomics J,* 1**,** 40-45.

JANES, S. M. & WATT, F. M. 2006. New roles for integrins in squamous-cell carcinoma. *Nat Rev Cancer,* 6**,** 175-83.

JENSEN, E. C. 2013. Quantitative analysis of histological staining and fluorescence using ImageJ. *Anat Rec (Hoboken),* 296**,** 378-81.

JIANG, X., JIANG, X., FENG, S., TIAN, R., YE, M. & ZOU, H. 2007. Development of efficient protein extraction methods for shotgun proteome analysis of formalin-fixed tissues. *J Proteome Res,* 6**,** 1038-47.

JONES, R., RUAS, M., GREGORY, F., MOULIN, S., DELIA, D., MANOUKIAN, S., ROWE, J., BROOKES, S. & PETERS, G. 2007. A CDKN2A mutation in familial melanoma that abrogates binding of p16INK4a to CDK4 but not CDK6. *Cancer Res,* 67**,** 9134-41.

JUNG, K., GANNOUN, A., SITEK, B., MEYER, H. E., STÜHLER, K. & URFER, W. 2005. Analysis of dynamic protein expression data. *RevStat-Statistical Journal,* 3**,** 99-111.

KALLURI, R. & WEINBERG, R. A. 2009. The basics of epithelial-mesenchymal transition. *J Clin Invest,* 119**,** 1420-8.

References

KANEHISA, M. & GOTO, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res,* 28**,** 27-30.

KAPLAN, R. N., RIBA, R. D., ZACHAROULIS, S., BRAMLEY, A. H., VINCENT, L., COSTA, C., MACDONALD, D. D., JIN, D. K., SHIDO, K., KERNS, S. A., ZHU, Z., HICKLIN, D., WU, Y., PORT, J. L., ALTORKI, N., PORT, E. R., RUGGERO, D., SHMELKOV, S. V., JENSEN, K. K., RAFII, S. & LYDEN, D. 2005. VEGFR1-positive haematopoietic bone marrow progenitors initiate the pre-metastatic niche. *Nature,* 438**,** 820-7.

KARAGAS, M. R., CUSHING, G. L., JR., GREENBERG, E. R., MOTT, L. A., SPENCER, S. K. & NIERENBERG, D. W. 2001. Non-melanoma skin cancers and glucocorticoid therapy. *Br J Cancer,* 85**,** 683-6.

KARAGAS, M. R., MCDONALD, J. A., GREENBERG, E. R., STUKEL, T. A., WEISS, J. E., BARON, J. A. & STEVENS, M. M. 1996. Risk of basal cell and squamous cell skin cancers after ionizing radiation therapy. For The Skin Cancer Prevention Study Group. *J Natl Cancer Inst,* 88**,** 1848-53.

KARAGAS, M. R., STANNARD, V. A., MOTT, L. A., SLATTERY, M. J., SPENCER, S. K. & WEINSTOCK, M. A. 2002. Use of tanning devices and risk of basal cell and squamous cell skin cancers. *Journal of the National Cancer Institute,* 94**,** 224-226.

KARAS, M. & HILLENKAMP, F. 1988. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem,* 60**,** 2299-301.

KARIA, P. S., JAMBUSARIA-PAHLAJANI, A., HARRINGTON, D. P., MURPHY, G. F., QURESHI, A. A. & SCHMULTS, C. D. 2014. Evaluation of American Joint Committee on Cancer, International Union Against Cancer, and Brigham and Women's Hospital tumor staging for cutaneous squamous cell carcinoma. *J Clin Oncol,* 32**,** 327-34.

KARP, N. A. & LILLEY, K. S. 2007. Design and analysis issues in quantitative proteomics studies. *Proteomics,* 7 Suppl 1**,** 42-50.

KASHANI-SABET, M. 2014. Molecular markers in melanoma. *Br J Dermatol,* 170**,** 31-5.

KATO, M., TAKAHASHI, M., AKHAND, A. A., LIU, W., DAI, Y., SHIMIZU, S., IWAMOTO, T., SUZUKI, H. & NAKASHIMA, I. 1998. Transgenic mouse model for skin malignant melanoma. *Oncogene,* 17**,** 1885-8.

KAVAKIOTIS, I., TSAVE, O., SALIFOGLOU, A., MAGLAVERAS, N., VLAHAVAS, I. & CHOUVARDA, I. 2017. Machine Learning and Data Mining Methods in Diabetes Research. *Comput Struct Biotechnol J,* 15**,** 104-116.

KAWAMURA, T., NOMURA, M., TOJO, H., FUJII, K., HAMASAKI, H., MIKAMI, S., BANDO, Y., KATO, H. & NISHIMURA, T. 2010. Proteomic analysis of laser-microdissected paraffin-embedded tissues: (1) Stage-related protein candidates upon non-metastatic lung adenocarcinoma. *J Proteomics,* 73**,** 1089-99.

KAZIANIS, S., MORIZOT, D. C., COLETTA, L. D., JOHNSTON, D. A., WOOLCOCK, B., VIELKIND, J. R. & NAIRN, R. S. 1999. Comparative structure and characterization of a CDKN2 gene in a Xiphophorus fish melanoma model. *Oncogene,* 18**,** 5088-99.

KELFKENS, G., DE GRUIJL, F. R. & VAN DER LEUN, J. C. 1991. Tumorigenesis by short-wave ultraviolet A: papillomas versus squamous cell carcinomas. *Carcinogenesis,* 12**,** 1377-82.

KHANNA, C., LINDBLAD-TOH, K., VAIL, D., LONDON, C., BERGMAN, P., BARBER, L., BREEN, M., KITCHELL, B., MCNEIL, E., MODIANO, J. F., NIEMI, S., COMSTOCK, K. E., OSTRANDER, E., WESTMORELAND, S. & WITHROW, S. 2006. The dog as a cancer model. *Nature Biotechnology,* 24**,** 1065-1066.

KHATRI, P. & DRAGHICI, S. 2005. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics,* 21**,** 3587-95.

KHATRI, P., SIROTA, M. & BUTTE, A. J. 2012. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol,* 8**,** e1002375.

KIM, A. L., LABASI, J. M., ZHU, Y., TANG, X., MCCLURE, K., GABEL, C. A., ATHAR, M. & BICKERS, D. R. 2005. Role of p38 MAPK in UVB-induced inflammatory responses in the skin of SKH-1 hairless mice. *J Invest Dermatol,* 124**,** 1318-25.

KIROVA, Y. M., CHEN, J., RABARIJAONA, L. I., PIEDBOIS, Y. & LE BOURGEOIS, J. P. 1999. Radiotherapy as palliative treatment for metastatic melanoma. *Melanoma Res,* 9**,** 611-3.

KLEIN, M., LOTEM, M., PERETZ, T., ZWAS, S. T., MIZRACHI, S., LIBERMAN, Y., CHISIN, R., SCHACHTER, J., RON, I. G., IOSILEVSKY, G., KENNEDY, J. A., REVSKAYA, E., DE KATER, A. W., BANAGA, E., KLUTZARITZ, V., FRIEDMANN, N., GALUN, E., DENARDO, G. L., DENARDO, S. J., CASADEVALL, A., DADACHOVA, E. & THORNTON, G. B. 2013. Safety and Efficacy of 188-Rhenium-Labeled Antibody to Melanin in Patients with Metastatic Melanoma. *Journal of Skin Cancer,* 2013**,** 8.

KOJIMA, K., BOWERSOCK, G. J., KOJIMA, C., KLUG, C. A., GRIZZLE, W. E. & MOBLEY, J. A. 2012. Validation of a robust proteomic analysis carried out on formalin-fixed paraffin-embedded tissues of the pancreas obtained from mouse and human. *Proteomics,* 12**,** 3393-402.

KONSTANTAKOU, E. G., VELENTZAS, A. D., ANAGNOSTOPOULOS, A. K., LITOU, Z. I., KONSTANDI, O. A., GIANNOPOULOU, A. F., ANASTASIADOU, E., VOUTSINAS, G. E., TSANGARIS, G. T. & STRAVOPODIS, D. J. 2017. Deep-proteome mapping of WM-266-4 human metastatic melanoma cells: From oncogenic addiction to druggable targets. *PLoS One,* 12**,** e0171512.

KOOPMANN, J., ZHANG, Z., WHITE, N., ROSENZWEIG, J., FEDARKO, N., JAGANNATH, S., CANTO, M. I., YEO, C. J., CHAN, D. W. & GOGGINS, M. 2004. Serum diagnosis of pancreatic adenocarcinoma using surface-enhanced laser desorption and ionization mass spectrometry. *Clinical Cancer Research,* 10**,** 860-868.

KOROTCOV, A., TKACHENKO, V., RUSSO, D. P. & EKINS, S. 2017. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol Pharm,* 14**,** 4462-4475.

KRAEMER, K. H., LEE, M. M., ANDREWS, A. D. & LAMBERT, W. C. 1994. The role of sunlight and DNA repair in melanoma and nonmelanoma skin cancer. The xeroderma pigmentosum paradigm. *Arch Dermatol,* 130**,** 1018-21.

KRAMER, M., STEIN, B., MAI, S., KUNZ, E., KONIG, H., LOFERER, H., GRUNICKE, H. H., PONTA, H., HERRLICH, P. & RAHMSDORF, H. J. 1990. Radiation-induced activation of transcription factors in mammalian cells. *Radiat Environ Biophys,* 29**,** 303-13.

KRIPKE, M. L. 1974. Antigenicity of murine skin tumors induced by ultraviolet light. *J Natl Cancer Inst,* 53**,** 1333-6.

KRIPKE, M. L. 1977. Latency, histology, and antigenicity of tumors induced by ultraviolet light in three inbred mouse strains. *Cancer Res,* 37**,** 1395-400.

References

KUME, H., MURAOKA, S., KUGA, T., ADACHI, J., NARUMI, R., WATANABE, S., KUWANO, M., KODERA, Y., MATSUSHITA, K., FUKUOKA, J., MASUDA, T., ISHIHAMA, Y., MATSUBARA, H., NOMURA, F. & TOMONAGA, T. 2014. Discovery of colorectal cancer biomarker candidates by membrane proteomic analysis and subsequent verification using selected reaction monitoring (SRM) and tissue microarray (TMA) analysis. *Mol Cell Proteomics,* 13**,** 1471-84.

KUSEBAUCH, U., HERNANDEZ-CASTELLANO, L. E., BISLEV, S. L., MORITZ, R. L., RONTVED, C. M. & BENDIXEN, E. 2018. Selected reaction monitoring mass spectrometry of mastitis milk reveals pathogen-specific regulation of bovine host response proteins. *J Dairy Sci,* 101**,** 6532-6541.

LAI, C., AUGUST, S., ALBIBAS, A., BEHAR, R., CHO, S. Y., POLAK, M. E., THEAKER, J., MACLEOD, A. S., FRENCH, R. R., GLENNIE, M. J., AL-SHAMKHANI, A. & HEALY, E. 2016. OX40+ Regulatory T Cells in Cutaneous Squamous Cell Carcinoma Suppress Effector T-Cell Responses and Associate with Metastatic Potential. *Clin Cancer Res,* 22**,** 4236-48.

LAI, C., AUGUST, S., BEHAR, R., POLAK, M., ARDERN-JONES, M., THEAKER, J., AL-SHAMKHANI, A. & HEALY, E. 2015. Characteristics of immunosuppressive regulatory T cells in cutaneous squamous cell carcinomas and role in metastasis. *Lancet,* 385 Suppl 1**,** S59.

LANGE, V., PICOTTI, P., DOMON, B. & AEBERSOLD, R. 2008. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol,* 4**,** 222.

LAO, J., CHEN, Y., LI, Z. C., LI, Q., ZHANG, J., LIU, J. & ZHAI, G. 2017. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Sci Rep,* 7**,** 10353.

LARRANAGA, P., CALVO, B., SANTANA, R., BIELZA, C., GALDIANO, J., INZA, I., LOZANO, J. A., ARMANANZAS, R., SANTAFE, G., PEREZ, A. & ROBLES, V. 2006. Machine learning in bioinformatics. *Brief Bioinform,* 7**,** 86-112.

LE QUESNE, J. P., SPRIGGS, K. A., BUSHELL, M. & WILLIS, A. E. 2010. Dysregulation of protein synthesis and disease. *J Pathol,* 220**,** 140-51.

LEFFELL, D. J. 2000. The scientific basis of skin cancer. *J Am Acad Dermatol,* 42**,** 18-22.

LEVINE, N., MOON, T. E., CARTMEL, B., BANGERT, J. L., RODNEY, S., DONG, Q., PENG, Y. M. & ALBERTS, D. S. 1997. Trial of retinol and isotretinoin in skin cancer prevention: a randomized, double-blind, controlled trial. Southwest Skin Cancer Prevention Study Group. *Cancer Epidemiol Biomarkers Prev,* 6**,** 957-61.

LEWIS, J. M., BURGLER, C. D., FREUDZON, M., GOLUBETS, K., GIBSON, J. F., FILLER, R. B. & GIRARDI, M. 2015. Langerhans Cells Facilitate UVB-Induced Epidermal Carcinogenesis. *J Invest Dermatol,* 135**,** 2824-33.

LEY, R. D. 1984. Photorepair of pyrimidine dimers in the epidermis of the marsupial Monodelphis domestica. *Photochem Photobiol,* 40**,** 141-3.

LEY, R. D. 2002. Animal models of ultraviolet radiation (UVR)-induced cutaneous melanoma. *Frontiers in bioscience : a journal and virtual library,* 7**,** d1531-4.

LI, B., XU, W. W., LAM, A. K. Y., WANG, Y., HU, H. F., GUAN, X. Y., QIN, Y. R., SAREMI, N., TSAO, S. W., HE, Q. Y. & CHEUNG, A. L. M. 2017. Significance of PI3K/AKT signaling pathway in metastasis of esophageal squamous cell carcinoma and its potential as a target for anti-metastasis therapy. *Oncotarget,* 8**,** 38755-38766.

References

LI, L., PAN, X. Y., SHU, J., JIANG, R., ZHOU, Y. J. & CHEN, J. X. 2014. Ribonuclease inhibitor up-regulation inhibits the growth and induces apoptosis in murine melanoma cells through repression of angiogenin and ILK/PI3K/AKT signaling pathway. *Biochimie,* 103**,** 89-100.

LI, Y. Y., HANNA, G. J., LAGA, A. C., HADDAD, R. I., LORCH, J. H. & HAMMERMAN, P. S. 2015. Genomic analysis of metastatic cutaneous squamous cell carcinoma. *Clin Cancer Res,* 21**,** 1447-56.

LIEBLER, D. C. & ZIMMERMAN, L. J. 2013. Targeted quantitation of proteins by mass spectrometry. *Biochemistry,* 52**,** 3797-806.

LIN, C. S., PARK, T., CHEN, Z. P. & LEAVITT, J. 1993. Human plastin genes. Comparative gene structure, chromosome location, and differential expression in normal and neoplastic cells. *J Biol Chem,* 268**,** 2781-92.

LIU, H., SADYGOV, R. G. & YATES, J. R., 3RD 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem,* 76**,** 4193-201.

LOMAS, A., LEONARDI-BEE, J. & BATH-HEXTALL, F. 2012. A systematic review of worldwide incidence of nonmelanoma skin cancer. *Br J Dermatol,* 166**,** 1069-80.

LONG, G. V., FLAHERTY, K. T., STROYAKOVSKIY, D., GOGAS, H., LEVCHENKO, E., DE BRAUD, F., LARKIN, J., GARBE, C., JOUARY, T., HAUSCHILD, A., CHIARION-SILENI, V., LEBBE, C., MANDALA, M., MILLWARD, M., ARANCE, A., BONDARENKO, I., HAANEN, J., HANSSON, J., UTIKAL, J., FERRARESI, V., MOHR, P., PROBACHAI, V., SCHADENDORF, D., NATHAN, P., ROBERT, C., RIBAS, A., DAVIES, M. A., LANE, S. R., LEGOS, J. J., MOOKERJEE, B. & GROB, J. J. 2017. Dabrafenib plus trametinib versus dabrafenib monotherapy in patients with metastatic BRAF V600E/K-mutant melanoma: long-term survival and safety analysis of a phase 3 study. *Ann Oncol,* 28**,** 1631-1639.

LONG, G. V., STROYAKOVSKIY, D., GOGAS, H., LEVCHENKO, E., DE BRAUD, F., LARKIN, J., GARBE, C., JOUARY, T., HAUSCHILD, A., GROB, J. J., CHIARION-SILENI, V., LEBBE, C., MANDALA, M., MILLWARD, M., ARANCE, A., BONDARENKO, I., HAANEN, J. B., HANSSON, J., UTIKAL, J., FERRARESI, V., KOVALENKO, N., MOHR, P., PROBACHAI, V., SCHADENDORF, D., NATHAN, P., ROBERT, C., RIBAS, A., DEMARINI, D. J., IRANI, J. G., SWANN, S., LEGOS, J. J., JIN, F., MOOKERJEE, B. & FLAHERTY, K. 2015. Dabrafenib and trametinib versus dabrafenib and placebo for Val600 BRAF-mutant melanoma: a multicentre, double-blind, phase 3 randomised controlled trial. *Lancet,* 386**,** 444-51.

LORCH, M., MASON, J. M., CLARKE, A. R. & PARKER, M. J. 1999. Effects of core mutations on the folding of a beta-sheet protein: implications for backbone organization in the I-state. *Biochemistry,* 38**,** 1377-85.

LOSQUADRO, W. D. 2017. Anatomy of the Skin and the Pathogenesis of Nonmelanoma Skin Cancer. *Facial Plast Surg Clin North Am,* 25**,** 283-289.

LYDIATT, W. M., PATEL, S. G., O'SULLIVAN, B., BRANDWEIN, M. S., RIDGE, J. A., MIGLIACCI, J. C., LOOMIS, A. M. & SHAH, J. P. 2017. Head and Neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual. *CA Cancer J Clin,* 67**,** 122-137.

LYNCH, C. M., ABDOLLAHI, B., FUQUA, J. D., DE CARLO, A. R., BARTHOLOMAI, J. A., BALGEMANN, R. N., VAN BERKEL, V. H. & FRIEBOES, H. B. 2017. Prediction of lung cancer patient survival via supervised machine learning classification techniques. *Int J Med Inform,* 108**,** 1-8.

MACLEAN, B., TOMAZELA, D. M., SHULMAN, N., CHAMBERS, M., FINNEY, G. L., FREWEN, B., KERN, R., TABB, D. L., LIEBLER, D. C. & MACCOSS, M. J. 2010. Skyline: an open source document

References

editor for creating and analyzing targeted proteomics experiments. *Bioinformatics,* 26**,** 966-8.

MADAN, V., LEAR, J. T. & SZEIMIES, R. M. 2010. Non-melanoma skin cancer. *Lancet,* 375**,** 673-85.

MAES, E., BROECKX, V., MERTENS, I., SAGAERT, X., PRENEN, H., LANDUYT, B. & SCHOOFS, L. 2013. Analysis of the formalin-fixed paraffin-embedded tissue proteome: pitfalls, challenges, and future prospectives. *Amino Acids,* 45**,** 205-18.

MANI, I., SHARMA, V., TAMBOLI, I. & RAMAN, G. 2001. Interaction of Melanin with Proteins - The Importance of an Acidic Intramelanosomal pH. *Pigment Cell Research,* 14**,** 170-179.

MANOLA, J., ATKINS, M., IBRAHIM, J. & KIRKWOOD, J. 2000. Prognostic factors in metastatic melanoma: a pooled analysis of Eastern Cooperative Oncology Group trials. *J Clin Oncol,* 18**,** 3782-93.

MAO, J., LIGON, K. L., RAKHLIN, E. Y., THAYER, S. P., BRONSON, R. T., ROWITCH, D. & MCMAHON, A. P. 2006. A novel somatic mouse model to survey tumorigenic potential applied to the Hedgehog pathway. *Cancer Res,* 66**,** 10171-8.

MARSDEN, J. R., NEWTON-BISHOP, J. A., BURROWS, L., COOK, M., CORRIE, P. G., COX, N. H., GORE, M. E., LORIGAN, P., MACKIE, R., NATHAN, P., PEACH, H., POWELL, B., WALKER, C. & BRITISH ASSOCIATION OF DERMATOLOGISTS CLINICAL STANDARDS, U. 2010. Revised U.K. guidelines for the management of cutaneous melanoma 2010. *Br J Dermatol,* 163**,** 238-56.

MARTINCORENA, I. & CAMPBELL, P. J. 2015. Somatic mutation in cancer and normal cells. *Science,* 349**,** 1483-9.

MARTINEZ, J. C., OTLEY, C. C., STASKO, T., EUVRARD, S., BROWN, C., SCHANBACHER, C. F., WEAVER, A. L. & TRANSPLANT-SKIN CANCER, C. 2003. Defining the clinical course of metastatic skin cancer in organ transplant recipients: a multicenter collaborative study. *Arch Dermatol,* 139**,** 301-6.

MCARTHUR, G. A., CHAPMAN, P. B., ROBERT, C., LARKIN, J., HAANEN, J. B., DUMMER, R., RIBAS, A., HOGG, D., HAMID, O., ASCIERTO, P. A., GARBE, C., TESTORI, A., MAIO, M., LORIGAN, P., LEBBÉ, C., JOUARY, T., SCHADENDORF, D., O'DAY, S. J., KIRKWOOD, J. M., EGGERMONT, A. M., DRÉNO, B., SOSMAN, J. A., FLAHERTY, K. T., YIN, M., CARO, I., CHENG, S., TRUNZER, K. & HAUSCHILD, A. 2014. Safety and efficacy of vemurafenib in BRAFV600E and BRAFV600K mutation-positive melanoma (BRIM-3): extended follow-up of a phase 3, randomised, open-label study. *The Lancet Oncology,* 15**,** 323-332.

MCCULLOCH, W. S. & PITTS, W. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics,* 5**,** 115-133.

MCLAFFERTY, F. W. 1981. Tandem mass spectrometry. *Science,* 214**,** 280-287.

MCLENDON, R. E., WIKSTRAND, C. J., MATTHEWS, M. R., AL-BARADEI, R., BIGNER, S. H. & BIGNER, D. D. 2000. Glioma-associated antigen expression in oligodendroglial neoplasms. Tenascin and epidermal growth factor receptor. *J Histochem Cytochem,* 48**,** 1103-10.

MEIER, F., WILL, S., ELLWANGER, U., SCHLAGENHAUFF, B., SCHITTEK, B., RASSNER, G. & GARBE, C. 2002. Metastatic pathways and time courses in the orderly progression of cutaneous melanoma. *Br J Dermatol,* 147**,** 62-70.

References

MERAD, M., MANZ, M. G., KARSUNKY, H., WAGERS, A., PETERS, W., CHARO, I., WEISSMAN, I. L., CYSTER, J. G. & ENGLEMAN, E. G. 2002. Langerhans cells renew in the skin throughout life under steady-state conditions. *Nat Immunol,* 3**,** 1135-41.

METZ, B., KERSTEN, G. F., HOOGERHOUT, P., BRUGGHE, H. F., TIMMERMANS, H. A., DE JONG, A., MEIRING, H., TEN HOVE, J., HENNINK, W. E., CROMMELIN, D. J. & JISKOOT, W. 2004. Identification of formaldehyde-induced modifications in proteins: reactions with model peptides. *J Biol Chem,* 279**,** 6235-43.

MIAN, S., UGUREL, S., PARKINSON, E., SCHLENZKA, I., DRYDEN, I., LANCASHIRE, L., BALL, G., CREASER, C., REES, R. & SCHADENDORF, D. 2005. Serum proteomic fingerprinting discriminates between clinical stages and predicts disease progression in melanoma patients. *J Clin Oncol,* 23**,** 5088-93.

MICHAYLIRA, C. Z., WONG, G. S., MILLER, C. G., GUTIERREZ, C. M., NAKAGAWA, H., HAMMOND, R., KLEIN-SZANTO, A. J., LEE, J. S., KIM, S. B., HERLYN, M., DIEHL, J. A., GIMOTTY, P. & RUSTGI, A. K. 2010. Periostin, a cell adhesion molecule, facilitates invasion in the tumor microenvironment and annotates a novel tumor-invasive signature in esophageal cancer. *Cancer Res,* 70**,** 5281-92.

MIGDEN, M. R., RISCHIN, D., SCHMULTS, C. D., GUMINSKI, A., HAUSCHILD, A., LEWIS, K. D., CHUNG, C. H., HERNANDEZ-AYA, L., LIM, A. M., CHANG, A. L. S., RABINOWITS, G., THAI, A. A., DUNN, L. A., HUGHES, B. G. M., KHUSHALANI, N. I., MODI, B., SCHADENDORF, D., GAO, B., SEEBACH, F., LI, S., LI, J., MATHIAS, M., BOOTH, J., MOHAN, K., STANKEVICH, E., BABIKER, H. M., BRANA, I., GIL-MARTIN, M., HOMSI, J., JOHNSON, M. L., MORENO, V., NIU, J., OWONIKOKO, T. K., PAPADOPOULOS, K. P., YANCOPOULOS, G. D., LOWY, I. & FURY, M. G. 2018. PD-1 Blockade with Cemiplimab in Advanced Cutaneous Squamous-Cell Carcinoma. *N Engl J Med,* 379**,** 341-351.

MILLER, P. E. & DENTON, M. B. 1986. The Quadrupole Mass Filter - Basic Operating Concepts. *Journal of Chemical Education,* 63**,** 617-622.

MILLIKAN, L. E., BOYLON, J. L., HOOK, R. R. & MANNING, P. J. 1974. Melanoma in Sinclair swine: a new animal model. *J Invest Dermatol,* 62**,** 20-30.

MISSERO, C. & ANTONINI, D. 2014. Crosstalk among p53 family members in cutaneous carcinoma. *Exp Dermatol,* 23**,** 143-6.

MITCHELL WELLS, J. & MCLUCKEY, S. A. 2005. Collision‐Induced Dissociation (CID) of Peptides and Proteins. *Methods in Enzymology.* Academic Press.

MOCELLIN, S., PASQUALI, S., ROSSI, C. R. & NITTI, D. 2010. Interferon alpha adjuvant therapy in patients with high-risk melanoma: a systematic review and meta-analysis. *J Natl Cancer Inst,* 102**,** 493-501.

MODI, B. G., NEUSTADTER, J., BINDA, E., LEWIS, J., FILLER, R. B., ROBERTS, S. J., KWONG, B. Y., REDDY, S., OVERTON, J. D., GALAN, A., TIGELAAR, R., CAI, L., FU, P., SHLOMCHIK, M., KAPLAN, D. H., HAYDAY, A. & GIRARDI, M. 2012. Langerhans cells facilitate epithelial DNA damage and squamous cell carcinoma. *Science,* 335**,** 104-8.

MOISSOGLU, K., MCROBERTS, K. S., MEIER, J. A., THEODORESCU, D. & SCHWARTZ, M. A. 2009. Rho GDP dissociation inhibitor 2 suppresses metastasis via unconventional regulation of RhoGTPases. *Cancer Res,* 69**,** 2838-44.

MOON, R. T., KOHN, A. D., DE FERRARI, G. V. & KAYKAS, A. 2004. WNT and beta-catenin signalling: diseases and therapies. *Nat Rev Genet,* 5**,** 691-701.

References

MORRA, L. & MOCH, H. 2011. Periostin expression and epithelial-mesenchymal transition in cancer: a review and an update. *Virchows Arch,* 459**,** 465-75.

MORRIS, S., COX, B. & BOSANQUET, N. 2009. Cost of skin cancer in England. *Eur J Health Econ,* 10**,** 267-73.

MOTLEY, R., KERSEY, P., LAWRENCE, C., BRITISH ASSOCIATION OF, D. & BRITISH ASSOCIATION OF PLASTIC, S. 2003. Multiprofessional guidelines for the management of the patient with primary cutaneous squamous cell carcinoma. *Br J Plast Surg,* 56**,** 85-91.

MOTLEY, R., KERSEY, P., LAWRENCE, C., BRITISH ASSOCIATION OF, D., BRITISH ASSOCIATION OF PLASTIC, S. & ROYAL COLLEGE OF RADIOLOGISTS, F. O. C. O. 2002. Multiprofessional guidelines for the management of the patient with primary cutaneous squamous cell carcinoma. *Br J Dermatol,* 146**,** 18-25.

MOZURAITIENE, J., BIELSKIENE, K., ATKOCIUS, V. & LABEIKYTE, D. 2015. Molecular alterations in signal pathways of melanoma and new personalized treatment strategies: Targeting of Notch. *Medicina-Lithuania,* 51**,** 133-145.

MURRAY, C. J., VOS, T., LOZANO, R., NAGHAVI, M., FLAXMAN, A. D., MICHAUD, C., EZZATI, M., SHIBUYA, K., SALOMON, J. A., ABDALLA, S., ABOYANS, V., ABRAHAM, J., ACKERMAN, I., AGGARWAL, R., AHN, S. Y., ALI, M. K., ALVARADO, M., ANDERSON, H. R., ANDERSON, L. M., ANDREWS, K. G., ATKINSON, C., BADDOUR, L. M., BAHALIM, A. N., BARKER-COLLO, S., BARRERO, L. H., BARTELS, D. H., BASANEZ, M. G., BAXTER, A., BELL, M. L., BENJAMIN, E. J., BENNETT, D., BERNABE, E., BHALLA, K., BHANDARI, B., BIKBOV, B., BIN ABDULHAK, A., BIRBECK, G., BLACK, J. A., BLENCOWE, H., BLORE, J. D., BLYTH, F., BOLLIGER, I., BONAVENTURE, A., BOUFOUS, S., BOURNE, R., BOUSSINESQ, M., BRAITHWAITE, T., BRAYNE, C., BRIDGETT, L., BROOKER, S., BROOKS, P., BRUGHA, T. S., BRYAN-HANCOCK, C., BUCELLO, C., BUCHBINDER, R., BUCKLE, G., BUDKE, C. M., BURCH, M., BURNEY, P., BURSTEIN, R., CALABRIA, B., CAMPBELL, B., CANTER, C. E., CARABIN, H., CARAPETIS, J., CARMONA, L., CELLA, C., CHARLSON, F., CHEN, H., CHENG, A. T., CHOU, D., CHUGH, S. S., COFFENG, L. E., COLAN, S. D., COLQUHOUN, S., COLSON, K. E., CONDON, J., CONNOR, M. D., COOPER, L. T., CORRIERE, M., CORTINOVIS, M., DE VACCARO, K. C., COUSER, W., COWIE, B. C., CRIQUI, M. H., CROSS, M., DABHADKAR, K. C., DAHIYA, M., DAHODWALA, N., DAMSERE-DERRY, J., DANAEI, G., DAVIS, A., DE LEO, D., DEGENHARDT, L., DELLAVALLE, R., DELOSSANTOS, A., DENENBERG, J., DERRETT, S., DES JARLAIS, D. C., DHARMARATNE, S. D., et al. 2012. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet,* 380**,** 2197-223.

NAGAHARU, K., ZHANG, X., YOSHIDA, T., KATOH, D., HANAMURA, N., KOZUKA, Y., OGAWA, T., SHIRAISHI, T. & IMANAKA-YOSHIDA, K. 2011. Tenascin C induces epithelial-mesenchymal transition-like change accompanied by SRC activation and focal adhesion kinase phosphorylation in human breast cancer cells. *Am J Pathol,* 178**,** 754-63.

NAGANO, M., HOSHINO, D., KOSHIKAWA, N., AKIZAWA, T. & SEIKI, M. 2012. Turnover of focal adhesions and cancer cell migration. *Int J Cell Biol,* 2012**,** 310616.

NAIDOO, K., JONES, R., DMITROVIC, B., WIJESURIYA, N., KOCHER, H., HART, I. R. & CRNOGORAC-JURCEVIC, T. 2012. Proteome of formalin-fixed paraffin-embedded pancreatic ductal adenocarcinoma and lymph node metastases. *J Pathol,* 226**,** 756-63.

NAIRN, R. S., KAZIANIS, S., MCENTIRE, B. B., DELLA COLETTA, L., WALTER, R. B. & MORIZOT, D. C. 1996. A CDKN2-like polymorphism in Xiphophorus LG V is associated with UV-B-induced melanoma formation in platyfish–swordtail hybrids. *Proceedings of the National Academy of Sciences of the United States of America,* 93**,** 13042-13047.

References

NARAYANAN, D. L., SALADI, R. N. & FOX, J. L. 2010. Ultraviolet radiation and skin cancer. *Int J Dermatol,* 49**,** 978-86.

NASSAR, D., LATIL, M., BOECKX, B., LAMBRECHTS, D. & BLANPAIN, C. 2015. Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma. *Nat Med,* 21**,** 946-54.

NAZARIAN, J., SANTI, M., HATHOUT, Y. & MACDONALD, T. J. 2008. Protein profiling of formalin fixed paraffin embedded tissue: Identification of potential biomarkers for pediatric brainstem glioma. *Proteomics Clin Appl,* 2**,** 915-24.

NEDELEC, B., FORGET, N. J., HURTUBISE, T., CIMINO, S., DE MUSZKA, F., LEGAULT, A., LIU, W. L., DE OLIVEIRA, A., CALVA, V. & CORREA, J. A. 2016. Skin characteristics: normative data for elasticity, erythema, melanin, and thickness at 16 different anatomical locations. *Skin Res Technol,* 22**,** 263-75.

NELDER, J. A. & WEDDERBURN, R. W. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society Series a-General,* 135**,** 370-+.

NIESSEN, W. M. A. 2006. *Liquid chromatography-mass spectrometry*, CRC Press.

NIKOLAEV, S. I., RIMOLDI, D., ISELI, C., VALSESIA, A., ROBYR, D., GEHRIG, C., HARSHMAN, K., GUIPPONI, M., BUKACH, O., ZOETE, V., MICHIELIN, O., MUEHLETHALER, K., SPEISER, D., BECKMANN, J. S., XENARIOS, I., HALAZONETIS, T. D., JONGENEEL, C. V., STEVENSON, B. J. & ANTONARAKIS, S. E. 2011. Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. *Nat Genet,* 44**,** 133-9.

NIRMALAN, N. J., HUGHES, C., PENG, J., MCKENNA, T., LANGRIDGE, J., CAIRNS, D. A., HARNDEN, P., SELBY, P. J. & BANKS, R. E. 2011. Initial development and validation of a novel extraction method for quantitative mining of the formalin-fixed, paraffin-embedded tissue proteome for biomarker investigations. *J Proteome Res,* 10**,** 896-906.

NISHIGORI, C., YAROSH, D. B., ULLRICH, S. E., VINK, A. A., BUCANA, C. D., ROZA, L. & KRIPKE, M. L. 1996. Evidence that DNA damage triggers interleukin 10 cytokine production in UV-irradiated murine keratinocytes. *Proceedings of the National Academy of Sciences of the United States of America,* 93**,** 10354-10359.

NOBLE, W. S. 2009. How does multiple testing correction work? *Nat Biotech,* 27**,** 1135-1137.

NOONAN, F. P., ZAIDI, M. R., WOLNICKA-GLUBISZ, A., ANVER, M. R., BAHN, J., WIELGUS, A., CADET, J., DOUKI, T., MOURET, S., TUCKER, M. A., POPRATILOFF, A., MERLINO, G. & DE FABO, E. C. 2012. Melanoma induction by ultraviolet A but not ultraviolet B radiation requires melanin pigment. *Nat Commun,* 3**,** 884.

O'DEA, D. 2000. The Costs of Skin Cancer to New Zealand. *A Report to the Cancer Society.* Wellington School of Medicine, University of Otago: University of Otago.

O'DELL, B. L., JESSEN, R. T., BECKER, L. E., JACKSON, R. T. & SMITH, E. B. 1980. Diminished immune response in sun-damaged skin. *Arch Dermatol,* 116**,** 559-61.

ORTEGA-MARTINEZ, I., GARDEAZABAL, J., ERRAMUZPE, A., SANCHEZ-DIEZ, A., CORTES, J., GARCIA-VAZQUEZ, M. D., PEREZ-YARZA, G., IZU, R., LUIS DIAZ-RAMON, J., DE LA FUENTE, I. M., ASUMENDI, A. & BOYANO, M. D. 2016. Vitronectin and dermcidin serum levels predict the metastatic progression of AJCC I-II early-stage melanoma. *Int J Cancer,* 139**,** 1598-607.

References

OSAKI, M., OSHIMURA, M. & ITO, H. 2004. PI3K-Akt pathway: its functions and alterations in human cancer. *Apoptosis,* 9**,** 667-76.

OSKARSSON, T., ACHARYYA, S., ZHANG, X. H. F., VANHARANTA, S., TAVAZOIE, S. F., MORRIS, P. G., DOWNEY, R. J., MANOVA-TODOROVA, K., BROGI, E. & MASSAGUE, J. 2011. Breast cancer cells produce tenascin C as a metastatic niche component to colonize the lungs. *Nature Medicine,* 17**,** 867-U256.

OSTASIEWICZ, P., ZIELINSKA, D. F., MANN, M. & WISNIEWSKI, J. R. 2010. Proteome, phosphoproteome, and N-glycoproteome are quantitatively preserved in formalin-fixed paraffin-embedded tissue and analyzable by high-resolution mass spectrometry. *J Proteome Res,* 9**,** 3688-700.

PALMER-TOY, D. E., KRASTINS, B., SARRACINO, D. A., NADOL, J. B., JR. & MERCHANT, S. N. 2005. Efficient method for the proteomic analysis of fixed and embedded tissues. *J Proteome Res,* 4**,** 2404-11.

PAS, J., WYSZKO, E., ROLLE, K., RYCHLEWSKI, L., NOWAK, S., ZUKIEL, R. & BARCISZEWSKI, J. 2006. Analysis of structure and function of tenascin-C. *Int J Biochem Cell Biol,* 38**,** 1594-602.

PASCUTTI, P. G. & ITO, A. S. 1992. EPR study of melanin-protein interaction: photoinduced free radicals and progressive microwave power saturation. *J Photochem Photobiol B,* 16**,** 257-66.

PATEL, V., HOOD, B. L., MOLINOLO, A. A., LEE, N. H., CONRADS, T. P., BRAISTED, J. C., KRIZMAN, D. B., VEENSTRA, T. D. & GUTKIND, J. S. 2008. Proteomic analysis of laser-captured paraffin-embedded tissues: a molecular portrait of head and neck cancer progression. *Clin Cancer Res,* 14**,** 1002-14.

PAULITSCHKE, V., GERNER, C., HOFSTATTER, E., MOHR, T., MAYER, R. L., PEHAMBERGER, H. & KUNSTFELD, R. 2015. Proteome profiling of keratinocytes transforming to malignancy. *Electrophoresis,* 36**,** 564-76.

PAULO, J. A., LEE, L. S., BANKS, P. A., STEEN, H. & CONWELL, D. L. 2012. Proteomic analysis of formalin-fixed paraffin-embedded pancreatic tissue using liquid chromatography tandem mass spectrometry. *Pancreas,* 41**,** 175-85.

PEINADO, H., LAVOTSHKIN, S. & LYDEN, D. 2011. The secreted factors responsible for pre-metastatic niche formation: old sayings and new thoughts. *Semin Cancer Biol,* 21**,** 139-46.

PENG, B., LIU, S., GUO, C., SUN, X. & SUN, M. Z. 2016. ANXA5 level is linked to in vitro and in vivo tumor malignancy and lymphatic metastasis of murine hepatocarcinoma cell. *Future Oncol,* 12**,** 31-42.

PENG, J. & GYGI, S. P. 2001. Proteomics: the move to mixtures. *J Mass Spectrom,* 36**,** 1083-91.

PICKERING, C. R., ZHOU, J. H., LEE, J. J., DRUMMOND, J. A., PENG, S. A., SAADE, R. E., TSAI, K. Y., CURRY, J. L., TETZLAFF, M. T., LAI, S. Y., YU, J., MUZNY, D. M., DODDAPANENI, H., SHINBROT, E., COVINGTON, K. R., ZHANG, J., SETH, S., CAULIN, C., CLAYMAN, G. L., EL-NAGGAR, A. K., GIBBS, R. A., WEBER, R. S., MYERS, J. N., WHEELER, D. A. & FREDERICK, M. J. 2014. Mutational landscape of aggressive cutaneous squamous cell carcinoma. *Clin Cancer Res,* 20**,** 6582-92.

PICOTTI, P. & AEBERSOLD, R. 2012. Selected reaction monitoring–based proteomics: workflows, potential, pitfalls and future directions. *Nature Methods,* 9**,** 555.

PICOTTI, P., BODENMILLER, B., MUELLER, L. N., DOMON, B. & AEBERSOLD, R. 2009. Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell,* 138**,** 795-806.

PINHO, S. S. & REIS, C. A. 2015. Glycosylation in cancer: mechanisms and clinical implications. *Nat Rev Cancer,* 15**,** 540-55.

PRIETO, D. A., HOOD, B. L., DARFLER, M. M., GUIEL, T. G., LUCAS, D. A., CONRADS, T. P., VEENSTRA, T. D. & KRIZMAN, D. B. 2005. Liquid Tissue™: proteomic profiling of formalin-fixed tissues. *Mass Spectrometry***,** 32-35.

PROKSCH, E., BRANDNER, J. M. & JENSEN, J. M. 2008. The skin: an indispensable barrier. *Exp Dermatol,* 17**,** 1063-72.

PROTA, G. 2000. Melanins, melanogenesis and melanocytes: Looking at their functional significance from the chemist's viewpoint. *Pigment Cell Research,* 13**,** 283-293.

RAIMONDI, S., SERA, F., GANDINI, S., IODICE, S., CAINI, S., MAISONNEUVE, P. & FARGNOLI, M. C. 2008. MC1R variants, melanoma and red hair color phenotype: a meta-analysis. *Int J Cancer,* 122**,** 2753-60.

RANGWALA, S. & TSAI, K. Y. 2011. Roles of the immune system in skin cancer. *Br J Dermatol,* 165**,** 953-65.

RASHBASH, J. 2016. *RE: director for cancer registry modernisation, UK Trend*

RATUSHNY, V., GOBER, M. D., HICK, R., RIDKY, T. W. & SEYKORA, J. T. 2012. From keratinocyte to cancer: the pathogenesis and modeling of cutaneous squamous cell carcinoma. *The Journal of Clinical Investigation,* 122**,** 464-472.

REELFS, O., TYRRELL RM FAU - POURZAND, C. & POURZAND, C. 2004. Ultraviolet a radiation-induced immediate iron release is a key modulator of the activation of NF-kappaB in human skin fibroblasts.

REIFENBERGER, J., KNOBBE, C. B., WOLTER, M., BLASCHKE, B., SCHULTE, K. W., PIETSCH, T., RUZICKA, T. & REIFENBERGER, G. 2002. Molecular genetic analysis of malignant melanomas for aberrations of the WNT signaling pathway genes CTNNB1, APC, ICAT and BTRC. *Int J Cancer,* 100**,** 549-56.

REJON, C., AL-MASRI, M. & MCCAFFREY, L. 2016. Cell Polarity Proteins in Breast Cancer Progression. *J Cell Biochem,* 117**,** 2215-23.

REYMOND, M. A. & SCHLEGEL, W. 2007. Proteomics in Cancer. *Advances in Clinical Chemistry.* Elsevier.

REZVANI, H. R., CARIO-ANDRE, M., PAIN, C., GED, C., DEVERNEUIL, H. & TAIEB, A. 2007. Protection of normal human reconstructed epidermis from UV by catalase overexpression. *Cancer Gene Ther,* 14**,** 174-86.

REZVANI, H. R., MAZURIER, F., CARIO-ANDRE, M., PAIN, C., GED, C., TAIEB, A. & DE VERNEUIL, H. 2006. Protective effects of catalase overexpression on UVB-induced apoptosis in normal human keratinocytes. *J Biol Chem,* 281**,** 17999-8007.

ROBINSON, S. J. & HEALY, E. 2002. Human melanocortin 1 receptor (MC1R) gene variants alter melanoma cell growth and adhesion to extracellular matrix. *Oncogene,* 21**,** 8037-46.

References

ROBOTI, P. & HIGH, S. 2012. The oligosaccharyltransferase subunits OST48, DAD1 and KCP2 function as ubiquitous and selective modulators of mammalian N-glycosylation. *J Cell Sci,* 125**,** 3474-84.

ROGENHOFER, N., NIENABER, L. R. M., AMSHOFF, L. C., BOGDANOVA, N., PETROFF, D., WIEACKER, P., THALER, C. J. & MARKOFF, A. 2018. Assessment of M2/ANXA5 haplotype as a risk factor in couples with placenta-mediated pregnancy complications. *J Assist Reprod Genet,* 35**,** 157-163.

ROGERS, H. W., WEINSTOCK, M. A., FELDMAN, S. R. & COLDIRON, B. M. 2015. Incidence Estimate of Nonmelanoma Skin Cancer (Keratinocyte Carcinomas) in the U.S. Population, 2012. *JAMA Dermatol,* 151**,** 1081-6.

ROGERS, H. W., WEINSTOCK, M. A., HARRIS, A. R., HINCKLEY, M. R., FELDMAN, S. R., FLEISCHER, A. B. & COLDIRON, B. M. 2010. Incidence estimate of nonmelanoma skin cancer in the United States, 2006. *Arch Dermatol,* 146**,** 283-7.

RONG, Y., ZUO, L., SHANG, L. & BAZAN, J. G. 2015. Radiotherapy treatment for nonmelanoma skin cancer. *Expert Rev Anticancer Ther,* 15**,** 765-76.

ROSCHER, I., FALK, R. S., VOS, L., CLAUSEN, O. P. F., HELSING, P., GJERSVIK, P. & ROBSAHM, T. E. 2018. Validating 4 Staging Systems for Cutaneous Squamous Cell Carcinoma Using Population-Based Data: A Nested Case-Control Study. *JAMA Dermatol,* 154**,** 428-434.

ROSENGREN PIELBERG, G., GOLOVKO, A., SUNDSTROM, E., CURIK, I., LENNARTSSON, J., SELTENHAMMER, M. H., DRUML, T., BINNS, M., FITZSIMMONS, C., LINDGREN, G., SANDBERG, K., BAUMUNG, R., VETTERLEIN, M., STROMBERG, S., GRABHERR, M., WADE, C., LINDBLAD-TOH, K., PONTEN, F., HELDIN, C. H., SOLKNER, J. & ANDERSSON, L. 2008. A cis-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse. *Nat Genet,* 40**,** 1004-9.

ROWE, D. E., CARROLL, R. J. & DAY, C. L., JR. 1992. Prognostic factors for local recurrence, metastasis, and survival rates in squamous cell carcinoma of the skin, ear, and lip. Implications for treatment modality selection. *J Am Acad Dermatol,* 26**,** 976-90.

RUNGER, T. M. & KAPPES, U. P. 2008. Mechanisms of mutation formation with long-wave ultraviolet light (UVA). *Photodermatol Photoimmunol Photomed,* 24**,** 2-10.

SALLAM, R. M. 2015. Proteomics in cancer biomarkers discovery: challenges and applications. *Dis Markers,* 2015**,** 321370.

SALZBERG, S. L. 1994. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning,* 16**,** 235-240.

SANDER, C. S., CHANG, H., HAMM, F., ELSNER, P. & THIELE, J. J. 2004. Role of oxidative stress and the antioxidant network in cutaneous carcinogenesis. *Int J Dermatol,* 43**,** 326-35.

SARIDAKI, Z., LILOGLOU, T., ZAFIROPOULOS, A., KOUMANTAKI, E., ZORAS, O. & SPANDIDOS, D. A. 2003. Mutational analysis of CDKN2A genes in patients with squamous cell carcinoma of the skin. *Br J Dermatol,* 148**,** 638-48.

SCALA, S., OTTAIANO, A., ASCIERTO, P. A., CAVALLI, M., SIMEONE, E., GIULIANO, P., NAPOLITANO, M., FRANCO, R., BOTTI, G. & CASTELLO, G. 2005. Expression of CXCR4 predicts poor prognosis in patients with malignant melanoma. *Clin Cancer Res,* 11**,** 1835-41.

SCHMIDT, H., JOHANSEN, J. S., GEHL, J., GEERTSEN, P. F., FODE, K. & VON DER MAASE, H. 2006. Elevated serum level of YKL-40 is an independent prognostic factor for poor survival in patients with metastatic melanoma. *Cancer,* 106**,** 1130-9.

SCHUFFLER, P. J., FUCHS, T. J., ONG, C. S., WILD, P. J., RUPP, N. J. & BUHMANN, J. M. 2013. TMARKER: A free software toolkit for histopathological cell counting and staining estimation. *J Pathol Inform,* 4**,** S2.

SCHWARZ, A., NOORDEGRAAF, M., MAEDA, A., TORII, K., CLAUSEN, B. E. & SCHWARZ, T. 2010. Langerhans cells are required for UVR-induced immunosuppression. *J Invest Dermatol,* 130**,** 1419-27.

SCHWEITZER, A. D., RAKESH, V., REVSKAYA, E., DATTA, A., CASADEVALL, A. & DADACHOVA, E. 2007. Computational model predicts effective delivery of 188-Re-labeled melanin-binding antibody to metastatic melanoma tumors with wide range of melanin concentrations. *Melanoma Research,* 17**,** 291-303.

SCICCHITANO, M. S., DALMAS, D. A., BOYCE, R. W., THOMAS, H. C. & FRAZIER, K. S. 2009. Protein extraction of formalin-fixed, paraffin-embedded tissue enables robust proteomic profiles by mass spectrometry. *J Histochem Cytochem,* 57**,** 849-60.

SENESCHAL, J., CLARK, R. A., GEHAD, A., BAECHER-ALLAN, C. M. & KUPPER, T. S. 2012. Human epidermal Langerhans cells maintain immune homeostasis in skin by activating skin resident regulatory T cells. *Immunity,* 36**,** 873-84.

SETLOW, R. B., GRIST, E., THOMPSON, K. & WOODHEAD, A. D. 1993. Wavelengths effective in induction of malignant melanoma. *Proc Natl Acad Sci U S A,* 90**,** 6666-70.

SETLOW, R. B., REGAN, J. D., GERMAN, J. & CARRIER, W. L. 1969. EVIDENCE THAT XERODERMA PIGMENTOSUM CELLS DO NOT PERFORM THE FIRST STEP IN THE REPAIR OF ULTRAVIOLET DAMAGE TO THEIR DNA. *Proceedings of the National Academy of Sciences of the United States of America,* 64**,** 1035-1041.

SETLOW, R. B. & SETLOW, J. K. 1962. EVIDENCE THAT ULTRAVIOLET-INDUCED THYMINE DIMERS IN DNA CAUSE BIOLOGICAL DAMAGE. *Proceedings of the National Academy of Sciences of the United States of America,* 48**,** 1250-1257.

SHAIN, A. H., BAGGER, M. M., YU, R., CHANG, D., LIU, S., VEMULA, S., WEIER, J. F., WADT, K., HEEGAARD, S., BASTIAN, B. C. & KIILGAARD, J. F. 2019. The genetic evolution of metastatic uveal melanoma. *Nat Genet,* 51**,** 1123-1130.

SHARMA, S., WAGH, S. & GOVINDARAJAN, R. 2002. Melanosomal proteins--role in melanin polymerization. *Pigment Cell Res,* 15**,** 127-33.

SHI, S. R., LIU, C., BALGLEY, B. M., LEE, C. & TAYLOR, C. R. 2006. Protein extraction from formalin-fixed, paraffin-embedded tissue sections: quality evaluation by mass spectrometry. *J Histochem Cytochem,* 54**,** 739-43.

SHLIAHA, P. V., BOND, N. J., GATTO, L. & LILLEY, K. S. 2013. Effects of traveling wave ion mobility separation on data independent acquisition in proteomics studies. *J Proteome Res,* 12**,** 2323-39.

SHULL, A. Y., LATHAM-SCHWARK, A., RAMASAMY, P., LESKOSKE, K., OROIAN, D., BIRTWISTLE, M. R. & BUCKHAULTS, P. J. 2012. Novel somatic mutations to PI3K pathway genes in metastatic melanoma. *PLoS One,* 7**,** e43369.

References

SIEGEL, R., NAISHADHAM, D. & JEMAL, A. 2013. Cancer statistics, 2013. *CA Cancer J Clin,* 63**,** 11-30.

SILVA, J. C., GORENSTEIN, M. V., LI, G. Z., VISSERS, J. P. & GEROMANOS, S. J. 2006. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics,* 5**,** 144-56.

SIRIWARDENA, B. S., KUDO, Y., OGAWA, I., KITAGAWA, M., KITAJIMA, S., HATANO, H., TILAKARATNE, W. M., MIYAUCHI, M. & TAKATA, T. 2006. Periostin is frequently overexpressed and enhances invasion and angiogenesis in oral cancer. *Br J Cancer,* 95**,** 1396-403.

SLUYTER, R. & HALLIDAY, G. M. 2001. Infiltration by inflammatory cells required for solar-simulated ultraviolet radiation enhancement of skin tumor growth. *Cancer Immunol Immunother,* 50**,** 151-6.

SOOD, A., WYKES, J., ROSHAN, D., WANG, L. Y., MCGUINNESS, J., FOWLER, A. & EBRAHIMI, A. 2019. A critical analysis of the prognostic performance of the 8th edition American Joint Committee on Cancer staging for metastatic cutaneous squamous cell carcinoma of the head and neck. *Head Neck*.

SOONG, S. J., HARRISON, R. A., MCCARTHY, W. H., URIST, M. M. & BALCH, C. M. 1998. Factors affecting survival following local, regional, or distant recurrence from localized melanoma. *J Surg Oncol,* 67**,** 228-33.

SOUSA, J. F., HAM, A. J., WHITWELL, C., NAM, K. T., LEE, H. J., YANG, H. K., KIM, W. H., ZHANG, B., LI, M., LAFLEUR, B., LIEBLER, D. C. & GOLDENRING, J. R. 2012. Proteomic profiling of paraffin-embedded samples identifies metaplasia-specific and early-stage gastric cancer biomarkers. *Am J Pathol,* 181**,** 1560-72.

SOUTH, A. P., PURDIE, K. J., WATT, S. A., HALDENBY, S., DEN BREEMS, N. Y., DIMON, M., ARRON, S. T., KLUK, M. J., ASTER, J. C., MCHUGH, A., XUE, D. J., DAYAL, J. H., ROBINSON, K. S., RIZVI, S. M., PROBY, C. M., HARWOOD, C. A. & LEIGH, I. M. 2014. NOTCH1 mutations occur early during cutaneous squamous cell carcinogenesis. *J Invest Dermatol,* 134**,** 2630-8.

SPARROW, L. E., SOONG, R., DAWKINS, H. J., IACOPETTA, B. J. & HEENAN, P. J. 1995. p53 gene mutation and expression in naevi and melanomas. *Melanoma Res,* 5**,** 93-100.

SPRUNG, R. W., BROCK, J. W. C., TANKSLEY, J. P., LI, M., WASHINGTON, M. K., SLEBOS, R. J. C. & LIEBLER, D. C. 2009. Equivalence of Protein Inventories Obtained from Formalin-fixed Paraffin-embedded and Frozen Tissue in Multidimensional Liquid Chromatography-Tandem Mass Spectrometry Shotgun Proteomic Analysis. *Molecular & Cellular Proteomics,* 8**,** 1988-1998.

SPRUNG, R. W., MARTINEZ, M. A., CARPENTER, K. L., HAM, A. J., WASHINGTON, M. K., ARTEAGA, C. L., SANDERS, M. E. & LIEBLER, D. C. 2012. Precision of multiple reaction monitoring mass spectrometry analysis of formalin-fixed, paraffin-embedded tissue. *J Proteome Res,* 11**,** 3498-505.

SRINIVAS, P. R., VERMA, M., ZHAO, Y. & SRIVASTAVA, S. 2002. Proteomics for cancer biomarker discovery. *Clin Chem,* 48**,** 1160-9.

STEIN, A. L. & TAHAN, S. R. 1994. Histologic correlates of metastasis in primary invasive squamous cell carcinoma of the lip. *J Cutan Pathol,* 21**,** 16-21.

STEINBERG, D. 2009. CART: Classification and Regression Trees. *Top Ten Algorithms in Data Mining,* 9**,** 179-201.

References

STRATIGOS, A., GARBE, C., LEBBE, C., MALVEHY, J., DEL MARMOL, V., PEHAMBERGER, H., PERIS, K., BECKER, J. C., ZALAUDEK, I., SAIAG, P., MIDDLETON, M. R., BASTHOLT, L., TESTORI, A., GROB, J. J., EUROPEAN DERMATOLOGY, F., EUROPEAN ASSOCIATION OF, D.-O., EUROPEAN ORGANIZATION FOR, R. & TREATMENT OF, C. 2015. Diagnosis and treatment of invasive squamous cell carcinoma of the skin: European consensus-based interdisciplinary guideline. *Eur J Cancer,* 51**,** 1989-2007.

STRICKLAND, P. T. 1986. Photocarcinogenesis by near-ultraviolet (UVA) radiation in Sencar mice. *J Invest Dermatol,* 87**,** 272-5.

SU, F., VIROS, A., MILAGRE, C., TRUNZER, K., BOLLAG, G., SPLEISS, O., REIS-FILHO, J. S., KONG, X., KOYA, R. C., FLAHERTY, K. T., CHAPMAN, P. B., KIM, M. J., HAYWARD, R., MARTIN, M., YANG, H., WANG, Q., HILTON, H., HANG, J. S., NOE, J., LAMBROS, M., GEYER, F., DHOMEN, N., NICULESCU-DUVAZ, I., ZAMBON, A., NICULESCU-DUVAZ, D., PREECE, N., ROBERT, L., OTTE, N. J., MOK, S., KEE, D., MA, Y., ZHANG, C., HABETS, G., BURTON, E. A., WONG, B., NGUYEN, H., KOCKX, M., ANDRIES, L., LESTINI, B., NOLOP, K. B., LEE, R. J., JOE, A. K., TROY, J. L., GONZALEZ, R., HUTSON, T. E., PUZANOV, I., CHMIELOWSKI, B., SPRINGER, C. J., MCARTHUR, G. A., SOSMAN, J. A., LO, R. S., RIBAS, A. & MARAIS, R. 2012. RAS mutations in cutaneous squamous-cell carcinomas in patients treated with BRAF inhibitors. *N Engl J Med,* 366**,** 207-15.

SUN, C. B., ZHAO, A. Y., JI, S., HAN, X. Q., SUN, Z. C., WANG, M. C. & ZHENG, F. C. 2017. Expression of annexin A5 in serum and tumor tissue of patients with colon cancer and its clinical significance. *World J Gastroenterol,* 23**,** 7168-7173.

SUN, X., LIU, S., WANG, J., WEI, B., GUO, C., CHEN, C. & SUN, M. Z. 2018. Annexin A5 regulates hepatocarcinoma malignancy via CRKI/II-DOCK180-RAC1 integrin and MEK-ERK pathways. *Cell Death Dis,* 9**,** 637.

SUPEK, F., BOSNJAK, M., SKUNCA, N. & SMUC, T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One,* 6**,** e21800.

SZKLARCZYK, D., FRANCESCHINI, A., WYDER, S., FORSLUND, K., HELLER, D., HUERTA-CEPAS, J., SIMONOVIC, M., ROTH, A., SANTOS, A., TSAFOU, K. P., KUHN, M., BORK, P., JENSEN, L. J. & VON MERING, C. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res,* 43**,** D447-52.

TADOKORO, T., KOBAYASHI, N., ZMUDZKA, B. Z., ITO, S., WAKAMATSU, K., YAMAGUCHI, Y., KOROSSY, K. S., MILLER, S. A., BEER, J. Z. & HEARING, V. J. 2003. UV-induced DNA damage and melanin content in human skin differing in racial/ethnic origin. *FASEB J,* 17**,** 1177-9.

TAKADATE, T., ONOGAWA, T., FUJII, K., MOTOI, F., MIKAMI, S., FUKUDA, T., KIHARA, M., SUZUKI, T., TAKEMURA, T., MINOWA, T., HANAGATA, N., KINOSHITA, K., MORIKAWA, T., SHIRASAKI, K., RIKIYAMA, T., KATAYOSE, Y., EGAWA, S., NISHIMURA, T. & UNNO, M. 2012. Nm23/nucleoside diphosphate kinase-A as a potent prognostic marker in invasive pancreatic ductal carcinoma identified by proteomic analysis of laser micro-dissected formalin-fixed paraffin-embedded tissue. *Clin Proteomics,* 9**,** 8.

TAKADATE, T., ONOGAWA, T., FUKUDA, T., MOTOI, F., SUZUKI, T., FUJII, K., KIHARA, M., MIKAMI, S., BANDO, Y., MAEDA, S., ISHIDA, K., MINOWA, T., HANAGATA, N., OHTSUKA, H., KATAYOSE, Y., EGAWA, S., NISHIMURA, T. & UNNO, M. 2013. Novel prognostic protein markers of resectable pancreatic cancer identified by coupled shotgun and targeted proteomics using formalin-fixed paraffin-embedded tissues. *Int J Cancer,* 132**,** 1368-82.

References

TANG, J., QIN, Z., HAN, P., WANG, W., YANG, C., XU, Z., LI, R., LIU, B., QIN, C., WANG, Z., TANG, M. & ZHANG, W. 2017. High Annexin A5 expression promotes tumor progression and poor prognosis in renal cell carcinoma. *Int J Oncol,* 50**,** 1839-1847.

TAS, F. 2012. Metastatic behavior in melanoma: timing, pattern, survival, and influencing factors. *J Oncol,* 2012**,** 647684.

THOMPSON, A. K., KELLEY, B. F., PROKOP, L. J., MURAD, M. H. & BAUM, C. L. 2016. Risk Factors for Cutaneous Squamous Cell Carcinoma Recurrence, Metastasis, and Disease-Specific Death: A Systematic Review and Meta-analysis. *JAMA Dermatol,* 152**,** 419-28.

THOMPSON, J. F., SCOLYER, R. A. & KEFFORD, R. F. 2005. Cutaneous melanoma. *The Lancet,* 365**,** 687-701.

THORPE, L. M., YUZUGULLU, H. & ZHAO, J. J. 2015. PI3K in cancer: divergent roles of isoforms, modes of activation and therapeutic targeting. *Nature reviews. Cancer,* 15**,** 7-24.

TING, W., SCHULTZ, K., CAC, N. N., PETERSON, M. & WALLING, H. W. 2007. Tanning bed exposure increases the risk of malignant melanoma. *Int J Dermatol,* 46**,** 1253-7.

TSAO, H., NADIMINTI, U., SOBER, A. J. & BIGBY, M. 2001. A meta-analysis of reverse transcriptase-polymerase chain reaction for tyrosinase mRNA as a marker for circulating tumor cells in cutaneous melanoma. *Arch Dermatol,* 137**,** 325-30.

UGUREL, S., UTIKAL, J. & BECKER, J. C. 2009. Tumor biomarkers in melanoma. *Cancer Control,* 16**,** 219-24.

ULRICH, C., KANITAKIS, J., STOCKFLETH, E. & EUVRARD, S. 2008. Skin cancer in organ transplant recipients--where do we stand today? *Am J Transplant,* 8**,** 2192-8.

VALLEJO-TORRES, L., MORRIS, S., KINGE, J. M., POIRIER, V. & VERNE, J. 2014. Measuring current and future cost of skin cancer in England. *J Public Health (Oxf),* 36**,** 140-8.

VALVERDE, P., HEALY, E., JACKSON, I., REES, J. L. & THODY, A. J. 1995. Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nat Genet,* 11**,** 328-30.

VAN DE MERBEL, A. F., VAN DER HORST, G., BUIJS, J. T. & VAN DER PLUIJM, G. 2018. Protocols for Migration and Invasion Studies in Prostate Cancer. *Methods Mol Biol,* 1786**,** 67-79.

VAN DER WEYDEN, L., PATTON, E. E., WOOD, G. A., FOOTE, A. K., BRENN, T., ARENDS, M. J. & ADAMS, D. J. 2016. Cross-species models of human melanoma. *J Pathol,* 238**,** 152-65.

VAPNIK, V. & CHERVONENKIS, A. 1974. Theory of pattern recognition. Nauka, Moscow.

VENESS, M. J. 2006. Defining patients with high-risk cutaneous squamous cell carcinoma. *Australas J Dermatol,* 47**,** 28-33.

VENNING, F. A., WULLKOPF, L. & ERLER, J. T. 2015. Targeting ECM Disrupts Cancer Progression. *Front Oncol,* 5**,** 224.

VIZCAINO, J. A., CSORDAS, A., DEL-TORO, N., DIANES, J. A., GRISS, J., LAVIDAS, I., MAYER, G., PEREZ-RIVEROL, Y., REISINGER, F., TERNENT, T., XU, Q. W., WANG, R. & HERMJAKOB, H. 2016. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res,* 44**,** D447-56.

References

VLAYKOVA, T., TALVE, L., HAHKA-KEMPPINEN, M., HERNBERG, M., MUHONEN, T., COLLAN, Y. & PYRHONEN, S. 2002. Immunohistochemically detectable Bcl-2 expression in metastatic melanoma: Association with survival and treatment response. *Oncology,* 62**,** 259-268.

VON MERING, C., HUYNEN, M., JAEGGI, D., SCHMIDT, S., BORK, P. & SNEL, B. 2003. STRING: a database of predicted functional associations between proteins. *Nucleic acids research,* 31**,** 258-261.

WABNITZ, G. H., KOCHER, T., LOHNEIS, P., STOBER, C., KONSTANDIN, M. H., FUNK, B., SESTER, U., WILM, M., KLEMKE, M. & SAMSTAG, Y. 2007. Costimulation induced phosphorylation of L-plastin facilitates surface transport of the T cell activation molecules CD69 and CD25. *Eur J Immunol,* 37**,** 649-62.

WAKAMATSU, K. & ITO, S. 2002. Advanced chemical methods in melanin determination. *Pigment Cell Res,* 15**,** 174-83.

WANG, D., KHOSLA, A., GARGEYA, R., IRSHAD, H. & BECK, A. H. 2016. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.

WANG, N. J., SANBORN, Z., ARNETT, K. L., BAYSTON, L. J., LIAO, W., PROBY, C. M., LEIGH, I. M., COLLISSON, E. A., GORDON, P. B., JAKKULA, L., PENNYPACKER, S., ZOU, Y., SHARMA, M., NORTH, J. P., VEMULA, S. S., MAURO, T. M., NEUHAUS, I. M., LEBOIT, P. E., HUR, J. S., PARK, K., HUH, N., KWOK, P. Y., ARRON, S. T., MASSION, P. P., BALE, A. E., HAUSSLER, D., CLEAVER, J. E., GRAY, J. W., SPELLMAN, P. T., SOUTH, A. P., ASTER, J. C., BLACKLOW, S. C. & CHO, R. J. 2011. Loss-of-function mutations in Notch receptors in cutaneous and lung squamous cell carcinoma. *Proc Natl Acad Sci U S A,* 108**,** 17761-6.

WANG, Q., ZHANG, M., TOMITA, T., VOGELSTEIN, J. T., ZHOU, S., PAPADOPOULOS, N., KINZLER, K. W. & VOGELSTEIN, B. 2017. Selected reaction monitoring approach for validating peptide biomarkers. *Proc Natl Acad Sci U S A,* 114**,** 13519-13524.

WARD, D. G., SUGGETT, N., CHENG, Y., WEI, W., JOHNSON, H., BILLINGHAM, L. J., ISMAIL, T., WAKELAM, M. J., JOHNSON, P. J. & MARTIN, A. 2006. Identification of serum biomarkers for colon cancer by proteomic analysis. *Br J Cancer,* 94**,** 1898-905.

WASSERSTEIN, R. L. & LAZAR, N. A. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. *American Statistician,* 70**,** 129-131.

WATSON, A. D., GUNNING, Y., RIGBY, N. M., PHILO, M. & KEMSLEY, E. K. 2015. Meat Authentication via Multiple Reaction Monitoring Mass Spectrometry of Myoglobin Peptides. *Anal Chem,* 87**,** 10315-22.

WATSON, M., GARNETT, E., GUY, G. P. & HOLMAN, D. M. 2014. The Surgeon General's call to action to prevent skin cancer.

WEBB-ROBERTSON, B. J., WIBERG, H. K., MATZKE, M. M., BROWN, J. N., WANG, J., MCDERMOTT, J. E., SMITH, R. D., RODLAND, K. D., METZ, T. O., POUNDS, J. G. & WATERS, K. M. 2015. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res,* 14**,** 1993-2001.

WEINBERG, A. S., OGLE, C. A. & SHIM, E. K. 2007. Metastatic cutaneous squamous cell carcinoma: an update. *Dermatol Surg,* 33**,** 885-99.

WEINSTEIN, D., LEININGER, J., HAMBY, C. & SAFAI, B. 2014. Diagnostic and prognostic biomarkers in melanoma. *J Clin Aesthet Dermatol,* 7**,** 13-24.

References

WEISS, S. A., HANNIFORD, D., HERNANDO, E. & OSMAN, I. 2015. Revisiting determinants of prognosis in cutaneous melanoma. *Cancer,* 121**,** 4108-23.

WEISSER, J., LAI, Z. W., BRONSERT, P., KUEHS, M., DRENDEL, V., TIMME, S., KUESTERS, S., JILG, C. A., WELLNER, U. F., LASSMANN, S., WERNER, M., BINIOSSEK, M. L. & SCHILLING, O. 2015. Quantitative proteomic analysis of formalin-fixed, paraffin-embedded clear cell renal cell carcinoma tissue using stable isotopic dimethylation of primary amines. *BMC Genomics,* 16**,** 559.

WENDT, M. K., SMITH, J. A. & SCHIEMANN, W. P. 2010. Transforming growth factor-beta-induced epithelial-mesenchymal transition facilitates epidermal growth factor-dependent breast cancer progression. *Oncogene,* 29**,** 6485-98.

WHITE-GILBERTSON, S., KURTZ, D. T. & VOELKEL-JOHNSON, C. 2009. The role of protein synthesis in cell cycling and cancer. *Mol Oncol,* 3**,** 402-8.

WISNIEWSKI, J. R. 2013. Proteomic sample preparation from formalin fixed and paraffin embedded tissue. *J Vis Exp***,** 50589.

WISNIEWSKI, J. R., OSTASIEWICZ, P. & MANN, M. 2011. High recovery FASP applied to the proteomic analysis of microdissected formalin fixed paraffin embedded cancer tissues retrieves known colon cancer markers. *J Proteome Res,* 10**,** 3040-9.

WISZTORSKI, M., FATOU, B., FRANCK, J., DESMONS, A., FARRE, I., LEBLANC, E., FOURNIER, I. & SALZET, M. 2013. Microproteomics by liquid extraction surface analysis: application to FFPE tissue to study the fimbria region of tubo-ovarian cancer. *Proteomics Clin Appl,* 7**,** 234-40.

WOLNICKA-GLUBISZ, A., STRICKLAND, F. M., WIELGUS, A., ANVER, M., MERLINO, G., DE FABO, E. C. & NOONAN, F. P. 2015. A melanin-independent interaction between Mc1r and Met signaling pathways is required for HGF-dependent melanoma. *Int J Cancer,* 136**,** 752-60.

WORLD HEALTH ORGANISATION. 2016. *How Common is Skin Cancer?* [Online]. Available: http://www.who.int/uv/faq/skincancer/en/index1.html [Accessed 17/02/2016 2016].

WU, M., CHEN, X., LOU, J., ZHANG, S., ZHANG, X., HUANG, L., SUN, R., HUANG, P., WANG, F. & PAN, S. 2016. TGF-beta1 contributes to CD8+ Treg induction through p38 MAPK signaling in ovarian cancer microenvironment. *Oncotarget,* 7**,** 44534-44544.

XIAO, Z., LI, G., CHEN, Y., LI, M., PENG, F., LI, C., LI, F., YU, Y., OUYANG, Y., XIAO, Z. & CHEN, Z. 2010. Quantitative proteomic analysis of formalin-fixed and paraffin-embedded nasopharyngeal carcinoma using iTRAQ labeling, two-dimensional liquid chromatography, and tandem mass spectrometry. *J Histochem Cytochem,* 58**,** 517-27.

XU, H., YANG, L., WANG, W., SHI, S. R., LIU, C., LIU, Y., FANG, X., TAYLOR, C. R., LEE, C. S. & BALGLEY, B. M. 2008. Antigen retrieval for proteomic characterization of formalin-fixed and paraffin-embedded tissues. *J Proteome Res,* 7**,** 1098-108.

XU, I. M., LAI, R. K., LIN, S. H., TSE, A. P., CHIU, D. K., KOH, H. Y., LAW, C. T., WONG, C. M., CAI, Z., WONG, C. C. & NG, I. O. 2016. Transketolase counteracts oxidative stress to drive cancer development. *Proc Natl Acad Sci U S A,* 113**,** E725-34.

XUE, G., HAO, L. Q., DING, F. X., MEI, Q., HUANG, J. J., FU, C. G., YAN, H. L. & SUN, S. H. 2009. Expression of annexin a5 is associated with higher tumor stage and poor prognosis in colorectal adenocarcinomas. *J Clin Gastroenterol,* 43**,** 831-7.

References

YAMADA, Y., BANNO, Y., YOSHIDA, H., KIKUCHI, R., AKAO, Y., MURATE, T. & NOZAWA, Y. 2006. Catalytic inactivation of human phospholipase D2 by a naturally occurring Gly901Asp mutation. *Arch Med Res,* 37**,** 696-9.

YAMAMOTO, M. & SUGIMOTO, T. 2016. Advanced Glycation End Products, Diabetes, and Bone Strength. *Curr Osteoporos Rep,* 14**,** 320-326.

YAMANO, Y., UZAWA, K., SAITO, K., NAKASHIMA, D., KASAMATSU, A., KOIKE, H., KOUZU, Y., SHINOZUKA, K., NAKATANI, K., NEGORO, K., FUJITA, S. & TANZAWA, H. 2010. Identification of cisplatin-resistance related genes in head and neck squamous cell carcinoma. *Int J Cancer,* 126**,** 437-49.

YAMASHITA, S. & KATSUMATA, O. 2017. Heat-Induced Antigen Retrieval in Immunohistochemistry: Mechanisms and Applications. *Methods Mol Biol,* 1560**,** 147-161.

YANG, F., SHEN, Y., CAMP, D. G., 2ND & SMITH, R. D. 2012. High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis. *Expert Rev Proteomics,* 9**,** 129-34.

YAO, M., SHANG, Y. Y., ZHOU, Z. W., YANG, Y. X., WU, Y. S., GUAN, L. F., WANG, X. Y., ZHOU, S. F. & WEI, X. 2017. The research on lapatinib in autophagy, cell cycle arrest and epithelial to mesenchymal transition via Wnt/ErK/PI3K-AKT signaling pathway in human cutaneous squamous cell carcinoma. *J Cancer,* 8**,** 220-226.

YATES, J. R., RUSE, C. I. & NAKORCHEVSKY, A. 2009. Proteomics by mass spectrometry: approaches, advances, and applications. *Annu Rev Biomed Eng,* 11**,** 49-79.

YOCUM, A. K., KHAN, A. P., ZHAO, R. & CHINNAIYAN, A. M. 2010. Development of selected reaction monitoring-MS methodology to measure peptide biomarkers in prostate cancer. *Proteomics,* 10**,** 3506-14.

YOKOYAMA, H. & MIZUTANI, R. 2014. Structural biology of DNA (6-4) photoproducts formed by ultraviolet radiation and interactions with their binding proteins. *Int J Mol Sci,* 15**,** 20321-38.

YOSHIDA, A., OKAMOTO, N., TOZAWA-ONO, A., KOIZUMI, H., KIGUCHI, K., ISHIZUKA, B., KUMAI, T. & SUZUKI, N. 2013. Proteomic analysis of differential protein expression by brain metastases of gynecological malignancies. *Hum Cell,* 26**,** 56-66.

YU, H. S., LIAO, W. T. & CHAI, C. Y. 2006. Arsenic carcinogenesis in the skin. *J Biomed Sci,* 13**,** 657-66.

YU, K. H., ZHANG, C., BERRY, G. J., ALTMAN, R. B., RE, C., RUBIN, D. L. & SNYDER, M. 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun,* 7**,** 12474.

YU, Y. Q., GILAR, M., LEE, P. J., BOUVIER, E. S. & GEBLER, J. C. 2003. Enzyme-friendly, mass spectrometry-compatible surfactant for in-solution enzymatic digestion of proteins. *Anal Chem,* 75**,** 6023-8.

ZENG, Q., CHEN, S., YOU, Z., YANG, F., CAREY, T., SAIMS, D. & WANG, C.-Y. 2002. *Hepatocyte Growth Factor Inhibits Anoikis in Head and Neck Squamous Cell Carcinoma Cells by Activation of ERK and Akt Signaling Independent of NF??B*.

ZHANG, M., BISWAS, S., QIN, X., GONG, W., DENG, W. & YU, H. 2016. Does Notch play a tumor suppressor role across diverse squamous cell carcinomas? *Cancer Med,* 5**,** 2048-60.

References

Appendix 1

Appendix 1: A Table of all known FFPE mass spectrometry based proteomic studies.

Type of FFPE tissue	Sample processing and buffers	LC/MS method	Total proteins (unique)	Notes	Reference
Breast cancer; Stage 0 (n = 7), stage II, 2 y recur (n = 5) stage II, non-recur (n = 4)), stage III (n = 9)	LMD, Liquid Tissue™	RP-HPLC LTQ	9437 total		(Bateman et al., 2011)
Colon adenoma (n = 4), FFPE	LMD. 0.1 M Tris –HCl pH 8.0/0.1 M DTT/0.5% (w/v) PEG 20,000/4 % SDS, 99 °C. MED-FASP, SAX	RP-HPLC Ion Trap/Orbitrap	Average 8481 per sample.		(Wisniewski, 2013)
Colon cancer and matched normal (n = 3)	0.1 M Tris-HCl pH 8.0/0.1 M DTT/0.5% (w/v) PEG 20,000/4% SDS at 99 °C 1 h, FASP, SAX	HPLC Ion Trap/Orbitrap	Cancer = 5985 ± 54 Normal = 5868 ± 110	from 6 SAX fractions	(Wisniewski et al., 2011)
SILAC labelled mouse liver	Xylene with a graded ethanol series. LMD or needle MD. 0.1 M Tris HCl pH 8.0/ 0.1 M DTT, blender, sonifier, SDS to 4%, 99 °C 1 h, clarification, FASP	RP-HPLC Ion Trap/ Orbitrap	FFPE = 5203	Frozen = 5426 91% identical between FFPE and Frozen. No significant difference observed in subcellular location. No storage time effect of FFPE. LMD had no effect.	(Ostasiewicz et al., 2010)
Mouse liver	Xylene with a graded ethanol series, 50 mM Tris HCL pH 7/ 2% SDS, sonication, 100 °C 20 min, 80 °C 2 h. 40,000 psi applied	RP-HPLC LTQ	Without pressure = 3449 With pressure = 5192	Extracted FFPE mouse liver with heat, augmented by elevated hydrostatic pressure. Found that extended storage time had no effect on pressure assisted antigen retrieval but huge effect on just heat. Frozen = 4932 Frozen with heat = 4451	(Fowler et al., 2012)
cSCC, Bowen's disease, actinic keratosis	LCM, 0.1% RapiGest/50mM TEAB, 95˚C 30 mins	RP-HPLC SWATH	3574		(Azimi et al., 2019)

Appendices

colon mucosal biopsies directly frozen (DF) RNA later immediate FFPE (iFFPE) FFPE are 30mins room temp (sFFPE)	xylene with graded ethanol series, 300mM Tris-HCL/12mM SDS/ 12mM SDC/pH9.0, 95 °C 60 mins, sonication	RP-HPLC Ion Trap/Orbitrap	iFFPE = 3384 sFFPE = 3328	DF = 3840 RNA later = 3718	(Bennike et al., 2016)
Human liver (n = 3)	Octane. 20 mM Tris pH9, dialysed against 100 mM Tris HCL pH8.2. Denatured via 8 M Urea, reduction, alkylation, diluted 100 mM ammonium acetate	capillary isotachophoresis (CITP)RP-HPLC Ion Trap	Average 3287	n = 3 - (3209, 3302, 3350)	(Xu et al., 2008)
Clear cell renal carcinoma	xyelen with graded ethanol series, 100mM HEPES pH7.5/4% SDS/ 50mM DTT 1hr 95°C, acetone precipitation, resuspended at 100mM NaOH, sonication, 40mM formaldehyde added after digestion and then quenched in 20mM glycine 20min 22°C	RP-HPLC-orbitrap	2938	1307 found in 3 replicates	(Weisser et al., 2015)
Glioblastoma	Octane. LMD. 20 mM Tris pH 9/2% SDS, 100 °C 20 min, 60 °C 2 h, Dialysed against 100 mM Tris HCL pH 8.2, reduction alkylation	CIEF RP-HPLC Ion Trap	FFPE = 2733	Frozen soluble = 2380 Frozen pellet = 3110	(Guo et al., 2007)
Normal renal kidney tissue (n = 16), clear cell renal cell carcinoma (n = 8)	Xylene with a graded ethanol series. 62.5 mM Tris–HCl pH 6.8/4% SDS/10% glycerol/100 mM DTT, 105 °C 45 min, FASP digestion.	RP-HPLC Ion Trap/Orbitrap	Normal = 2663 total Carcinoma = 2516 total	Normal mean = 2121 Carcinoma mean = 1671 no difference in the number of proteins identified between tissue blocks of different ages	(Craven et al., 2013)
Intestinal-type gastric cancer (n = 10), metaplasia (n = 10), normal mucosa (n = 10)	Microdissection, Sub-X with a graded ethanol series. 100 mM ABC pH8, 80 °C 2 h, trifluoroethanol, sonication, 60 °C 1 h, sonication, Carboxyethylphosphine, reduction, alkylation, trypsin.	Peptide IEF (IPG strips), RP-HPLC Ion Trap/Orbitrap	Average 2350 protein/sample		(Sousa et al., 2012)
Colon adenoma	Sub-X with a graded ethanol series. 100 mM ABC alone, and 1mM EDTA, or 100 mM pyridoxamine, 80 °C 2 h	RP-HPLC Ion Trap	FFPE = 2302.	Frozen = 2554 FFPE had less lysine C-terminal to arginine C-terminal peptides (little effect on IDs), no difference in sub-cellular location, increased methionine oxidation; 17% at 1 year, 25% at 10	(Sprung et al., 2009)

				years. EDTA and pyridoxamine decreased protein identifications. No effect of storage time on protein yield after 10 years.	
cutaneous squamous cell carcinoma (n=24)	deparaffinised, 0.1% RapiGest/0.1m HEPES pH8/0.2mM DTT, 95 °C 4 h.	RP-HPLC Orbitrap	2102		(Foll et al., 2017)
Clear cell renal cell carcinoma, FFPE and fresh frozen	Sub-X with a graded ethanol series, 100 mM ABC pH 8, 80 °C 2 h.	MRM RP-HPLC Triple Quad	FFPE = 1982	Frozen = 2154	(Sprung et al., 2012)
Human renal carcinoma	Octane and methanol, 20 mM Tris HCL pH 9/2% SDS, 100 °C 20 min, 60 °C 2 h, dialysis for SDS removal	CIEF RP-HPLC Ion Trap	FFPE heat induced AR 1 = 1830 FFPE heat induced AR 2 = 1962 FFPE no heat = 962	Frozen = 1341 Average amount protein extracted from FFPE = 10.0 mg/ml, fresh control tissue = 11.5 mg/ml.	(Shi et al., 2006)
temporal cortex from Alzheimer's patients (N=3, n=3)	100mM ABC/20% acetonitrile, 95 °C 1 hour, 65 °C 2 hours, then either: 1. 20mM DTT 57 °C 1 hour, 50mM IAA 2. 0.2% RapiGest/ 50mM ABC, 20mM DTT 57 °C 1 hour, 50mM IAA 3. 50mM Tris-HCL/pH7.4/ 1% triton X-100/ 0.5% SDC/ 0.1% SDS/ 150mM NaCl/ 2mM EDTA	RP-HPLC Ion Trap/Orbitrap	1. = 1715 2. =1773 3. = 1598		(Drummond et al., 2015)
benign nevi (n = 25), primary melanoma (n = 12), metastatic melanoma (n = 24)	Xylene with a graded ethanol series, microdissection, 20 mM Tris pH 7/2% SDS, 90 °C for 1 h. Bioruptor; 5 min, 65 °C for 4 h.	RP HPLC Ion Trap/Orbitrap	Total 1528		(Byrum et al., 2013)
Pancreatic ductal adenocarcinoma, primary tumours, matched lymph nodes (n = 7)	Xylene with a graded ethanol series, LMD, Liquid Tissue™, 95 °C 1.5 h.	MudPIT, RP-HPLC Ion Trap/Orbitrap	1504 unique proteins	854 common to all samples	(Naidoo et al., 2012)
Endometrial cancer, malignant areas compared to stromal areas	Xylene with a graded ethanol series, LMD, 100 mM ABC/20% ACN, 95 °C 1 h, 65 °C 2 h.	RP-HPLC Ion Trap/Orbitrap	Average 1500 proteins ID per run.		(Alkhas et al., 2011)
cSCC (n=5)	LCM, 0.1% RapiGest/50mM TEAB, 95˚C 30 mins	RP-HPLC Ion Trap/Orbitrap	1310		(Azimi et al., 2016)
Pancreatic cancer, poor prognosis (n = 4), better prognosis (n = 4) noncancerous pancreatic ductal tissues (n = 5)	Xylene with a graded ethanol series, LMD, Liquid Tissue™, 95 °C 1.5 h.	RP-HPLC Ion Trap/Orbitrap	1229 total	Poor = 924 Better = 845 Normal = 730	(Takadate et al., 2013)
Wilm's tumour and healthy renal tissue (n=7)	Heptane, 2% Tris/2% SDS, 100 °C for 20 min,	RP-HPLC Ion Trap/Orbitrap	FFPE normal = 1168		(Hammer et al., 2014)

	80 °C for 2 h. Centrifugation 14,000×g. Methanol-chloroform precipitation. 25 mM ABC/1% RapiGest				
mouse liver **Prostate cancer - PCa** **Benign prostate hyperplasia - BPH**	58°C 60 mins, SubX deparaffinization and graded ethanol, haematoxylin, dehydrated ethanol series, rehydrated in 50% glycerol, LCM	RP-HPLC Ion Trap	mouse liver= 684 PCa= 1156 BPH= 702		(Prieto et al., 2005)
ovarian cancer	toluene with graded ethanol series, 20mM Tris-HCL/pH 9, 97 °C 30 mins, Trypsin digestion on tissue using chemical inkjet printer. Solvent was then applied and aspirated to a low bind tube	RP-HPLC Ion Trap/Orbitrap	total 1109	792 in cancerous region, 983 in benign.	(Wisztorski et al., 2013)
Pancreatic cancer, poor prognosis (n = 4), better prognosis (n = 4)	Xylene with a graded ethanol series, LMD, Liquid Tissue™, 95 °C 1.5 h.	RP-HPLC Ion Trap/Orbitrap	Total 1099 unique proteins	845 better prognostic 924 poor prognostic	(Takadate et al., 2012)
lung cancer (n = 2), squamous cell carcinoma (n = 1), hepatic metastasized colorectal cancer (n = 3)	Xylene with a graded ethanol series. 100 mM HEPES pH 7.5/4 % SDS/50 mM DTT, 95 °C 1 h. Centrifugation 14,000 rpm. Acetone precipitation. 100 mM NaOH pH 8	RP-HPLC Ion Trap/Orbitrap	1003		(Bronsert et al., 2014)
Mouse liver	1. Frozen - 40 mM Tris/6 M guanidine-HCl/65 mM DTT pH8.2, sonication, clarification, reduction, alkylation, 1M guanidine-HCL 2. 40 mM Tris/6 M guanidine-HCl/65 mM DTT pH8.2, sonication, clarification, reduction, alkylation, 1M guanidine-HCL 3. 40mM Tris/2% SDS pH8.2, sonication, 100ºC 20mins, 60ºC 2hrs, reduction, alkylation, 0.1% SDS 4. 40 mM Tris/6 M guanidine-HCl/65 mM DTT pH8.2, sonication, reduction, alkylation, 1M guanidine-HCL 5. 40 mM Tris/6 M guanidine-HCl/65 mM DTT pH8.2, sonication, 100ºC 30mins, clarification	RP-HPLC Ion Trap	1 = 976 2= 130 3= 820 4= 331 5= 827 6= 526		(Jiang et al., 2007)

	6. Pellet from 5. - 90% formic acid 5mins, cyanogens bromide 1g/ml overnight dark, ph8.5, lyophilised, 40mM Tris/ 6M guanidine-HCL				
Melanoma melanocytic nevus (n = 1), metastatic melanoma (n = 1)	LMD, Liquid Tissue™. 1DE, band excision, trypsin digestion.	HPLC LTQ-XL	888		(Byrum et al., 2011)
mouse liver (N=6, n=3) mouse colon (N=3, n=3) Human colon (N=3, n=3)	Xylene with graded ethanol series, suspended in: 1. 20mM Tris-HCL/ 2% SDS/200mM DTT/ 20% glycerol/ 1% protease inhibitor/pH8.8 2. 40mM Tris-HCL/6M guanidine-HCL/ 65mM DTT/ pH8.2 3.25mM Tris-HCL/ 150mM NaCl/ 1% NP-40/ 1% SDS/ 0.1% SDS/ pH 7.6 4.5mM DTT/ 0.2% RapiGest/ pH 8.4 5. 2% SDS/ pH 8 6. 100mM ABC/ 30% acetonitrile/ pH8.4 7. 50% 100mM ABC/ 50% triflouroethanol 8. 20mM Tris-HCL/ 0.5% SDS/ 1.5% CHAPS/ 200mM DTT/ 10% glycerol/ pH 8.8	RP-HPLC Ion Trap/Orbitrap	mouse liver: 1,5,8 = 887, 737, 693 mouse colon: 1,5,8 = 772, 223, 185 human colon: 1, 5, 8= 681, 463, 554	all methods tried on mouse liver, best three picked thereafter	(Broeckx et al., 2016)
Nasopharyngeal carcinoma, WHO type I (n = 10), II (n = 10), III (n = 10), and normal (n = 10)	Octane and methanol. 20 mM Tris/2% SDS pH7, 100 °C 20 min, 60 °C 2 h. Centrifugation 12,000×g. TCA acetone precipitation. iTRAQ.	HPLC - SCX, RP-HPLC Q-TOF	730 total	Compared FFPE with previously analysed fresh frozen tissue. Found all 730 proteins in FF dataset. Compared subcellular localization and molecular function groups; distribution of proteins similar between FFPE and FF.	(Xiao et al., 2010)
Colorectal cancer and paired adjacent control colon mucosa (n = 3)	Xylene with a graded ethanol series, 20 mM TrisHCl pH 8.8/200 mM DTT/2% SDS/1% protease inhibitor, 98 °C for 20 min, 80 °C for 2 h. Centrifugation 14,000×g.	RP-HPLC Ion Trap/Orbitrap	713 total		(Maes et al., 2013)
irradiated C57BL/6 mice heart tissue (n = 3)	Xylene with a graded ethanol series, 20 mM Tris–HCl pH 8.8/2%	RP-HPLC Ion Trap/Orbitrap	Total 544		(Azimzadeh et al., 2012)

Appendices

	SDS/1% beta-octylglucoside/200 mM DTT/200 mM glycine. 100 °C 2, Centrifugation 14,000×g, Precipitated, suspended Tris buffer.				
Normal pancreas (n = 3), chronic pancreatitis (n = 3), pancreatic cancer (n = 3), FFPE	heptane, methanol, Centrifugation 20,000×g, dried, resuspended 250 µL 6 M guanidine-HCl/50 mM ABC/20mM DTT, pH 8.5, 70 °C for 1 h. IAA, DTT.	RP-HPLC Ion Trap/Orbitrap	Total 525	(Proteins identified in at least 2 of 3 specimens).	(Paulo et al., 2012)
1. Lung cancer without lymph node involvement (n=7) 2. Lung cancer with lymph node involvement (n=6) 3. lymph nodes involved	LCM, xylene with a graded ethanol series. LMD. Liquid Tissue™.	RP-HPLC LTQ	1. 449 2. 438 3. 233	649 total unique proteins	(Kawamura et al., 2010)
Human Aorta, Unfixed (n = 3), Fresh frozen 3 month (n=3) Fresh frozen 15 years (n=3) FFPE 15 years (n=2)	Xylene with a graded ethanol series, 100 mM Tris–HCl pH 8.0/4% SDS/100 mM DTT, extracted at either: 24 °C 1 h 14.7 psi, 95 °C 1 h 14.7 psi, or 95 °C 1 h 40,000 psi using a NEP 2320 Barocycler,	RP-HPLC Ion Trap/Orbitrap	Average FFPE = 370	Average unfixed = 283 Average frozen 3 month = 564 Average frozen 15 y = 20	(Fu et al., 2013)
Non-alcoholic steatohepatitis, FFPE and fresh frozen	xylene and graded ethanol series, LMD, Liquid Tissue™, heated 95 °C 1.5 h.	RP-HPLC LTQ	FFPE = 367 total	225 common between FFPE and frozen, 142 unique FFPE, 493 unique to frozen. Frozen = 718 total	(Bell et al., 2011)
Head and neck cSCC; normal (n = 4), well differentiated (WD; n = 4), moderately differentiated (MD; n = 4), poorly differentiated(PD; n = 4),	SafeClear II. LMD. Liquid Tissue™, heat 95 °C 1.5 h	RP-HPLC Ion Trap	Averages: N = 147.5 WD = 351.5 MD = 274.5 PD = 244.3		(Patel et al., 2008)
Colon cancer	SubX deparaffinization and graded ethanol, microdisected, Liquid Tissue, 95°C for 90 mins, reduced in 10mM DTT at 95°C for 5 mins	RP-HPLC Ion Trap	350	Also did SELDI-TOF and MALDI-TOF-TOF but did not report protein yield	(Prieto et al., 2005)
Nephrectomy tumour and normal (n = 4)	Xylene with a graded ethanol series. RapiGest buffer, 105 °C 30 min, cool, vortex, 80 °C 2 h, reduction, alkylation	RP-HPLC Q-TOF (MS^E)	FFPE = 283	Frozen = 268	(Nirmalan et al., 2011)
Rat spinal cords, healthy and experimental autoimmune encephalomyelitis (EAE)	Xylene with 70% ethanol. Needle dissection, Liquid Tissue kit, 95°C 90 mins, iTRAQ	HPLC SCX, MALDI TOF/TOF	FFPE = 262 unique proteins	Frozen = 500	(Jain et al., 2012)
Mouse pancreatic tissues (n = 8), FFPE and matched frozen for method developed, Human pancreatic cancer (n = 11)	Qproteome kit, Xylene with a graded ethanol series, Extraction buffer, 100 °C 20 min, 80 °C for 2 h.	RP-HPLC Ion Trap	Mouse FFPE = 237 Control = 178 cancer = 198	Mouse Frozen = 271,	(Kojima et al., 2012)

and uninvolved tissue (n = 8), FFPE	Centrifugation 14,000×g.				
Mouse heart	All methods were deparaffinised and rehydrated in Xylene and graded ethanol, 1. Laemmli buffer/2% SDS/ protease inhibitor cocktail 2. 2% CHAPS/ protease inhibitor cocktail 3. 0.2% Tween 20/ protease inhibitor cocktail 4. RIPA buffer/2% SDS/ 1% NP40/ protease inhibitor cocktail 5. 20mM Tris-HCL pH 8.8/ 2% SDS/ 1% beta-octylglucoside/ 200mM DTT/ 200mM glycine/ protease inhibitor cocktail all methods incubated at 100°C 20mins, 80°C 2hrs,	RP-HPLC LTQ Orbitrap	192 total proteins	A mix of the non-ionic detergent 1% beta-octylglucoside plus 2% SDS gave optimal protein release from FFPE sections. Increasing amounts of SDS beyond 4% did not enhance the protein yield further. 17 fractions from SDS page separation	(Azimzadeh et al., 2010)
Paediatric brainstem gliomas (n = 2)	Xylene with a graded ethanol series, 100 mM ABC/30% ACN, 95 °C 30 min, 65 °C 3 h, 18O proteolytic labelling.	RP-HPLC Ion Trap 18O proteolytic labelling	188 total		(Nazarian et al., 2008)
mouse liver	1. LCM - scraped into tube, 60°C for 30mins, deparaffinization reagent, graded ethanol, liquid tissue buffer + 0.5% RapiGest, 95°C for 90mins 2. non-LCM - 60°C for 30 mins, xylene washes, graded ethanol, LCM, Liquid Tissue™ + 0.5% RapiGest, 95°C for 90mins both reduced in 5mM DTT for 1 hour 60°C, alkylated 15mM IAA room temp 1 hour, trypsin digest in 75mM ABC overnight	RP-HPLC Ion Trap	185	LCM = 170 non-LCM= 132	(Scicchitano et al., 2009)
rat liver	1. Deparaffinised and rehydrated in graded ethanol, sonication 2% SDS/ 100mM ABC, reduction, alkylation 2. 40mM Tris/ 6M guanidine-HCL/ 65mM DTT pH 8.2, 100°C 30 mins, 50mM ABC, alkylation 3. Qproteome kit, 80°C for 2 hrs	RP-HPLC Q-TOF (MS^E)	1. Not reported 2. Not reported 3. Not reported 4. 173 5. 166	Each done in triplicate, only proteins in 2/3 included.	(Aarnisalo et al., 2010)

Appendices

	4. Liquid tissue kit, 95°C for 90 mins, 5. Liquid tissue kit, 95°C for 90 mins, reduction, alkylation				
Adenoma parathyroid (n = 5) from sporadic primary hyperparathyroidism(PHPT) patients	Xylene with a graded ethanol series. 20 mM Tris–HCl pH 4/0.2 M glycine/2% SDS, sonication, 4 °C 1 h, 100 °C 20 min, 60 °C 2 h. clarification,	RP-HPLC Ion Trap/Orbitrap	163		(Donadio et al., 2011)
Non-small cell lung tumours (NSCLC) Renal cell carcinomas (RCC)	xylene and graded ethanol, 40mM Tris/ 6M Gdn HCL pH 8.2, 100°C 20 mins, 80°C 2 hrs	RP-HPLC Ion Trap/Orbitrap	NSCLC = 151 RCC = 154	NSCLC FF = 166 phosphoproteome carried out on FF and FFPE: NSCLC FF = 56 NSCLC FFPE = 49 RCC FFPE = 42	(Gamez-Pozo et al., 2011)
Primary gynaecological tumour and matched brain metastases (n = 15)	Xylene with a graded ethanol series, Qproteome™ FFPE Tissue kit, 100 °C 20 min, 80 °C for 2 h. Centrifugation 14,000×g.	RP-HPLC Ion Trap	129 total	Primary = 76 Metastatic = 101,	(Yoshida et al., 2013)
Temporal bone	Deparaffinization in heptane for 1 hour, methanol to remove insoluble heptane layer, 2% SDS/ 100mM ABC/ 20mM DTT/ pH 8.5, sonicated, 70°C 1 hour, reduction and alkylation	RP-HPLC Ion Trap	FFPE=123	Frozen= 94	(Palmer-Toy et al., 2005)
Oral HPV lesions, HIV positive (n = 5), negative (n = 5)	Liquid Tissue™, 95 °C 1.5, iTRAQ	RP-HPLC MALDI-TOF/TOFiTRAQ	114		(Jain et al., 2008)
Sheep skeletal muscle and liver	Xylene with a graded ethanol series, 20 mM Tris HCL pH 8.8/2% SDS/200 mM DTT, 100 °C 20 min, 80 °C 2 h	HPLC Q-TOF	FFPE = 66	Frozen = 85	(Addis et al., 2009)

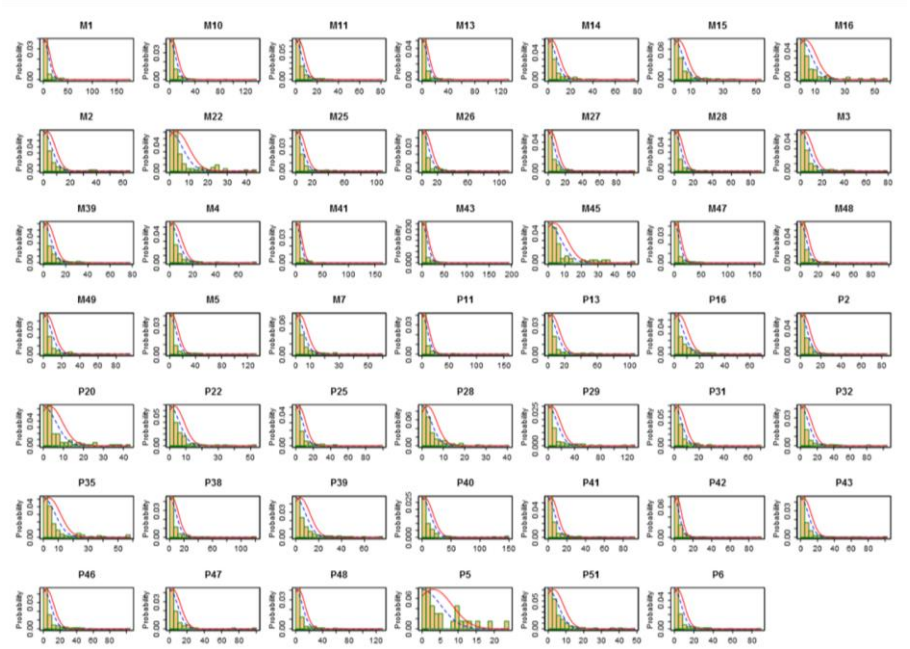
RP, reverse phase. HPLC, high pressure liquid chromatography. LTQ, linear ion trap. LMD, laser microdissection. FASP, filter aided separation protocol. SAX, strong anion exchange. CIEF, capillary isoelectric focusing. IEF, isoelectric focusing.

Appendix 2

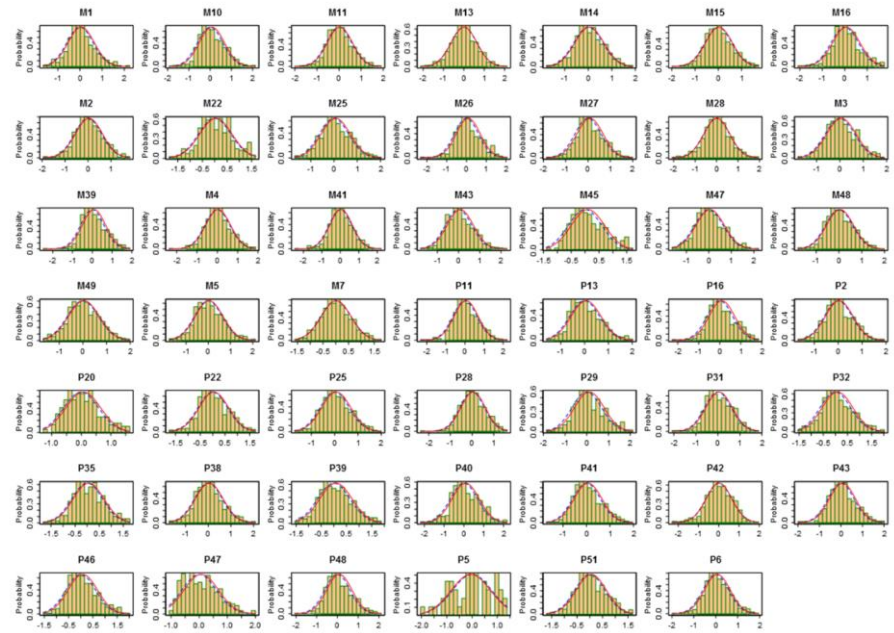
Appendix 2: Histograms of 1D and 2D SCC proteomic quantification data.

Histograms of proteomic quantification data for each cSCC was plotted in the R package Inferno. The distribution was evidently a mixture of normal (Gaussian) and non-normal data. Consequently, a conservative non-parametric approach was used in the subsequent statistical analyses.

1D Raw data

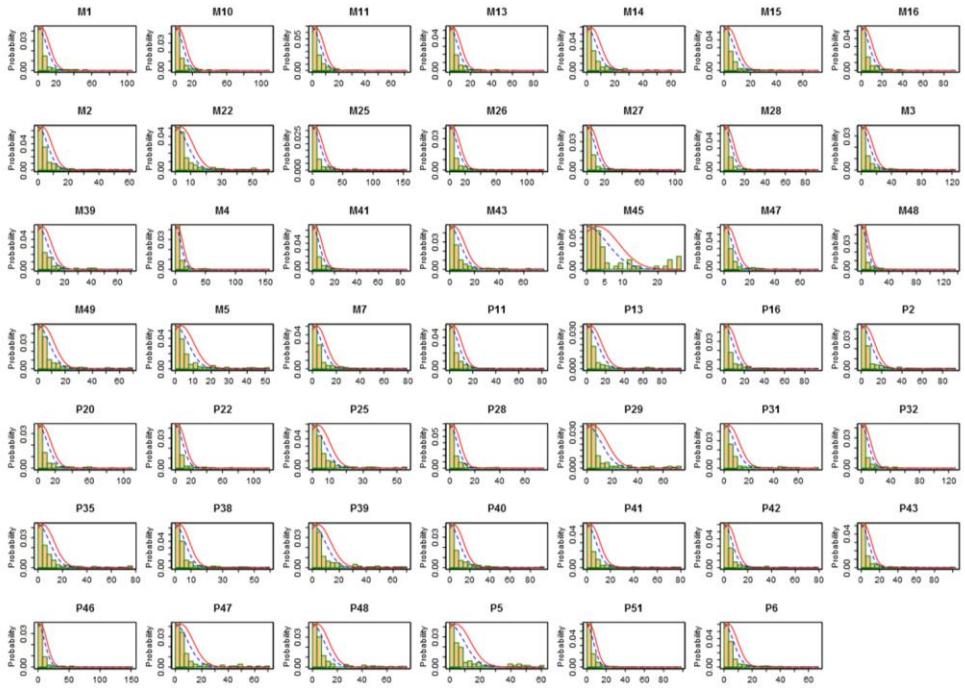


1D Log10 data

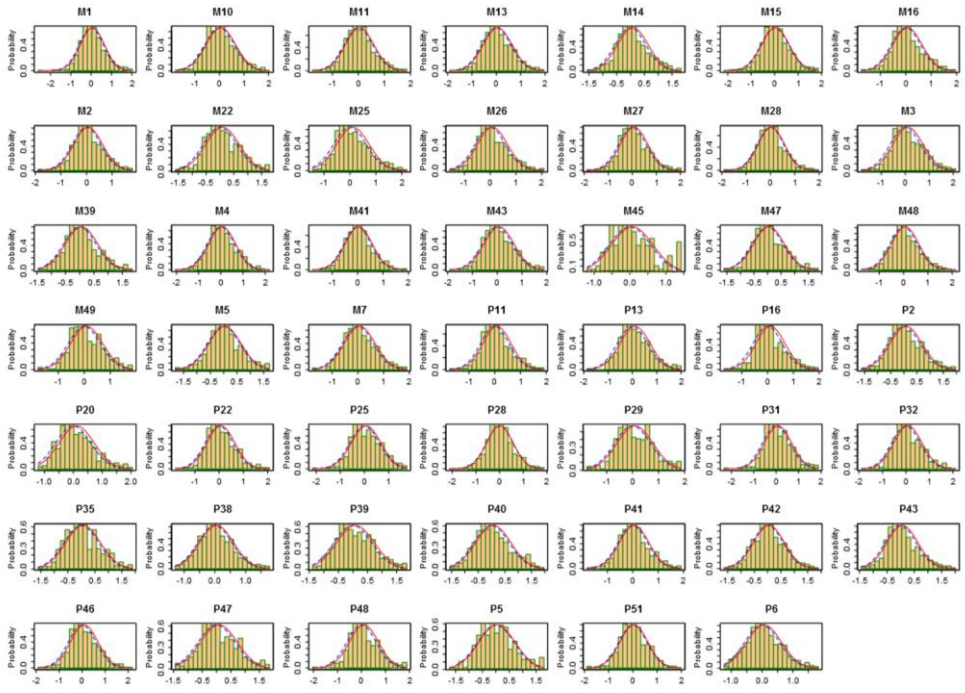


Appendices

2D Raw data



2D Log10 data

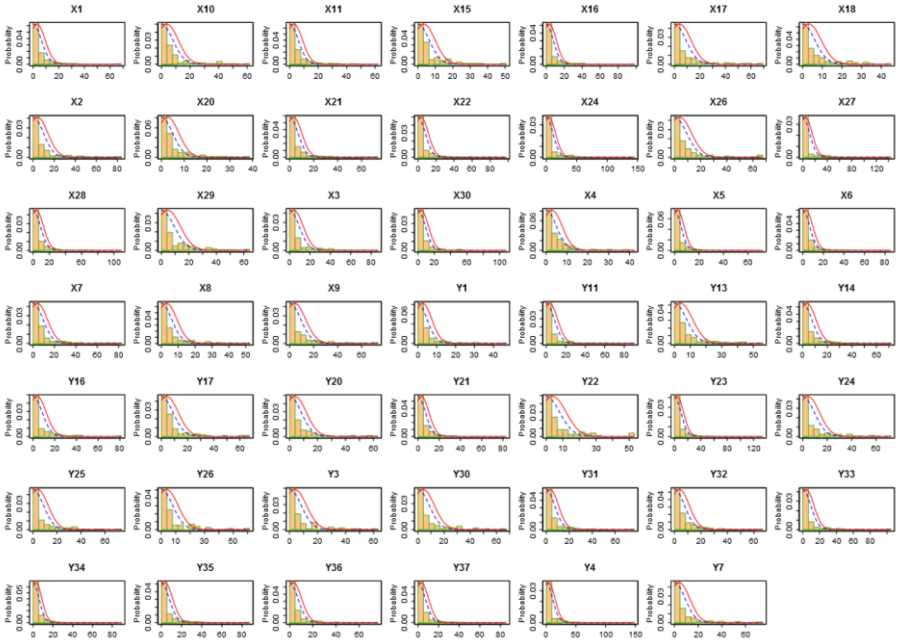


Appendix 3

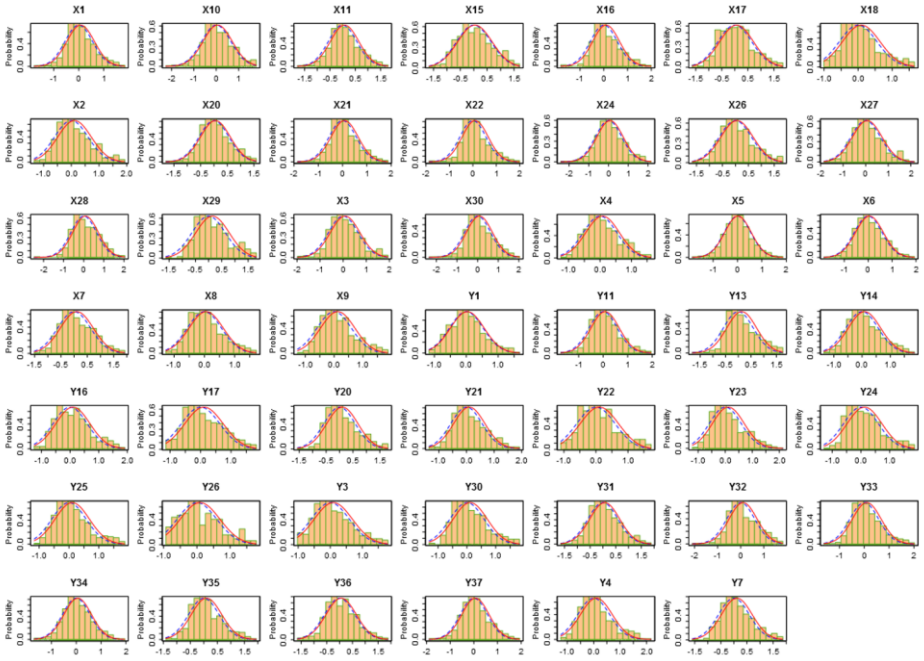
Appendix 3: Histograms of 1D and 2D melanoma proteomic quantification data.

Histograms of proteomic quantification data for each melanoma was plotted in the R package Inferno. The distribution was normal in some cases but non-normal in others. Consequently, a conservative non-parametric approach was used for statistical analyses.

1D Raw Data

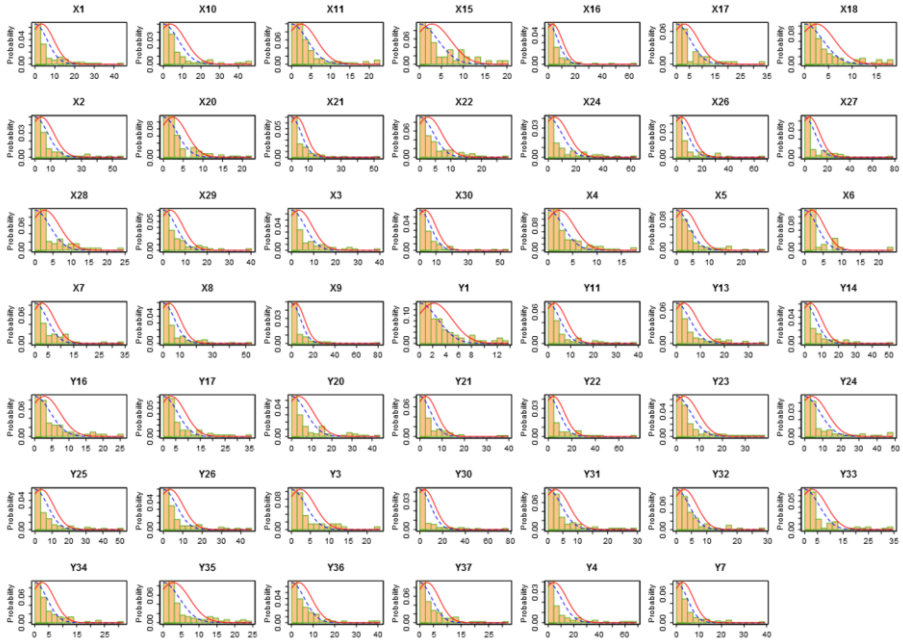


1D Log10 Data

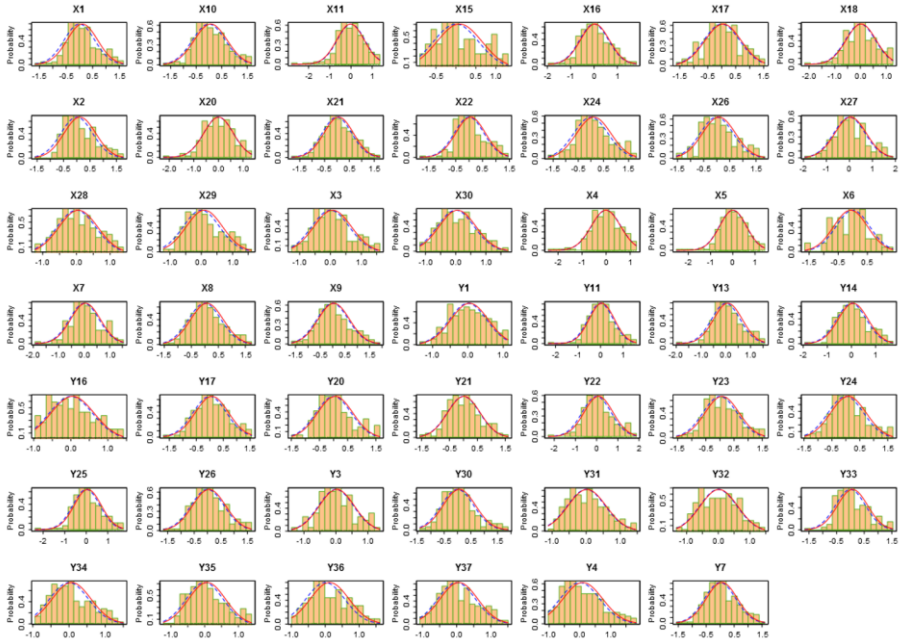


Appendices

2D Raw Data



2D Log10 Data



Appendix 4

Appendix 4: Machine learning algorithms used in this thesis

ALGORITHM	DESCRIPTION	REFERENCE
ADAPTIVE BOOSTING (ADABOOST)	Boosting algorithm that uses very weak learners to create voting weights and ultimately a strong learner. Changes weights of votes after each iteration to try and include all sample measurements. The more each iteration contributes to overall success the more weight it is given in strong learner	(Freund, 1996)
BOOTSTRAP AGGREGATORS (BAGGING)	Decision tree based method. Uses bootstrapping to create resamples of training data where each resample will contain a variety of the initial data. These resamples are then aggregated. Doing this reduces the chance of individual trees/branches overfitting the data as a tree is created for each resample, therefore only the important branches are maintained.	(Breiman, 1996)
C5.0	Builds decision trees where each node splits classes based on information gained. The attribute with the highest information gain is used to split it further.	(Salzberg, 1994)
CLASSIFICATION AND REGRESSION TREES (CART)	Creates both classification and regression trees where applicable.	(Breiman et al., 1984, Steinberg, 2009)
EXTREME GRADIENT BOOSTING: DROPOUTS MEET MULTIPLE ADDITIVE REGRESSION TREES (XGBDART)	Very similar to gbm except is designed for speed and to reduce overfitting. Uses weak learners and their loss function to create strong learner. However, employs occasional randomisation throughout creation of weak learners which reduces the correlation between learners, ultimately reducing overfitting. DART specifically uses dropouts to reduce over fitting	{Rashmi, 2015 #598}
EXTREME GRADIENT BOOSTING TREE (XGBTREE)	Very similar to gbm except is designed for speed and to reduce overfitting. Uses weak learners and their loss function to create strong learner. However, employs occasional randomisation throughout creation of weak learners which reduces the correlation between learners, ultimately reducing overfitting. DART specifically uses dropouts to reduce over fitting	(Chen and Guestrin, 2016)

Appendices

GENERLISED LINEAR MODEL (GLM)	Fits multiple linear regression models for continuous response variable to a discreet or continuous predictor. In this case the response variable is binary and so glm creates a logistic regression model.	(Nelder and Wedderburn, 1972)
GLMNET	Similar to above in creating logistic regression but uses penalised maximum likelihood to reduce complexity of model and reduce overfitting	(Nelder and Wedderburn, 1972)
GRADIENT BOOSTING MACHINE (GBM)	Boosting algorithm that uses weak learners to optimise gradient of loss function. Each weak learner attempts to reduce gradient of loss function subsequent to previous learner. Can result in overfitting	(Breiman, 1997)
K'S NEAREST NEIGHBOUR (KNN)	Uses number of neighbours to determine unknown data points	(Cover and Hart, 1967)
LINEAR DISCRIMINANT ANALYSIS (LDA)	Similar to a PCA plot whereby the algorithm looks for linear combinations of variables which best explain the data. LDA however, attempts to explicitly model the difference between the classes of data	(Cohen et al., 2014)
NAÏVE BAYES (NB)	uses Bayes theorem to calculate the probability of something given a training set where probabilities are learnt	(Duba and Hart, 1973)
NEURAL NETWORK (NNET)	based off of the animal neuronal network, an artificial neural network is a predictive tool that utilised several hidden layers, all interconnected, to predict an outcome based off of training data	(McCulloch and Pitts, 1943)
RANDOM FOREST (RF)	Several decision trees combined to gain the mode of the classes (if classification) or mean (if regression).	(Breiman, 1999)
REGULARISED RANDOM FOREST (RRF)	Uses regularisation to establish if a split in a branch achieves any additional information. If it does not then the prior node is weighted more	(Breiman, 1999)
SUPPORT VECTOR MACHINE (SVM)	Represents the data in points in space to separate the categories with a wide a margin as possible. Test data is then mapped onto this space to determine its predicted category	(Vapnik and Chervonenkis, 1974)