NMR-Based Metabolomic Studies of Breast Cancer

Warren Yabsley

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

School of Physics and Astronomy

School of Chemistry

February 2013

The candidate confirms that the work submitted is his/her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Publication: Nuclear Magnetic Resonance Spectroscopy in the Detection and Characterisation of Cardiovascular Disease: Key Studies

Yabsley W, Homer-Vanniasinkam S and Fisher J. *ISRN Vascular Medicine*, 2012, 2012: 11.

Parts of Chapter 4 are based on the work from this publication. W. Yabsley conducted the research for the review and wrote various drafts. S. Homer-Vanniasinkam identified the need for work in this area. J. Fisher edited the review's drafts. S. Homer-Vanniasinkam and J. Fisher commented on the final draft of the review.

Acknowledgements

I would like to thank my supervisor Dr. Julie Fisher. It is difficult to find words to express my gratitude for everything she has done throughout my PhD. I am very grateful for the guidance, time and support that she has provided as well as putting up with my 'unconventional' style (it sounds better than odd!). I would also like to thank my second supervisors Val Speirs (breast cancer) and Shervanthi Homer-Vanniasinkam (arterial disease) for their guidance and support.

There have been many other people and organisations that I would like to thank that have helped me at various points throughout my project: Michalis Charalambous for supplying plasma and urine samples; BCCTB for supplying tissue samples; Michael Twigg for supplying arterial disease samples; Cassey McRae for writing the MATLAB script for SHY analysis and Tamas Bansagi for modifying it; Jennie Dickinson for much time and patience with the GC work; and the White Rose Doctoral Training Centre for funding my PhD.

The people who I have worked with have made so much difference to my enjoyment of the PhD so I would to thank all past and present members of the Fisher group and everyone who was in Office 2.23, you know who you are. I would especially like to thank Cassey McRae for many fruitful discussions regarding both work and non-work!

In addition to those who I have worked with I have met many other great people through various organisations. This has enhanced my time in Leeds and always kept my motivation high. Thank you to Leeds CAMRA and the University Hiking, Real Ale and Pool and Snooker Clubs/Societies. I have put a lot into them all but have got so much back. Thank you Katie for the brilliant 2½ years we had together.

I would also like to thank Mother who has always believed in me and let me make my own decisions and to Margaret for always being there as well. I hope this makes you proud.

I have been extremely fortunate to have had such a wonderful experience during the whole time of my PhD, thank you all so much and I look forward to my travelling and who knows what else that life has to offer. That's enough drivel for now, well, almost!

Explore. Engage. Learn.

Abstract

Breast cancer is the most frequently diagnosed form of cancer and leading cause of cancer death in females. Current screening techniques, such as mammography, are inadequate. Screening reduces the number of deaths caused by breast cancer but also results in a large number of women with non-life threatening forms of the disease being treated unnecessarily. Initial screening can produce a false positive result, which causes much anxiety. Clearly there is a need for a more reliable approach; nuclear magnetic resonance (NMR)-based metabolomics has been used to this end. Plasma and urine from female patients with breast cancer or abnormal but non-cancerous breast state and extracts from tumour and adjacent normal tissue obtained from those afflicted with the disease have been analysed in an attempt to elucidate a biomarker of disease presence or tumour grade.

Aqueous extracts of tumour tissue compared to healthy adjacent (normal) tissue revealed increased levels of lactate, alanine, creatine, glutamate, glutamine, glycerophosphocholine (GPC), glycerophospholipids, phosphocholine (PCho), taurine, uridine-5'-diphosphate-N-acetylgalactosamine, (UDP-GaINAc) and uridine-5'-diphosphate-N-acetylglucosamine (UDP-GlcNAc) and decreased levels of glucose. Increased lactate and decreased glucose is in agreement with the Warburg effect whereby cancer cells predominantly produce energy by a high rate of glycolysis. Analysis of lipophilic extracts did not reveal a difference between tumour and normal samples. Metabolite levels did not indicate the tumour grade using either type of extracts.

Singly, plasma and urine data did not identify changing metabolite levels with breast cancer or tumour grade but correlations between the two biofluids were established using statistical heterospectroscopy (SHY). Through an unidentified species it was implied that a negative correlation existed between glycerol and certain lipids for patients with breast cancer.

These studies have provided useful insights into tumour metabolism and correlations of metabolites in different biofluids.

Contents

Figures

Tables

Abbreviations

1D	One dimensional
2D	Two dimensional
3D	Three dimensional
α-ILE	Allo-Isoleucine
β-AIB	β-Aminoisobutyric acid
AA	Amino acid
AAA	α-Aminoadipic acid
ABA	α-Aminobutyric acid
ABI	Ankle brachial index
ALA	Alanine
ASN	Asparagine
ASP	Aspartic acid
ATP	Adenosine 5´-triphosphate
B_0	Static field
BC	Breast cancer
BCCTB	Breast Cancer Campaign Tissue Bank
BMI	Body Mass Index
C-C	Cystine
CHD	Coronary heart disease
CI	Critical ischaemia
CoA	Coenzyme A
CPMG	Carr-Purcell-Meiboom-Gill
CPU	Central processing unit
cum	Cumulative
CVD	Cardiovascular disease
DAGPL	Diacylglycerophospholipid
DCIS	Ductal carcinoma *in situ*
DModX	Distance to model
$DMSO_2$	Dimethyl sulphone
DNA	Deoxyribonucleic acid
ECG	Electrocardiogram
EDTA	Ethylenediaminetetraacetic acid
ER	Oestrogen receptor
ER-	Oestrogen receptor negative
ER+	Oestrogen receptor positive
ERETIC	Electronic reference to access in vivo concentrations
FDR	False discovery rate
FID	Flame ionisation detector

FID	Free induction decay
g	Relative centrifugal force
GC	Gas chromatography
GC-MS	Gas chromatography-mass spectrometry
GC-TOF-MS	Gas chromatography coupled to time-of-flight mass spectrometry
GDH	Glutamate dehydrogenase
GLN	Glutamine
GLU	Glutamic acid
GLY	Glycine
GMD	Golm Metabolome Database
GPC	Glycerophosphocholine
GPE	Glycerophosphoethanolamine
H_0	Null hypothesis
H_1	Alternative hypothesis
HDL	High density lipoprotein
HER-	Human epidermal growth factor receptor negative
HER+	Human epidermal growth factor receptor positive
HER2	Human epidermal growth factor receptor
HIS	Histidine
HMDB	Human Metabolome Database
HMW	High molecular weight
HRT	Hormone replacement therapy
HYP	Hydroxyproline
Hz	Hertz
ILE	Isoleucine
IS	Internal standard
k	1024
KEGG	Kyoto Encyclopaedia of Genes and Genomes
kg	Kilogram
L	Litre
LC	Liquid chromatography
LCIS	Lobular carcinoma *in situ*
LC-MS	Liquid chromatography-mass spectrometry
LDL	Low-density lipoprotein
LEU	Leucine
LLOD	Lower limit of detection
LMW	Low molecular weight
LYS	Lysine
m	Metre
M_0	Equilibrium magnetisation
M_z	+*z*-magnetisation

MAS	Magic angle spinning
MET	Methionine
MHz	Megahertz
MI	Myocardial infarction
mm	Millimetre
ms	Millisecond
MS	Mass spectrometry
MSI	Metabolomics Standards Initiative
MVA	Multivariate analysis
N	Normal
NAD^+	Nicotinamide adenine dinucleotide
NADH	Reduced nicotinamide adenine dinucleotide
NMR	Nuclear magnetic resonance
NOESY	Nuclear Overhauser effect spectroscopy
NSTEACS	Non-ST-elevation acute coronary syndrome
OAA	Oxaloacetate
O-GlcNAc	O-linked β-N-acetylglucosamine
O-GlcNAcylation	O-linked β-N-acetylglucosamine glycosylation
OPLS-DA	Orthogonal partial least squares-discriminant analysis
ORN	Ornithine
PAD	Peripheral arterial disease
PC	Principal component
PCA	Principal components analysis
PCr	Phosphocreatine
PDC	Phosphatidylcholine
PDE	Phosphatidylethanolamine
PDH	Pyruvate dehydrogenase
PE	Phosphoethanolamine
PHE	Phenylalanine
PLS-DA	Partial least squares-discriminant analysis
ppm	Parts per million
PQ	Probabilistic quotient
PQN	Probabilistic quotient normalisation
PR	Progesterone receptor
PR-	Progesterone receptor negative
PR+	Progesterone receptor positive
PRESAT	Presaturation
PRESS	Predictive residual sum of squares
PRO	Proline
RAM	Random access memory
RD	Relaxation delay

RF	Radio frequency
RNA	Ribonucleic acid
RNase	Ribonuclease
RSD	Residual standard deviation
RSS	Residual sum of squares
s	Second
SAR	Sarcosine
SER	Serine
SHY	Statistical heterospectroscopy
SOM	Self organising map
SSX	Sum of squares of X
STOCSY	Statistical total correlation spectroscopy
T	Tumour
T_1	Longitudinal relaxation time constant
T_2	Transverse relaxation time constant
TCA	Tricarboxylic acid
THR	Threonine
TIA	Transient ischaemic attack
TMS	Tetramethylsilane
TRP	Tryptophan
TSP	3-trimethylsilylpropionic acid
TYR	Tyrosine
UDP-GaINAc	Uridine-5'-diphosphate -N-acetylgalactosamine
UDP-GlcNAc	Uridine-5'-diphosphate-N-acetylglucosamine
UK	United Kingdom
USA	United States of America
UV	Unit variance
VAL	Valine
v	Versus
VLDL	Very low density lipoprotein
WHO	World Health Organisation

Chapter 1. Introduction

1.1 Breast Cancer

1.1.1 Prevalence of Breast Cancer

Breast cancer caused over 450,000 deaths and was diagnosed in 1.38 million women worldwide in the latest year for which statistics are available (2008).[1] This equates to 14% of total female cancer deaths and 23% of female diagnoses.[1] Using the GLOBOCAN 2008 online tool[2] it is estimated that the number of cases diagnosed in the United Kingdom (UK) will rise from 46,000 in 2008 to 56,000 in 2030 with deaths rising by 25% from 12,000 to 15,000. Worldwide figures are estimated to continue to rise with nearly 750,000 deaths and 2.15 million cases diagnosed in 2030.

Although recorded incidences increased in many Westernised countries in the last two decades of the previous millennium, a substantial proportion will have been due to increased use of testing, *e.g.* through national breast screening programmes to identify the presence of breast cancer, therefore it is difficult to estimate any potential change to the prevalence in the population.[1] However, incidence and mortality rates continue to rise in many developing countries where screening is less common. In these areas the rises have been mainly attributed to changes in reproductive patterns, physical inactivity and obesity.[1]

Increased risk of developing breast cancer has been linked with alcohol consumption and reproductive factors that include a long menstrual history, nulliparity, recent use of postmenopausal hormone therapy or oral contraceptives and late age at first birth.[1] Incidence rates have started to decrease in some Westernised countries in the last few years, largely due to reduced use of postmenopausal hormone therapy, but breast cancer remains a leading cause of

death in women.[1] This is why further studies and alternative strategies are needed in the attempt to reduce the burden of this worldwide disease.

1.1.2 Types and Grade of Breast Cancer

Breast cancer, like other cancers, occurs due to mutations in genes, which results in alterations to cell proliferation, differentiation and death.[3] Certain proteins enhance resistance of cells to apoptosis (programmed cell death).[4] Breast cancer is a multifaceted disease comprised of distinct biological subtypes that have varied clinical, pathologic and molecular features with different prognostic and therapeutic implications.[5] The female reproductive hormones, oestrogen and progesterone, have a major impact on breast cancer and control postnatal mammary gland development.[6]

Oestrogen receptors (ER) are cellular proteins that bind to the most biologically active type of oestrogen. DNA synthesis, cell division and production of proteins including progesterone receptor (PR) proteins result from the interaction.[7] Approximately 50-80% of breast cancers have elevated levels of ER[7-9] compared to normal breasts for which the level is extremely low.[7] Presence of oestrogen or progesterone receptors results in breast cancers being referred to as ER+ or PR+, respectively, and similarly ER- and PR- refers to non-presence.

Breast cancers that are hormone receptor positive (either ER or PR, or both) are receptive to treatment by endocrine (anti-oestrogen) therapies. For post-menopausal women anti-aromatase agents can be administered that inhibit the final step of oestrogen synthesis[10] whereas tamoxifen, the primary treatment in pre-menopausal women, binds to the ER.[11] The anti-aromatase agents are designed to inactivate the aromatase in postmenopausal women, which is different to the prevalent aromatase form in premenopausal women.[9] Approximately 50-60% of the patients with ER+ tumours respond to endocrine therapy.[11]

Approximately 25-30% of breast cancer cases have increased levels of human epidermal growth factor receptor 2 (HER2) and the associated protein is present in abnormally high levels in the cells.[11] Breast cancers that overexpress HER2 are more aggressive forms of the disease[12] but trastuzumab (Herceptin) binds to HER2 and causes an increase in a protein that halts cell proliferation.[13] Triple negative breast cancers (ER-, PR- and HER2-) do not respond to endocrine therapy or other available targeted agents and chemotherapy is the current treatment therapy.[14] This form of the disease is associated with a lower survival rate.[14] The status of the three receptors can be combined to indicate breast cancer sub-types (luminal A, luminal B and basal like).[5]

Presence of the different receptors can be determined by immunohistochemistry staining of biopsy samples whereby specific antibodies bind to the proteins of interest.[15,16] A common method of assessing this is the Allred method[17] where separate scores are assigned to the overall stain intensity (0-3) and the percentage of tissue that is stained (0-5). A score of 0 indicates undoubtedly that the receptor is not present whereas undoubtedly positive status is assigned for scores of 8 with scores in between indicating moderate positivity.[18,19]

Breast cancer can be of an invasive or non-invasive nature. Ductal carcinoma is the most common invasive form and lobular carcinoma is the second most common form accounting for 4-15% of all breast cancers.[20] In addition to invasive breast cancer, non-invasive forms can manifest with ductal carcinoma *in situ* (DCIS) the most common type.[21] Lobular carcinoma *in situ* (LCIS) is less common than DCIS, approximate ratio 1:8, and is viewed as a marker of an increased risk of invasive breast cancer rather than as a true precursor lesion.[21] Carcinoma *in situ* can also be present for invasive breast cancers.

A grade can be calculated for invasive breast cancers to indicate the severity. Rakha *et al.*[22] state that histologic grading is now part of the minimum data set for breast cancer pathology reporting produced by the United Kingdom Royal College of Pathologists and European Commission and is endorsed by the World Health

Organisation (WHO) and the College of American Pathologists. The Nottingham grading system[23] is the most widely used histologic grading system.[22] Pathologists assess three morphological criteria (tubule formation, mitotic count and nuclear polymorphism), which are given a score 1-3, with 3 the most severe, and combined to give an overall score. Scores of 3-5, 6-7 and 8-9 are classified as grade 1 (low), 2 (intermediate) and 3 (high), respectively.[24]

1.1.3 Cell Proliferation

Given sufficient oxygen, most normal cells metabolise glucose *via* glycolysis into pyruvate that is oxidised in the tricarboxylic acid (TCA) cycle and CO_2 is produced. Reduced nicotinamide adenine dinucleotide (NADH, the reduced form of NAD^+) is produced in the TCA cycle, which fuels oxidative phosphorylation that allows adenosine 5´-triphosphate (ATP) production. NAD^+ results from oxidative phosphorylation and 36 ATP molecules are produced per glucose molecule *via* this process.[25,26] When there is insufficient oxygen, *i.e.* under anaerobic conditions, oxidative phosphorylation is not possible so ATP and NAD^+ cannot be produced. For glycolysis to proceed NAD^+ is required and can only be formed when pyruvate is converted to lactate, hence the large amount of lactate that is produced through anaerobic glycolysis. This produces two molecules of ATP per glucose molecule.[25,26] Cancer cells convert most glucose to lactate with only approximately 5% of pyruvate directed to the TCA cycle resulting in four molecules of ATP per glucose molecule.[25] As first observed by Warburg, aerobic glycolysis proceeds irrespective of whether oxygen is present and is often referred to as the Warburg effect.[25,26]

Through mutation, metabolism of cancer cells is adapted to facilitate the uptake and incorporation of nutrients, *e.g.* nucleotides, amino acids and lipids, for growth and to produce a new cell. Aerobic glycolysis, fatty acid synthesis and mitochondrial glutamine metabolism are three pathways that are considered important in cell growth and proliferation.[27] Glucose, glutamine and lipids are in near-constant supply and, unlike differentiated cells, there is no need to optimise ATP production for maximum energy so aerobic glycolysis can proceed. Lactate, produced by

aerobic glycolysis, may provide carbon, which is one of the components required for rapid cell division and some specialised non-proliferating cells could recycle lactate to glucose.[25] Glucose provides most of the carbon required for fatty acid synthesis, being initially converted to acetyl-coenzyme A (acetyl-CoA), which is used to produce citrate in the TCA cycle. Citrate production is also upheld by glutamine because it too supplies carbon.[28] Acetyl-CoA is also required for the biosynthesis of amino acids.[29] Changes to nutrient levels or usage pathways might predispose those with certain metabolic diseases, such as obesity, to cancer. Further knowledge of regulation of proliferating cell metabolism is required along with how tumour and whole body metabolisms interact to enhance cancer prevention strategies.[25]

This can be investigated by metabolomic studies that incorporate nuclear magnetic resonance (NMR) spectroscopy as the data acquisition method and multivariate analysis (MVA) to determine whether differences exist in the levels of metabolites. Determination of whether metabolic changes occur due to presence of breast cancer or severity of the tumour will aid the information and knowledge required to reduce the effects of this disease.

1.2 Nuclear Magnetic Resonance Spectroscopy

1.2.1 Relaxation

When a static magnetic field, $\boldsymbol{B_0}$, is applied individual nuclei precess around the *z*-axis at the Larmor frequency. Together, the vectors for individual nuclei are represented by the bulk magnetisation vector, which at equilibrium is aligned along the +*z*-axis.[30,31] When a perpendicular radio frequency (RF) pulse is applied the populations of the nuclear spins are perturbed and the bulk magnetisation is rotated towards the *x*-*y* plane (Figure 1.1). If the RF pulse is applied for the correct length of time the bulk magnetisation will transfer 90^{o} because the populations of α and β states of the nuclei that make up the bulk magnetisation are equalised.[30,31] Following this, through the process of relaxation, the system returns to equilibrium

through the spins dissipating energy, thus allowing the population of the energy levels to return to the Boltzmann distribution.[30,31]

The dissipation of energy from the spins to the surrounding lattice and subsequent return of the bulk magnetisation to its equilibrium state on the +*z*-axis is known as longitudinal, or spin lattice, relaxation.[30,31] Oscillating magnetic fields at the correct Larmor frequency are not abundant but are required for excited nuclei to off-load energy and regain the ground state. The fields are produced by the motion of neighbouring nuclei and the process of regaining the ground state in this way is characterised by a time constant, T_1, the longitudinal relaxation time. Despite longitudinal relaxation being an enthalpic process, the change in temperature is undetectable.[30,31]

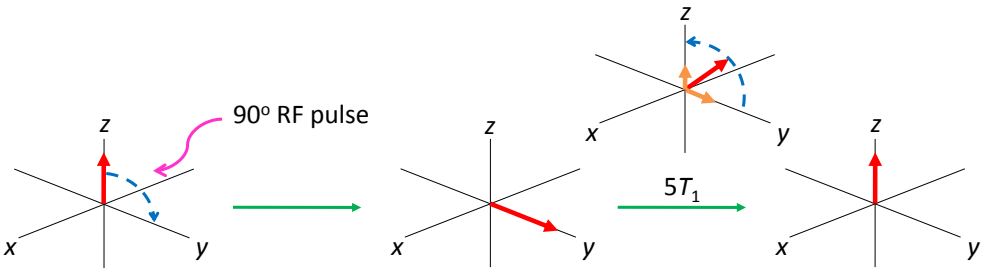


Figure 1.1 The effect of a 90° RF pulse on bulk magnetisation and recovery to the *z*-axis.

By assuming equilibrium is approached exponentially, the recovery can be explained by the Bloch equation:

$$\frac{d\boldsymbol{M_z}}{dt} = \frac{\boldsymbol{M_0}-\boldsymbol{M_z}}{T_1} \qquad (1.1)$$

where $\boldsymbol{M_0}$ is the equilibrium magnetisation, $\boldsymbol{M_z}$ is the +*z*-magnetisation and T_1 is the longitudinal relaxation time constant (though often referred to as the longitudinal relaxation time).[30,31] By solving the above equation, the longitudinal magnetisation at time *t* can be obtained:

$$\boldsymbol{M_z} = \boldsymbol{M_0}(1-e^{-t/T_1}) \qquad (1.2)$$

To achieve essentially full recovery of $\boldsymbol{M_z}$ magnetisation a period of $5T_1$ is required.[30,31]

After a 90° RF pulse the net magnetisation is in the *x*-*y* plane. If all spins experienced an identical magnetic field, precession at the same frequency would occur. However, each spin will experience a different magnetic field due to inhomogeneity of $\boldsymbol{B_0}$ and local magnetic fields that result from intramolecular and intermolecular interactions.[30,31] If the sample is thought to be divided into extremely small regions whereby the field was uniform within each region, magnetisation vectors would precess at the same frequency within each region. These regions are known as isochromats.[30,31] Spins subject to greater magnetic fields will precess more quickly and move ahead of the bulk magnetisation vector whilst those spins experiencing weaker magnetic fields will precess more slowly and fall behind the bulk magnetisation vector.[30,31] Given sufficient time, this fanning-out of the isochromats will lead to zero net magnetisation in the *x*-*y* plane (Figure 1.2). This type of relaxation is known as transverse, or spin-spin, relaxation, which follows an exponential decay represented by the time constant T_2^*:

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_{2(\Delta \boldsymbol{B_0})}} \tag{1.3}$$

where local magnetic fields are characterised by T_2, inhomogeneity of the static field is represented by $T_{2(\Delta \boldsymbol{B_0})}$ and the combined relaxation time characterised by T_2^*.[31]

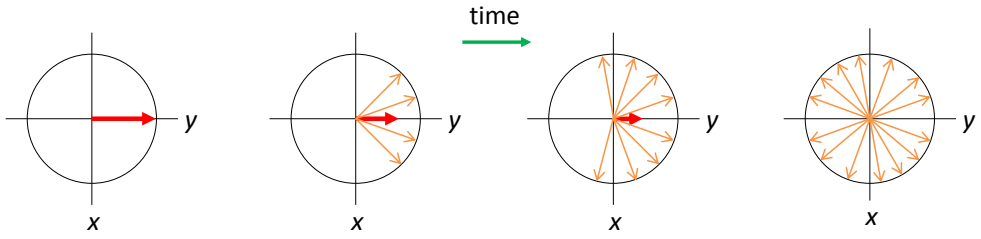


Figure 1.2 Individual spins fanning out causing zero net magnetisation in the *x*-*y* plane (transverse relaxation).[31]

Transverse relaxation is an entropic process because energy is transferred between spins rather than lost to the surrounding lattice as per longitudinal relaxation.[30,31] Due to longitudinal relaxation being the return of the magnetisation to equilibrium and the +*z*-axis it follows that $T_2 \leq T_1$.[31]

Vectors that dephase quickly have short T_2^* times and because of large frequency differences between the vectors, broad lineshapes result as characterised by large half-height linewidths of the peaks for that molecule.[31] Long T_2^* times correspond to slow transverse relaxation and result in narrower lineshapes. High molecular weight (HMW) molecules typically display broader resonances than low molecular weight (LMW) molecules but these can be attenuated and resonances from LMW molecules are observed when using a spin-echo pulse sequence, such as the Carr-Purcell-Meiboom-Gill (CPMG) sequence.[30,31]

1.2.1.1 Spin-Echo Pulse Sequence

The spin-echo pulse sequence was devised to measure an accurate T_2 value of a sample by attempting to remove the effect of field inhomogeneity.[30,31] The initial 90^o_x pulse pushes the magnetisation onto the *y*-axis where the inhomogeneity of the static field causes isochromats to fan-out during the time period τ. A second pulse rotates all the isochromats by 180° around the *x*-axis to the -*y*-axis allowing the differently precessing isochromats to catch up with the average position, hence the magnetisation vector is refocussed[30,31] (Figure 1.3).

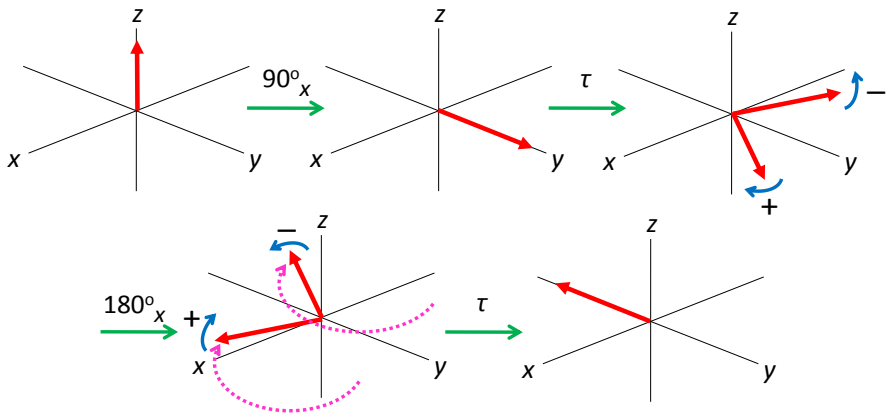


Figure 1.3 Magnetisation during the spin-echo pulse sequence and refocusing of the magnetic vectors dephased by field inhomogeneity.[31]

The signal decays during the initial τ delay then increases to an echo peak after the 180° pulse and second τ delay, therefore reappearing after 2τ. The second part of the echo is an exponential decay and is Fourier transformed to produce a spectrum

which is free from field inhomogeneity effects.[30,31] However, local magnetic fields that result from intramolecular and intermolecular interactions cause diffusion of molecules which is not refocussed so the intensity of the echo is reduced by T_2.[30,31]. The Carr-Purcell pulse sequence uses many repetitions of short τ delays and 180^{o}_{x} pulses resulting in the intensity of the echoes decaying according to T_2 but because of the large number of echoes, the loss between each echo due to diffusion is small. However, if the 90^{o}_{x} pulse is incorrectly measured, any errors will become cumulative when the 180^{o}_{x} pulses are performed leading to less transverse magnetisation and resulting in a less intense signal.[30,31] The CPMG pulse sequence overcomes this problem by replacing the rotation of the magnetisation around the *x*-axis with that around the *y*-axis, thus making the error in the pulse width non-cumulative and refocussing of the vectors is achieved after an even number of 180^{o}_{y} pulses[30,31] (Figure 1.4).

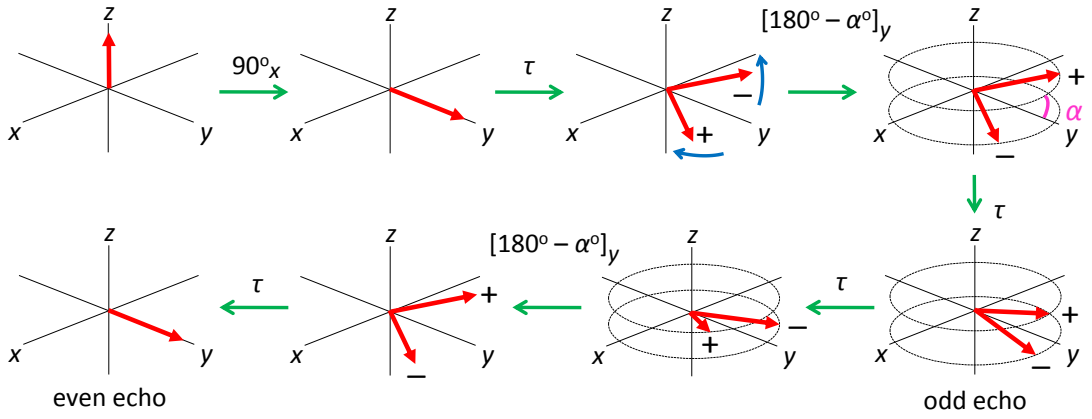


Figure 1.4 The operation of the Carr-Purcell-Meiboom-Gill (CPMG) sequence in the presence of pulse imperfections.[31]

NMR spectra can be edited according to molecular size due to differences in T_2 values[30]. Large molecules have short T_2 times so during the τ delay the isochromats will have fanned-out. Small molecules, which have long T_2 times, will only be starting to fan-out so will be refocussed after the 180^{o}_{y} pulse and detected due to being in the *x*-*y*-plane[30,31]. The intensities of signals from large molecules in biofluids, such as proteins and lipoproteins will thus be reduced (Figure 1.5).

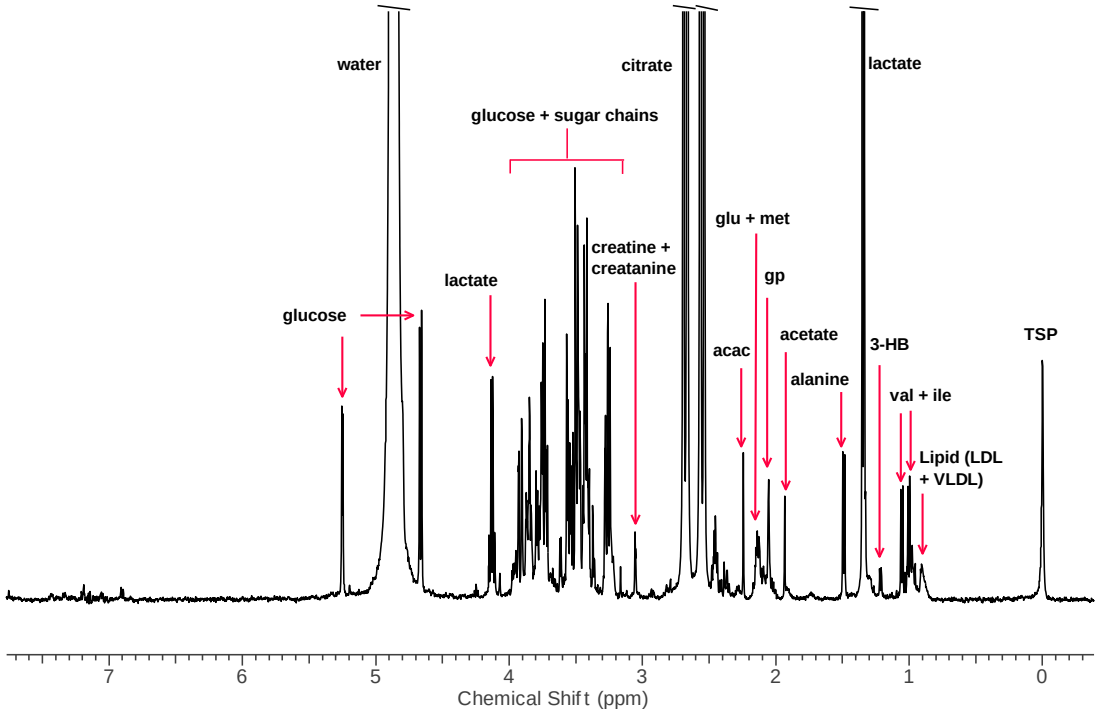


Figure 1.5 500 MHz ^{1}H-NMR spectrum of human blood plasma. Abbreviations: 3-HB, 3-hydroxybutyrate; acac, acetoacetate; glu, glutamate; gp, glycoprotein; ile, isoleucine; LDL, low density lipoprotein; met, methionine; TSP, 3-trimethylsilylpropionic acid; val, valine; VLDL, very low density lipoprotein. Citrate is present because it is the anti-coagulant used in the collection tubes.

1.3 Metabolomics

Living organisms will attempt to maintain homeostasis when confronted with disruptive influences such as disease and drug treatment. These stimuli can cause alterations to body conditions that disrupt the normal ratio of metabolites and hence alter the biofluid and tissue profile.[32] The normal ratio will be specific to each individual because of influences such as genetic composition, diet and lifestyle but loss of homeostasis is the initial step towards disease.[33]

Metabonomics was first described in 1999 by Nicholson *et al.*[34] but the underlying principles and methodologies were developed in the previous decades whilst the concept of studying perturbations to body fluids is centuries old. Over 3,000 years ago disease indication was recognised as being linked to changes in biological fluids[33,35] but in the 1940s analytical techniques were used to profile these fluids to investigate normality. During the 1970s and early 1980s pattern recognition started to be used in conjunction with analytical techniques, mainly gas chromatography (GC) and GC coupled with mass spectrometry (gas chromatography-mass spectrometry; GC-MS).[33] Nuclear magnetic resonance (NMR) spectroscopy developed greatly during the latter decade and became used for profiling studies.[36] The basis of metabolomics had been established.

The total collection of endogenous metabolites is known as the metabolome and is estimated there are at least 3,000 metabolites in the human metabolome.[37,38] Metabolomics and metabonomics are both terms for metabolome studies and in these approaches sampling provides a picture or snapshot of the metabolome at one point in time.[39] Each has a unique definition but the methods and approaches of metabonomics and metabolomics are now highly convergent[32,40] with the distinctions being mainly historical.[41] Metabolomics has been defined as the comprehensive analysis of all measurable metabolites in the metabolome under a given set of conditions, and metabonomics as the measurement of the fingerprint of biochemical perturbations caused by disease, drugs and toxins.[42] Metabolomics will be used throughout the report because this term is becoming more commonly used amongst research groups.

Proteins, transcripts and genes can be measured in a similar way to metabolites by the corresponding ‘omics’ (Figure 1.6). The number of metabolites is less than the number of any of the three aforementioned body species but metabolite levels can be regarded as the final downstream product of biological systems,[43,44] thus providing an excellent insight into the result of influences on the system. Additionally, running costs associated with metabolome studies are generally lower

than with the other 'omics'.[42] The most common biological samples used for metabolomics studies are urine and blood plasma and serum, with cell and tissue extracts, tissue, seminal fluid, amniotic fluid, cerebrospinal fluid, saliva, synovial fluid, digestive fluids, blister and cyst fluids, lung aspirates and dialysis fluids also having been used.[32,45,46] The more common biological samples are simple to collect and easily provide sufficient volume for analysis.

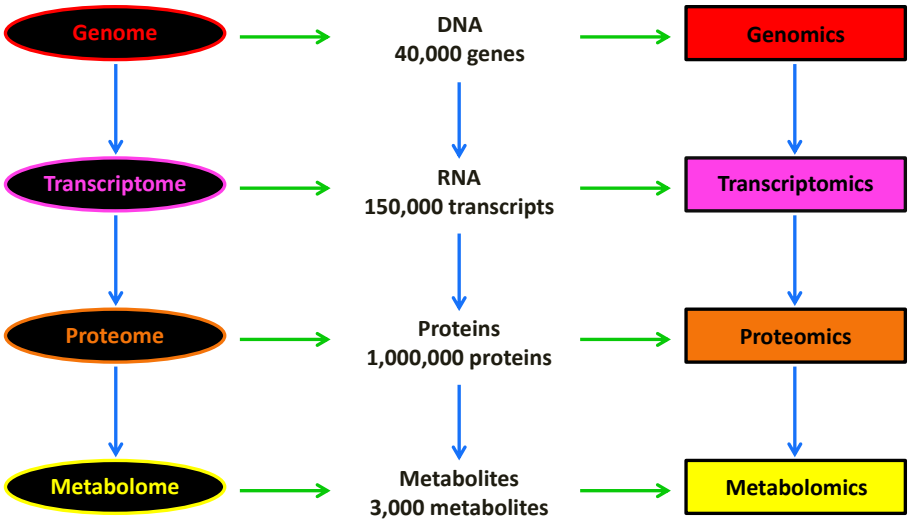


Figure 1.6 The relationship between the 'omics' technologies. DNA, deoxyribonucleic acid; RNA, ribonucleic acid.

Despite providing a picture of the metabolome at a unique time, observation over a period can be achieved and monitored through dynamic modelling when samples from multiple time points are obtained. Every subject acts as their own control thus potentially negating the effect of every individual having a different 'standard' metabolite profile allowing variation around the individual 'standard' profile to be the focus of the investigation.[47] However, it is often not clinically feasible to obtain suitable samples at multiple time points; progression of slow forming disease is more difficult to monitor than for shorter term reactions, for example, in the days following organ transplantation when the patient is already in a clinical environment and it is likely there will be less variables in sample handling.

Today, the two main platforms used for data acquisition in metabolomics are NMR spectroscopy and mass spectrometry (MS). The latter is combined with a chromatography method, commonly liquid chromatography (LC) or GC, to enable separation of the metabolites prior to detection. Advantages and disadvantages are associated with each technique so combination of the two provides the most thorough analysis.[48] NMR spectroscopy was the more commonly employed method in early metabolomics studies[34,49] and has the advantages of high reproducibility,[50] requiring little preparation for most sample media[51] and being sample non-destructive. MS has emerged as an alternative method that is now frequently applied due to its high sensitivity[45,52] but metabolite identification is not as universal as for NMR spectroscopy[52,53] and different separation techniques are required for different substance classes.[53,54]

Sensitivity is the largest weakness of NMR spectroscopy, which presents itself as a detection limit in the micromolar range for common biological samples rather than the nanomolar range as for MS.[37] A large water peak, which can obscure metabolites, is always observed for biofluids. Additionally, low molecular weight metabolites can be obscured by the broad envelope of high molecular weight resonances of proteins. Both problems can be substantially addressed by application of appropriate pulse sequences.[55,56]

1.4 Data Analysis

1.4.1 Chemometrics

Chemometrics uses statistics and pattern recognition techniques to analyse chemical numerical data.[57] Metabolomics studies generate many more measured variables than there are observables (samples). Large and complex data tables are constructed that are not able to be summarised by traditional statistics.[58] The first stage of chemometrics involves data reduction through the process of spectral binning, whereby the spectrum is divided into smaller regions (bins) that can be

compared to the bins of other spectra in the data set. For each bin, the area under the spectral curve is calculated and analysed using multivariate analysis (MVA).[57] Additional data processing steps, namely normalisation and scaling, are required between data reduction and MVA to ensure fair comparisons are made.[42,59] An overview of the process is given in Figure 1.7.

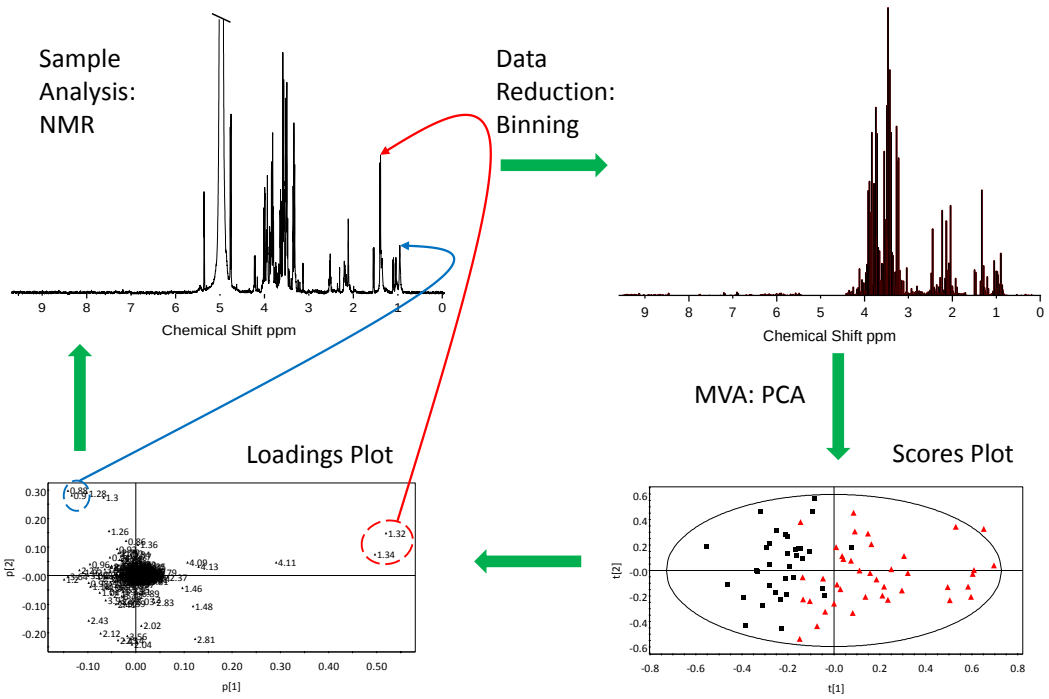


Figure 1.7 Overview of chemometrics. For this example, plasma is analysed using NMR spectroscopy followed by binning as the data reduction method. Principal components analysis (PCA) is the MVA employed producing a scores plot and a loadings plot, which can be used to identify trends across the set of patients.

1.4.1.1 Data Reduction

Data reduction is commonly achieved through binning (alternatively known as bucketing), whereby the spectrum is divided into smaller regions (bins) and the area under the signals contained in each bin is obtained. The binning step is required to take account of small changes in chemical shifts between samples so the same signals can be compared across spectra.

Traditionally, spectra have been divided into equal sized bins of 0.04 parts per million (ppm).[59,60] This bin width has been used for urine because it is a good

compromise between resolution and variation in peak positions.[59] All binning techniques result in a loss of information due to a reduction in the number of data points,[60] typically a 64k point NMR spectrum is reduced to about 250 variables (bins).[61] Chemical shift variations can occur for some components, such as citrate, due to ionic strength and pH variation across the samples.[59] Variations in these factors are less prevalent for other biofluids, such as plasma and serum, but the same bin width has been conventionally adopted in many studies involving these biofluids. Some of the first studies using MVA in metabolomics used different bin widths: 0.02 ppm[62] and 0.05 ppm.[63]

Smaller bin widths have been used in more recent studies[64,65] because many computational limitations on data matrix sizes have been removed allowing improved interpretation of the results by relating peaks with individual metabolites but the size is limited by peak shifting. Variable bin widths have also been employed[66] whereby the bin width is allowed to vary by a percentage each way of a set bin width. This is aimed at accounting for small chemical shift differences by identifying local minima in the spectra and integrating a peak in a single bin[67] which extends the advantages, and reduces the disadvantages, of smaller bin widths. However, even with adoption of variable smaller bin widths, a bin can integrate two peaks that change oppositely with, for example, disease versus control, resulting in the bin amplitude not accurately reflecting the intensity of either peak.[68]

An alternative method to binning to ensure comparable signals are analysed is to perform peak alignment on the original data thus eliminating the need for data reduction and allowing full resolution data to be used in further analysis. However, the magnitude of misalignments within a single sample may not be consistent so that a simple alignment correction across an entire spectrum cannot be performed and more sophisticated computer programmes are required.[68]

1.4.1.2 Normalisation

For each spectrum the reduced data are then normalised to accommodate concentration changes that are unrelated to the factors being investigated. Urinary volume, and hence concentration, can vary greatly whereas greater regulation occurs for plasma and serum.[59] Compared to normal urine, drug effects and food deprivation can cause dilution by a factor of ten.[69] Two common normalisation procedures are constant sum and constant peak.[59] The former is the most widely employed and consists of normalising to the sum of the integrals of the whole spectrum. For every spectrum, each bin integral is divided by the total spectrum integral and the sum of all the new integrals is set to one.[70] For all biofluids and tissues, the influence of down-regulation of certain metabolites should be approximately balanced by the up-regulation of other metabolites.[69] However, this approximation can fail if certain metabolite levels change vastly, for example, the addition of a metabolite from a drug or ethanol from alcohol consumption, will cause the other peaks in the spectrum to appear to decrease because the total spectrum integral is greater.[59] Constant peak normalisation requires an internal reference compound or metabolite to be present at constant concentration. Proteins, including albumin, in plasma and serum for example, can bind non-specifically to certain reference compounds, such as 3-trimethylsilylpropionic acid (TSP),[71] and cause variation in their free concentration. An alternative method is probabilistic quotient normalisation[69] whereby for each test spectrum every bin integral is divided by the integral of the same bin of a reference spectrum. The median of these quotients is calculated and the test spectrum bin integrals are divided by this value.[69] This negates potential problems with highly variable metabolites that would affect constant sum normalisation.

1.4.1.3 Scaling

Scaling regulates the relative importance of each variable. Metabolites that have a high concentration are not always the most informative, and without scaling, variation in lower concentration metabolites would be overlooked.[72] Scaling occurs

for the integral of one bin (variable) throughout the whole sample series and occurs for the variables of the whole spectrum.[59]

All of the subsequent scaling methods firstly centre the data to allow differences in concentrations to occur around zero instead of around the concentration mean of the metabolite resulting in the only variation being that between the samples without any offset.[73] Mean-centred scaling is just the centring process; each value of the normalised variable has the mean of the whole sample set for that normalised variable subtracted from it. Without further scaling, high concentration metabolites have a large contribution.[74] Unit variance (UV) incorporates standard deviation into the scaling factor, whereby centred data is divided by the standard deviation of the whole sample set for that normalised variable.[72,73] All variables have equal potential to influence the model but as a result the noise level of spectra can contribute strongly.[74,75] The effect of pareto scaling is between that of mean centring and UV, whereby the influence of changes in the concentration of highly abundant metabolites is decreased more than that for changes in less abundant metabolites.[72] For pareto scaling, centred data is divided by the square root of the standard deviation of the whole sample set for that normalised variable.[73] Pareto scaling has been recommended for metabolomics data[58] but is still prone to enhancing the contribution of variables with high variance[76] and alternatives have been used.[77]

1.4.1.4 Multivariate Analysis

MVA considers several related variables simultaneously.[78] Multivariate statistical methods can be unsupervised, such as principal components analysis (PCA), or supervised, as for partial least squares-discriminant analysis (PLS-DA). Unsupervised models are used to reveal grouping within a data set without previous knowledge of the class of individual samples, whereas supervised models use class information to produce models, from which the class of further samples can be predicted.[79-81] All multivariate statistical analyses were performed using SIMCA-P+ software.

1.4.1.4.1 PCA

PCA represents a multivariate data table (matrix ***X***) as a low dimensional plane, providing an overview of the data that may reveal groups of observations, trends and outliers.[58] For *n* observations (samples), a *k* dimensional space is constructed where *k* is the number of variables (bins). Each sample is represented by one point. When *k* is greater than three, *k* dimensional space is difficult to visualise but mathematically this space is similar to two dimensional (2D) or three dimensional (3D) space.[82,83] The first principal component (PC) is fitted, which is the linear combination of the original variables, and is the line describing maximum variation between the points in multidimensional space, hence the line best accounts for the shape of the swarm of points. Following, the second PC, which accounts for the next largest amount of variation, is fitted, then the third PC and so on.[58] Each PC is orthogonal to all of the other PCs, thus ensuring there is no correlation between the PCs.[84] Most of the information in the original matrix is contained in the first few PCs, with two to five PCs usually providing a representative overview.[85,86] Any two PCs can form a plane, onto which each observation can be projected (Figure 1.8a). A scores value, which is the coordinate value along the PC line, can now be obtained and the resultant scores plot created.[86]

Each PC is related to the original variables through the angles between the two, α_1, α_2 etc. where the number relates to the variable. The loadings vector of each PC, ***p***, is related to all of the cos α values between that PC and the original variables (Figure 1.8b).

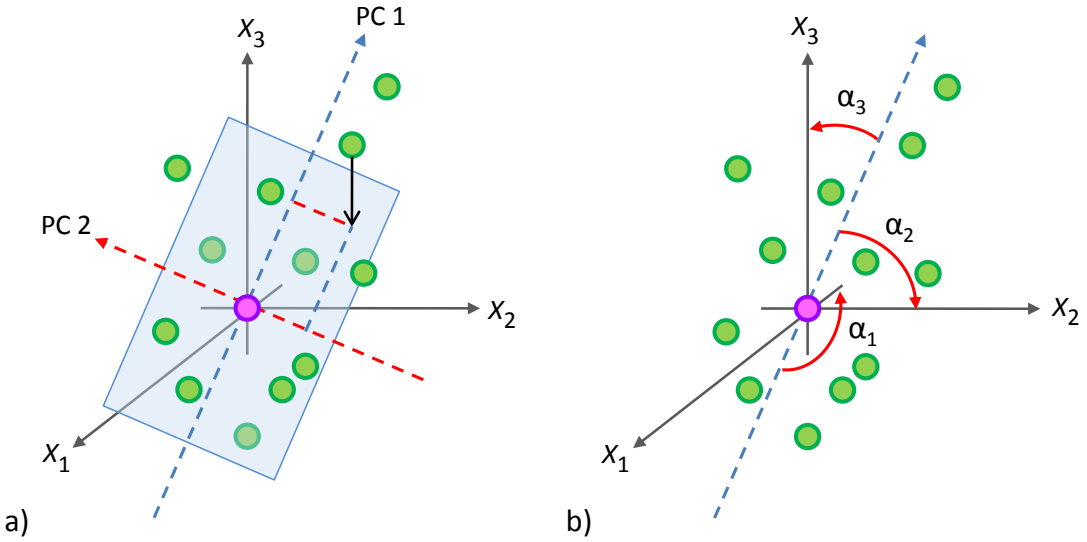


Figure 1.8: Overview of PCA. a) Formation of PCs and b) obtaining loadings vectors.

If a PC lines up with a variable axis the angle will be small so the loading will be close to 1 indicating strong influence, whereas if the PC is nearly orthogonal to the variable axis the loading will be close to 0, hence showing little influence. A PC opposite to a variable axis provides a loading close to -1 and hence a strong negative influence[86]. However, $\sum \boldsymbol{p}^2 = 1$ so the greater number of variables, the lower numerically the loadings will be.[78] Loading values can be plotted, in an analogous manner to the score values, to create a loadings plot. Directions in the scores plot correspond to directions in the loadings plot which allows the bins to be identified that cause separation between samples.[58]

The original matrix, ***X***, is summarised by ***T*** (scores) and ***P*** (loadings) matrices:

$$\boldsymbol{X} = \boldsymbol{T} * \boldsymbol{P} + \boldsymbol{E} \qquad (1.4)$$

where ***E*** is the residual matrix and contains the unexplained variation.[83] The better the model, the smaller the value of ***E***.[86]

Scores plots allow strong outliers to be visualised and can be detected by Hotelling's T^2, a multivariate generalisation of Student's *t*-distribution, by defining a 'normal' area and providing a tolerance region for the data in a two-dimensional scores plot.[86] Distance to model (DModX) is the detection tool used to establish moderate

outliers. Geometrically, the greater the distance of projection of the sample onto the plane of PCs, the greater the DModX value and the less well represented the point is by the two PCs.[86]

The number of PCs in a model is determined by the difference in degree of fit and predictive ability. The goodness of fit of a model is estimated by R^2, the explained variation:

$$R^2 = 1 - (\text{RSS}/\text{SSX}) \tag{1.5}$$

where RSS (residual sum of squares) = $\sum(\text{observed} - \text{fitted})^2$ and SSX (sum of squares of $\boldsymbol{X}$) represents the total variation in the $\boldsymbol{X}$ matrix after mean centring. When the number of components increases, R^2 tends towards 1 but at this value everything is predicted, including noise, which is detrimental.[86]

The predictive power of a model is summarised by $Q^2\boldsymbol{X}$, which uses cross validation, whereby the data is divided into groups and the model is generated devoid of one group. This model is used to predict the deleted group resulting in a partial predictive residual sum of squares (PRESS) value and this process is repeated many times to ensure the process occurs for each group.[81] SIMCA-P+ software (Umetrics, Umeå, Sweden) uses seven groups as default, though this can cause problems if samples are collected over a week because the samples could be dependent on the day, thus non-random groups would be removed. Removal of groups occurs firstly in rows (observations) followed by columns (variables).[86] The partial PRESS values are summed[81]:

$$\text{PRESS} = \sum(\text{observed} - \text{predicted})^2 \tag{1.6}$$

and

$$Q^2 = 1 - (\text{PRESS}/\text{SSX}) \tag{1.7}$$

If a new PC enhances the predictive power compared to the previous PRESS value then the new component is retained.[86]

Certain aspects of the usefulness of the model are associated with $R^2\boldsymbol{X}$ and $Q^2\boldsymbol{X}$.[81] $R^2\boldsymbol{X}$ is always greater than $Q^2\boldsymbol{X}$, high $R^2\boldsymbol{X}$ and $Q^2\boldsymbol{X}$ values are desirable, the difference between $R^2\boldsymbol{X}$ and $Q^2\boldsymbol{X}$ should not exceed 0.3 and $R^2\boldsymbol{X}$ greater than 0.5 is considered good and typical for metabolomics.[86]

1.4.1.4.2 PLS-DA

Unlike PCA, partial least squares-discriminant analysis (PLS-DA) incorporates class identifiers to improve separation. In addition to an ***X***-matrix of observations (samples) and variables (bins) that is used in PCA, a ***Y***-matrix is created that consists of the same observations but the variables are classes,[86] *e.g.* case or control. Observations belonging to the class have a value of one and observations not belonging to the class have a value of zero. Geometrically, for both matrices a swarm of points is present in *k*-dimensional space where *k* is the number of variables. For the first component a line is added in each swarm that best describes the swarm, in a similar way to PCA, but additionally provides good correlation between the points along the line. This maximises the separation between observations belonging to the different classes.[86] Projecting the observations onto the first component gives scores values: $\boldsymbol{t}_1$ and $\boldsymbol{u}_1$ for ***X*** and ***Y*** swarms, respectively. The contribution of each variable in the ***X***-matrix to the modelling of ***Y*** is reflected by the weights ($\boldsymbol{w}^*_1$) with $\boldsymbol{c}_1$ the equivalent for the *Y*-variables. Further components can be added to improve the approximation of, and correlation between, ***X*** and ***Y***. Similarly to PCA, the quality of PLS-DA models can be initially determined by R^2 and Q^2 values, however, the explained and predicted variation of the ***Y***-data ($R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$, respectively) are considered; the values indicative of good models for PCA models also apply to PLS-DA models.[86] The number of components is assessed in a similar manner to PCA.[86]

The aim of PLS-DA models is to predict class membership of samples from the ***X***-data. Models can overfit the data resulting in separation between classes arising by chance. Ideally, to validate the model new samples that were not used in the building of the model would have their class predicted. This is not always possible

for smaller sample sets. Alternative cross-validation methods are available.[79,87] Two methods have been employed in this study for prediction of samples depending on the sample numbers available. One-third of samples can be randomly excluded from generation of a new model and the classes of these samples predicted. This is repeated twice so that all samples have been predicted. For smaller groups 'leave one out' cross-validation builds a model devoid of one sample, the class of which is then predicted. This is repeated until every sample has been predicted thus the number of PLS-DA models generated for validation is the same as the number of samples. Permutation testing is an alternative validation method.[79,86]

Permutation testing plots show how models in which the *Y*-variables, *i.e.* group classification, are randomised compare to the original PLS-DA model.[86] The *y*-axis represents the $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ values of all models and the *x*-axis the correlation coefficients between permuted and original variables. For the original model the correlation coefficient is 1.0. Plots showing 5% or more of permuted models that outperform the original model, *i.e.* higher $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ values, indicate a poorly modelled response.[88] Regression lines indicate model validity with the intercept values being interpretable as measures of 'background' $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ values obtained by fit to random data. An $R^2\boldsymbol{Y}$ intercept value below 0.3-0.4 and a $Q^2\boldsymbol{Y}$ intercept value less than 0.05 indicate a valid model.[86]

1.4.2 Univariate Statistics

1.4.2.1 Hypothesis Testing and *p*-values

A hypothesis is a statement about the population that is to be tested on a sample of the population. The null hypothesis, H_0, is that there is no difference between, for example, two groups of the population and the alternative hypothesis, H_1, is that there is a difference. Within the context of this work, the following is an example:
H_0: The mean integral for a peak in the NMR spectrum is the same for patients that do not have breast cancer as for those who do.

H_1: The mean integral for a peak in the NMR spectrum is not the same for patients that do not have breast cancer as for those who do.

Two tailed statistical analysis is required because H_1 states that the integral is not the same meaning it can be either greater or less than the integral in H_0. If there was reason to suspect the integral would be higher one-tailed analysis could be performed and the same would apply if there was evidence for a lower integral value.

Acceptance or rejection of H_0 is determined by a *p*-value that is related to the means of the two groups.[89] The *p*-value is the probability of obtaining a test value at least as extreme as the one that was observed, assuming that H_0 was true.[89] For example, if a *p*-value of 0.001 was obtained after comparing the means of two groups and H_0 was true, the likelihood of the occurrence of this observation by chance is once every thousand times. There is strong evidence to suggest that there is a difference between the means of the two groups and H_0 should be rejected. In this case the *p*-value is statistically significant.[89]

The significance level determines whether a *p*-value is statistically significant. Conventionally, the 5% significance level is used so if a *p*-value is $\leq$0.05 it is concluded to be statistically significant and that there is a difference between the means of the two groups resulting in the rejection of H_0.[89] The significance level is important because it governs the number and type of errors. A type I error (false positive) occurs if H_0 is rejected when it is true; the likelihood of this event is the same as the significance level whereas a type II error (false negative) results when H_0 is accepted but it is false.[89]

1.4.2.2 Tests for Statistical Significance

The appropriate test to be used for generation of a *p*-value depends on the distribution of the data in the two groups that are compared. The Shapiro-Wilk test ascertains whether the values of an integral are normally distributed within each

group.[90] For this test, H_0 is that the data distribution is normal and H_1 is that the data distribution is not normal.

If both groups are independent and the data distribution is normal the Student's *t*-test can be used. This test assumes that the variances of the two groups are equal, which can be determined by the Levene's test.[89] If the variances are not equal the Welch-Aspin test[91] is used, a modification of the Student's *t*-test. If the data distribution is non-normal for one or both independent groups the Mann-Whitney U test can be applied. The median integral values are compared instead of the means to reduce the effect of any outliers that cause the data to be non-normally distributed.[89]

If the two groups are not independent, for example if the groups are measurements of the same patients taken from two different sample types or at two different time points, the paired samples *t*-test is required for normally distributed data and the Wilcoxon Matched-Pair test for non-normally distributed data.[89]

1.4.2.3 Tests for Multiple Comparisons

For multiple testing, false positives can be a problem. For 200 tests (approximately the traditional number of bins used in metabolomics) it would be expected, implementing a significance level of 0.05, for 10 false positives to occur hence why univariate analysis alone cannot be employed.

To overcome the multiple testing problem the assignment of an adjusted *p*-value to each test is performed. The Bonferroni method is a conservative technique where the *p*-values are multiplied by the number of tests and compared to the significance level chosen for the single test. While the number of false positives is reduced so is the number of true discoveries.[92] A more powerful alternative is the false discovery rate (FDR) correction. The FDR is the expected proportion of false positive findings among all rejected hypotheses, for example an FDR of 0.05 means that

among all findings called significant, 5% of these are truly not significant on average.[93]

1.4.3 Alternative Pattern Recognition Methods

Statistical total correlation spectroscopy (STOCSY) enhances peak identification and highlights metabolites linked *via* metabolic pathways by identifying correlations between peaks.[94] Correlation coefficients are calculated between the columns of two matrices, $\boldsymbol{X_1}$ (n x v_1) and $\boldsymbol{X_2}$ (n x v_2), where n is the number of samples and v_1 and v_2 are the number of variables in each matrix. A correlation matrix $\boldsymbol{C}$ (v_1 x v_2) is produced with the strongest correlations indicating the peaks emanate from the same metabolite and weaker positive or negative correlations indicative of linked metabolites.[94] If $\boldsymbol{X_1} = \boldsymbol{X_2}$ the 2D map will have the correlation of the peak with itself along the diagonal.[94]

Statistical heterospectroscopy (SHY) is an extension of STOCSY whereby $\boldsymbol{X_2}$ could be generated from analysis of the same samples but using a different analytical platform,[95] a different sample type, *e.g.* $\boldsymbol{X_1}$ = plasma data and $\boldsymbol{X_2}$ = urine data, from the same subjects[96] or, again from the subjects, a sample at a different time point.

1.4.4 Standardisation and Knowledge Sharing

The metabolomics community has aimed to increase knowledge sharing, both in terms of methodology and results, and standardisation of reporting to allow successful data dissemination.[97] This has been attempted in part through the establishment of the Metabolomics Standards Initiative (MSI) and its subsequent publications regarding reporting requirements.[98-100] Some groups have freely allowed access to data, either raw or pre-processed, upon publication.[50] Although research groups will have individual optimum experimental procedures, protocol publications[51,101,102] have allowed wider distribution of successful procedures. Online databases such as Human Metabolome Database[103] (HMDB), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Golm Metabolome Database

(GMD) provide access to detailed information about metabolites including compound identification by way of mass and NMR spectra libraries and metabolite pathways.

1.5 Gas Chromatography

A liquid sample is vaporised and injected into the head of chromatographic column. The mobile phase, an inert gas, flows through the column transporting the analytes.[104] Each analyte leaves the column at a different time depending on how strongly it is adsorbed by the stationary phase, which is a liquid fixed on the surface of an inert solid. The partition ratio for an analyte is equal to the ratio of its concentration in the stationary phase to its concentration in the mobile phase.[104] Therefore, a greater partition ratio results in a longer retention time.

Flame ionisation detectors (FIDs) are extensively used in GC to produce chromatograms.[105] Upon exiting the column, the analytes pass through the nearby hydrogen/air flame and pyrolysis occurs (thermochemical decomposition of organic substances at elevated temperatures without the participation of oxygen). Electrons are formed that generates a current between the electrodes positioned by the flame, which is translated as a peak in the chromatogram.[105]

1.6 Previous NMR and Metabolomics Studies of Breast Cancer

Tissue has been examined by NMR spectroscopy for over 40 years.[106,107] Initial investigation of breast cancer tissue concentrated on T_1 and T_2 relaxation times for the water resonance and an increase in the values for tumour tissue compared to healthy tissue was attributed to an increase in the motional freedom of the water molecules of tumours.[106,108] Similar analyses followed for serum[109] including an extension into lipid spectra investigation.[110] Some of the first studies that investigated metabolite levels in serum, plasma and urine were performed by Nicholson *et al*.[111,112] A number of breast cancer metabolomic and metabolic

profiling studies have since been performed on biofluids, tissue and tissue extracts with most having used pattern recognition in the analyses, whereas some studies have targeted specific metabolites for investigation. Breast cancer research employing NMR spectroscopy pre-dates the time of a definition for metabolomics[34] but has been included because similar concepts and practices have been used. A summary of metabolomic and NMR studies of breast cancer are shown in Table 1.1. Reviews concerning metabolomics and cancer,[113-118] or specifically breast cancer,[119,120] have been published.

Serum from breast cancer patients and healthy controls, with pre-, post- or unknown menopausal status, has been compared by the complementary techniques of NMR spectroscopy and MS.[121] PCA analysis of NMR data indicated lactate and lipids were by far the most influential species and increased levels were present in case samples. Conversely, taurine and glucose levels were decreased for this group. However, no scaling was performed so this may be the reason why the larger signals appear dominant. MS data was not separated by PCA or PLS-DA.

Further studies have analysed serum in relation to breast cancer. For patients who had breast cancer, metabolic changes in serum due to recurrence (of cancer in any location) were analysed[122] and Oakman *et al*.[123] investigated differences between patients with early stage breast cancer and metastatic disease. Keun *et al*.[124] examined weight gain during chemotherapy because it had been suggested to lead to increased chance of cancer recurrence and possibly decreased overall survival. Lactate increased in patients who experienced weight gain (classed as >1.5 kilogram [kg]). For the same group valine, tyrosine and alanine mean levels increased but the difference was not statistically significant. A follow-on study by the group included investigation into response to treatment for breast cancer.[125] Concentrations of 12 metabolites previously identified[124] were calculated but individually none showed statistically significant associations with treatment outcome. However, jointly high glucose, high lactate and low alanine were associated with progression of disease despite treatment.[125]

No metabolomics studies utilising plasma for breast cancer investigation could be identified but one of the earliest NMR spectroscopy studies of plasma samples was conducted by Fossel *et al*.[126] A statistically significantly narrower average line width of the methyl and methylene resonances of lipids in patients with breast cancer than in healthy subjects or those with benign tumours was concluded. Lipid alterations associated with breast cancer may change the distribution of the lipoprotein density profile and thereby influence the line shape of those signals. However, using similar methods, Holmberg disagreed with the findings stating it was not possible to distinguish between the three groups[127] though lipoprotein subclass profiling of plasma or serum has been used extensively in a commercial capacity to assess risk of cardiovascular disease (CVD).[128] The data handling, however, has been criticised with regards to its accuracy in resolving the lipoprotein signal into individual component signals.[129]

Slupsky *et al*.[130] analysed urine from breast cancer patients and healthy controls and 26 metabolites were found to decrease in case samples. The lower metabolite levels have been attributed to the Warburg effect and consequential results. With less circulating glucose and amino acids, due to use by tumours, an overall decrease in energy metabolism elsewhere in the body has been postulated. This could reduce other pathways, such as the urea cycle, resulting in lower urea and creatine concentrations and possibly affect the gut microbial metabolism and population.[130]

Menopausal status was shown to be a confounding factor in MS analysis of urine from patients with breast, ovarian or cervical cancer.[131] For patients with breast cancer compared to controls the concentrations of 5-hydroxymethyl-2-deoxyuridine and 8-hydroxy-2-deoxyguanosine were significantly increased. 5α-tetrahydrocorticosterone, as identified by PLS-DA loadings, was not significantly increased.[131]

Pattern recognition was first incorporated into analysis of breast cancer tissue extracts by Gribbestad *et al.*[132] Tumour and healthy breast tissue was obtained from 16 patients and perchloric acid extraction performed, allowing polar

metabolites to be analysed. PCA showed a decrease of glucose and an increase of choline-containing compounds in tumour tissue.[132] Earlier studies using the same extraction procedure concluded a higher content of lactate, taurine, succinate, phosphocholine (PCho), uridine-5'-diphosphate-N-acetylglucosamine (UDP-GlcNAc) and uridine-5'-diphosphate-N-acetylgalactosamine (UDP-GaINAc) and lower levels of glucose, *myo*-inositol and phosphocreatine (PCr) were present in tumour tissue as compared to healthy tissue.[133,134] However, no statistical significance was attributed to the findings.

Beckonert *et al*.[135] have analysed extracts from breast tissue. The extracts were prepared using a dual extraction method, which afforded both polar and lipophilic extracts. *Myo*-inositol and glucose were reduced in tumour tissue whereas taurine was increased. Alanine, PCho, phosphoethanolamine (PE) and UDP-hexoses (UDP-GlcNAc and UDP-GaINAc) were raised in tumour tissue compared to adjacent healthy tissue. Following histopathological analysis of control tissue, samples that predominantly consisted of fatty tissue had a high content of *myo*-inositol whereas glucose was raised in samples with high amounts of connective tissue. The aromatic region of the spectra contained many more signals, such as UDP-hexoses, in tumour tissue compared with control; it was concluded water-soluble metabolites are more prevalent in tumour tissue,[135] in agreement with Sitter *et al*.[136] Lipid metabolites, including phosphatidylethanolamine (PDE), unsaturated and saturated fatty acids, cholesterol esters and sphingomyelin-like substances, showed concentration increases in grade 3 samples.[135] The statistical significance of the findings was not discussed.

One of the first studies to use magic angle spinning (MAS) NMR spectroscopy to analyse intact breast tissue was performed by Cheng *et al*.[137] MVA was not used but levels of previously reported metabolites associated with breast cancer, stated within the paper, were investigated using a small number of samples. The relative intensity of phosphocholine to choline and lactate to choline increased for both intermediate grade 2-3 and grade 3 samples compared to grade 2 samples. The relative intensities involving lactate were statistically significant.

Further MAS NMR spectroscopy studies have been performed but using biopsy tissue samples.[138-140] Tumour and healthy tissue samples were separated using PCA but separation based on grade did not result.[138] Identification of metabolites was not the primary aim of the study instead class prediction of samples was the main interest. Using PLS-DA, Giskeodegard *et al*.[139] revealed separation between ER+/ER- samples and partial separation between PR+/PR- samples but not according to lymph node status. Loadings plots revealed ER- samples were associated with increased glycine, glycerophosphocholine (GPC), choline and alanine and decreased ascorbate, creatine, taurine and PCho whereas PR- samples have increased ascorbate, creatine, PCho, lactate, glycine, GPC, choline and alanine.

For biopsy tissue samples, Li *et al*.[140] identified that taurine and choline-containing compounds were elevated in cancer tumours compared to non-cancerous samples (benign tumours and the adjacent normal tissue) as was the signal tentatively assigned to aspartate. The PR status (PR+ or PR-) of 10 from 13 (77%) cancer patients was correctly predicted by 'leave one out' cross-validation of an orthogonal partial least squares-discriminant analysis (OPLS-DA) model. Axillary lymph node status could not be predicted when all samples were used.[140]

In a different study, PCA analysis of data acquired by MAS NMR spectroscopy from surgically removed tissue did not clearly differentiate between patients with good and poor prognosis of breast cancer.[136] PC 5, accounting for 6% of the variation, provided a tendency to separate samples based on prognosis with glycine raised in poor prognosis patients. Patients who subsequently died within five years of sample donation were separated in PC 2 (18% of the variation) from those who were not deceased, irrespective of cancer recurrence status. High levels of taurine, GPC and creatine combined with low levels of glycine and PCho characterised those patients alive five years after surgery.[136] However, the statistical significance of the change in these metabolite levels was not discussed. Ratios of metabolite levels were significantly different between patients with good and poor prognosis and also between those who either had recurring cancer or were deceased and "healthy" patients five years after sample donation. Taurine/glycine, GPC/glycine

and total cholines/glycine ratios were higher for good prognosis patients and taurine/glycine and GPC/glycine ratios were raised in “healthy” patients. However, although mentioned, a correction for multiple tests was not employed for the 45 ratio tests. Collagen in connective tissue contains much glycine and histopathological analysis revealed a higher percentage of connective tissue was present in samples from poor prognosis patients.[136] This was forwarded as the reason for higher glycine levels in poor prognosis patients and the subsequent lower ratios that contained glycine.

The composition of tissue samples can vary with different percentages of cancer cells, connective tissue, fat and healthy tissue.[136] The level of metabolites could depend on the composition of the tissue.[136] Concentrations of glycine, GPC, PCho and total choline were positively correlated with the cancer cell fraction whereas taurine, GPC, choline and total choline-containing metabolites correlated negatively with fat tissue as did PCho with connective tissue. *Myo*-inositol positively correlated with healthy tissue. The authors concluded the majority of signals from low molecular weight metabolites arise from cancer cells whilst fat tissue contains minor amounts of these metabolites.[136] Further evidence of metabolite level variation with type of tissue was presented by Sitter *et al*.[141] The fraction of cancer cells within the tissue sample separated patients with grade 2 cancer using PCA: a higher percentage of cancer cells correlated with greater glycine and PCho levels. The cancer cell percentage varied from 0% to over 50%. The same study also indicated GPC/PCho and GPC/choline ratios were higher and the PCho/choline ratio lower in adjacent healthy tissue samples.[141]

Table 1.1 Summary of metabolomics and NMR studies of breast cancer.

Author [Year]	Bio-material	Study	Data acquisition method	Region of variation identification method	Changing metabolites
Gu *et al*.[121] [2011]	Serum	BC (27) v. control (30)	NMR	PCA	(+) lac, lip (-) tau, glc
Asiago *et al*.[122] [2010]	Serum	Cancer recurrence (20) v. non-recurrence (36)	NMR	From own previous study	(-) for, his, pro, cho
Asiago *et al*.[122] [2010]	Serum	Cancer recurrence (20) v. non-recurrence (36)	GCxGC/MS	From own previous study	(-) glu, NAG, 3-H-2-MBA
Oakman *et al*.[123] [2011]	Serum	Metastatic BC (51) v. early BC (44)	NMR	OPLS	(+) phe, glc, pro, lys, NAC (-) lip
Keun *et al*.[124] [2009]	Serum	Weight gain (10) v. no weight gain (11)	NMR	PLS-DA	(+) lac
Stebbing *et al*.[125] [2012]	Serum	With treatment: Progression of BC v. no progression	NMR	From previous study[124]	Combined (+) glc, lac and (-) ala
Slupsky *et al*.[130] [2010]	Urine	BC (38) v. control (62)	NMR	Specialist software	(-) cre, ace, suc, lac, pyr, for, ile, sur, leu, asn, ure, glc, eta, dim, 4-HPA, crt, ala, hip, ura, val, aco, unknowns at 4.34, 3.94, 3.35, 2.60, 2.36 ppm

Table 1.1 Continued.

Author [Year]	Bio-material	Study	Data acquisition method	Region of variation identification method	Changing metabolites
Woo *et al*.[131] [2009]	Urine	BC (10) v. control (22)	GC-MS	PLS-DA	(+) 5-HM-2-DOU, 8-H-2-DOG
Gribbestad *et al.*[132] [1999]	Tissue extracts	Tumour (16) v. healthy (16)	NMR	PCA	(+) cho compounds (-) glc
Gribbestad *et al.*[133] [1993]	Tissue extracts	Tumour (11) v. healthy (7)	NMR	No PR	(+) lac, suc, PCho (-) glc, ino
Gribbestad *et al.*[134] [1994]	Tissue extracts	Tumour (unstated) v. healthy (unstated)	NMR	No PR	(+) lac, suc, tau, PCho (-) glc, myo, PCr
Beckonert *et al*.[135] [2003]	Tissue extracts	Tumour (49) v. healthy (39)	NMR	SOM	(-) myo, glc
Beckonert *et al*.[135] [2003]	Tissue extracts	Grade severity (3 = 22, 2= 26, 0 = 41)	NMR	SOM	(+) ala, UDP-H, PCho, PE Grade 3 (+) PDE, UFA, SFA, CE, sph
Sitter *et al*.[136] [2010]	Tissue	Good prognosis (13) v. poor prognosis (16)	NMR	PCA	See text
Cheng *et al*.[137] [1998]	Tissue	Grade 3 or intermediate grade 2-3 v. grade 2	NMR	From previous study	(+) PCho/cho, lac/cho
Bathen *et al*.[138] [2007]	Biopsy tissue	Tumour (91) v. healthy (48)	NMR	PCA	Not identified

Table 1.1 Continued.

Author [Year]	Bio-material	Study	Data acquisition method	Region of variation identification method	Changing metabolites
Bathen *et al*.[138] [2007]	Biopsy tissue	Grade 3 (36) v. grade 2 (37)	NMR	PLS - no separation	
Bathen *et al*.[138] [2007]	Biopsy tissue	Positive hormone (ER+/PR+, ER+/PR- and ER-/PR+) (63) v. negative hormone (ER-/PR-) (13)	NMR	PLS	(+) tau, GPC (-) lac, gly, scy, PCho
Bathen *et al*.[138] [2007]	Biopsy tissue	Positive lymph node (36) v. negative lymph mode (43)	NMR	PLS	(+) gly, PCho (-) tau
Giskeodegard *et al*.[139] [2010]	Biopsy tissue	ER- (39) v. ER+ (118)	NMR	PLS-DA	(+) gly, GPC, cho, ala (-) asc, cre, tau, PCho
Giskeodegard *et al*.[139] [2010]	Biopsy tissue	PR- (60) v. PR+ (94)	NMR	PLS-DA	(+) asc, lac, gly, GPC, PCho, cho, cre ala
Giskeodegard *et al*.[139] [2010]	Biopsy tissue	Positive lymph node (64) v. negative lymph mode (88)	NMR	PLS-DA - no separation	
Li *et al*.[140] [2011]	Biopsy tissue	Cancer (13) v. non-cancer (18)	NMR	OPLS-DA	(+) tau, cho, asp
Sitter *et al*.[141] [2006]	Tissue	Tumour (85) v. healthy (18)	NMR	PCA	Not identified

Table 1.1 Continued.

For the study column, values in parenthesis are sample numbers. 3-H-2-MBA, 3-hydroxy-2-methyl-butanoic acid; 4-HPA, 4-hydroxyphenylacetate; 5-HM-2-DOU, 5-hydroxymethyl-2-deoxyuridine; 8-H-2-DOG, 8-hydroxy-2-deoxyguanosine ace, acetate; aco, *trans*-aconitate; ala, alanine; asc, ascorbate; asn, asparagine; asp, aspartate; CE, cholesterol esters; cho, choline; cre, creatine; crt, creatinine; dim, dimethylamine; eta, ethanolamine; for, formate; glc, glucose; glu, glutamic acid; gly, glycine; GPC, glycerophosphocholine; hip, hippurate; his, histidine; ile, isoleucine; ino, inositol; lac, lactate; leu, leucine; lip, lipids; lys, lysine; myo, *myo*-inositol; NAC, *N*-acetyl-cysteine; NAG, *N*-acetyl-glycine; PCho, phosphocholine; PCr, phosphocreatine; PDE, phosphatidylethanolamine; PE, phosphoethanolamine; phe, phenylalanine; pro, proline; pyr, pyroglutamate; scy, *scyllo*-inositol; SOM, self organising map; SFA, saturated fatty acids; sph, sphingomyelin-like substances; suc, succinate; sur, sucrose; tau, taurine; UDP-H, uridine-5'-diphosphate-hexose; UFA, unsaturated fatty acids; ura, uracil; ure, urea; and val, valine.

Chapter 2. NMR Analysis of Plasma

The following chapter will describe ^{1}H-NMR spectroscopy analysis of plasma obtained from 69 women using CPMG pulse sequence. Of these, 40 had been collected from women who had been diagnosed as having a form of breast cancer, either non-invasive or an invasive tumour with associated grade 1, 2 or 3. For the purposes of classification in this study, cases of DCIS without invasive cancer were assigned as grade 0. Samples were available from 29 patients from a control group that had been referred to the clinic because of abnormal breast state, which was subsequently diagnosed as non-cancerous abnormalities such as cysts or fat necrosis.

MVA of the spectra was performed as described in Section 1.4.1.4 in an attempt to identify possible biomarkers of breast cancer occurrence and progression. A number of confounding factors were considered either singly or in combination during the data interrogation.

Samples were prepared as per Section 8.1.1.1 and data collected as detailed in Section 8.2.1. Section 8.3 applies for spectral processing with dark regions listed in Table 8.3. Constant sum normalisation was used. MVA was performed as detailed in Section 8.4.

2.1 Results

Each plasma sample was given a four digit identifier during collection. The first digit establishes breast cancer status: 1 for control and 2 for case. The subsequent digits indicate the order in which the samples were collected. Sample 2041 is an exception; it was saved prior to cancer status being confirmed and was initially given a control identifier but upon breast cancer diagnosis it was reclassified from

control to case. A summary of the demographic details for the participants in the study is provided in Table 2.1.

Table 2.1 Selected demographics for the 40 case and 29 control samples. The range, average and median for cigarettes smoked refers to current and former smokers combined.

		BMI ($kg\ m^{-2}$)	Age (Years)	Cigarettes Smoked (Thousand)	Smoking Status
Case	Range	18.9-50.2	51-97	4-365	Current 7
	Average	28.4	67	167	Former 15
	Median	27.4	63	183	Never 18
Control	Range	17.6-50.7	51-87	2-329	Current 3
	Average	30.2	63	129	Former 13
	Median	29.2	62	110	Never 13

BMI = Body Mass Index.

2.1.1 Analysis of Whole Spectrum

2.1.1.1 Initial Analysis

A PCA model was built for the data and produced five PCs; the scores plot of PC 2 versus PC 1 is presented in Figure 2.1. The goodness of fit ($R^2\boldsymbol{X}$(cum), the fraction of the sum of squares of all the *X*-variables that are explained by the model) is 0.731. The predictive ability ($Q^2\boldsymbol{X}$(cum), the fraction of the total variation of the *X*-variables that can be predicted by the model) is 0.384. With a difference of more than 0.3 between $R^2\boldsymbol{X}$ and $Q^2\boldsymbol{X}$ there is a greater likelihood that too much noise or outlying data points have been incorporated in the model.[86] Figure 2.1 supports this because three samples are indicative of being strong outliers being located a considerable distance outside of Hotelling's T^2. This defines the normal area corresponding to, in this case, 95% confidence, the extent of which is bound by the ellipse.[86] The complementary loadings plot is shown in Figure 2.2, which enabled identification of the plasma components responsible for the high scores. Bins in the range 1.163-1.200 ppm and 3.627-3.681 ppm are consistent with the presence of ethanol[103] in the samples and the bins centred at 0.892 and 1.288 ppm correspond

to lipid species.[103,142] In addition to the three suspected outliers a fourth sample, 1003, was not positioned with the majority of samples, which exhibited tight clustering. The difference in signal intensities of lipids and ethanol between the four non-clustered samples and the remaining 65 is easily visible in the original spectra as shown by Figure 2.3. A large triplet signal (from the methyl group of ethanol) is displayed at around 1.18 ppm by the samples from patients 1003, 2003 and 2036. Plasma does not usually display signals in this region as illustrated by the red trace of sample 1017 in Figure 2.3. Sample 1003 has a much lower level of ethanol relative to the two other ethanol containing samples so is not as far removed from the cluster of samples in the scores plot (Figure 2.1). An extremely large lipid signal is present for the fourth outlier, sample 1002. Although the signal is present for all samples in the study the level is many times greater, as illustrated in Figure 2.3, and as such exerts a strong influence on the statistical analysis process. The medical records of the patients were consulted but no obvious reason for the presence of ethanol or excess lipids could be identified. Unrecorded consumption of alcohol was attributed for causing the ethanol signals.

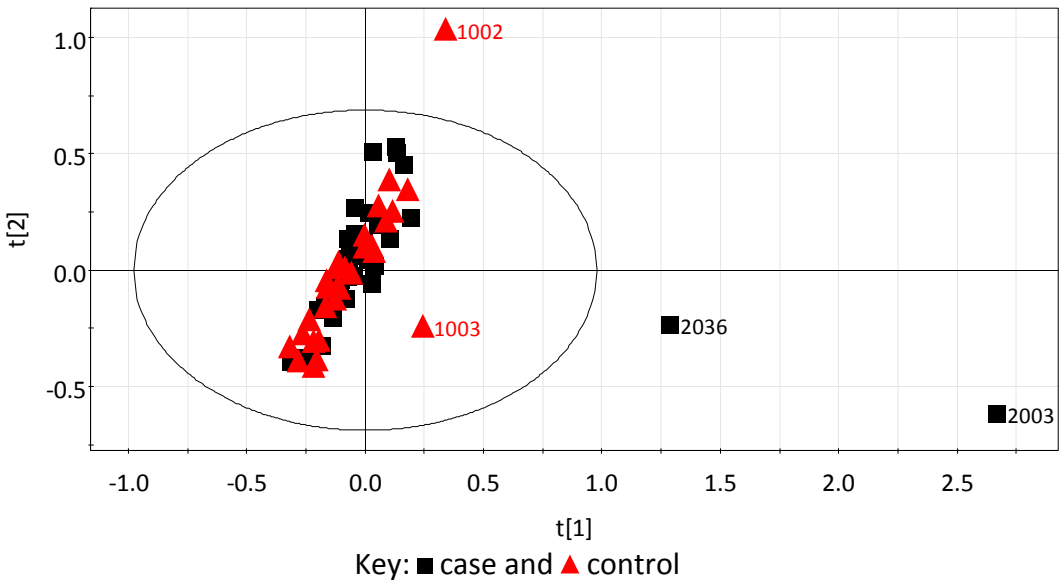


Figure 2.1 PCA scores plot of whole ^{1}H-NMR plasma chemical shift data for all 40 case and 29 control samples showing the first two model components. R^2X = 0.307 and 0.154, and Q^2X = 0.240 and 0.115 for PC 1 and PC 2, respectively.

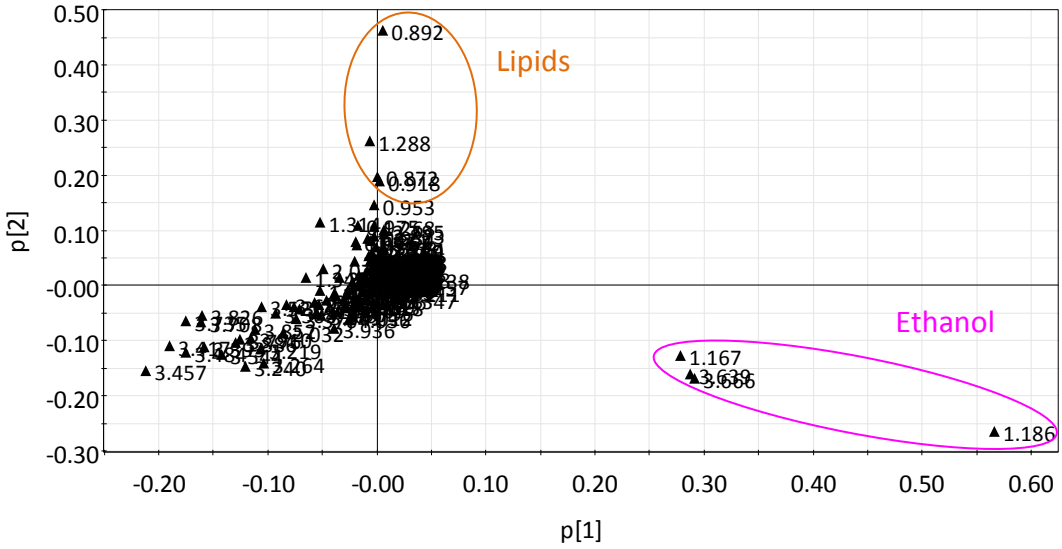


Figure 2.2 PCA loadings plot corresponding to the model displayed in Figure 2.1.

Two options present themselves on how to address the 'problematic' regions. Either the regions can be removed from analysis by adding them to the 'dark regions', as described in Table 8.3, or the samples in question can be removed from further statistical analysis procedures. The latter option was chosen for two reasons. Regarding the ethanol signals, although the triplet centred at 1.18 ppm, resulting from the methyl protons, does not occur in a region where other signals are present, the signal from methylene protons, a quartet centred at 3.65 ppm, does and overlaps with glucose signals. By making the ethanol regions 'dark regions' unrelated signals would also be excluded. Secondly, due to the intensity of the lipid signals of sample 1002, wide spectral regions, in excess of 0.10 ppm, are occupied and incorporate underlying signals such as valine and lactate (Figure 2.3). Too broad a region would have to be excluded if sample 1002 was to be included in analysis. The four samples were therefore removed and PCA analysis was repeated. A four PC model was generated with an $R^2\boldsymbol{X}$(cum) value of 0.592 and 0.412 for $Q^2\boldsymbol{X}$(cum). Scores and loadings plots for PC 1 versus PC 2 are shown in Figure 2.4 and Figure 2.5, respectively.

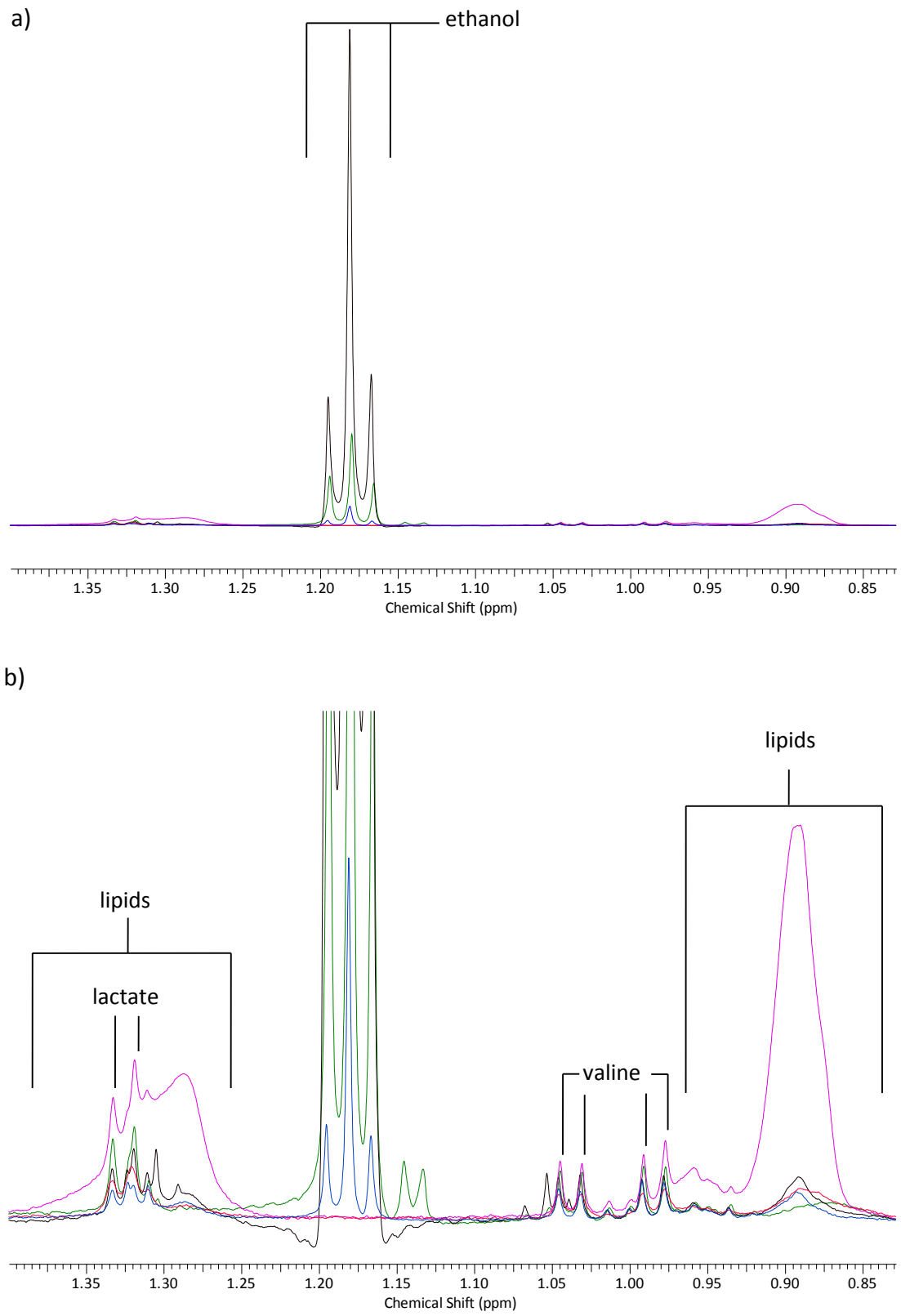


Figure 2.3 Spectral region of four samples excluded from analysis plus a retained sample emphasising the size of a) ethanol methyl proton peaks (top) and b) lipid peaks (above). Trace colours: pink = 1002; blue = 1003; red = 1017 (retained); black = 2003; green = 2036.

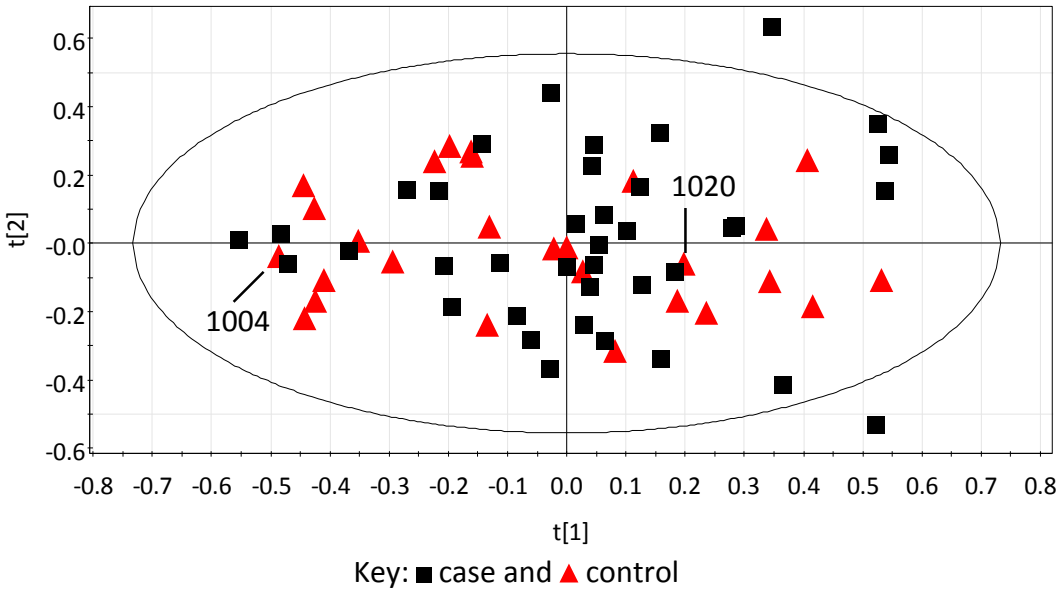


Figure 2.4 PCA scores plot of plasma data for 38 case and 27 control samples excluding ethanol or atypical lipid level containing samples showing the first two model components. R^2X = 0.278 and 0.160, and Q^2X = 0.221 and 0.119 for PC 1 and PC 2, respectively. Samples 1004 and 1020 were from patients with a different ethnicity, *i.e.* not white British; their importance is discussed in further analysis.

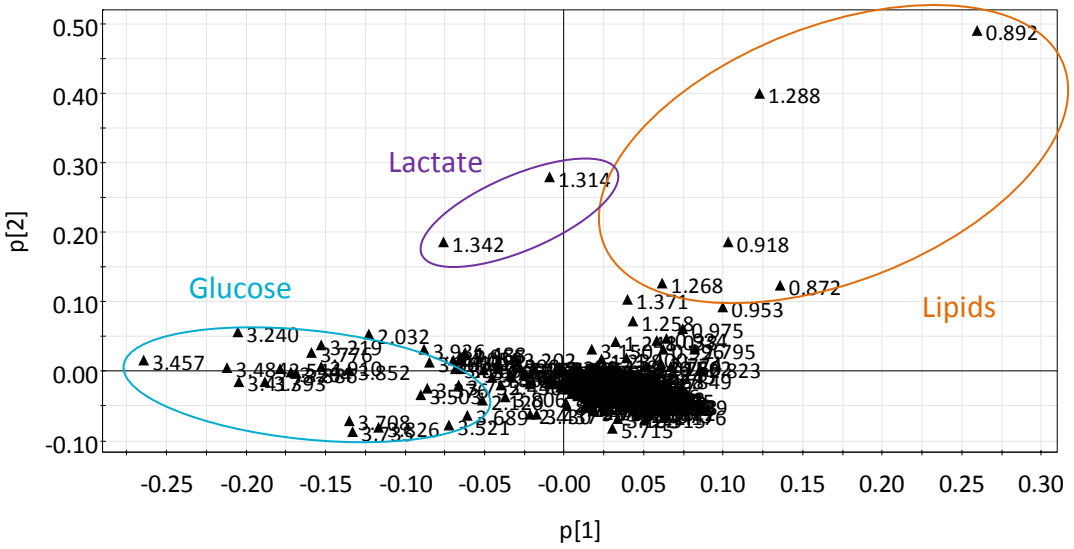


Figure 2.5 PCA loadings plot corresponding to the model displayed in Figure 2.4.

Following removal of the four aforementioned samples the scores range is much smaller and there are no apparent outliers. However, there is still no discrimination between case and control samples. There is a very slight tendency for some case samples to occupy positive PC 1 and PC 2 scores space, which from the loadings plot (Figure 2.5) is indicative of higher lipid levels. This is supported by study

findings that for post-menopausal patients with breast cancer compared to those not afflicted there was a significant increase in total lipid levels in serum[143] and low density lipoprotein cholesterol and triglycerides in plasma.[144] The plasma findings were attributed to oxidative stress caused by an imbalance between reactive oxygen species and the antioxidant capacity of the cell, possibly due to the susceptibility of the lipoprotein to undergo oxidation. Overall, this can result in cellular damage leading to conversion to malignant cells.[144] Additionally, in the scores plot it can be noted that the majority of samples that have the highest negative scores value in PC 2 (>-0.35) are controls despite being fewer in number. Glucose is the species that occupies the corresponding space in the loadings plot therefore this indicates that the level of glucose is elevated in these samples. In tissue extracts the observation that glucose concentration is higher in unaffected tissue compared to malignant tissue has been reported with more active glycolysis in tumour tissue forwarded as the explanation.[135]

At this stage close attention was paid to demographic details of the patients. It was noted that all but two of the women were British and white. As ethnicity is a known confounding factor in metabolomics,[92] highlighted by metabolomic analysis of urine that showed four separate clusters in PCA scores consisting of northern Chinese, southern Chinese, Japanese, and UK and American samples,[145] it was deemed best practice to remove the two samples from the study cohort for further analysis. The study was therefore reduced to 63 patients, consisting of 38 case and 25 control samples, and PCA repeated. Note, however, that the metabolic profile of the two samples were not outliers (samples labelled in Figure 2.4) and the resultant scores and loadings plots with the two samples removed (data not shown) were almost identical to those where the two samples had been included. The summarised demographics for the 63 patients (table not shown) were very similar to that for the 69 patients (Table 2.1). Ranges for BMI, age and number of cigarettes smoked for case and control samples were unaffected.

Four PCs were again generated, this time for the remaining 63 samples with $R^2\boldsymbol{X}$(cum) and $Q^2\boldsymbol{X}$(cum) values of 0.591 and 0.386, respectively (data not shown).

The above discussion, which related to when the two samples from patients whose ethnicity was not white British were included, is equally applicable to the new data with the two samples excluded. Given the slight tendency of grouping of some case and control samples in different scores space, PLS-DA was applied in an attempt to connect the information in the data matrix and properties of the samples, in this case breast cancer status. It was not possible to generate a model indicating that the data did not correlate with breast cancer status.

2.1.1.2 Ductal Type and Cancer Grade Investigation

Breast cancer is not a single disease, rather it is very heterogeneous because it represents multiple diseases,[146] therefore a reduction in diversity within case group classification could enhance separation between case and control groups *via* MVA. Case samples were separately categorised according to presence of single occurrence invasive ductal carcinoma and breast cancer grade. Two of the 38 case samples were diagnosed as having more than one instance of graded cancer resulting from multiple invasive carcinomas. Although these samples are included in case versus control analysis they have been excluded from investigation into differences between grades. Table 2.2 summarises selected parameters for samples with a single occurrence of graded tumour.

Table 2.2 Selected demographics for the 36 samples with a single occurrence of breast cancer.

Cancer Grade	Zero 8						One 6						Two 12						Three 10					
Ductal Type	No 8			Yes 0			No 1			Yes 5			No 2			Yes 10			No 4			Yes 6		
Smoking Status	N 5	F 1	C 2	N 0	F 0	C 0	N 0	F 1	C 0	N 2	F 3	C 0	N 0	F 1	C 1	N 5	F 4	C 1	N 1	F 2	C 1	N 3	F 1	C 2

Smoking status: N = never, F = former and C = current.

Single occurrence invasive ductal carcinoma was identified for 21 samples; the other 15 samples exhibited, for example, lobular or metaplastic carcinoma. Single

occurrence invasive ductal carcinoma samples acted as the case group, whilst the control group remained as previously. A similar range of positions in scores space was apparent for both groups hence no further separation was observed in the three component model ($R^2\boldsymbol{X}$ = 0.553 and $Q^2\boldsymbol{X}$ = 0.400) (Figure 2.6).

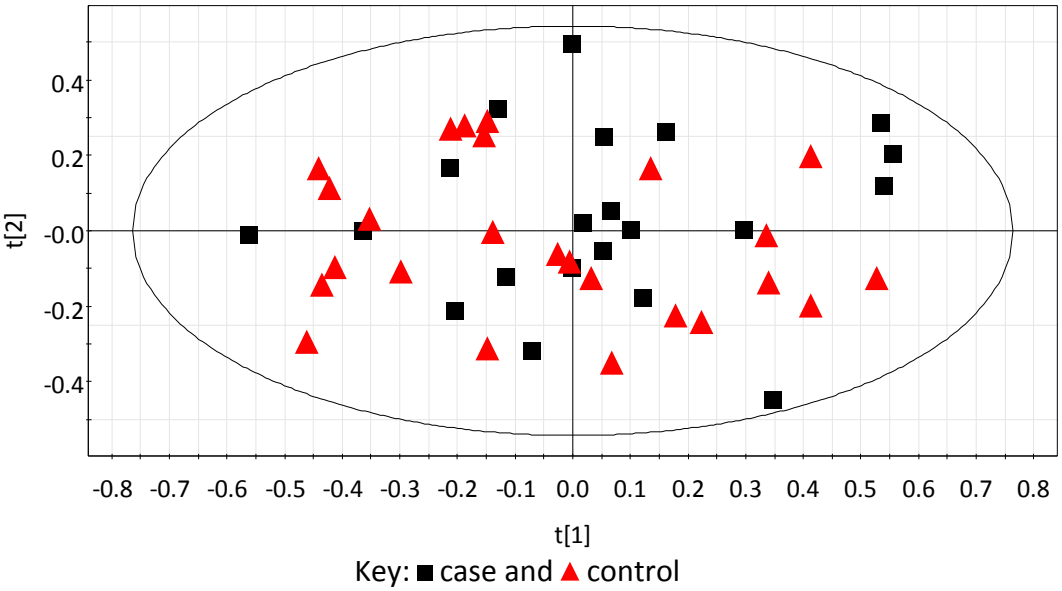


Figure 2.6 PCA scores plot of plasma data for 21 single occurrence invasive ductal carcinoma case and 25 control samples showing the first two model components. $R^2\boldsymbol{X}$ = 0.307 and 0.154, and $Q^2\boldsymbol{X}$ = 0.240 and 0.115 for PC 1 and PC 2, respectively.

In addition to case versus control investigation, potential markers of cancer grade can be explored. Figure 2.7 shows the scores plot whereby case samples are coloured according to breast cancer grade. The loadings plot (not shown) is very similar to Figure 2.5. The three component model had $R^2\boldsymbol{X}$(cum) and $Q^2\boldsymbol{X}$(cum) values of 0.548 and 0.333, respectively. Given that lower levels of glucose are associated with cancer cells[147] it is surprising that a selection of grade 2 samples is clustered in the area corresponding to the glucose region in loadings space. Five other grade 2 samples are positioned very close to the centre of the scores plot indicating, comparatively, neither glucose nor lactate is elevated or reduced, whilst only one grade 2 sample has an elevated lactate level. Grade 3 samples do not show increased glucose levels but only four are indicative of having elevated lactate levels with three having low levels of glucose and lactate. The sample that is most influenced by lipids (high PC 1 and 2 scores values) has an

associated grade of zero. Higher lipid levels are associated with greater BMI[148] so it is surmised that the patient's BMI is high given that this is one of two patients for whom BMI data is not available.

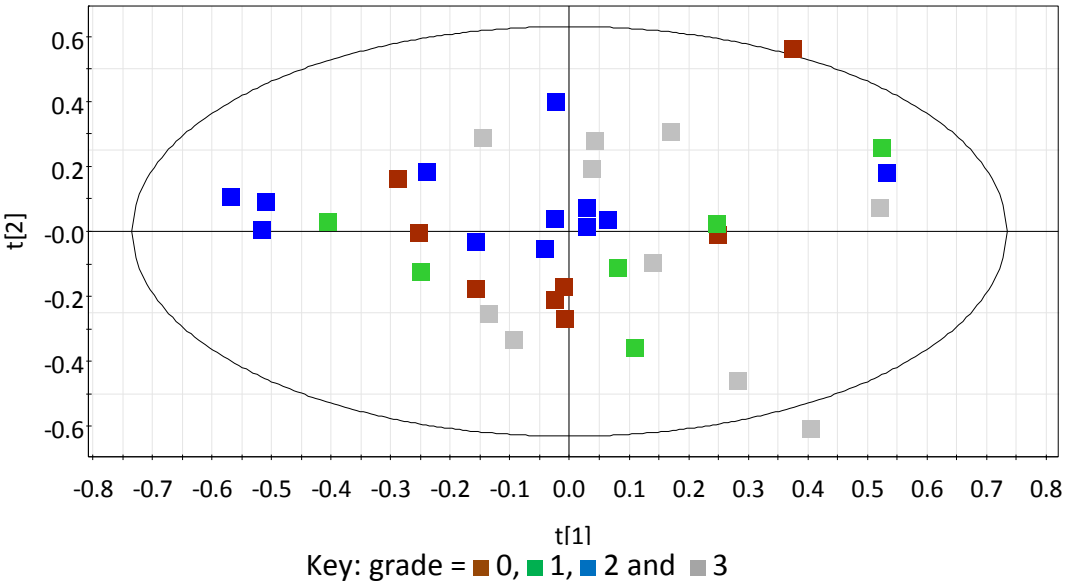


Figure 2.7 PCA scores plot showing the first two model components of plasma data for 36 single occurrence case samples coloured according to tumour grade. R^2X = 0.260 and 0.192, and Q^2X = 0.153 and 0.149 for PC 1 and PC 2, respectively.

Given the heterogeneity and the non-discrete nature of the disease it is difficult to draw conclusions from the data especially when small samples numbers are involved. However, a compromise needs to be made between sample heterogeneity and number of samples, thus orchestrating scores to be determined for single occurrence ductal carcinoma case samples with cancer grade as the basis for identification of clustering. PC 2 versus PC 1 is shown in the scores plot Figure 2.8 but the loadings plot of the five PC model ($R^2\boldsymbol{X}$(cum) = 0.729 and $Q^2\boldsymbol{X}$(cum) = 0.306) is not displayed.

Many of the observations described for single occurrence case samples are not present for single occurrence ductal carcinoma case samples, in part due to the smaller number of samples. For both models, the information content of the data did not reflect cancer grade. The positioning of the data points displayed in the relevant PCA scores plots was not able to be explained by the grade of all single

occurrence case samples or just single occurrence invasive ductal carcinoma cases. Models were not able to be generated for either data set using PLS-DA.

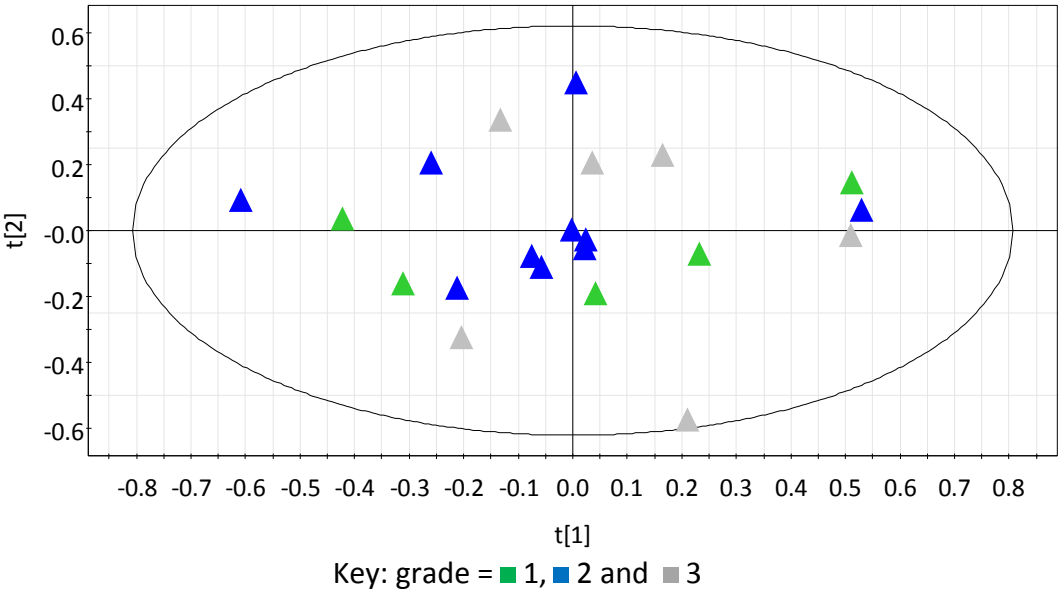
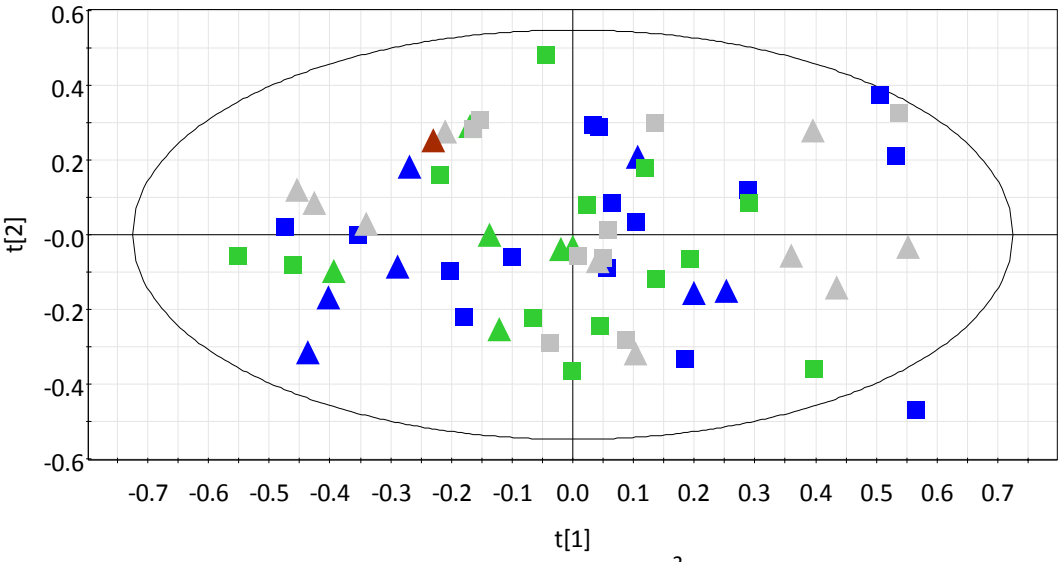


Figure 2.8 PCA scores plot showing the first two model components of plasma data for 21 single occurrence ductal carcinoma case samples coloured according to tumour grade. R^2X = 0.317 and 0.187, and Q^2X = 0.170 and 0.124 for PC 1 and PC 2, respectively.

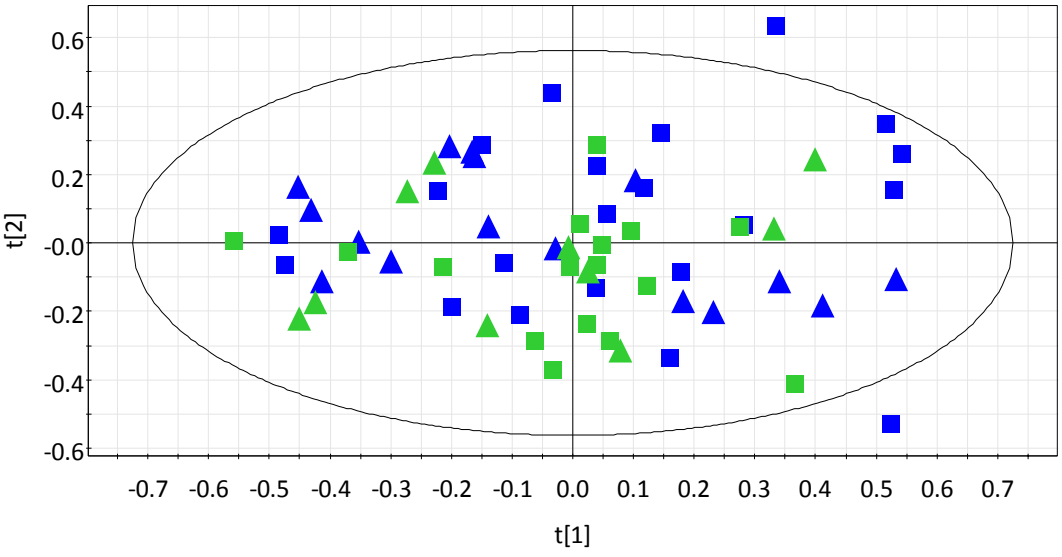
2.1.1.3 Influence of Potential Confounding Factors

Given that there is no clear discrimination between case and control samples or case grades, exploration of potential data correlation with factors that could be expected to influence sample content ensued, including BMI classification (Figure 2.9; four component model, $R^2\boldsymbol{X}$(cum) = 0.581 and $Q^2\boldsymbol{X}$(cum) = 0.368) and smoking status (Figure 2.10; four component model, $R^2\boldsymbol{X}$(cum) = 0.591 and $Q^2\boldsymbol{X}$(cum) = 0.386). BMI was classified according to values of <18.5, 18.5-24.9, 25-29.9 and ≥30 kg m^{-2}, suggestive of underweight, healthy weight, overweight and obese patients, respectively.[149] Smoking status was defined as 'exposed' (current or former smoker) or 'never exposed' (never-smoked).



Key: shape - □ case and △ control; colour - BMI (kg m^{-2}) = ● <18.5, ● 18.5-24.9, ● 25-29.9 and ● ≥30

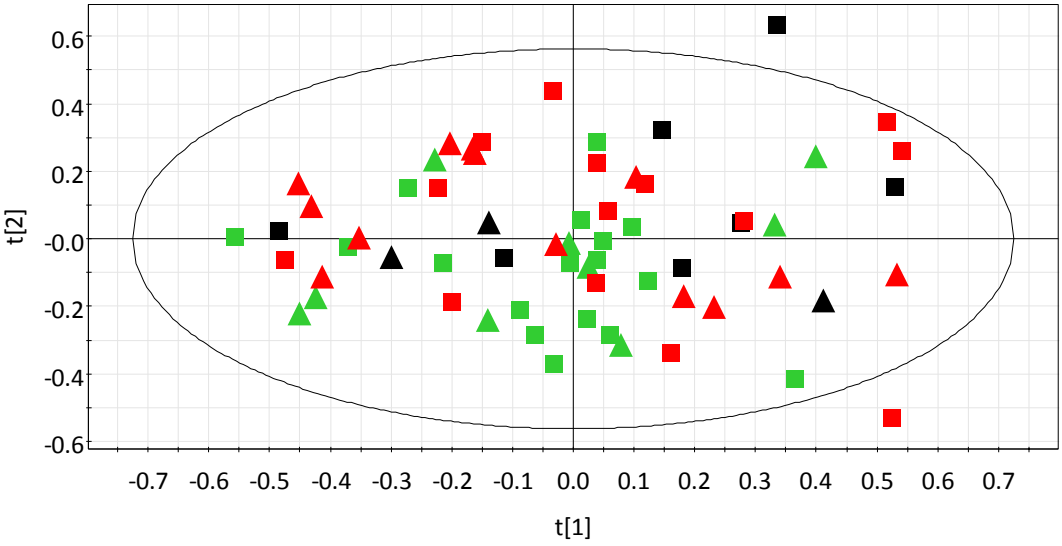
Figure 2.9 PCA scores plot showing the first two model components of plasma data for 37 case and 24 control samples coloured according to BMI (kg m^{-2}) categories (<18.5 [case = 0 and controls = 1], 18.5-24.9 [13 and 6], 25-29.9 [16 and 6] and ≥30 [8 and 11], suggestive of underweight, healthy weight, overweight and obese patients, respectively). Two samples were excluded because BMI data was not available. R^2X = 0.273 and 0.156, and Q^2X = 0.211 and 0.099 for PC 1 and PC 2, respectively.



Key: shape - □ case and △ control; colour - smoking status = ● 'never exposed', ● 'exposed'

Figure 2.10 PCA scores plot showing the first two model components of plasma data for 38 case and 25 control samples coloured according to smoking status: 26 'never exposed' [never smoked] (case = 17 and control = 9) and 37 'exposed' [current or former smoker] (21 and 16). R^2X = 0.272 and 0.163, and Q^2X = 0.197 and 0.122 for PC 1 and PC 2, respectively.

PCA did not show that data was reflective of the status of either factor and when PLS-DA was applied a model was unable to be built for BMI as the class identifier. A one component model (data not shown) was generated for smoking status but clustering of the two groups was not observed and additionally the $R^2\boldsymbol{X}$ and $R^2\boldsymbol{Y}$ values (the fraction of the sum of the squares of the *X*-variables and *Y*-variables explained by the model, respectively) were 0.165 and 0.237, respectively. The $Q^2\boldsymbol{Y}$ value (the fraction of the total variation in the *Y*-variables that could be predicted by the model) was 0.093; all three values were too low for the model to be considered good. However, in PCA models there is a slight tendency for the samples that have the highest scores values in both PCs 1 and 2 to have an associated BMI of 25.5-29.9 or ≥30 kg m^{-2}, suggestive of overweight or obese patients, respectively, and originate from those whose smoking status is 'exposed'. It is widely accepted that higher lipid levels are associated with greater BMI,[148] which the slight trend in Figure 2.9 supports in conjunction with the associated loadings plot (not shown but extremely similar to Figure 2.5). Figure 2.10 shows there could be a slight tendency for 'exposed' samples to have higher levels of lipids. In metabolomic analysis of serum by Wang-Sattler *et al.*[150] higher lipid levels were identified for current smokers compared to former or non-smokers whilst former smokers were found to be separated from non-smokers suggesting that the influence of cigarette smoke in human blood remains for years. Consequently, the 'exposed' category was divided into two categories: current and former smokers. The PCA scores plot Figure 2.11 and loadings plot (not shown but extremely similar to Figure 2.5) clearly shows that samples from current smokers do not correspond with higher levels of lipids. However, given that patients who only recently ceased smoking could be classed as a former smoker, irrespective of the number of cigarettes smoked per day and length of exposure, the smoking habits of former smokers were investigated.



Key: shape - □ case and △ control; colour - smoking status = ● never, ● former and ● current.

Figure 2.11 PCA scores plot showing the first two model components of plasma data for 38 case and 25 control samples coloured according to smoking status: 26 never smoked (17 and 9), 27 former smokers (14 and 13) and 10 current smokers (7 and 3). Values in parentheses relate to number of case and control samples, respectively. Model descriptors as per Figure 2.10.

Of the seven samples from former smokers that had positive PC 1 and 2 scores values, for the two with the highest PC 1 scores value (and most separated) cigarette consumption was not high relative for the study. The rank was 24th and 33rd out of 37 for number of cigarettes smoked, plus the cessation periods were 10 and 45 years, respectively. This further reduces the possibility that the slight trend is due to higher lipid levels in smokers. As an extension the number of cigarettes smoked was modelled for all 'exposed' samples but no trend in scores space was observed (data not shown; three component model, $R^2\boldsymbol{X}$(cum) = 0.539 and $Q^2\boldsymbol{X}$(cum) = 0.392).

Neither the two ('exposed' or 'never exposed') nor three groups (never smoked, former smoker or current smoker) used for smoking status categorisation is optimum. There are three variables associated with smoking: number of cigarettes smoked per day, number of years smoked (together the number of cigarettes smoked can be calculated) and time since last cigarette, in terms of years. The former smokers pose the most problems for classification. It would be difficult to

differentiate between, for example, a patient who smoked heavily for a number of years but ceased shortly before sample donation from a current, light smoker of a few years. Additionally, it could be argued some former smokers could be classed as 'never exposed' in terms of the current effect of smoking given the low number of cigarettes smoked and long period of cessation, *e.g.* a patient who smoked 5 cigarettes per day for 3 years and ceased 40 years ago. Any values for number of cigarettes smoked and cessation length used for classification would be arbitrary, hence the categories used.

Although this study has not revealed any clustering of samples based on smoking status, cancer status was applied to samples from only patients whose smoking status was 'never exposed' (scores plot shown in Figure 2.12, loadings plot not shown; four component model, $R^2\boldsymbol{X}$(cum) = 0.662 and $Q^2\boldsymbol{X}$(cum) = 0.373) because it has been documented that smoking exposure could be a potential confounding factor.[150] Cancer status did not separate the 17 case and 9 control samples.

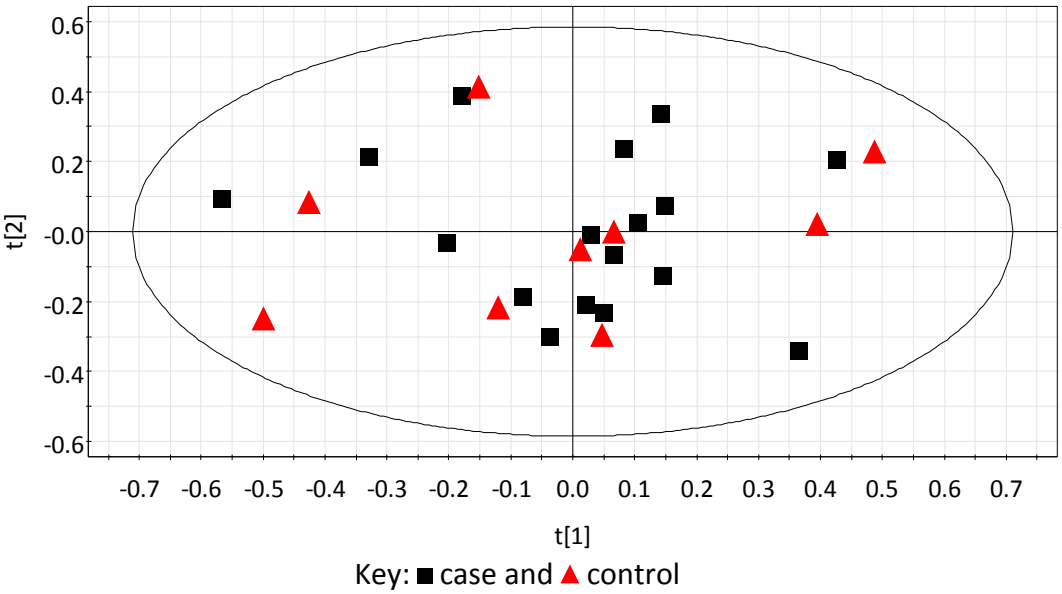


Figure 2.12 PCA scores plot of plasma data for case and control samples from patients who have never smoked. R^2X = 0.278 and 0.188, and Q^2X = 0.147 and 0.077 for PC 1 and PC 2, respectively.

Further investigation related to BMI was performed. Each BMI classification, with the exception of <18.5 kg m^{-2} because there was only one sample in this class, was

subjected to application of cancer status for MVA. Separation in scores space was not observed according to cancer status for BMI ≥30 kg m^{-2} (Figure 2.13; five component model, $R^2\boldsymbol{X}$(cum) = 0.785 and $Q^2\boldsymbol{X}$(cum) = 0.456) or for the other two BMI classifications (data not shown). Unlike for BMI classifications of 18.5-24.9 and 25-29.9 kg m^{-2} where PLS-DA models were unable to be fitted, a two component model ($R^2\boldsymbol{X}$(cum) = 0.454, $R^2\boldsymbol{Y}$(cum) = 0.669 and $Q^2\boldsymbol{Y}$(cum) = 0.238) was created for BMI ≥30 kg m^{-2} (Figure 2.14). The loadings plot (Figure 2.15) shows bins containing lactate and glucose are driving the separation between case and control samples.

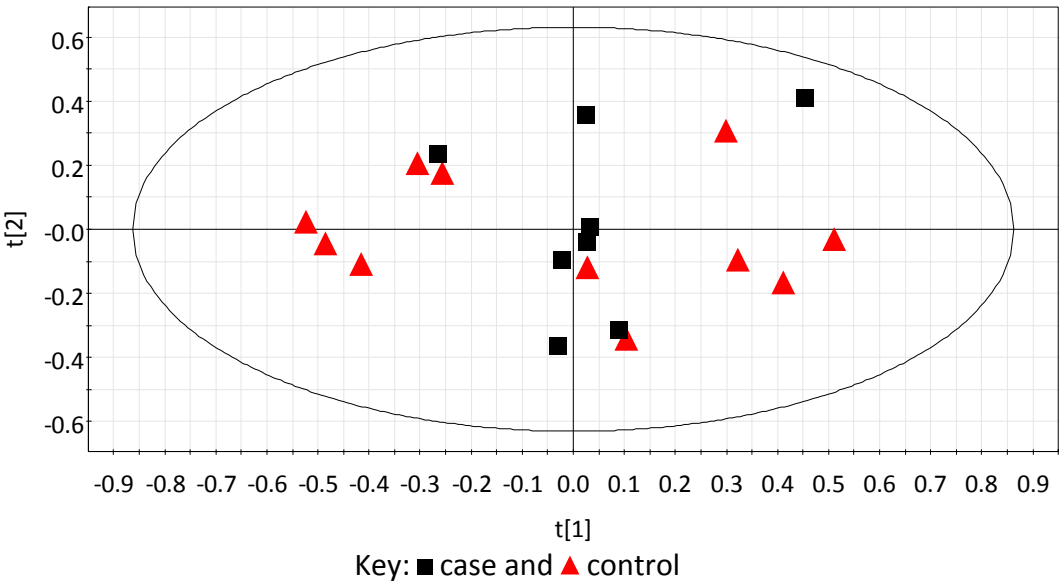


Figure 2.13 PCA scores plot showing the first two model components of plasma data for 8 case and 11 control samples from patients who had a BMI ≥30 kg m^{-2}. R^2X = 0.364 and 0.195, and Q^2X = 0.261 and 0.138 for PC 1 and PC 2, respectively.

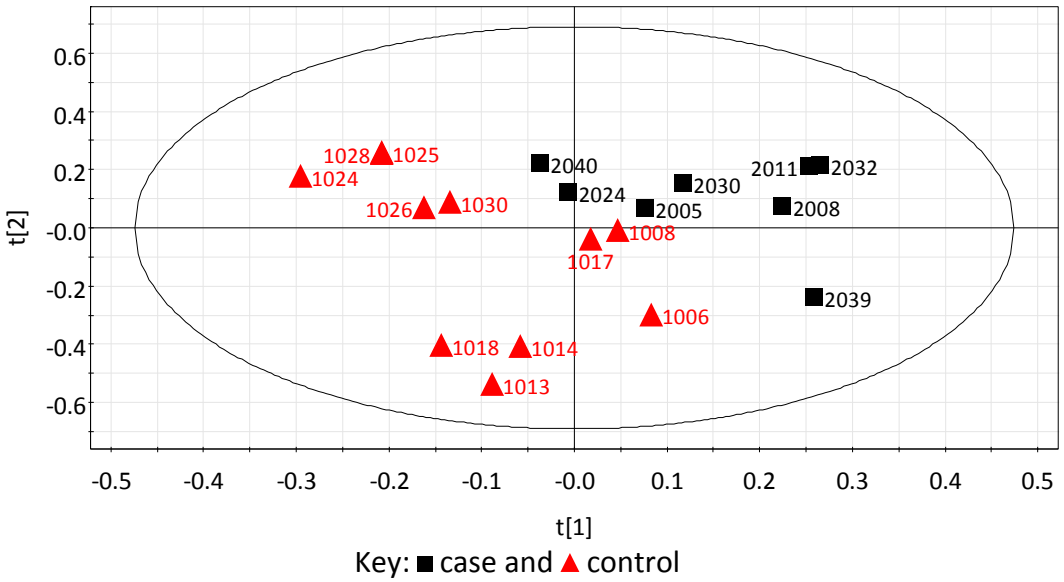


Figure 2.14 PLS-DA scores plot of plasma data for 8 case and 11 control samples from patients who had a BMI ≥30 kg m^{-2}.

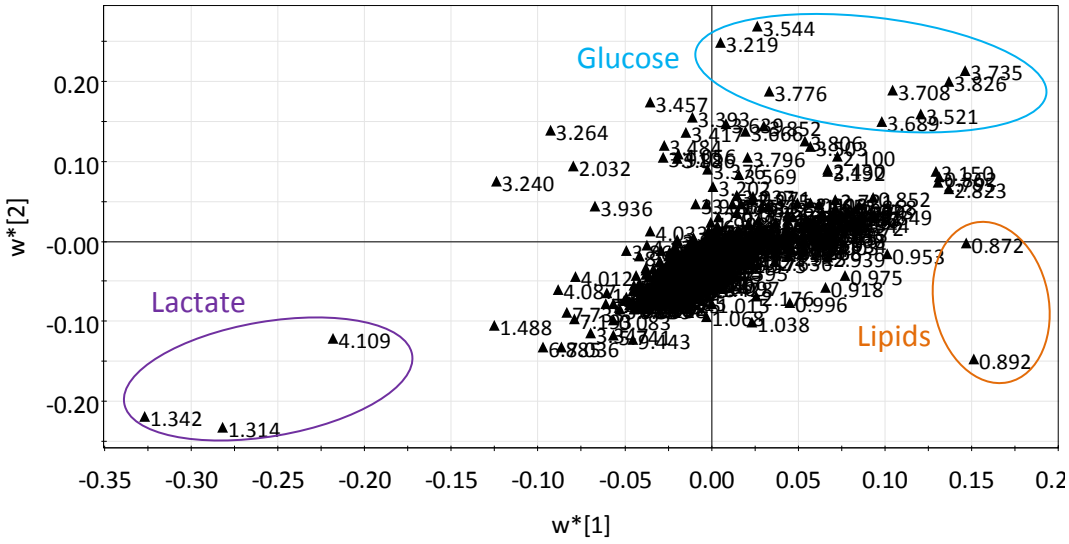


Figure 2.15 PLS-DA loadings plot corresponding to the model displayed in Figure 2.14.

It is well known that glucose uptake and formation of lactate increases as normal cells transform to malignant cells[147] but this is not reflected in the plasma data set for these samples. The area occupied in the scores plot by most of the control samples is influenced by lactate and the equivalent metabolite for some of the case samples is glucose. All samples have an associated BMI of ≥30 kg m^{-2}, and with some greatly above this value (the maximum BMI for case and control samples is 50.2 and 50.7 kg m^{-2}, respectively) the samples represented do not have typical

associated BMIs. Higher blood lactate levels have been observed with obesity and in patients who gained weight during chemotherapy.[124] None of the 63 patients received neoadjuvant therapy (chemotherapy prior to an operation) and because samples were taken before surgery, adjuvant therapy (chemotherapy after an operation) would not apply, therefore it cannot be postulated that obese patients were more likely to gain weight whilst receiving chemotherapy and hence have greater plasma levels of lactate. However, the validity of the model has to be considered. Table 2.3 shows the leave-one-out cross validation parameters.

Table 2.3 'Leave-one-out' cross-validation parameters of the PLS-DA model shown in Figure 2.14.

Sample Excluded	Number of Components	R^2X(cum)	R^2Y(cum)	Q^2Y(cum)	*Y*-Predicted*	
					1	2
1006	2	0.467	0.693	0.257	0.307	0.693
1008	2	0.462	0.723	0.199	0.281	0.719
1013	2	0.432	0.680	0.176	1.097	-0.097
1014	2	0.452	0.671	0.098	0.896	0.204
1017	2	0.474	0.735	0.205	0.282	0.718
1018	0	/	/	/	/	/
1024	1	0.110	0.619	0.052	0.990	0.010
1025	1	0.116	0.603	0.065	0.739	0.261
1026	1	0.140	0.576	0.163	0.753	0.247
1028	1	0.112	0.613	0.140	0.814	0.186
1030	1	0.140	0.570	0.213	0.672	0.328
2005	1	0.165	0.563	0.231	0.697	0.303
2008	0	/	/	/	/	/
2011	0	/	/	/	/	/
2024	1	0.179	0.569	0.107	0.847	0.153
2030	0	/	/	/	/	/
2032	0	/	/	/	/	/
2039	1	0.111	0.634	0.079	0.619	0.381
2040	2	0.467	0.735	0.108	0.892	0.108

*A sample was regarded as belonging to a grade by having a ***Y***-predicted value >0.50. Incorrect classification is represented by red shading and correct by orange or green, corresponding to a ***Y***-predicted value of 0.60-0.70 or >0.70, respectively.

Of the 11 control samples, seven (63%) were predicted correctly whilst the predictive ability of case samples was 0% due to having none of the eight samples correctly predicted. In total five of the 19 models were unable to be built, four controls and one case. Together, this indicates that the data could be overfitted by the model. Application of permutation testing allowed the model to be further scrutinised.

Figure 2.16 shows many permuted models for the case class to have higher $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ values than the original model. The intercept values of the regression lines for $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ are 0.622 and -0.117, respectively. The $R^2\boldsymbol{Y}$ value is much greater than 0.4 so it is indicated that the model has overfitted the data thus furthering the conclusion from 'leave-one-out' cross validation. The permutation testing plot for the control class is not shown but with $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ intercept values of 0.619 and -0.119, respectively, the same conclusion was made.

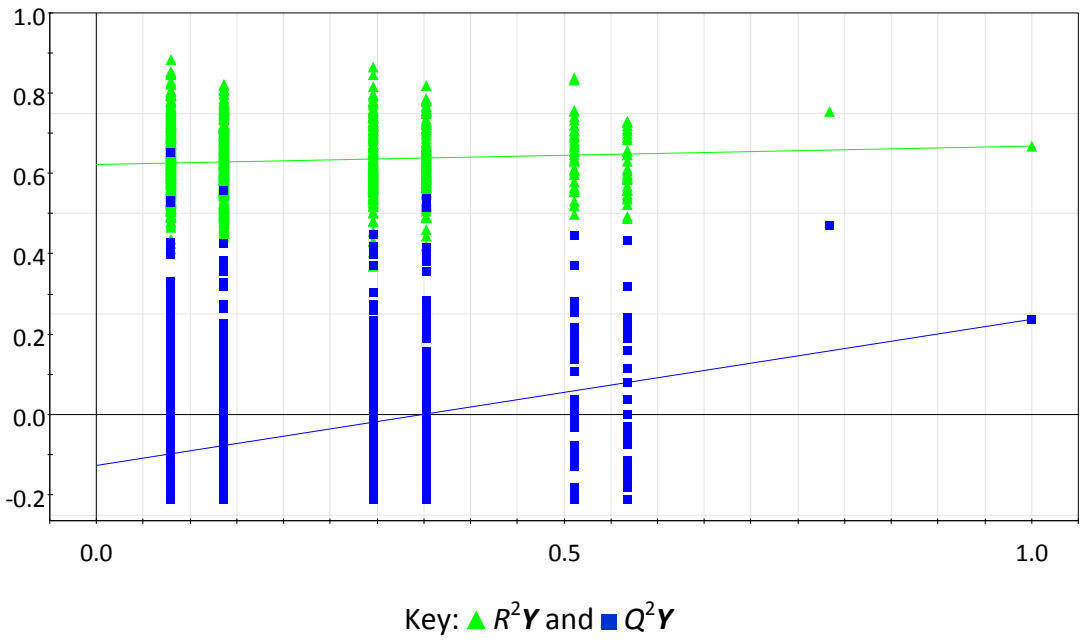


Figure 2.16 Permutation testing plots for the case class in the PLS-DA model shown in Figure 2.14. The R^2Y and Q^2Y intercept values of the regression lines are 0.622 and -0.117, respectively.

It is possible that there are distinct metabolic markers but these may be masked by potential confounding factors other than smoking status and BMI. Consequently,

other patient parameters were interrogated for potential data correlation (Table 2.4). Parameters were classified in a number of ways dependent on the medical record information. For some parameters, such as hormone replacement therapy (HRT) usage, there was a positive or negative class whereas for others, for example PR value, a fixed number of values, and hence classes, were attributed. For many parameters where a range of values existed, three classes were devised to each accommodate one-third of the samples. Individual values were also modelled if appropriate. Some parameters were only applicable to case samples.

Table 2.4 Parameters, with type of classification, used to investigate potential correlation with observed data in multivariate analysis.

Parameter (units)	Classification
Age [at time of sample donation]	Individual values
Height (m)	Individual values
Weight (kg)	Individual values
BMI (kg m^{-2})	Individual values; 3 percentile groups
Family history of breast cancer	Positive or negative
Menarche age	Individual values
Oral contraception usage	Positive or negative
Oral contraception usage for over one year	Positive or negative
Number of pregnancies	Individual values
Number of children	Individual values
Breastfed at least one child	Positive or negative
Breastfed more than one child	Positive or negative
Breastfed at least one child for two months or longer	Positive or negative
Breastfed more than one child for two months or longer	Positive or negative
Total number of months breastfed to children	Individual values
Menopause age	Individual values
HRT usage	Positive or negative
HRT usage within six weeks of breast cancer diagnosis	Positive or negative
Current smoker	Positive or negative
Never smoked	Positive or negative
Never smoked or stopped more than 10 years ago	Positive or negative
Number of cigarettes smoked	Individual values
Cancer grade	1, 2 or 3
DCIS [simplified grade]	Low, intermediate, high and N/A
ER alpha score	0/8 - 8/8 and N/A
PR score	0/8 - 8/8 and N/A
HER2 status	Yes, no or N/A
LCIS	Yes, no or N/A
Lympho-vascular invasion	Yes, no or N/A
Nottingham prognostic index	Individual values
Invasive cancer size (mm)	Individual values
Overall cancer size (mm)	Individual values

Table 2.4 Continued.

Prothrombin time	Individual values
White blood cell count ($x10^9$ L^{-1})	Individual values; 3 percentile groups
Platelet count ($x10^9$ L^{-1})	Individual values; 3 percentile groups
Haemoglobin (g dL^{-1})	Individual values; 3 percentile groups
Red blood cell count ($x10^{12}$ L^{-1})	Individual values; 3 percentile groups
Eosinophil count ($x10^9$ L^{-1})	Individual values; 3 percentile groups
Monocyte count ($x10^9$ L^{-1})	Individual values; 3 percentile groups
Lymphocyte count ($x10^9$ L^{-1})	Individual values; 3 percentile groups
Basophil count ($x10^9$ L^{-1})	Individual values; 3 percentile groups
Neutrophil count ($x10^9$ L^{-1})	Individual values; 3 percentile groups
International normalised ratio	Individual values; 3 percentile groups
Activated partial thromboplastin time (s)	Individual values; 3 percentile groups
Glucose (mmol L^{-1})	Individual values; 3 percentile groups
Alanine transaminase (units L^{-1})	Individual values; 3 percentile groups
Bilirubin (µmol L^{-1})	Individual values; 3 percentile groups
Phosphate (mmol L^{-1})	Individual values; 3 percentile groups
Alkaline phosphatase (units L^{-1})	Individual values; 3 percentile groups
Albumin (g L^{-1})	Individual values; 3 percentile groups
Albumin-adjusted calcium (mmol L^{-1})	Individual values; 3 percentile groups
Sodium (mmol L^{-1})	Individual values; 3 percentile groups
Potassium (mmol L^{-1})	Individual values; 3 percentile groups
Creatinine (µmol L^{-1})	Individual values; 3 percentile groups
Urea (mmol L^{-1})	Individual values; 3 percentile groups

BMI, body mass index; DCIS, ductal carcinoma in situ; ER, oestrogen receptor; HER2, human epidermal growth factor 2; HRT, hormone replacement therapy; LCIS, lobular carcinoma in situ; and PR, progesterone receptor.

The observed data did not correlate with any of the 55 parameters (plots not shown), which would indicate that *via* the data acquisition and analysis methods used none of the parameters can be identified as individual confounding factors. Given this and the number of parameters, it was not feasible to combine two or more parameters to investigate the cumulative effect of potentially confounding factors or exploration using parameter matched samples, with the exception of BMI, age and smoking status because these are commonly identified as confounding factors.[(151)]

Ten case and ten control samples were matched singly according to BMI or age. For both groups the ten samples closest to the median of the 63 samples, 27.3 kg m^{-2} for BMI and 63 years for age, were chosen. Furthermore, selecting only five case and five control matched samples from patients who had never smoked ensured the range for BMI and age was kept as small as possible whilst incorporating another matched factor. Although using a larger sample number would be better practice for MVA, the range for the two parameters of BMI and age would increase greatly, otherwise former smokers would have to be included in an attempt to limit the range expansion in both groups. Table 2.5 summarises the demographics for the parameters age, BMI and smoking status.

Table 2.5 BMI and/or age range for parameter matched samples.

Cancer Status	**Matched Parameter(s) Range [Never Smoked Cases/Controls]**			
	BMI [4/4]	Age [3/3]	BMI and Age [5/5]	
			BMI	Age
Case	25.8 - 28.7	60 - 64	24.7 - 29.1	56 - 68
Control	24.4 - 31.4	60 - 66	22.0 - 30.0	53 - 75

Units: BMI = kg m^{-2}.

The PCA scores plots for the BMI matched (Figure 2.17; three component model, $R^2\boldsymbol{X}$(cum) = 0.571 and $Q^2\boldsymbol{X}$(cum) = 0.153) and age matched samples (Figure 2.18; two component model, $R^2\boldsymbol{X}$(cum) = 0.465 and $Q^2\boldsymbol{X}$(cum) = 0.235) did not exhibit two clusters of samples in scores space based on breast cancer status. Further evidence is therefore provided that metabolites that are indicative of breast cancer status cannot be identified from this data set when the variation of a single potential confounding factor is much reduced.

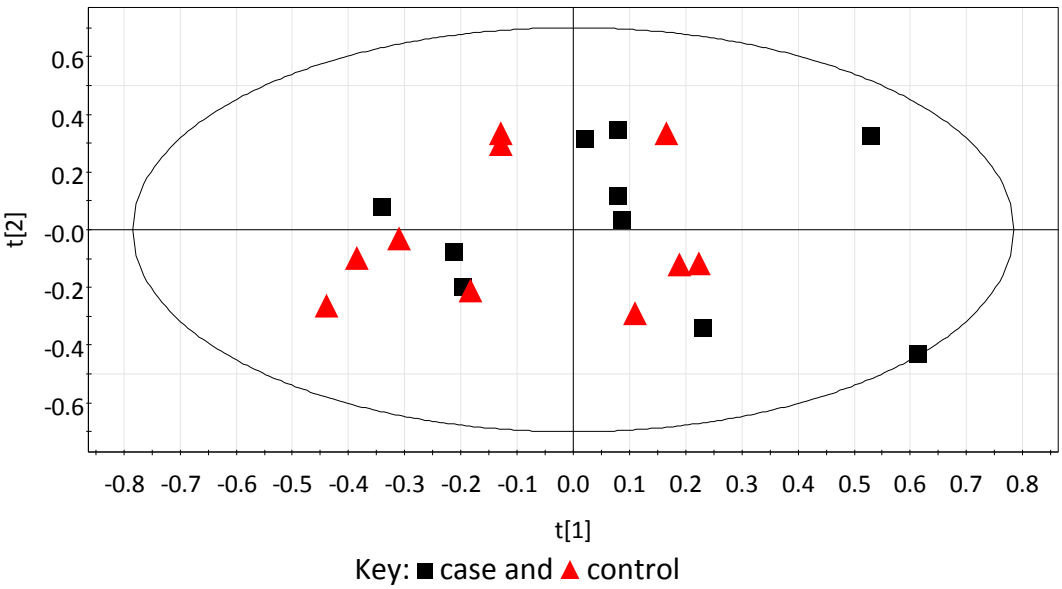


Figure 2.17 PCA scores plot showing the first two components of plasma data for best BMI matched 10 case and 10 control samples. R^2X = 0.257 and 0.205, and Q^2X = 0.035 and 0.099 for PC 1 and PC 2, respectively.

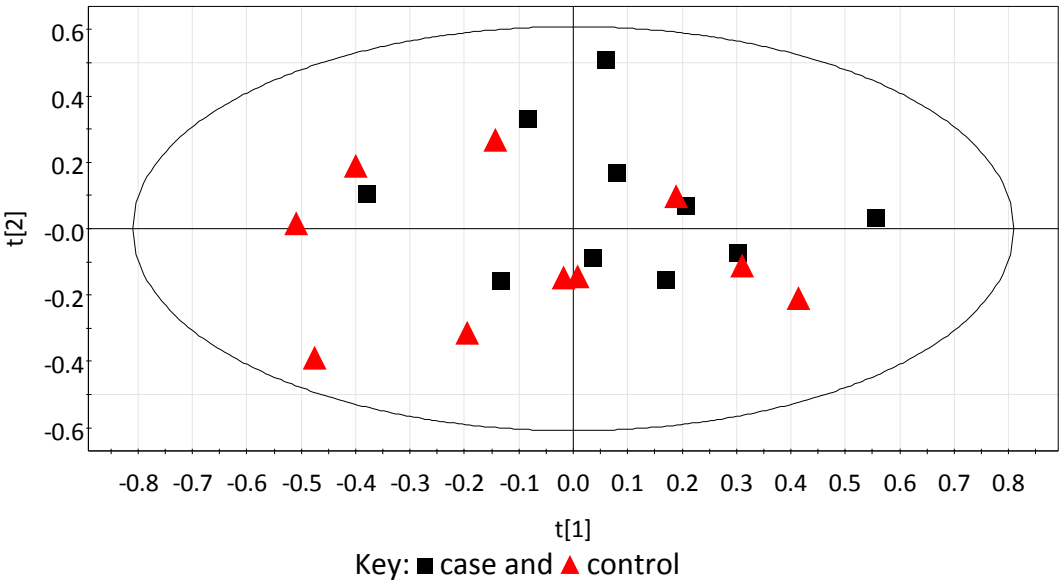


Figure 2.18 PCA scores plot of plasma data for best age matched 10 case and 10 control samples.

Figure 2.19 indicates a separation between case and control samples in scores space for the best age, BMI and smoking status matched five samples for case and control groups (two component model, $R^2\boldsymbol{X}$(cum) = 0.543 and $Q^2\boldsymbol{X}$(cum) = 0.089). The bins that have the most effect on the observed separation in the loadings plot (Figure 2.20) contain lactate, lipids and glucose. Positioning of these bins in loadings space is similar to that in the corresponding space when the full range of samples

has been included (Figure 2.5) with the exception of bins containing lactate (1.314 and 1.342 ppm), which are influential in PC 2 for the full sample range but in PC 1 for the matched samples.

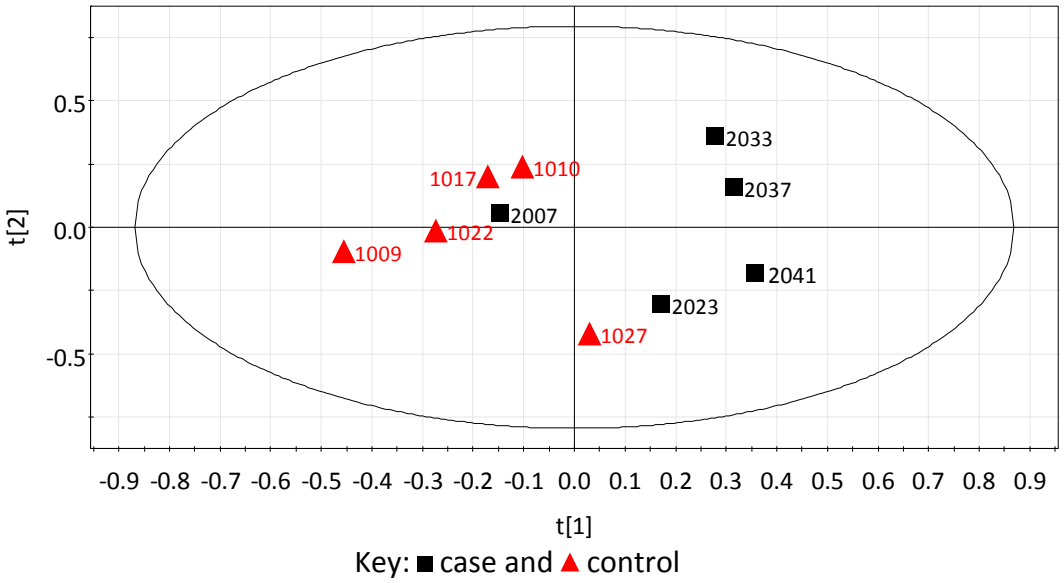


Figure 2.19 PCA scores plot of plasma data for best BMI, age and 'never smoked' status matched 5 case and 5 control samples.

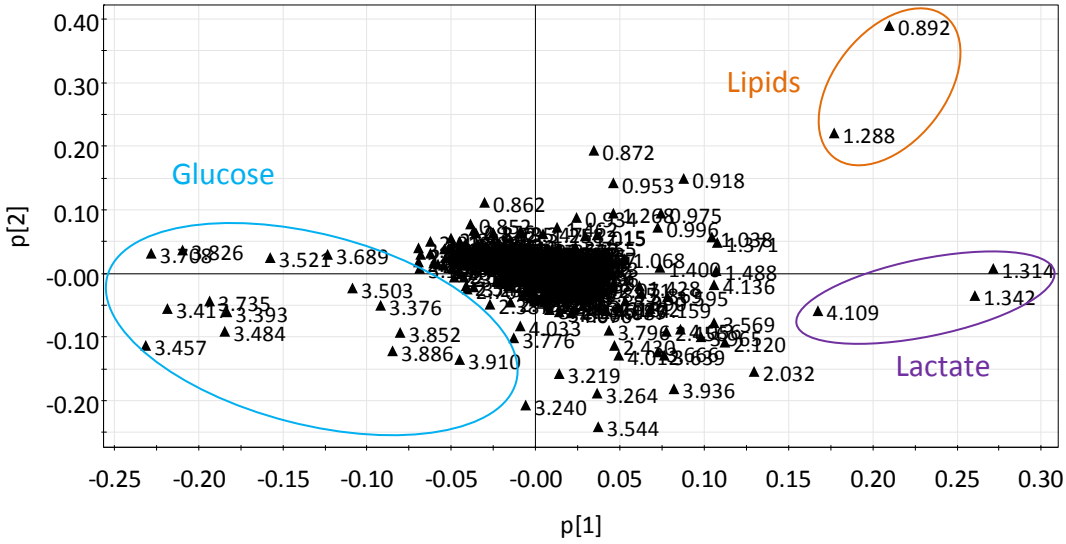


Figure 2.20 PCA loadings plot corresponding to the model displayed in Figure 2.19.

The scores plot (Figure 2.21) of the one component PLS-DA model ($R^2\boldsymbol{X}$ = 0.287, $R^2\boldsymbol{Y}$ = 0.733 and $Q^2\boldsymbol{Y}$ = 0.394) separated all samples according to class except for 2007, which was closer to control rather than case samples in scores space. The loadings

plot (not shown) identified the same bins contributing to the separation in scores space as those influencing PC 1 in the PCA loadings plot.

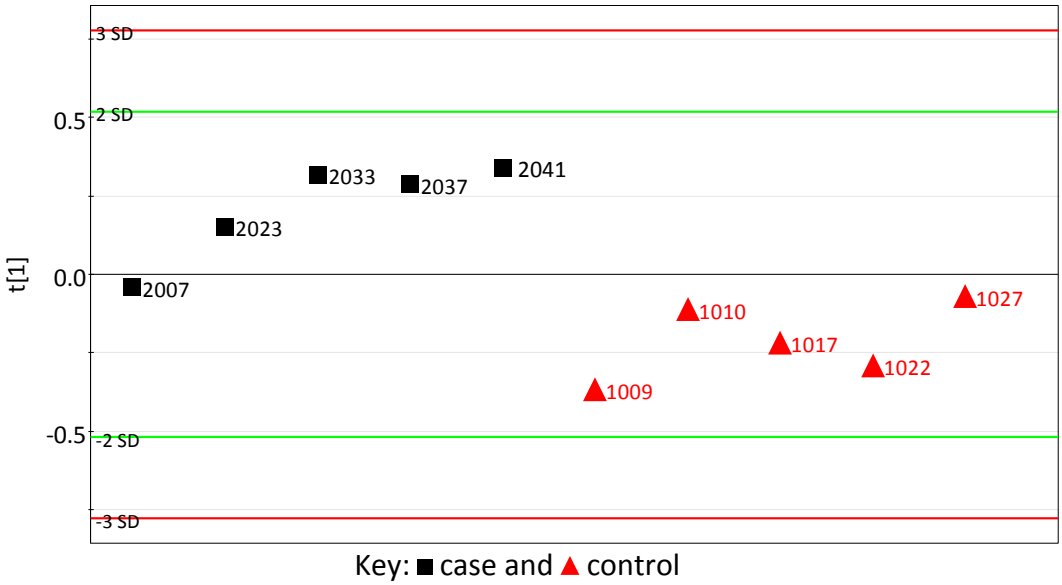


Figure 2.21 PLS-DA scores plot of plasma data for best BMI, age and smoking status matched 5 case and 5 control samples. Classed according to sample type.

Validation of the model was performed by 'leave-one-out' cross-validation (Table 2.6). Four out of five (80%) case and control samples were correctly predicted. Singly, this value is sufficiently high to suggest that the model has not overfitted the data although three of the eight samples had a ***Y***-predicted value between 0.50 and 0.60 so the classification of these should be treated with caution. Additionally, the $R^2\boldsymbol{Y}$ value of the first component was high for all models and for many, more than one component resulted but the $R^2\boldsymbol{Y}$ value approached one so predictions were made using a single component. Permutation testing was required to evaluate whether the model had overfitted the data (case class plot shown in Figure 2.22, control class plot not shown). The $R^2\boldsymbol{Y}$ intercept value of the regression line was 0.622 for the case class and 0.620 for the control class whilst the $Q^2\boldsymbol{Y}$ intercept value was -0.085 for both classes. Some permuted values had greater $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ values than the original model.

Taking all of this into consideration, it cannot be substantiated that case samples compared to control samples that are matched according to BMI, age and 'never

smoked' status have increased levels of lipids and lactate but reduced glucose as indicated by both the PCA and PLS-DA models. Although 'leave-one-out' cross-validation supported the findings of the PLS-DA model, permutation testing did not. This is possibly due to the small number of samples included, which resulted in a large R^2Y value for the first component and subsequent high fit of the data. A greater number of samples would be required to confirm whether levels of lipids, lactate and glucose were different in case and control samples.

Table 2.6 'Leave-one-out' cross-validation parameters for the PLS-DA model shown in Figure 2.21.

Sample Excluded	Number of Components	R^2X	R^2Y	Q^2Y	*Y*-Predicted*	
					1	2
1009	1	0.272	0.742	0.481	1.015	-0.015
1010	1	0.302	0.763	0.570	0.537	0.463
1017	1	0.314	0.726	0.540	0.575	0.425
1022	1	0.289	0.710	0.472	0.903	0.097
1027	1	0.321	0.800	0.606	0.313	0.687
2007	1	0.301	0.848	0.540	0.741	0.259
2023	1	0.302	0.757	0.438	0.440	0.560
2033	1	0.302	0.694	0.454	0.133	0.867
2037	1	0.282	0.735	0.519	0.193	0.807
2041	1	0.288	0.711	0.510	-0.015	1.015

*A sample was regarded as belonging to a grade by having a ***Y***-predicted value >0.50. Incorrect classification is represented by red shading and correct by pink, orange or green, corresponding to a ***Y***-predicted value of <0.60, 0.60-0.70 or >0.70, respectively.

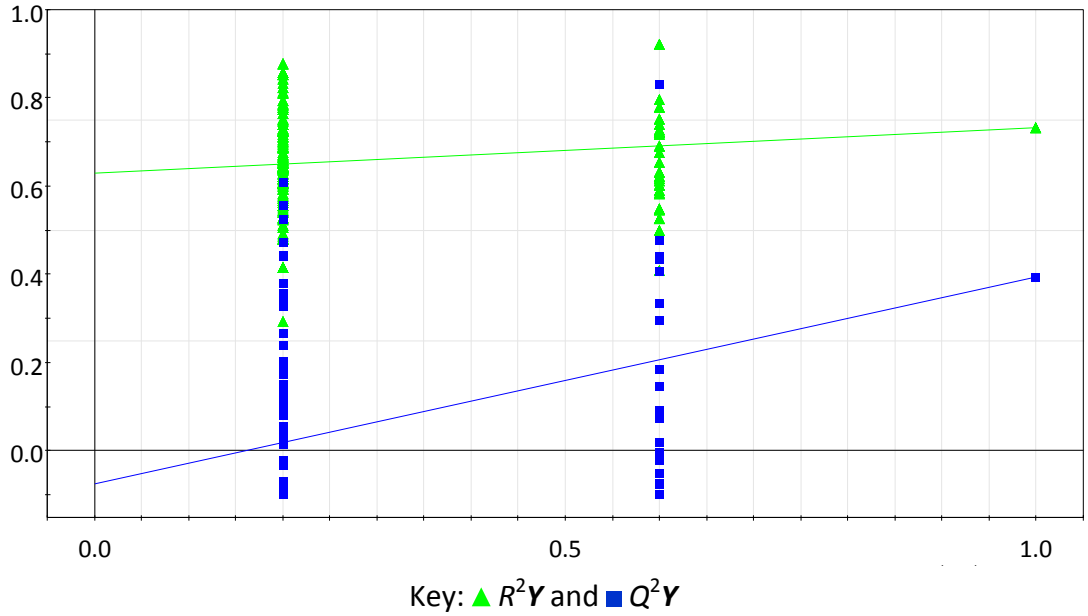


Figure 2.22 Permutation testing plots for the case class in the PLS-DA model shown in Figure 2.21. The R^2Y and Q^2Y intercept values of the regression lines are 0.622 and -0.085, respectively.

Sample 2007 exhibited high-grade DCIS rather than invasive ductal carcinoma meaning the case type is different to three other case samples, which could indicate why it is not positioned with the other case samples. However, 2041 is also a DCIS sample and is of intermediate grade with a smaller overall size, thus is not supportive of the above argument. The medical record for sample 1027 does not indicate any characteristics that could suggest a greater tendency to develop cancer, for example, there is no family history of breast cancer, and although none of the control samples had normal breasts, 1027 is not unique being one of five in total to exhibit breast cysts.

Using a greater number of samples would increase the reliability of any model findings but, for this study, would result in the samples being unable to be matched according to BMI, age and smoking status; the ranges of the two parameters and smoking status would be similar to that when all samples were used. With just five samples in each group the age range has already increased greatly and the BMI range has also expanded.

2.1.2 Analysis of Spectrum Excluding the Glucose Region

For whole spectra, the data distribution along PC 1 was shown to be strongly influenced by bins in the region containing glucose (3.180-3.940 ppm) for case and control samples (Figure 2.5) and as previously noted, a decreased level of glucose has been associated with breast cancer[135,147] though the level can be directly and rapidly affected by diet.[66,152,153] It was not possible to regulate dietary consumption before sample donation so as a result the region containing glucose signals was removed to reduce any potential effect of diet on analysis.

The same systematic procedures were followed as per whole spectrum analysis: PCA and PLS-DA were performed for case and control samples combined and classed according to various parameters. A summary of the models is shown in Table 2.7.

Table 2.7 Parameters used for classification and model descriptors excluding the glucose region (3.180-3.940 ppm). PLS-DA models were not able to be built if model descriptors are not listed.

Samples Included	Parameter (class)	Plot (PCA unless stated)		PCs	R^2X (cum)	R^2Y (cum)	Q^2X/ Q^2Y (cum)
		Scores	Loadings				
All	State (case and control)	Figure 2.23	Figure 2.24	3	0.517	/	0.371
All	BMI (<18.5, 18.5-24.9, 25-29.9 and ≥30)	Figure 2.25	/	3	0.507	/	0.329
All	Smoking status ('never exposed' and 'exposed')	Figure 2.26	Figure 2.24	3	0.517	/	0.371
All	Smoking status ('never exposed' and 'exposed')	PLS-DA Figure 2.27	PLS-DA Figure 2.28	2	0.316	0.360	0.112
Single occurrence invasive ductal carcinoma and control	State (case and control)	Figure 2.30	/	2	0.446	/	0.298
Never-smoked	State (case and control)	Figure 2.31	/	3	0.586	/	0.362
Single occurrence case	Grade (0, 1, 2, 3 and 4)	Figure 2.32	/	2	0.458	/	0.245
Single occurrence invasive ductal carcinoma tumour	Grade (1, 2 and 3)	/	/	0	/	/	/

Units: BMI = kg m^{-2}.

There is not clear separation between case and control samples in scores space (Figure 2.23) but a very slight tendency for some case samples to have higher lipid levels (bins centred at 0.892 and 1.288 ppm; Figure 2.24) is apparent as is the case when the region that includes glucose is retained (Figure 2.4). The bin centred at 2.032 ppm has a high PC 1 loadings value. The main signal contained within this bin is attributed to glycoproteins.[142] Lactate also has a high PC 1 loadings value so it is difficult to ascertain the contribution of each metabolite towards the positioning of samples in the scores plot but there does not seem to be correlation between

glycoproteins and case samples despite glycoprotein association with breast cancer along with various other cancers.[154-156]

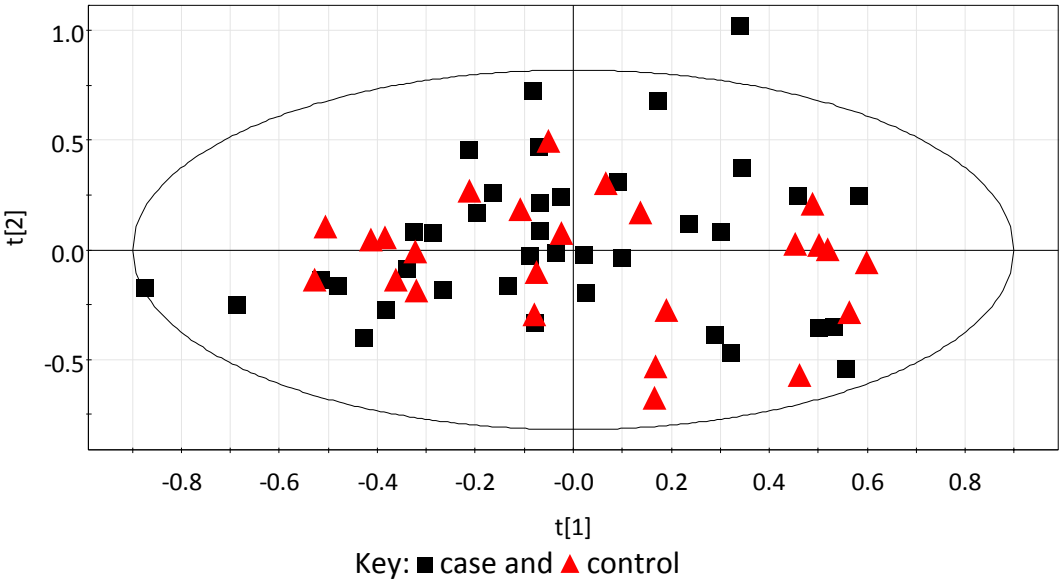


Figure 2.23 PCA scores plot showing the first two components of plasma data excluding the glucose region (3.18-3.94 ppm) for 38 case and 25 control samples. R^2X = 0.237 and 0.196, and Q^2X = 0.140 and 0.203 for PC 1 and PC 2, respectively.

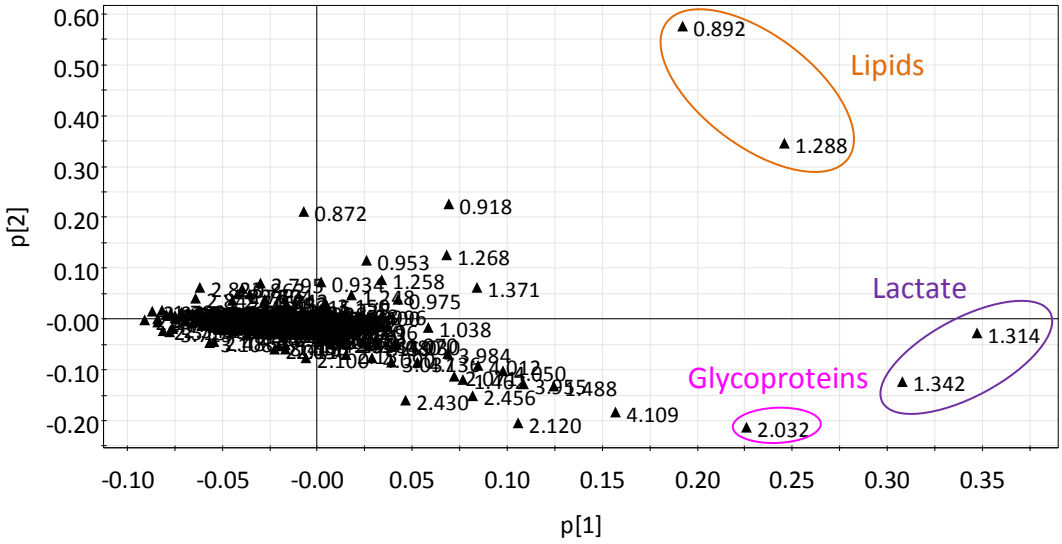
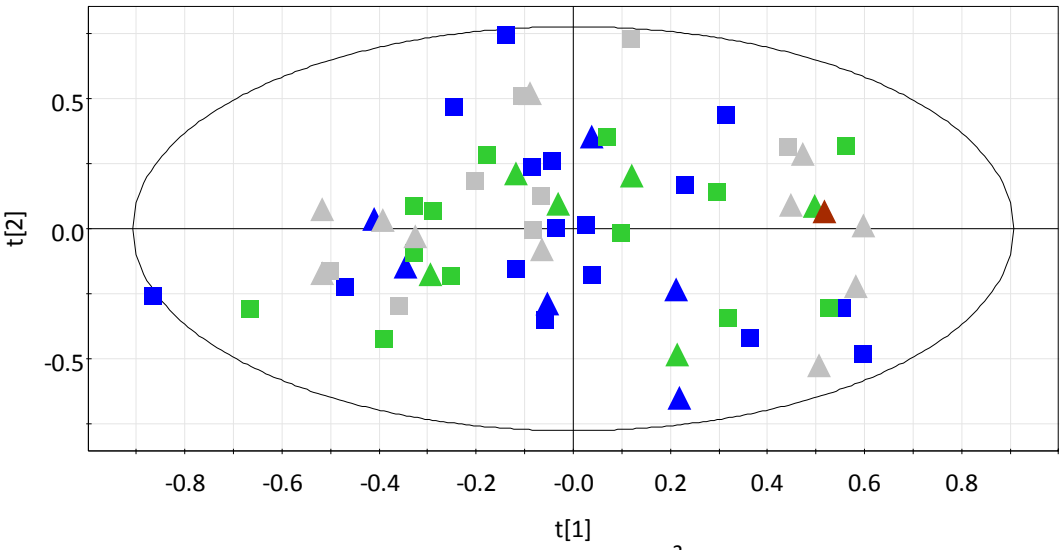


Figure 2.24 PCA loadings plot corresponding to the model displayed in Figure 2.23.

In Figure 2.25, of the 12 samples with the highest PC 1 scores value (>0.4), six have an associated BMI ≥30 kg m^{-2} and five of these are controls. The main corresponding bins in the loadings plot (not shown, extremely similar to Figure 2.24) contain lactate and glycoproteins. When PCA was employed, increased lactate was not

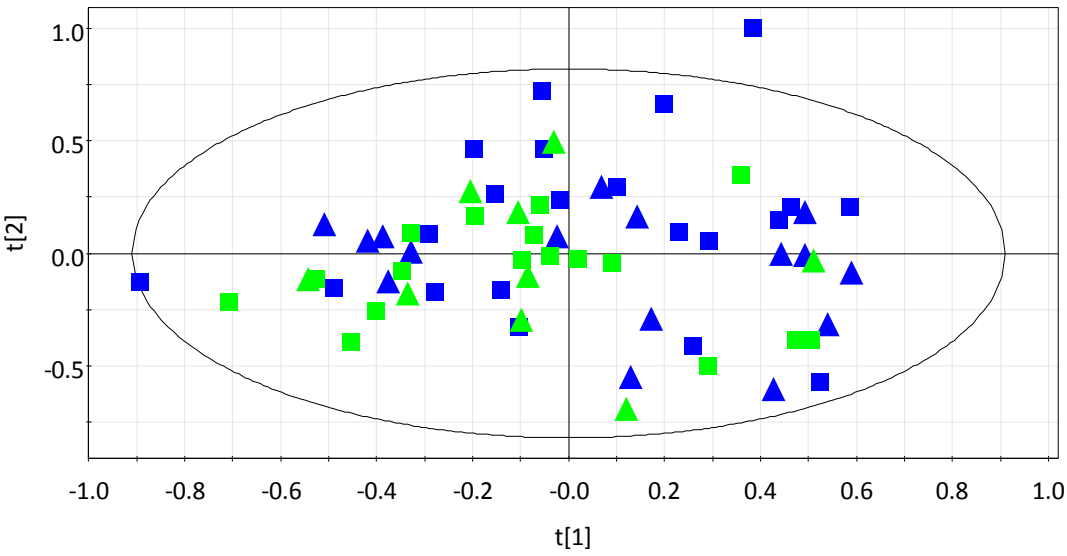
associated with controls from this BMI group when the glucose region was included (Figure 2.9) but was when PLS-DA was applied to only the case and control samples that had an associated BMI ≥30 kg m^{-2} (Figure 2.14). Unfortunately, due to the small number of samples and the trend not always being present, it is not possible to make substantiated conclusions regarding the lactate level for those with a BMI of ≥30 kg m^{-2} in relation to cancer status. There appears to be no association between glycoproteins and BMI.



Key: shape - □ case and △ control; colour - BMI (kg m^{-2}) = ● <18.5, ● 18.5-24.9, ● 25-29.9 and ● ≥30

Figure 2.25 PCA scores plot showing the first two components of plasma data excluding the glucose region (3.18-3.94 ppm) based on BMI. Group sample numbers from Figure 2.9 apply. R^2X = 0.243 and 0.178, and Q^2X = 0.141 and 0.170 for PC 1 and PC 2, respectively.

Regarding smoking status, the samples with the highest lipid levels are from patients who conform to the 'exposed' (current or former smoker) category (scores and loadings plots shown in Figure 2.26 and Figure 2.24, respectively), which is the same observation as when the glucose region was included. Additionally, of the 13 samples with the highest PC 1 scores value (>0.4), ten originated from 'exposed' patients.

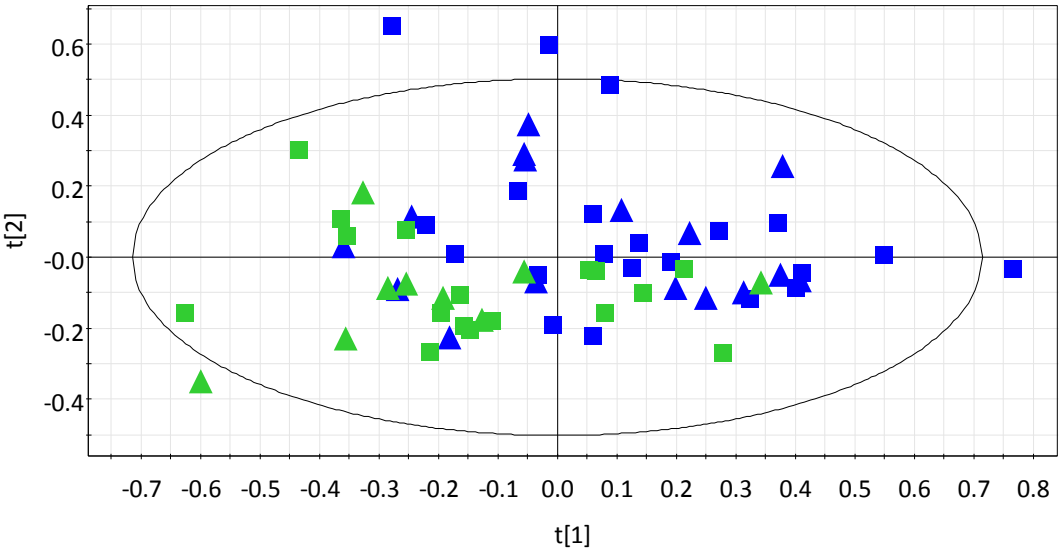


Key: shape - □ case and △ control; colour - smoking status = ● 'never exposed', ● 'exposed'

Figure 2.26 PCA scores plot showing the first two components of plasma data excluding the glucose region (3.18-3.94 ppm) based on smoking status. Group sample numbers from Figure 2.10 apply. R^2X = 0.237 and 0.196, and Q^2X = 0.140 and 0.203 for PC 1 and PC 2, respectively.

PLS-DA did not completely separate 'exposed' and 'never exposed' groups in scores space though a strong tendency towards separation could be observed (Figure 2.27). Much of the separation is along PC 1 with a contribution from PC 2. Lipids and lactate strongly influence PC 1 and levels are indicated to be lower in 'exposed' patients (Figure 2.28). However, the quality of the model is low (Table 2.7) so validation is even more imperative. The validity of the model was assessed through exclusion of a third of the samples and their class memberships predicted using models built from the remaining samples. For one instance, a model was not able to be built using two-thirds of the samples thus the maximum total number of samples for which class could be predicted correctly was 67%. One of the other models only had a Q^2Y value of 0.087 so predictions made based on this model would have to be treated with caution. Permutation testing ('exposed' class plot shown in Figure 2.29, 'never exposed' samples plot not shown) indicated the model had not overfitted the data because the R^2Y intercept value of the regression line was 0.323 for the 'exposed' class and 0.320 for the 'never exposed' class whilst the Q^2Y intercept values were -0.174 and -0.177, respectively. However, the model was poor due to

the high number of permuted models that had $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ values in excess of the values for the original model. Combined, this information leads to the conclusion, despite tentative indication from PLS-DA, that lower levels of lipids and lactate are not present in plasma from patients whose smoking status was 'exposed'.



Key: shape - □ case and △ control; colour - smoking status = ● 'never exposed', ● 'exposed'

Figure 2.27 PLS-DA scores plot of plasma data excluding the glucose region (3.18-3.94 ppm) based on smoking status. Group sample numbers from Figure 2.10 apply.

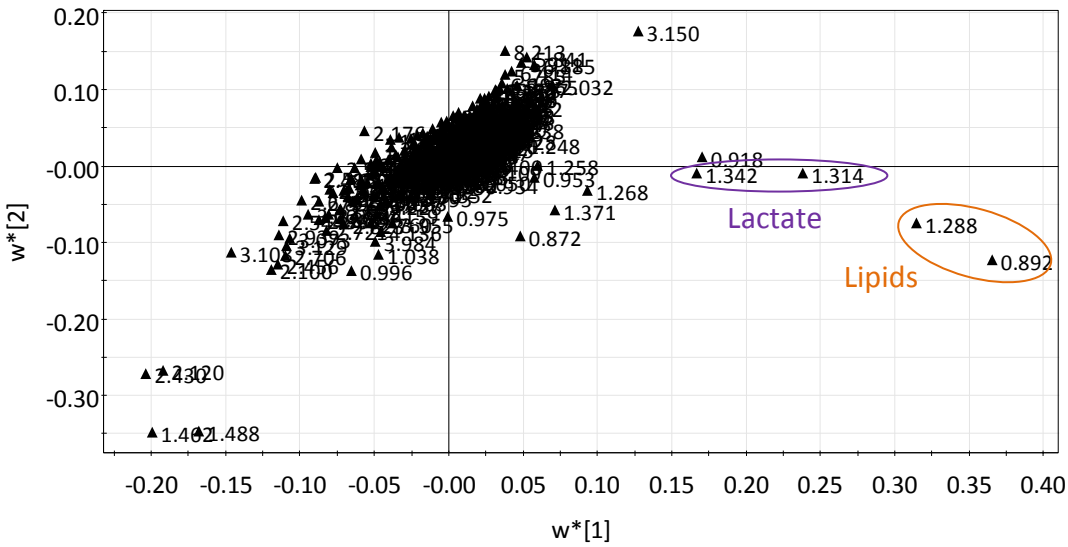


Figure 2.28 PLS-DA loadings plot corresponding to the model displayed in Figure 2.27.

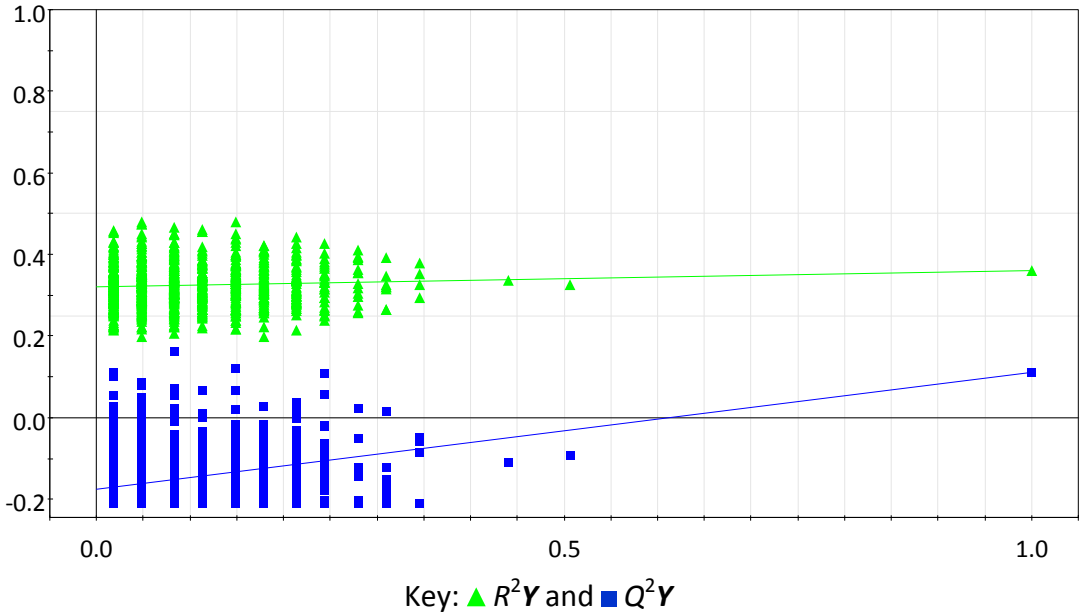


Figure 2.29 Permutation testing plots for the 'exposed' class in the PLS-DA model shown in Figure 2.27. The R^2Y and Q^2Y intercept values of the regression lines are 0.323 and -0.174, respectively.

No clear separation was observed between: single occurrence ductal cases and controls (Figure 2.30); cases and controls amongst the 'never exposed' smoking group (Figure 2.31); and single occurrence case samples based on tumour grade (Figure 2.32). It is interesting to note for tumour grade that although some grade 3 samples do exhibit higher lactate levels the three samples that have the greatest negative PC 1 value are all grade 3, which from the loadings plot (not shown but very similar to Figure 2.24) would indicate low levels of lactate. Although a similar observation is apparent with the glucose region retained (Figure 2.7) it is more visible with the glucose region removed. This is indicative that for plasma the lactate level, known to increase in malignant cells,[147] is not reflective of tumour grade.

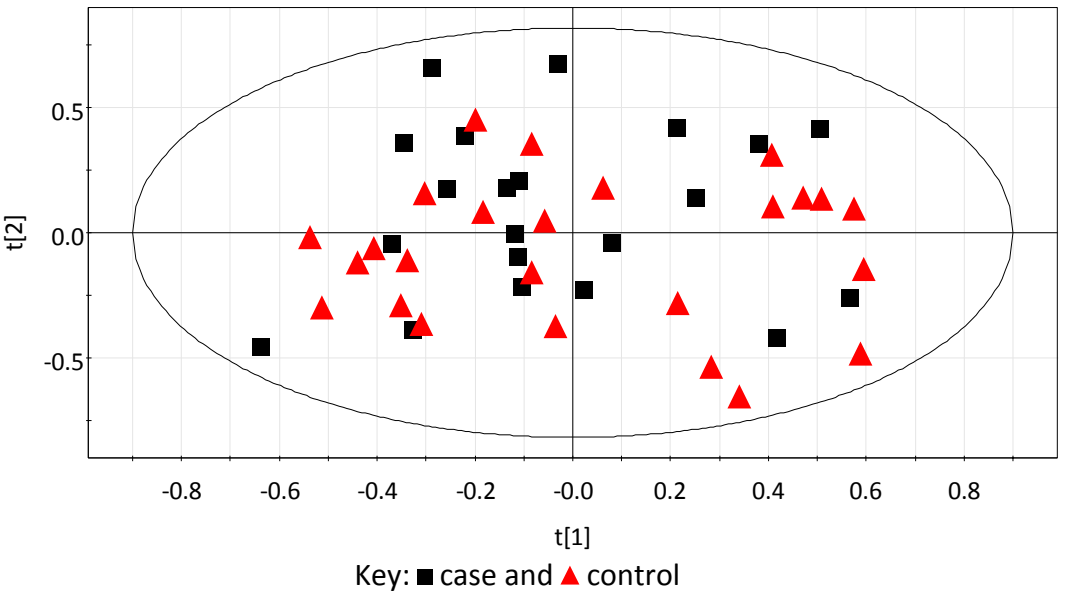


Figure 2.30 PCA scores plot of plasma data excluding the glucose region (3.18-3.94 ppm) for 21 single occurrence ductal case and 25 control samples.

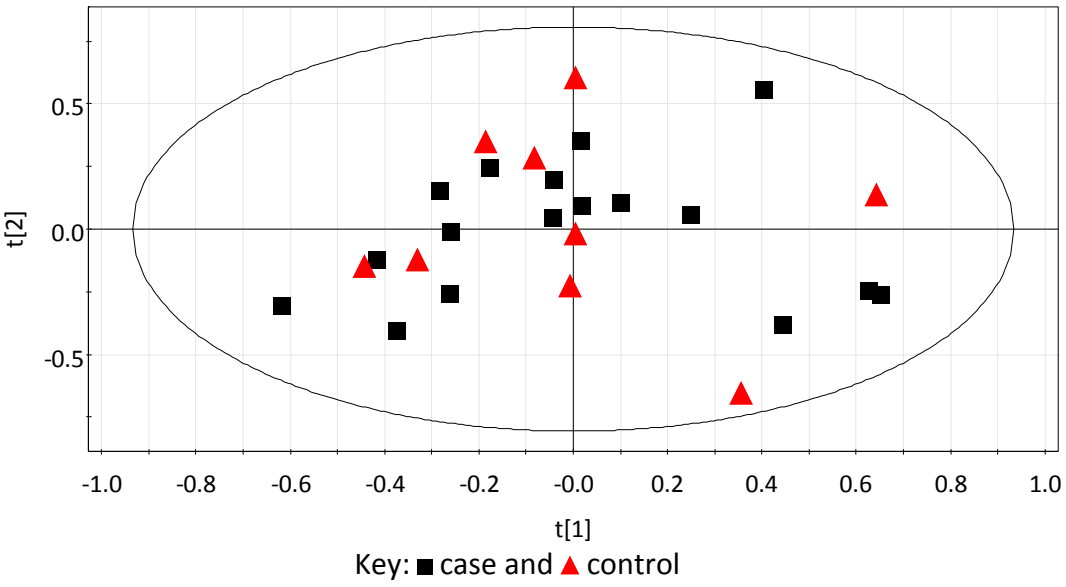


Figure 2.31 PCA scores plot showing the first two components of plasma data excluding the glucose region (3.18-3.94 ppm) for 17 case and 9 control samples from patients who have never smoked. R^2X = 0.271 and 0.202, and Q^2X = 0.133 and 0.151 for PC 1 and PC 2, respectively.

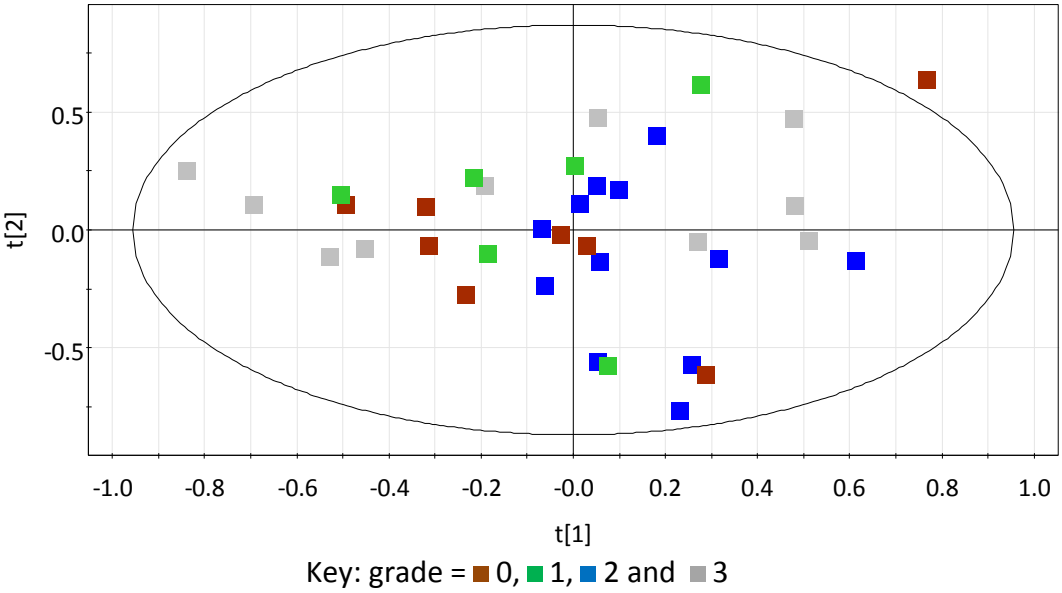


Figure 2.32 PCA scores plot of plasma data excluding the glucose region (3.18-3.94 ppm) for 36 single occurrence case samples based on tumour grade.

PLS-DA models were not created for any parameter with the exception of smoking status (Figure 2.27). Together, for all samples, this indicates that removal of glucose does not improve the possibility that the data could be related to cancer status, cancer grade, BMI classification or smoking status in this study. Given none of the PCA or PLS-DA models characterised the specific groups clearly it was deemed prudent to investigate best matched samples, thus eliminating some sources of variation.

Samples were again matched according to BMI, age or BMI, age and 'never smoked' status. The PCA scores plots for ten case and ten control BMI matched samples (Figure 2.33; two component model, $R^2\boldsymbol{X}$(cum) = 0.442 and $Q^2\boldsymbol{X}$(cum) = 0.065) or age matched samples (Figure 2.34; one component model, $R^2\boldsymbol{X}$ = 0.248 and $Q^2\boldsymbol{X}$ = 0.076) did not show separation between cases and controls in scores space. For both sets of samples, a PLS-DA model was not able to be built.

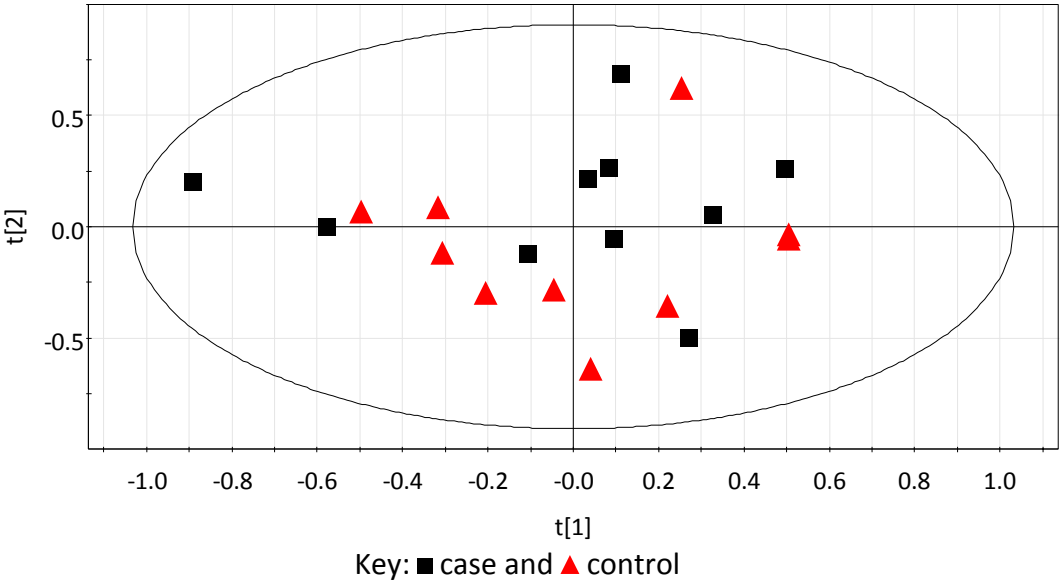


Figure 2.33 PCA scores plot of plasma data excluding the glucose region (3.18-3.94 ppm) for best BMI matched 10 cases and 10 controls.

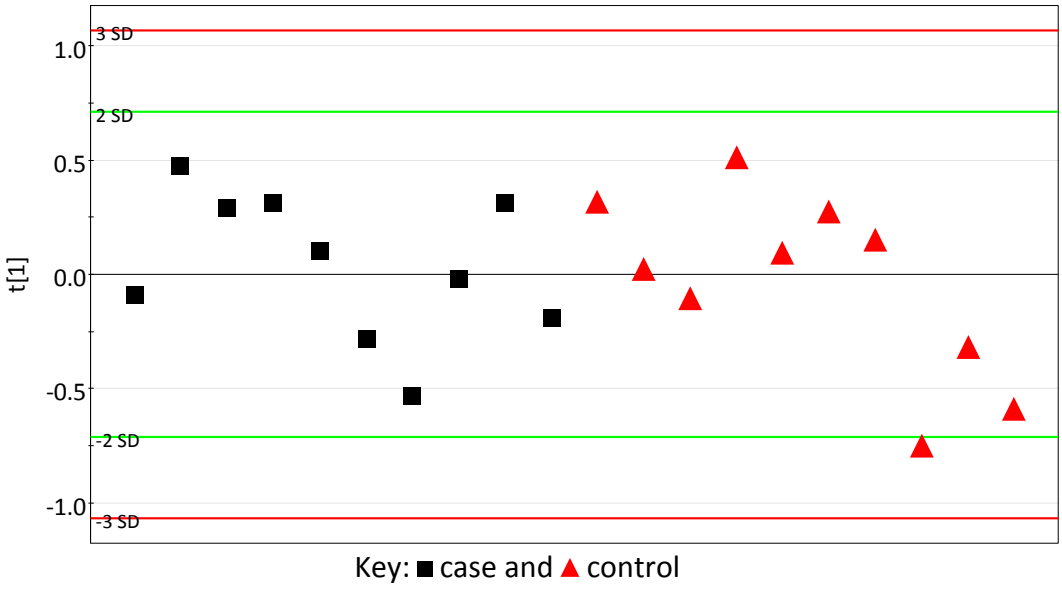


Figure 2.34 PCA scores plot of plasma data excluding the glucose region (3.18-3.94 ppm) for best age matched 10 cases and 10 controls.

A PCA model was unable to be created for the five case and five control samples that were matched according to BMI, age and 'never smoked' status but a one component PLS-DA model ($R^2\boldsymbol{X}$ = 0.247, $R^2\boldsymbol{Y}$ = 0.816 and $Q^2\boldsymbol{Y}$ = 0.422) was built. Scores and loadings plots are shown in Figure 2.35 and Figure 2.36, respectively. Lipids and lactate are indicated to be higher in case samples.

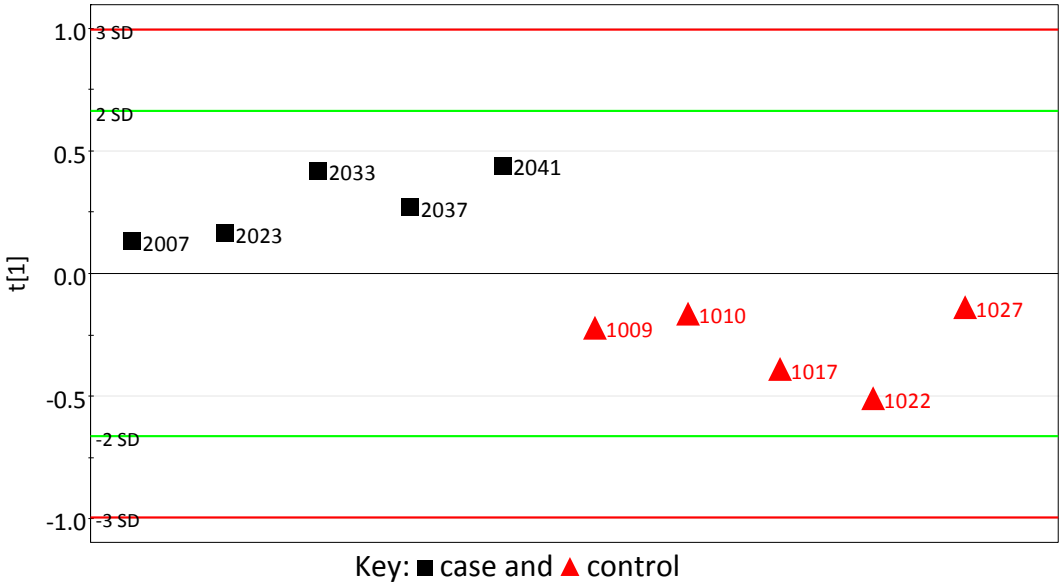


Figure 2.35 PLS-DA scores plot of plasma data excluding the glucose region (3.18-3.94 ppm) for best BMI, age and 'never smoked' status matched 5 case and 5 control samples. Classed according to cancer status.

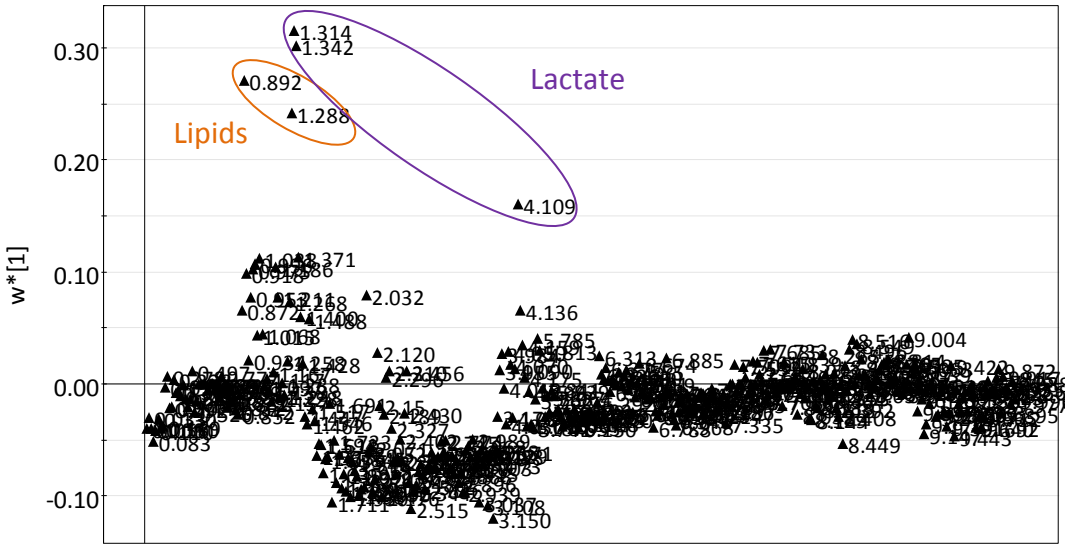


Figure 2.36 PLS-DA loadings plot corresponding to the model displayed in Figure 2.35.

Validation of the model was performed by 'leave-one-out' cross-validation (Table 2.8). Four out of five (80%) control samples and three out of five (60%) case samples were correctly predicted but two of the eight samples had a ***Y***-predicted value between 0.50 and 0.60 so the classification of these should be treated with caution. Only three samples had a correct ***Y***-predicted value of over 0.70. For many of the models created that had one sample excluded, more than one component resulted

but the R^2Y value approached one so predictions were made using a single component. Permutation testing (case samples plot shown in Figure 2.37, control samples plot not shown) indicated the model had overfitted the data because the R^2Y intercept value of the regression line was 0.601 for the case class and 0.606 for the control class whilst the Q^2Y intercept values were -0.074 and -0.096, respectively. It cannot be concluded that case samples compared to control samples that are matched according to BMI, age and 'never smoked' status have increased levels of lipids and lactate. The removal of glucose from the analysis did not enhance separation between case and control groups or validity of PLS-DA models.

Table 2.8 'Leave-one-out' cross-validation parameters for the PLS-DA model shown in Figure 2.35.

Sample Excluded	Number of Components	R^2X	R^2Y	Q^2Y	*Y*-Predicted*	
					1	2
1009	1	0.254	0.815	0.565	0.670	0.330
1010	1	0.265	0.847	0.659	0.530	0.470
1017	1	0.257	0.817	0.579	0.691	0.309
1022	1	0.230	0.830	0.552	1.218	-0.218
1027	1	0.294	0.862	0.675	0.287	0.713
2007	1	0.270	0.827	0.410	0.500	0.500[#]
2023	1	0.259	0.861	0.514	0.510	0.490
2033	1	0.265	0.781	0.502	0.095	0.905
2037	1	0.256	0.830	0.625	0.740	0.260
2041	1	0.248	0.807	0.590	-0.040	1.040

*A sample was regarded as belonging to a grade by having a ***Y***-predicted value >0.50. Incorrect classification is represented by red shading and correct by pink, orange or green, corresponding to a ***Y***-predicted value of <0.60, 0.60-0.70 or >0.70, respectively.

[#]***Y***-predicted value >0.50 when more than three decimal places retained.

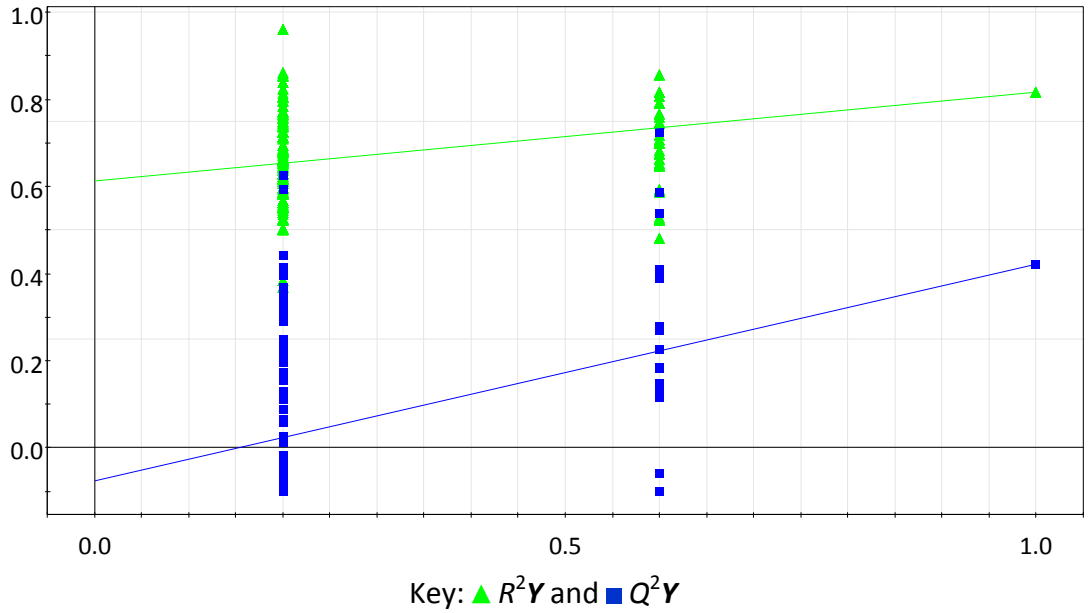


Figure 2.37 Permutation testing plots for the case class in the PLS-DA model shown in Figure 2.35. The R^2Y and Q^2Y intercept values of the regression lines are 0.601 and -0.074, respectively.

2.2 Conclusions

All parameters in the patients' medical records were treated as potential confounding factors but for whole spectra, when the effects of each were analysed singly, none were identified as such for this study. However, when the five case and five control samples that were best matched according to BMI, age and 'never smoked' status, PCA provided tentative separation between case and control samples based on increased levels of lactate and decreased glucose for case samples. This is in agreement with the Warburg effect: in malignant cells glucose uptake increases as does lactate formation.[147] 'Leave-one-out' cross validation supported the validity of the same finding from the PLS-DA model but permutation testing indicated the model had overfitted the data, possibly due to the small sample number. A larger cohort with well defined BMI and age ranges for the same smoking status, *i.e.* 'never smoked', would be needed to verify the observation.

The same models indicate a slight tendency for case samples to be associated with higher lipid levels, which is in accordance with oxidative stress caused by an imbalance between reactive oxygen species and the antioxidant capacity of the cell,

possibly due to the susceptibility of the lipoprotein to undergo oxidation. Overall, this can result in cellular damage leading to conversion to malignant cells.[144] However, greater prominence of case samples having a higher lipid level would be required to evidence this trend.

Clear grouping of case and control samples was not observed using PCA nor was there correlation between breast cancer status and data acquired when PLS-DA was utilised, indicating for this study there is no substantive evidence to suggest an association between breast cancer and a change in metabolism that can be determined by metabolomic analysis of NMR spectroscopy acquired data from plasma. This also applies to tumour grade, whether to all cases or just invasive ductal carcinoma cases. The same conclusions were made following analysis of data from whole spectra and spectra with the glucose region removed.

Despite glucose levels being rapidly affected by diet it is advised to retain the region in initial analysis of plasma. Dietary affects can be subsequently analysed. There is no evidence to suggest removing glucose enhances separation between groups in question and although the level of the metabolite has not been found in this study to determine breast cancer status or grade, the Warburg effect[147] highlights its importance in cancer studies.

Chapter 3. NMR Analysis of Urine

In an analogous manner to that described in Chapter 2, urine from the same patient cohort has been analysed by ^{1}H-NMR spectroscopy with the exception that 1D (one dimensional) Nuclear Overhauser Effect Spectroscopy (NOESY) pulse sequence was used. Chemical shift data was used in analysis.

Samples were prepared as per Section 8.1.1.2 and data collected as detailed in Section 8.2.2. Section 8.3 applies for spectral processing with dark regions listed in Table 8.3. Constant sum normalised data was used initially. MVA was performed as detailed in Section 8.4.

3.1 Results

Ethanol was found to be present in urine samples from the three patients previously identified through analysis of their plasma sample as likely to have consumed alcohol prior to sample donation. These samples were excluded from all analyses; no other samples contained ethanol signals. Additionally, patients' samples whose ethnicity was not 'white British' were excluded because they were not sufficiently represented (Section 2.1.1.1). The urine sample from patient 1002, whose plasma sample contained abnormal lipid levels, was included but the spectrum from sample 1008 was incorrectly saved thus leaving the same number of case and control samples: 38 and 25, respectively.

3.1.1 Initial Analysis of Spectrum and Comparison of Normalisation Methods

A PCA model was built for constant sum normalised data and produced two PCs; the scores plot is presented in Figure 3.1 and the corresponding loadings plot in Figure 3.2. $R^2\boldsymbol{X}$(cum) is 0.460, with PC 1 explaining 0.374 of the variation, whilst $Q^2\boldsymbol{X}$(cum) is 0.390.

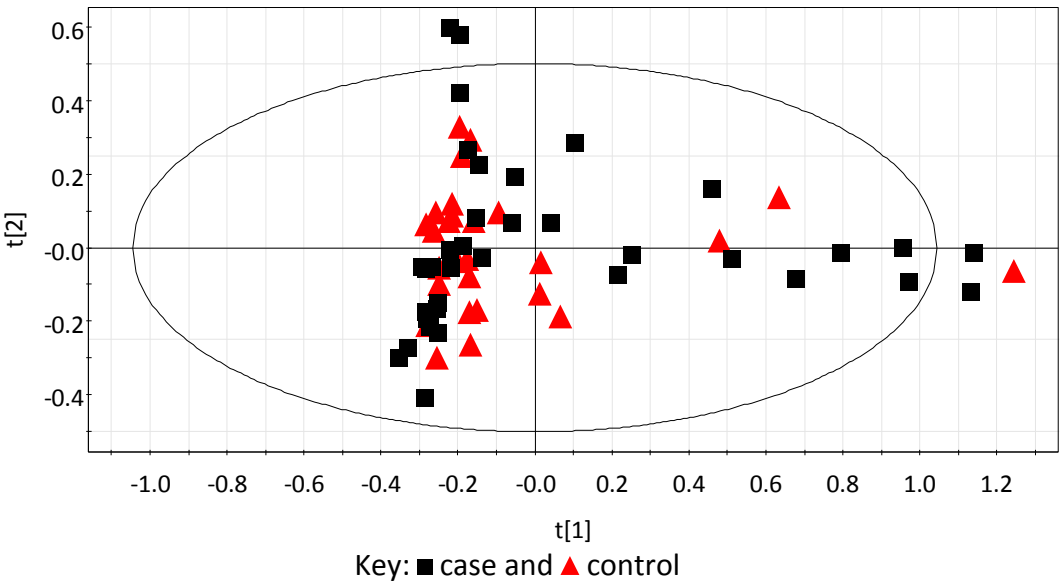


Figure 3.1 PCA scores plot of whole ^{1}H-NMR constant sum normalised urine chemical shift data for case and control samples.

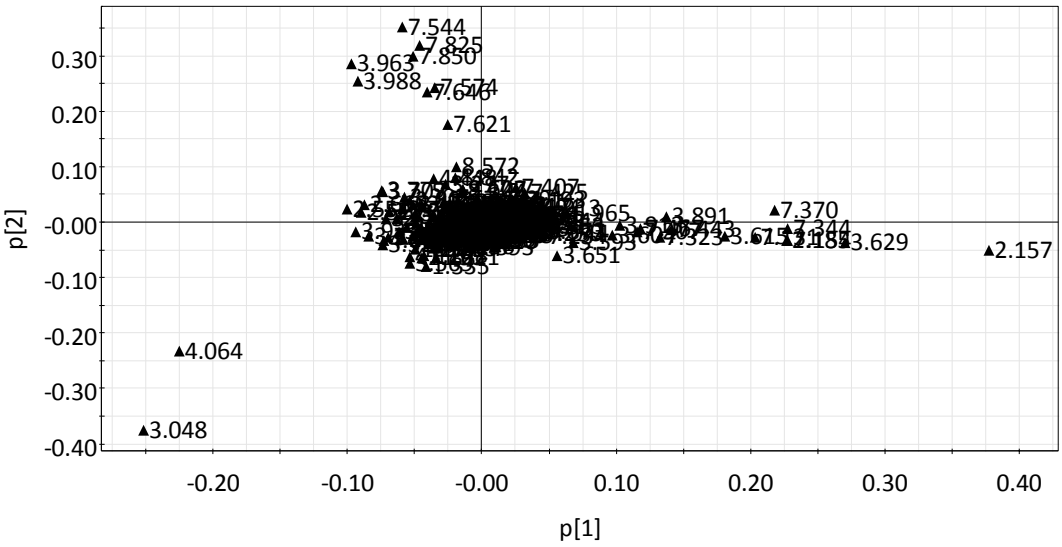


Figure 3.2 PCA loadings plot corresponding to the model displayed in Figure 3.1.

There were 13 samples, three controls and ten cases, which had positive PC 1 scores values that were isolated from the remainder of the samples in scores space. The loadings plot showed a number of bins with positive PC 1 loadings values greater than or equal to 0.10: 2.158, 2.186, 3.604, 3.616, 3.630, 3.892, 3.913, 7.131, 7.154, 7.307, 7.324, 7.344, 7.371, 7.446 and 7.467 ppm. The origin of these signals could not initially be identified through metabolite lists or databases[66,103,142,157-163] so the aforementioned 13 patients' last food and drink consumption details prior to sample donation, as well as drug and supplement history, were analysed for

recurrent substances. Additionally, records of other diseases and ailments were studied for frequent instances amongst patients; hypertension was the most diagnosed disease with eight patients being affected but seven of the other 50 patients were also afflicted. No food or drink item was common that might cause such distinct signals. Irrespective of dosage or strength, 70 different drugs or supplements were recorded for the 13 patients but no substances, including common 'over-the-counter' medication such as acetaminophen (paracetamol) and aspirin, were taken by more than three patients. Apart from a doublet at 7.32 ppm, signals in the spectrum did not correspond with acetaminophen signals,[103] however other sources suggested signals around 2.16 ppm were resultant from acetaminophen.[164,165] Signals from acetaminophen glucuronide, acetaminophen sulphate and *N*-acetylcysteine conjugate of acetaminophen were correlated with acetaminophen signals[165] and matched all of the unknown signals in the previously stated bin regions. It was concluded that acetaminophen and associated breakdown products were responsible for the unassigned signals in the spectra and patients had consumed acetaminophen on an *ad hoc* basis that was not reported. It was not possible to generate a PLS-DA model.

Whilst investigation proceeded to determine the origin of influential signals in PC 1, the impact of these signals on the normalisation procedure was considered. Although constant sum normalisation attempts to account for dilution differences between samples it assumes that changes in concentrations of single metabolites have a small influence on the total concentration change therefore a linear concentration series of samples would result in a linear series of integrals. This is in addition to the assumption that the influence of down-regulation of certain metabolites should be approximately balanced by the up regulation of other metabolites.[69] However, this approximation can fail if there is a large change in the concentration of a metabolite(s), in this instance those associated with acetaminophen. This is because other signals in the spectrum will appear to decrease due to lower normalised intensities as a result of the total spectrum integral being greater.[59] Dieterle *et al*.[69] proposed probabilistic quotient normalisation (PQN) as a solution. This method assumes that concentration

changes of a single metabolite only influences parts of the spectrum hence the intensity of the majority of signals, rather than all, is a function of dilution. A most probable quotient between the signals of each spectrum in the series and of a reference spectrum is calculated as the normalisation factor, effectively this is a number that is required to make the most number of data point or bin intensities the same as the reference spectrum. A full description of the method to obtain probabilistic quotient (PQ) normalised data is provided in Section 8.3.1.2 but in brief, perform constant sum normalisation to ensure the absolute magnitude is constant, calculate a reference spectrum, calculate the quotients of all variables of the test spectrum with those of the reference spectrum, calculate the median of these quotients, excluding any noise regions (Table 8.7 lists regions included in the calculation of the median quotient), and divide all variables of the test spectrum by this median.[69]

The samples with acetaminophen signals had the lowest median quotients; the eight samples with the highest positive PC 1 scores values had the lowest values, the range being 0.627-0.767. Median quotients less than 1 would be expected because given there are more signals present the relative intensity for each signal would be less using constant sum normalisation. Therefore, dividing all variables of the test spectrum by the associated median quotient would increase the intensity of signals, the majority to a similar level to those of the calculated median spectrum. As expected, the scores plot (Figure 3.3; loadings plot not shown) of the model created using PQ normalised data (two PCs, $R^2\boldsymbol{X}(\text{cum}) = 0.455$ and $Q^2\boldsymbol{X}(\text{cum}) = 0.362$) compared to constant sum normalised data exhibited higher PC 1 scores values for those samples containing acetaminophen. However, the relative positions of the samples remained very similar and separation between case and control samples was not apparent. It was not possible to generate a PLS-DA model.

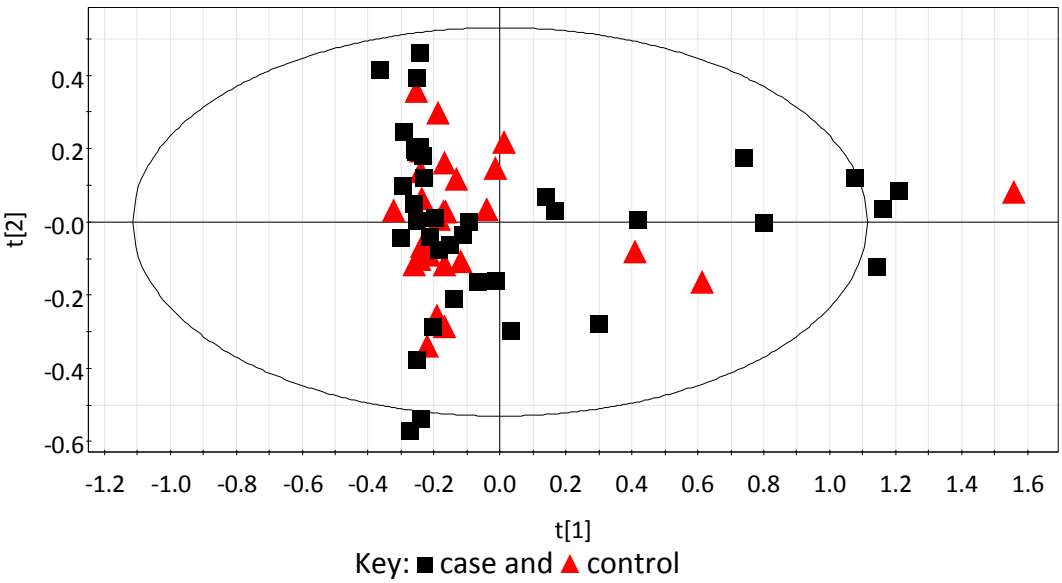


Figure 3.3 PCA scores plot of PQ normalised urine data for case and control samples.

3.1.2 Analysis of Spectrum Excluding Acetaminophen Regions Using Probabilistic Quotient Normalisation

The aforementioned bin regions of acetaminophen and associated breakdown products were excluded and the data re-normalised. The median quotients now ranged from 0.624 to 1.199. Separation was not present between case and control samples in PCA scores space (Figure 3.4; two PC model, $R^2\boldsymbol{X}$(cum) = 0.303 and $Q^2\boldsymbol{X}$(cum) = 0.186). Creatinine (bin regions 3.034-3.064 and 4.043-4.073 ppm) dominated PC 1 in the loadings plot whilst PC 2 was most heavily influenced by hippurate (bin regions 3.954-3.984, 7.530-7.590, 7.611-7.662 and 7.815-7.866 ppm) (Figure 3.5).

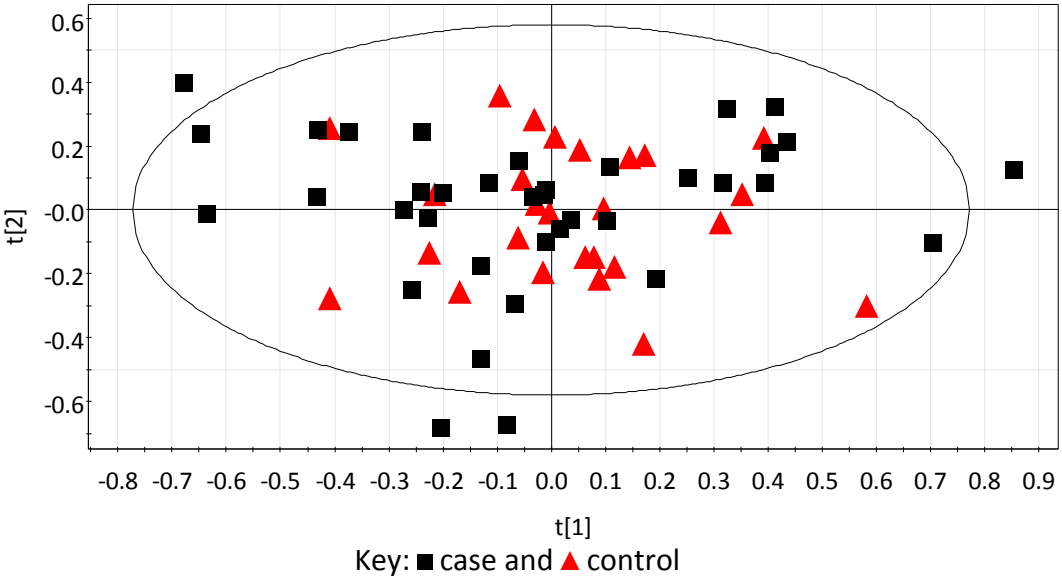


Figure 3.4 PCA scores plot of PQ normalised urine data excluding acetaminophen regions for case and control samples.

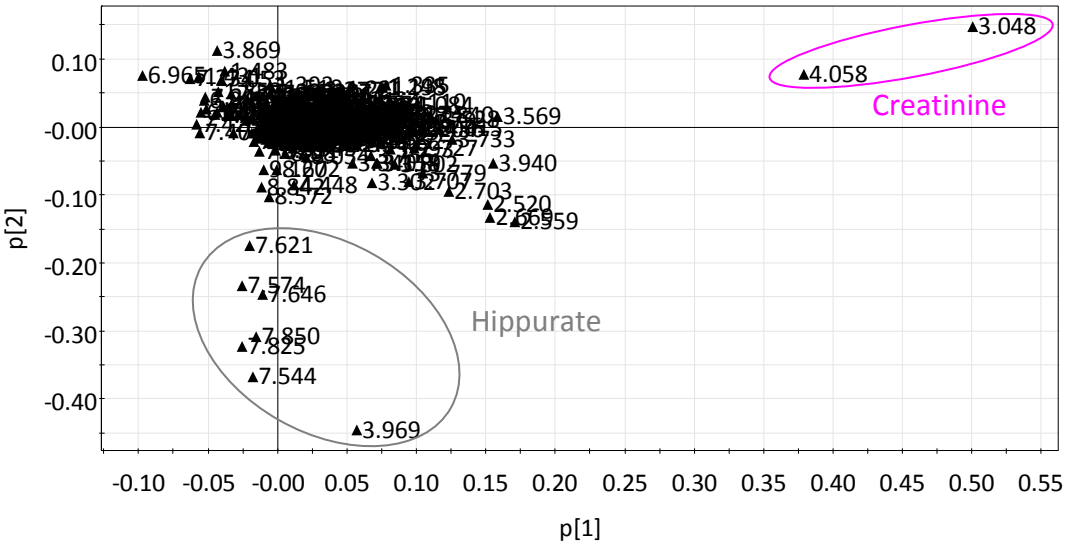


Figure 3.5 PCA loadings plot corresponding to the model displayed in Figure 3.4.

Given there was no clear separation between case and control samples subsequent investigation focused on breast cancer specific parameters (Table 3.1).

Table 3.1 Parameters used for classification in PCA and model descriptors excluding acetaminophen regions.

Parameter (Classification)	Sample Numbers (Total)	PCs	R^2X (cum)	Q^2X (cum)
Grade (0, 1, 2, 3, 1 and 2 or 2, 2, and 2)	7, 5, 12, 11, 1, 1 (37)	2	0.356	0.176
Grade (0, 1, 2 or 3)	7, 5, 12, 11 (35)	2	0.360	0.179
Grade (1, 2 or 3)	5, 12, 11 (28)	2	0.361	0.122
Ductal and grade (1, 2 or 3)	4, 10, 7 (21)	1	0.257	0.129
ER score (0/8 - 8/8)	11=0/8, 2=3/8, 1=6/8, 21=8/8 (35)	2	0.362	0.184
PR score (0/8 - 8/8)	13=0/8, 1=2/8, 1=3/8, 1=4/8, 1=5/8, 1=6/8, 4=7/8, 9=8/8 (31)	2	0.375	0.181
HER2 status (-ve or +ve)	27, 6 (33)	1	0.230	0.135
Combined ER/PR/HER2 class (1 - 8)	13=2, 2=4, 2=7, 6=8 (23)	1	0.220	0.058

With the exception of the model that included 37 samples, all models contained samples that were from patients with a single tumour occurrence; one patient exhibited two, and another three, separate occurrences.

In the clinical environment both ER and PR status can be represented by a score out of eight and reflects confidence of assignment: 8/8 is undoubtedly positive status whereas 0/8 is undoubtedly negative status and 1/8 is not possible[18,19]. Conventional practice of assigning ER and PR scores of 3/8 to 5/8 as positive status[19] has not been adopted (Table 3.2). This was to provide greater disparity between the two statuses whilst only eliminating two and three samples based on ER and PR scores, respectively. Only positive and negative status samples were included in the assignment of combined ER/PR/HER2 values (Table 3.3).

Table 3.2 ER and PR status used in this study based on respective scores.

ER or PR Score	0/8	2/8	3/8	4/8	5/8	6/8	7/8	8/8
Status Assigned	(-)	(-)	None	None	None	(+)	(+)	(+)

(-) = negative, (+) = positive

Table 3.3 Arbitrary classes assigned to ER/PR/HER2 statuses.

Class	1	2	3	4	5	6	7	8
Status of Descriptors	ER+ PR+ HER2+	ER- PR+ HER2+	ER+ PR- HER2+	ER+ PR- HER2-	ER- PR+ HER2-	ER+ PR+ HER2-	ER- PR- HER2+	ER- PR- HER2-
Subtype Indication	Luminal B	Luminal B	Luminal B	Luminal A	Luminal A	Luminal A	Basal like	Basal like

Clear separation in the PCA scores plots was not visible based on any of the eight parameters (data not shown). Only the scores plot classed according to grade of single occurrence tumours is shown (Figure 3.6). PCA was not performed solely for the samples that were assigned as ER+ or ER- because information could be inferred from the model created based on ER scores; the same applied to PR status. From the scores plot for combined ER/PR/HER2 values, evaluation based on tumour subtypes was conducted by combining classes 7 and 8 to form the basal like group: no separation was present based on luminal A, luminal B and basal like tumours.

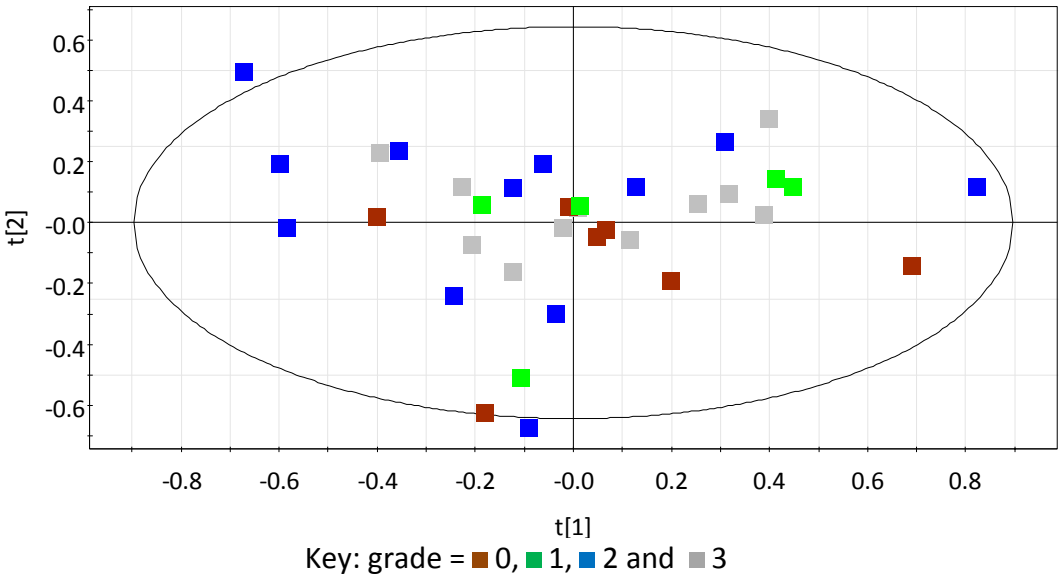


Figure 3.6 PCA scores plot of PQ normalised urine data excluding acetaminophen regions for single occurrence tumour grade. Coloured according to grade.

PLS-DA was performed based on all of the eight parameters as well as for ER+/- and PR+/- statuses but a model was only created based on ER/PR/HER2 class. The one PC model was poor ($R^2\boldsymbol{X}$ = 0.202, $R^2\boldsymbol{Y}$ = 0.202 and $Q^2\boldsymbol{Y}$ = 0.050) and no separation

was observed between classes (scores plot shown in Figure 3.7, loadings plot not shown). Of note though, in both PCA and PLS-DA scores plots, samples assigned as class 4 (ER+, PR- and HER2-; luminal A) had the greatest negative PC 1 scores values. Samples belonging to class 2 (luminal B) were also hormone receptor positive therefore it could be hypothesised that samples in classes 2 and 4 would be positioned in similar scores space but this is not observed. Additionally, there were only two samples in class 4.

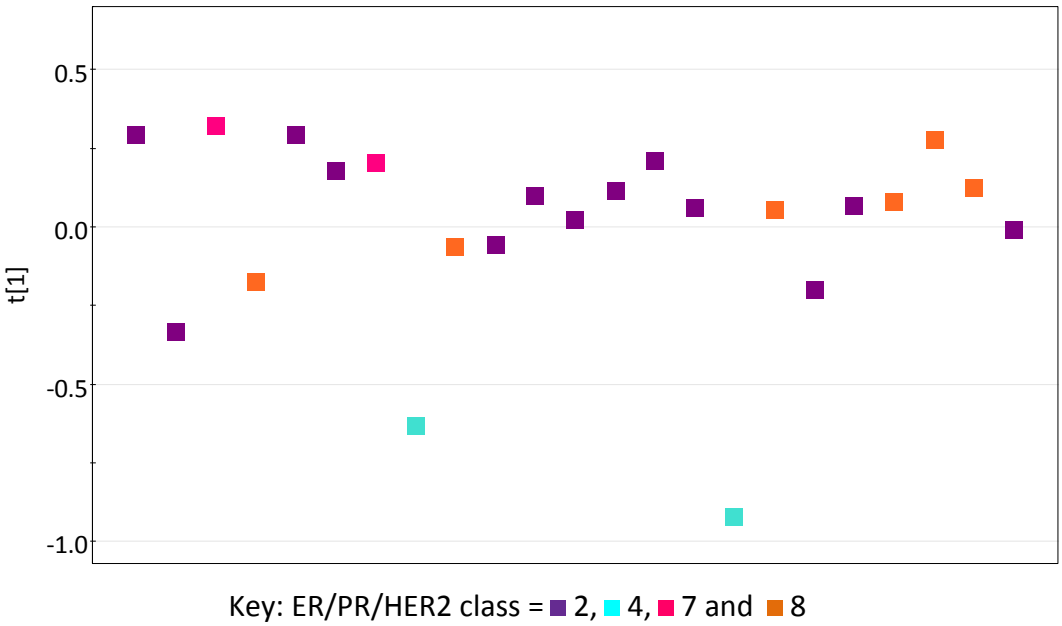


Figure 3.7 PLS-DA scores plot of PQ normalised urine data excluding acetaminophen regions for ER/PR/HER2 class. Descriptors related to classes shown in Table 3.3.

3.1.3 Analysis of Spectrum Excluding Acetaminophen, Creatinine and Hippurate Regions

Further analysis ensued whereby creatinine (3.034-3.064 and 4.043-4.073 ppm) and hippurate signals (3.954-3.984, 7.530-7.590, 7.611-7.662 and 7.815-7.867 ppm) were also removed and the remaining data re-normalised. The median quotient range was 0.775-1.249.

For the eight aforementioned parameters, PCA models could not be generated. Only single occurrence case samples graded as 0, 1, 2 or 3 were modelled by PLS-DA

(one PC; $R^2\boldsymbol{X}$ = 0.155, $R^2\boldsymbol{Y}$ = 0.142 and $Q^2\boldsymbol{Y}$ = 0.017) and separation of grades was not observed. PLS-DA models were not able to be generated based on ER+/- or PR+/- status.

3.1.4 Analysis of Spectrum Excluding Acetaminophen Samples

Due to the intensity of the acetaminophen signals, it was possible that underlying peaks were obscured that could influence separation between sample classes. To determine whether this scenario was applicable, samples that contained acetaminophen signals were excluded. Creatinine and hippurate regions were not excluded. The remaining 50 samples were re-normalised resulting in median quotients within the range 0.613-1.210. A summary of the PCA models is shown in Table 3.4.

Table 3.4 Parameters used for classification in PCA and model descriptors excluding acetaminophen samples.

Parameter (class)	Samples in class (total)	PCs	R^2X	Q^2X
State (case or control)	23, 27 (50)	1	0.147	0.034
Grade (0, 1, 2, 3 or 1 and 2)	7, 3, 7, 9, 1 (27)	1	0.196	0.041
Grade (0, 1, 2 or 3)	7, 3, 7, 9 (26)	1	0.200	0.057
Grade (1, 2 or 3)	3, 7, 9 (19)	1	0.224	0.051
Ductal and grade (1, 2 or 3)	3, 6, 5 (14)	1	0.253	0.041
ER status (0/8 - 8/8)	10=0/8, 1=3/8, 1=6/8, 14=8/8 (26)	1	0.199	0.051
PR status (0/8 - 8/8)	9=0/8, 1=2/8, 1=3/8, 1=4/8, 1=5/8, 1=6/8, 3=7/8, 5=8/8 (22)	1	0.206	0.053
PR status (-ve or +ve)	10, 9 (19)	0	/	/
HER2 status (-ve or +ve)	18, 5 (23)	1	0.205	0.042
Combined ER, PR and HER2 status (1 - 8)	8=2, 2=7, 5=8 (15)	0	/	/

All models generated were poor and no separation was observed irrespective of the parameter tested. PLS-DA was performed for all parameters but models did not result.

3.1.5 Analysis of Spectrum with Potential Confounding Factors Minimised

BMI, age and smoking are well known confounding factors[151] and analysis of a small plasma data set, where variation amongst the three factors had been minimised, resulted in partial separation between case and control samples (Section 2.1.1.3). As per Chapter 2, ten case and ten control samples were matched singly according to BMI or age whilst five samples of each type were matched according to BMI, age and 'never smoked' status. With the inclusion of sample 1002 and exclusion of sample 1008, figures related to BMI and age changed (Table 3.5). The demographics of the five best matched control samples according to BMI, age and never smoked status were unchanged because sample 1008 was not one of samples included and sample 1002 was sourced from an ex-smoker.

Table 3.5 Demographics of samples included in plasma and urine data sets.

Samples	Biofluid	Sample Exclusive to Biofluid	Age		BMI (kg m^{-2})	
			Average	Median	Average	Median
All	Plasma	1008	66.0	63.0	29.0	27.3
	Urine	1002	65.8	63.0	28.9	27.3
Best age matched	Plasma	1008	63.4	63.5	/	/
	Urine	1001	63.6	63.5	/	/
Best BMI matched	Plasma	1026	/	/	27.2	26.8
	Urine	1002	/	/	27.0	26.8

PCA and PLS-DA were applied to the data sets. Values associated with the models are shown in Table 3.6.

Separation between case and control samples was not visible when variation in age or BMI, age and 'never smoked' status was minimised. For BMI matched samples there is a slight tendency for case samples to have lower $\boldsymbol{t}$[1] values (Figure 3.8)

Table 3.6 Model descriptors for analysis using samples that were best matched according to a parameter(s).

Parameter(s) Matched	PCA			PLS-DA			
	PCs	R^2X	Q^2X	PCs	R^2X	R^2Y	Q^2Y
BMI	1	0.234	0.059	1	0.221	0.436	0.126
Age	1	0.246	0.103	0	/	/	/
BMI, age and 'never smoked'	1	0.314	0.044	0	/	/	/

Units: BMI = kg m^{-2}.

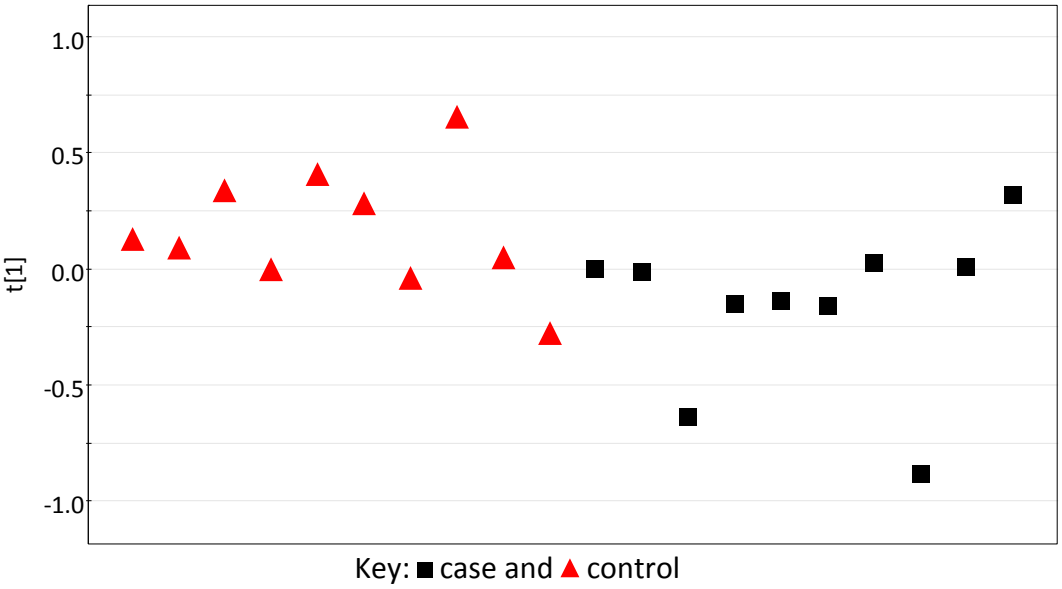


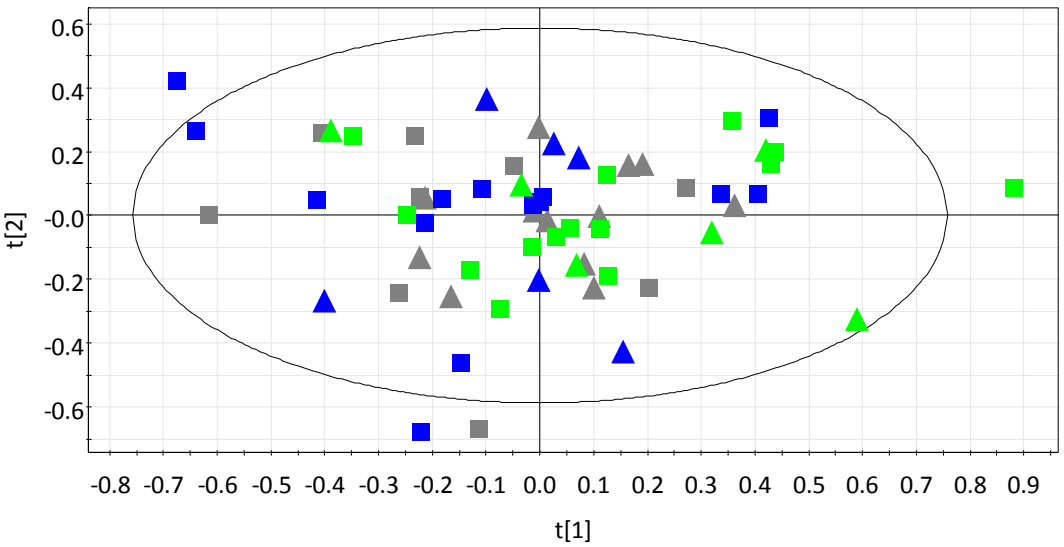
Figure 3.8 PCA scores plot of PQ normalised urine data for case and control samples matched according to BMI.

Creatinine, which is positively associated with BMI,[166] strongly dominated positive PC 1 loadings space (figure not shown) despite these samples being BMI matched with a range of 24.4-30.0 kg m^{-2}. This data set indicates creatinine could have a greater influence than that associated with BMI alone, however, the association between BMI and creatinine for all samples warrants further investigation.

Models were created based on two classifications of patients' BMI; data was available for 61 samples. The first classification was determined as follows: samples with the lowest third of BMI values were designated as class 1, class 2 for the middle third and the highest third was assigned class 3 status. The scores plot of the two PC model ($R^2\boldsymbol{X}$(cum) = 0.297 and $Q^2\boldsymbol{X}$(cum) = 0.156) is shown in Figure 3.9; the loadings

plot (not shown) is extremely similar to Figure 3.5. The other form of classification was directly dependent on BMI values being indicative of weight status: classes 1, 2, 3 and 4 corresponded to <18.5, 18.5-24.9, 25-29.9 and ≥30.0 kg m^{-2}, suggestive of underweight, healthy weight, overweight and obese patients, respectively (data not shown).[149] The BMI ranges when the data was split into three classes were similar to those for weight status: 1 = <24.9, 2 = 25.0-28.9 and 3 = ≥29.0 kg m^{-2}.

For both, no correlation was observed between BMI and creatinine levels. The seven samples with the greatest positive PC 1 scores values in Figure 3.9 did not belong to the class that contained the greatest BMI values. The same applied to ten samples when BMI was modelled as being an indicator of weight status.



Key: shape - □ case and △ control; colour - BMI class = ● lowest, ● middle and ● highest

Figure 3.9 PCA scores plot of PQ normalised urine data coloured according to three classes of BMI (kg m^{-2}) value (one-third of samples in each class).

Creatinine positively correlates with muscle mass[167,168] to a greater extent than with BMI[166] and because the latter is not a measure of muscle mass this could explain why correlation is not shown between creatinine and BMI.

Analysis was previously performed excluding samples that contained acetaminophen signals (Section 3.1.2) and sample 2023, which was included in the

best age matched samples, had these signals present. Given there was no separation between the ten case and ten control samples analysis using a replacement for this sample was not performed. Additionally, the age range would increase if 2023 (63 years) was substituted by the sample from the patient with the next closest age to the median (Table 3.7). A similar argument applies to the best BMI matched samples. Samples 2017, 2023 and 2026 would need to be replaced and there was not clear separation between the 10 control and seven other case samples. 2023 was also one of the five best BMI, age and 'never smoked' samples and, again, a model was not created using a replacement.

Table 3.7 Current and potential range of parameters with acetaminophen containing samples included (current range) or excluded (potential range).

Parameter(s) Matched	Samples to be Replaced	Current Range	Potential Range
Age	2023	60-64	60-67
BMI	2017, 2023, 2026	25.8-28.7	25.5-28.9
BMI, age and 'never smoked'	2023	24.7-29.1 and 56-68	23.9-29.1 and 56-78 or 24.7-32.4 and 56-68

Units: BMI = kg m^{-2}. For best matched BMI, age and 'never smoked' samples, all samples were from patients who had never smoked and the first range value refers to BMI and the second age; the potential range would depend on whether greater emphasis was placed on BMI or age.

3.2 Conclusions

Non-prescribed medication can exert a strong influence on metabolomics investigations as shown by the presence of signals originating from acetaminophen or acetaminophen breakdown products.

Exclusion of acetaminophen regions resulted in creatinine and hippurate dominating loadings space but discrimination based on case and control samples and numerous descriptors of tumours was not possible. Additional exclusion of

these two metabolites generally made building of models unfeasible and removal of samples that contained acetaminophen signals did not provide further information regarding separation of the remaining samples. Analysis of select samples matched according to age, BMI or BMI, age and ‘never smoked’ status did not differentiate between case and control samples.

Analysis of urine data did not provide discrimination between samples based on various descriptors for this sample set.

Chapter 4. SHY Analysis of Plasma and Urine Data

SHY is an extension of STOCSY and allows latent variables to be identified between data acquired using different platforms, such as NMR spectroscopy and MS, or between different biofluids and/or tissues.[96] Covariance between signal intensities in the same or related molecules is analysed.[95] If different technologies were used and a known signal from NMR spectroscopy data was shown to be correlated to an unknown ion from MS, identification of that ion could be possible. The reverse could also apply if the origin of the signal in NMR spectra was uncertain due to low intensity or it being poorly resolved.[169] Different metabolites that are not common to both of the data acquisition methods but are linked *via* metabolic pathways could also be highlighted. By using different sample types the same principles would apply. Additionally, if tissue was used as one of the sample types or one of the biofluids was more difficult to obtain than the other and covariances were identified, the need to collect the more invasive biofluid or tissue would be reduced.[96,170] This would have positive implications for patients.

4.1 Data Acquisition Methodology

Plasma (Chapter 2) and urine (Chapter 3) NMR spectroscopy data from the same patient set were analysed by SHY however full resolution data was used.[96] For both biofluids PQN was applied (Section 8.3.1.2) and Table 8.7 lists regions included in the calculation of the median quotient. Case and control samples were analysed separately to allow any potential differences in correlations to be visualised. Correlation coefficients were calculated between data point intensities of the two biofluids as detailed in Section 8.6. The script used in MatLab, version 7.12.0.635 (R2011a) (The MathWorks, Inc., Natick, Massachusetts, USA) to generate correlation coefficients was written in-house by C. McRae and modified by T. Bansagi. Pearson's[95,96,170,171] and Spearman's correlation coefficients[96] have been used in previous studies. The magnitude of correlation coefficients is indicated by the colourmap of plots: darkest red shows the strongest positive correlation and

darkest blue the strongest negative correlation. Numerically +1 indicates perfect correlation and -1 perfect anti-correlation. Correlation coefficients with *p*-value >0.001 were considered to be spurious[95,96,171] and set to zero, appearing as white in plots. The stringent *p*-value limit was implemented to reduce spurious correlations and is equivalent to a Bonferroni correction for 50 independent tests.[96] Other studies have not considered a correction factor implementing $p < 0.05$ as the cut-off value.[170] The same study argued that only correlations between the same signals from the same metabolites in different biofluids could be investigated but this eliminates many benefits of performing SHY. Even using a confidence level of 99.9% spurious correlations will result through random chance due to thousands of points present.[171] However, by using full resolution data the two-dimensional space occupied in the plot by each potential correlation is smaller and it is unlikely that multiple spurious points would result at regions of signal intensity in both biofluids. Additionally, it would be possible to observe potential correlation that involved a metabolite whose signal overlapped with that of a non-correlated metabolite, which could be obscured using binned data. Larger areas of correlation would be investigated to confirm whether the correlation was non-spurious by distinguishing points that are clustered or arranged in patterns that coincide with the spectra.[95]

Data points were used in analysis with conversion to chemical shift values implemented for clarity in this chapter including labelling of SHY plots. Processing of full resolution data is intensive in terms of computing power: a personal computer (Evesham Technology, Evesham, Worcestershire, UK; processor speed = 3.0 GHz, random access memory (RAM) = 2032 MB, processor type = Intel(R) Pentium(R) 4 central processing unit (CPU)) could only process data that resulted in approximately a 6,000 x 6,000 matrix and the highest specification personal computer (Dell Inc., Round Rock, Texas, USA; processor speed = 2.93 GHz, RAM = 16 GB, processor type = Intel(R) Core(TM) i7 CPU) accessible for which these analyses could be performed was unable to produce a square 37,815 matrix that included *p*-values. Removal of the 4.500-6.200 ppm region that had been set to zero because it contained water and additionally urea for urine, reduced each spectrum

to 30,852 data points but generation of a SHY plot was still not possible. Removing regions that had a value of zero for all samples resulted in 9,060 and 27,450 data points for plasma and urine, respectively, and allowed plot generation. PQN was implemented for both data sets.[96]

Due to the large number of data points in plots, subsequently points have been approximated to the nearest 25; this number of points equates to approximately 0.006 ppm. Within the text a capital letter and in some instances additionally a number, in parentheses refers to an area highlighted in Figures.

4.2 Results

4.2.1 Pearson's Analysis

Figure 4.1 shows the SHY plot for case samples obtained using full digitised data excluding regions that had an integral value of zero for all samples with Figure 4.2 the equivalent for control samples. Differences in covariance can be identified between case and control samples. For example, there is strong correlation for cases in areas A and B whereas for controls this is not apparent. The reverse applies to areas C and D. To assess the validity of correlation, determination of the cause(s) is required.

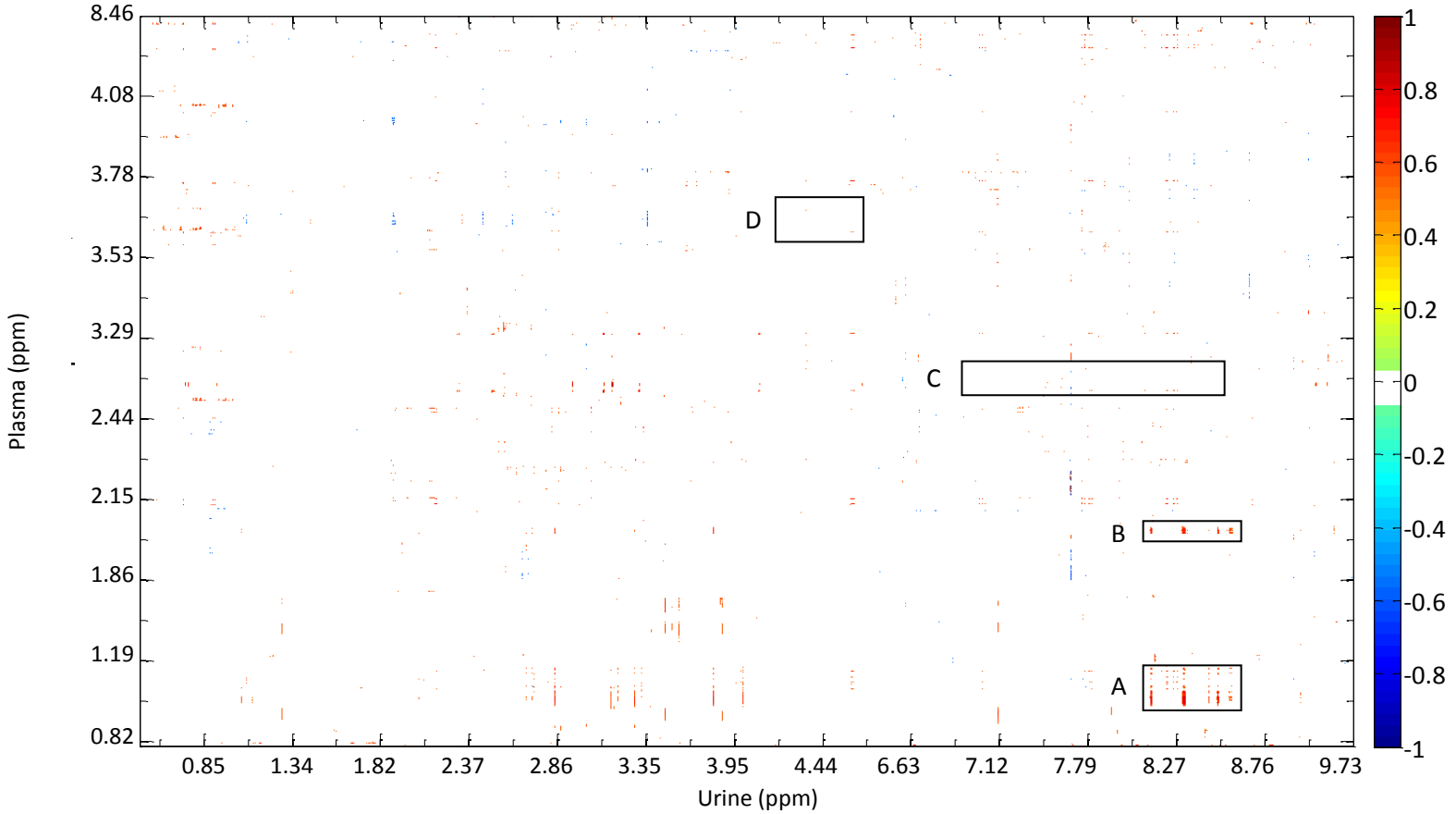


Figure 4.1 Output from Pearson's SHY analysis of urine and plasma data from case samples. The colour bar indicates the Pearson correlation coefficients. Boxed areas, with labels, highlight areas referred to in the text.

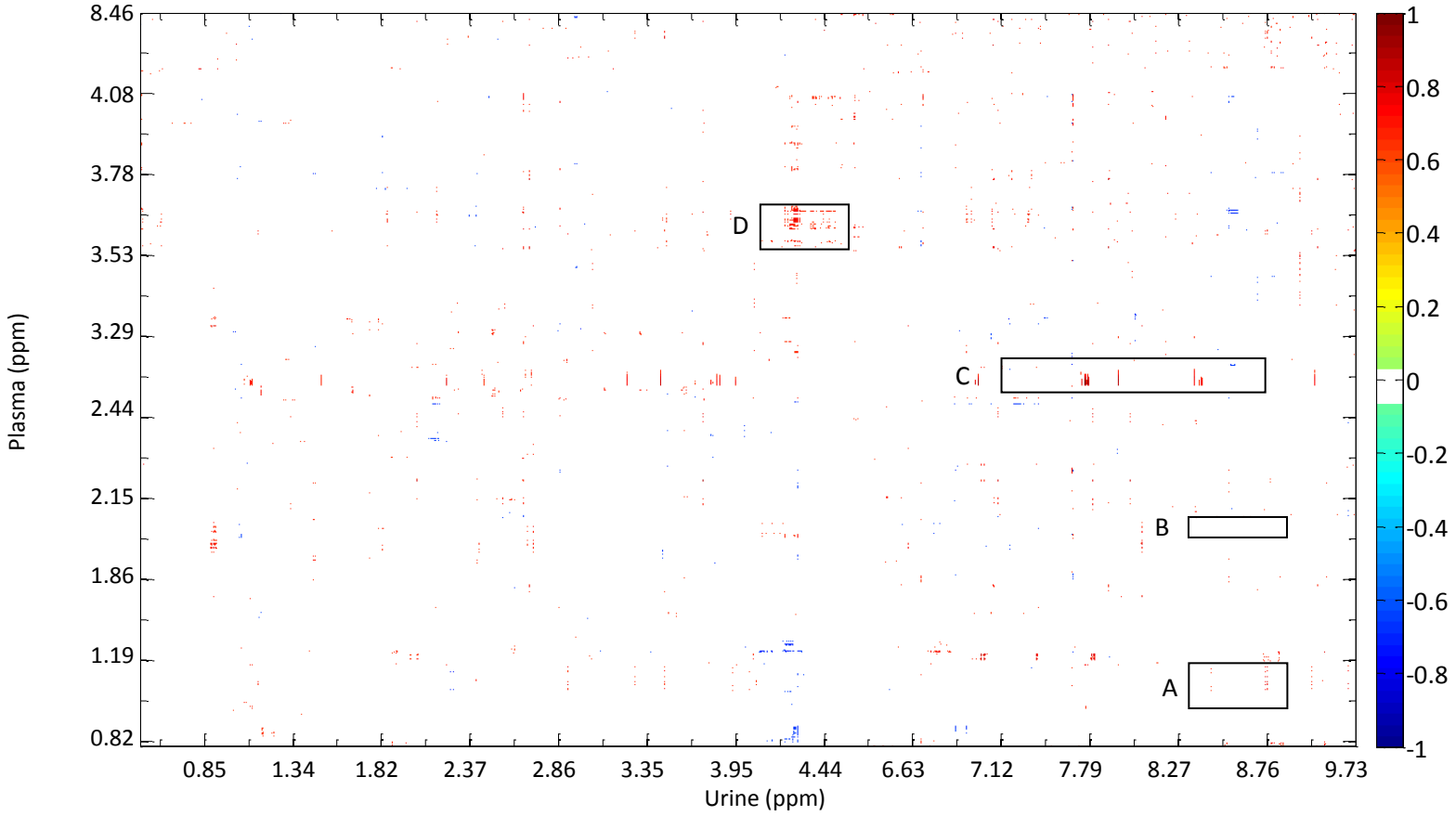


Figure 4.2 Output from Pearson's SHY analysis of urine and plasma data from control samples. The colour bar indicates the Pearson correlation coefficients. Boxed areas, with labels, highlight areas referred to in the text.

Areas of potential true correlations for case samples were investigated. Figure 4.3 is an expansion of Figure 4.1 that includes areas A and B. The Figure shows strongest correlation of plasma occurs at 0.956 ppm (P1). The correlated urine ppm values, with ranges in parenthesis, are 8.128 ppm (8.123-8.133 ppm; U1), 8.311 ppm (8.299-8.323 ppm; U2), 8.448 ppm (8.445-8.450 ppm; U3), 8.500 ppm (8.494-8.506 ppm; U4) and 8.567 ppm (8.555-8.579 ppm; U5). In plasma, strong correlation centred at 2.058 ppm (P3) is visible with the aforementioned urine areas (U1-U5) though the correlation area at 8.448 ppm (U3) is very small. The plasma correlation region centred at 2.058 ppm (P3) extends from 2.046 ppm to 2.070 ppm whereas for 0.956 ppm the continuous region is between 0.932 ppm and 0.981 ppm (P1) and for the non-continuous region the range is 0.932-1.042 ppm (P1+P2). The majority of an upfield signal of a valine doublet and likely a leucine triplet, possibly with an underlying isoleucine triplet, are contained between 0.932 ppm and 0.981 ppm (P1).[103] Extending the range to 1.042 ppm (P1+P2) incorporates the whole of the valine doublet and the upfield half of another valine doublet as well as an isoleucine doublet. The correlation region 2.046 ppm to 2.070 ppm (P3) consists of indistinct signals.

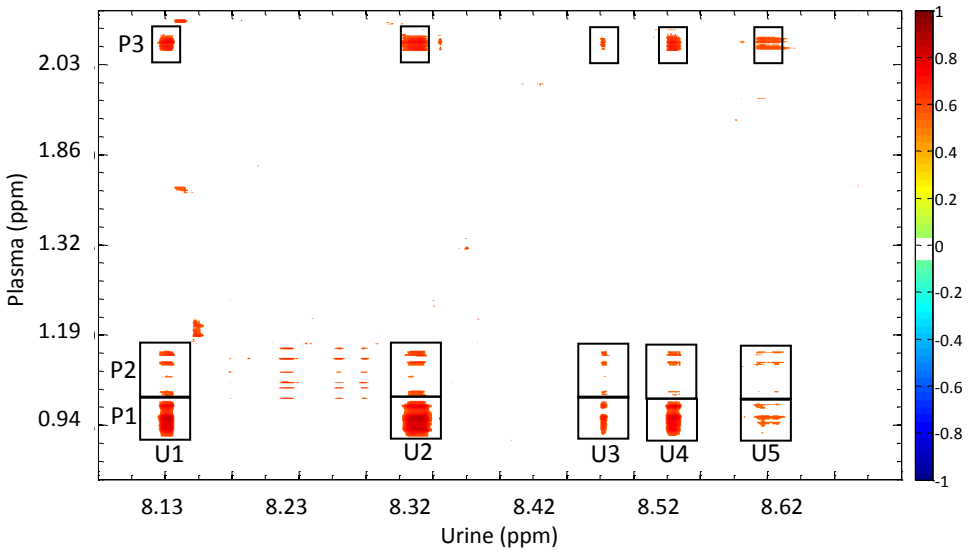


Figure 4.3 Expansion of output from Pearson's SHY analysis of urine and plasma data from case samples showing the urine 8.079-8.665 ppm and plasma 0.871-2.100 ppm area. The colour bar indicates the Pearson correlation coefficients. Boxed areas, with labels, highlight areas of correlation referred to in the text.

Sample 2039 contains a broad singlet in each of the aforementioned urine ranges whereas other samples only display noise. The exception is 8.555-8.579 ppm, which contains the upfield half of a very broad signal visible in most samples; this range has the least strong correlations. To determine whether the correlations are due to sample 2039 alone the urine and plasma signal intensities of all samples were plotted for the data point that has approximately the strongest correlation in each region. The plot related to 0.956 ppm and 8.311 ppm in plasma and urine spectra, respectively, is shown in Figure 4.4. One sample, 2039, is vastly different to all others and would heavily influence the correlation coefficients and *p*-values. The same applies to the equivalent plot (data not shown) using the same plasma value but 8.567 ppm for urine. Replacing 0.956 ppm with 2.058 ppm for plasma also resulted in a similar plot (data not shown) to that shown in Figure 4.4. The unique signals to sample 2039 were not able to be identified.

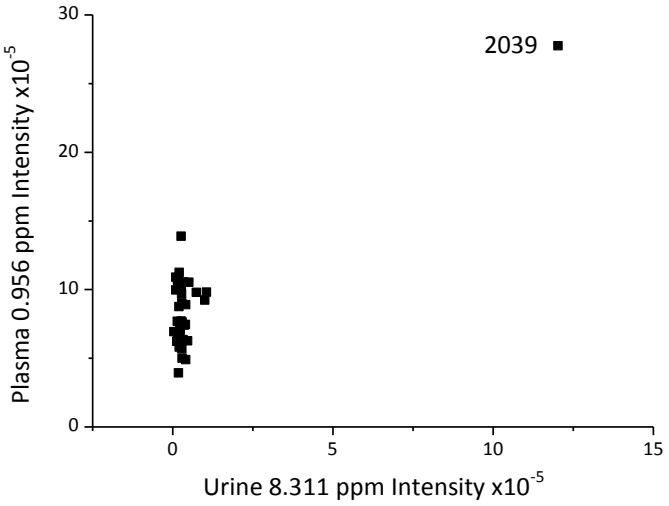


Figure 4.4 Scatter plot for case samples showing intensities of data points with maximum correlation within urine 8.299-8.323 ppm and plasma 0.932-0.981 ppm ranges.

A large rectangular area of positive correlation is present for controls (area D in Figure 4.2) between 4.222-4.497 ppm in urine spectra (U1-U9) and 3.555-3.677 ppm in plasma spectra (P1-P3) (Figure 4.5). The strongest correlation value is at 4.277 ppm (U2 centre) and 3.641 ppm (P3 centre) for urine and plasma, respectively. For plasma the area extends to the edge of non-overlapped glycerol

signals:[103,142] a complete doublet of doublets centred at 3.65 ppm is contained in P3 and two downfield peaks of another are located in P1, the other two peaks overlap with a glycine signal (Figure 4.6). Additionally, a threonine doublet[103,142] is present that spans 3.58-3.61 ppm but this is approximately the area in Figure 4.5 that correlations are not present (P2). Glycerol is a component of triglycerides and phospholipids.[103]

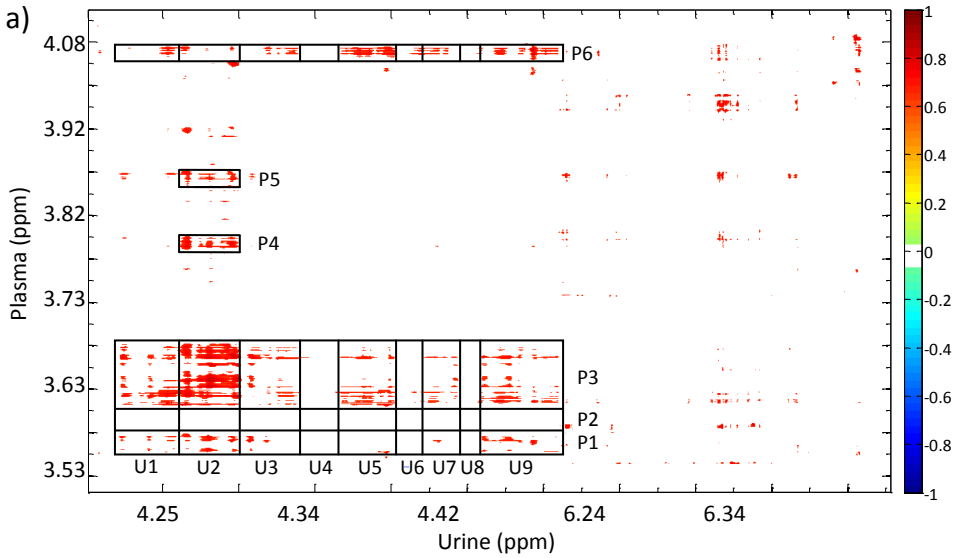


Figure 4.5 Expansion of output from Pearson's SHY analysis of urine and plasma data from control samples showing the urine 4.198-6.411 ppm and plasma 3.519-4.113 ppm area. The colour bar indicates the Pearson correlation coefficients. Boxed areas, with labels, highlight areas of correlation referred to in the text.

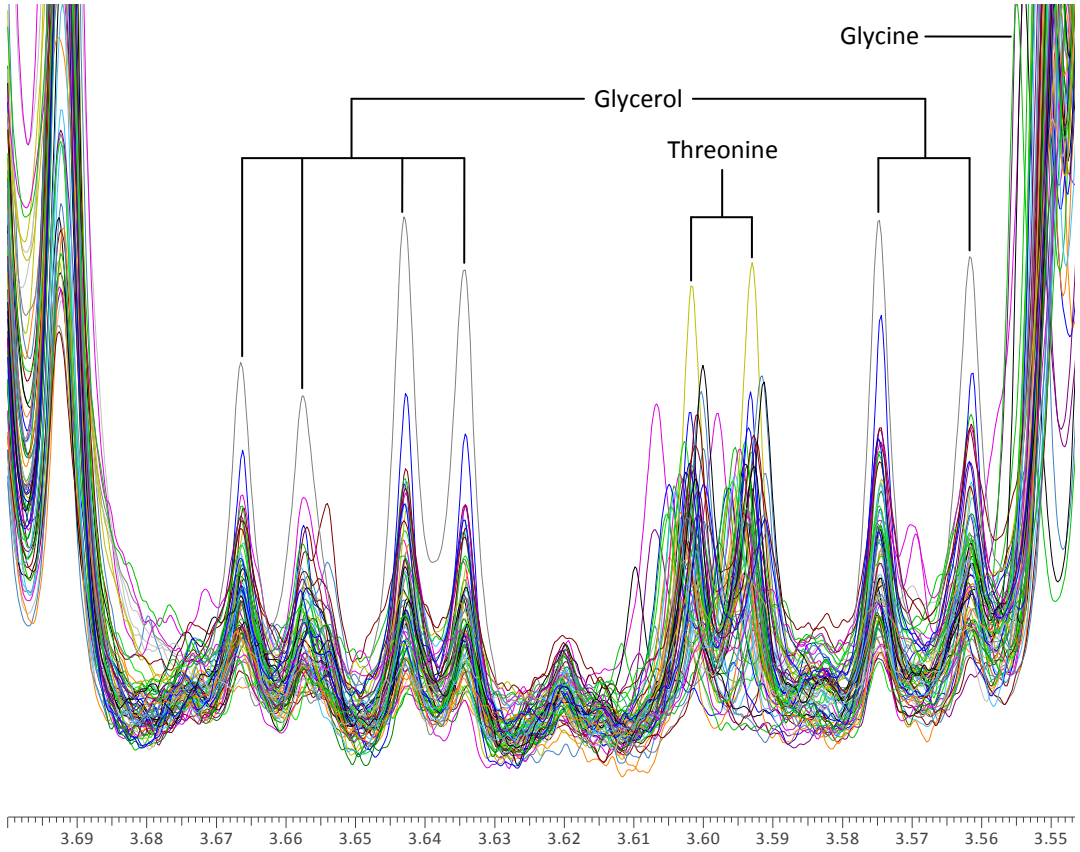


Figure 4.6 Expansion of plasma spectra that incorporates 3.555-3.677 ppm for which positive correlation is present with urine.

Generally, the urine region is complicated and contains a number of unassigned signals.[66] Distinguishable singlets at 4.34 ppm (U4), 4.40 ppm (U6) and 4.44 ppm (U8) do not correlate with plasma data. Despite no presence of a signal in the plasma region 4.064-4.070 ppm (P6) large areas of correlation are present for a similar urine region to previously except for around 4.277 ppm (U1, U3-U9). This indicates the correlations present are spurious. It is proposed that correlations between glycerol in plasma (P1 and P3) and 4.259-4.296 ppm (U2) in urine are not spurious.

Further areas of correlation are present for the region U2 and both 3.787-3.805 ppm (P4) and 3.860-3.873 ppm (P5) regions for plasma. The low signal intensities in the plasma regions make assignment difficult but the three most downfield signals of the glycerol triplet of triplets could be present in the former

range;[103,172] the remaining signals would be obscured by large glucose signals. Additionally, Nicholson *et al*.[142] reported glycerol presence at 3.87 ppm.

In Figure 4.7 the trend of signal intensities for the data point with the strongest correlation between urine and plasma (U2 centre, P3 centre) can be approximated as linear with a percentage increase in one mirrored by the other; an outlying sample is not causing correlation. Identification of the urinary metabolite responsible for correlation could not be established.

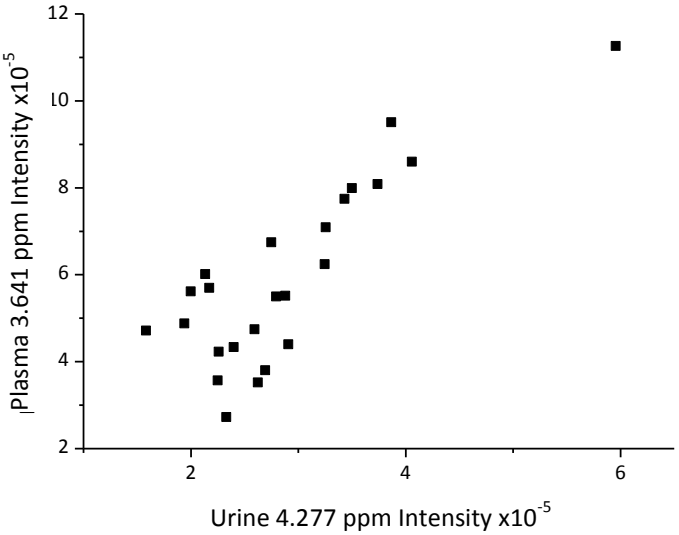


Figure 4.7 Scatter plot for control samples showing intensities of data points with maximum correlation within urine 4.222-4.497 ppm and plasma 3.555-3.677 ppm ranges.

Area C in Figure 4.2 is shown in Figure 4.8 and displays many of the correlation areas between urine regions, which include those centred at 1.092 (not shown), 6.973 (A), 6.985 (B), 7.756 (C), 7.774 (D), 7.933 (E), 8.360 (F) and 8.390 ppm (G), and plasma centred at 3.146 ppm. The signal in plasma spectra corresponds to dimethyl sulphone ($DMSO_2$) singlet.[103] Sample 1017's maximum normalised intensity for any data point of the signal is in excess of 50 times that of any other control sample (data not shown). $DMSO_2$ is a common dietary supplement, often in combination with glucosamine and chondroitin sulphate. Singly, or combined, $DMSO_2$ is popularly used for arthritic and rheumatic pain[173] but occurs naturally at low micromolar concentrations in plasma. It derives from dietary sources, intestinal bacterial

metabolism and endogenous methanethiol metabolism.[174] Medical records for the patient indicated no drug or supplement history but given acetaminophen was detected in a number of urine samples (Chapter 3) despite no record of consumption it is conceivable that the patient neglected to inform of $DMSO_2$ usage especially because it is a non-prescribed supplement. Most of the urine chemical shift values do not correspond with maximum intensities of signals and no clear intra-sample pattern is observed between signal intensities of the aforementioned urine regions. It is indicated that correlations resulted due to the effect of one sample.

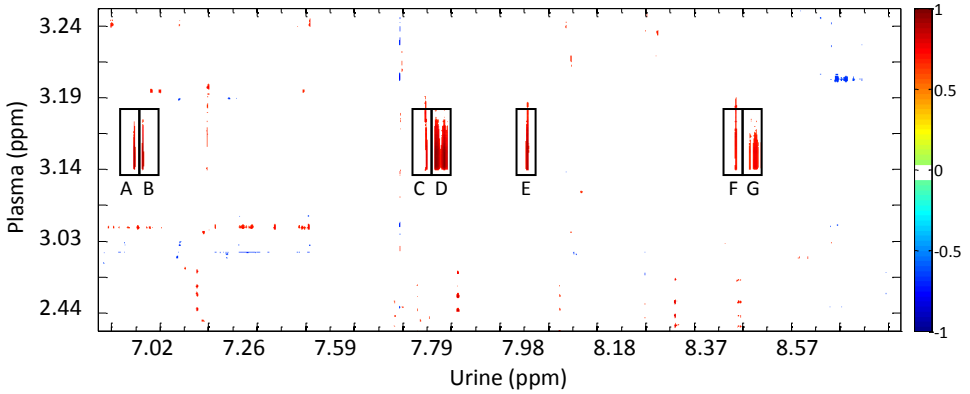


Figure 4.8 Expansion of output from Pearson's SHY analysis of urine and plasma data from control samples showing the urine 6.899-8.689 ppm and plasma 2.426-3.250 ppm area. The colour bar indicates the Pearson correlation coefficients. Boxed areas, with labels, highlight areas of correlation referred to in the text.

4.2.2 Spearman's Analysis

Pearson's correlation is sensitive to exogenous perturbations and potential outliers,[96] for example the unique signals of sample 2039 could be termed outliers as could the greatly increased $DMSO_2$ signal for sample 1017. Spearman's correlation is considered more robust and both approaches should be used in a comprehensive analysis.[96] Using the same parameters Spearman's correlation coefficients were calculated and plots shown in Figure 4.9 and Figure 4.10 for case and control samples, respectively.

Figure 4.9 indicates, through paucity of correlation in areas A and B, the correlations exhibited in Figure 4.1 in these areas were as the result of unique signals in sample 2039.

The largest area exhibiting strong (positive) correlation is in area E, centred at 1.275 ppm and 1.285 ppm for urine and plasma, respectively. The latter corresponds to presence of lipids[142] and is further verified through a smaller area of correlation (F) displayed in the same urine region but at 0.901 ppm for plasma, also characteristic of lipids.[103,142] A number of sources[103,158-160,163,175] were used in an attempt to identify the urine signal that was present in all samples but the singlet was not been able to be assigned. However, Engelke *et al.*[176] determined 3-hydroxyisovaleric acid as the species responsible for a singlet at 1.28 ppm.

The concentration of 3-hydroxyisovaleric acid is an indicator of reduced activity of the biotin-dependent enzyme 3-methylcrotonyl-CoA carboxylase.[177] The enzyme converts 3-methylcrotonyl-CoA, which is derived from leucine, to 3-methylglutaconyl-CoA but with reduced enzyme activity the pathway to 3-hydroxyisovaleryl-CoA is favoured. CoA and 3-hydroxyisovaleric acid are the degradation products of 3-hydroxyisovaleryl-CoA.[103] In summary, SHY indicates positive correlation between 3-hydroxyisovaleric acid, and hence negative correlation between biotin, and lipids; negative correlation between biotin levels and plasma lipids has been reported.[178,179] The most abundant biotin signal, a triplet at 2.21 ppm,[103] was not detected in either biofluid. A study reported 3-hydroxyisovaleric acid was greater in urine for smokers than control subjects and suggested smoking increased biotin catabolism.[180]

Smoking parameters were evaluated relative to the normalised value for the data point at 1.275 ppm in urine, corresponding to the tentatively assigned 3-hydroxyisovaleric acid signal. Different smoking statuses (‘current smoker’, ‘ex-smoker’ and ‘never smoked’ assigned as per Chapter 2 and Chapter 3) were not grouped relative to data point integral. When ranked on integral value the ‘ex-smoker’ group occupied the highest and lowest (36^{th}) positions whilst ‘current

smoker' group ranged between 3rd and 34th and 'never smoked' 5th to 35th. The number of cigarettes smoked also did not correlate with integral value. Of the 20 patients who did not form the 'never smoked' class, consumption was highest for a current smoker whose integral was ranked 34th. Ranks 1 to 4 corresponded to 18th, 16th, 6th and 19th greatest amount of cigarettes smoked. No evidence is presented to suggest a correlation between the signal area at 1.275 ppm and smoking.

Figure 4.10 shows similar correlation coefficients to Figure 4.2 for area D (urine 4.222-4.497 ppm and plasma 3.555-3.677 ppm). Due to the robust test and high confidence level used it is indicative that the correlation in the aforementioned area, for which responsible signals were assigned to glycerol in plasma, is not spurious. Within the aforementioned urine region negative correlation is shown with the plasma region 0.810-1.267 ppm (G and H) and the strongest (negative) correlation is located between 0.810 ppm and 0.883 ppm (G). The latter region corresponds to part of a lipid signal: the upfield side to approximately 0.005 ppm from the maximum intensity. The upfield side of another lipid signal corresponds to 1.267 ppm; the maximum intensity is located at approximately 1.285 ppm.

The urine metabolite shows opposite associations with glycerol and part of lipid signals. Glycerides are lipid esters of the glycerol molecule and fatty acids, with triglycerides having three fatty acids. Triglycerides are a major component of very low density lipoproteins (VLDL),[181] which of the lipoprotein types contributes most strongly to the downfield part of the aforementioned lipid signals whereas high density lipoproteins (HDL) influences the upfield region of lipid signals.[128] This would indicate negative association between glycerol and HDL in control patients but conversely negative association has been observed between triglycerides and HDL-cholesterol in patients with breast cancer compared to controls;[182] CVD is indicated by this triglyceride/HDL-cholesterol profile.[183] Although breast cancer is a risk factor for developing CVD[182] non-breast cancer medical conditions are not known for patients in this study so it is conceivable that control patients could have been afflicted by CVD.

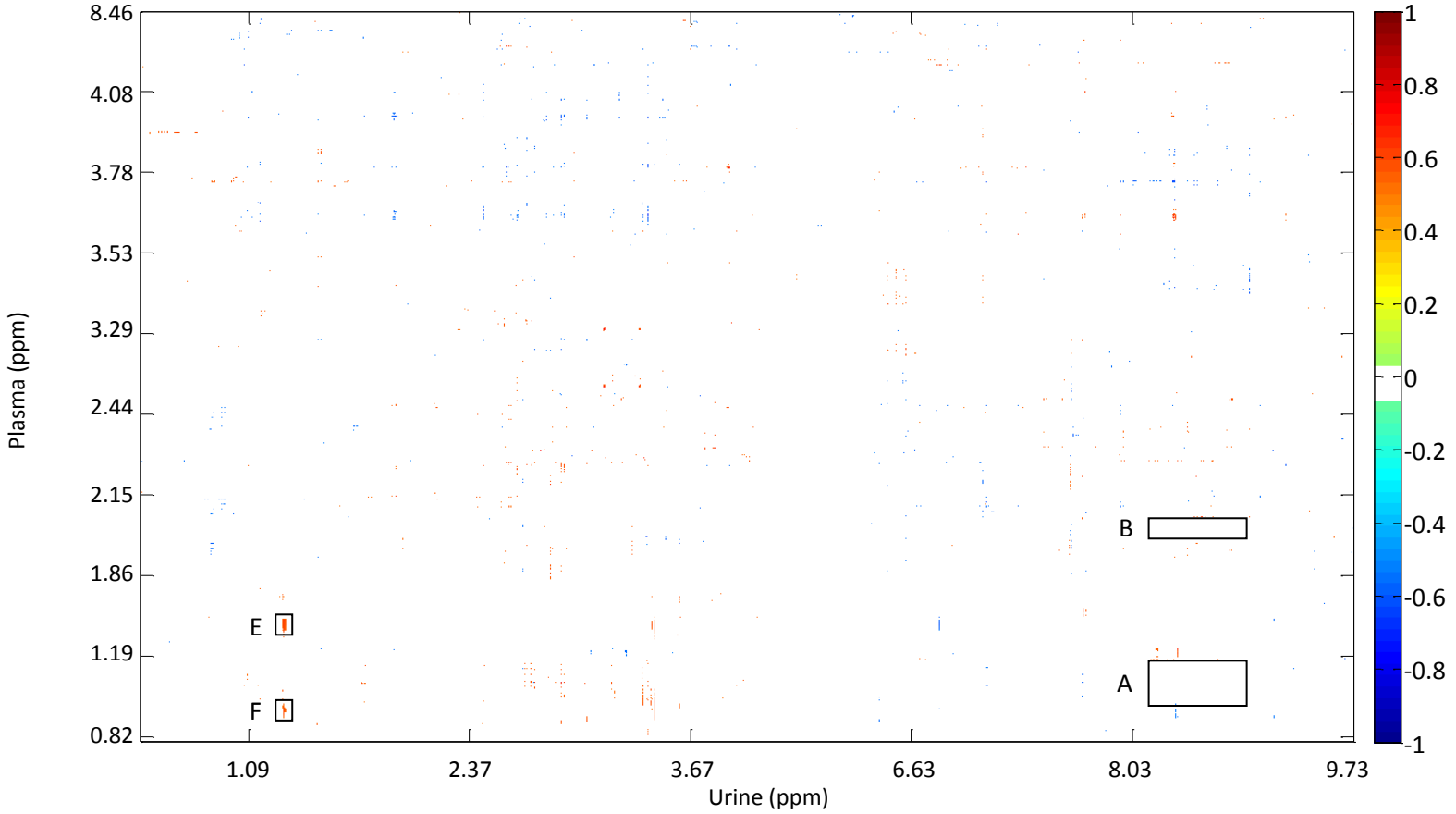


Figure 4.9 Output from Spearman's SHY analysis of urine and plasma data from case samples. The colour bar indicates the Spearman correlation coefficients. Boxed areas, with labels, highlight areas referred to in the text.

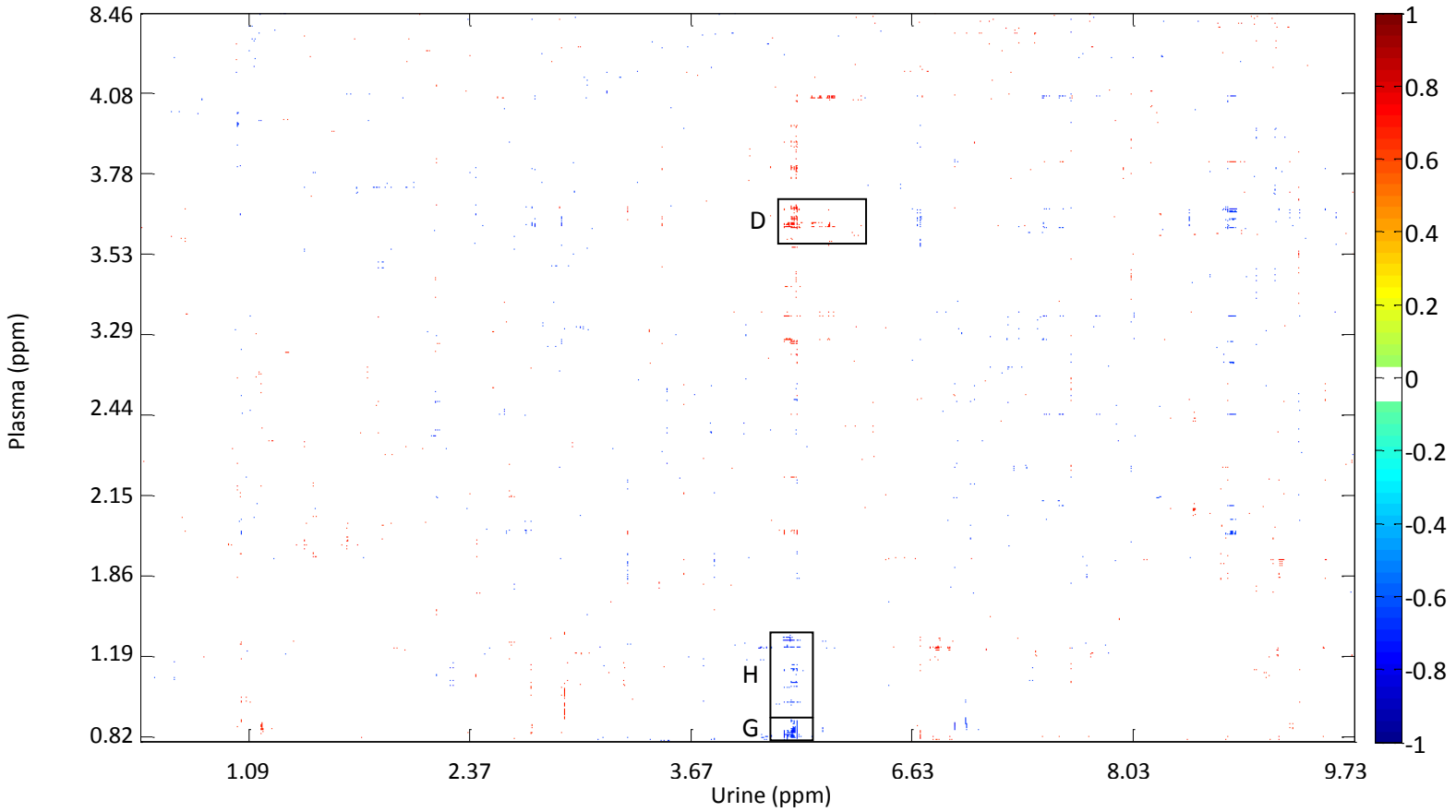


Figure 4.10 Output from Pearson's SHY analysis of urine and plasma data from control samples. The colour bar indicates the Spearman correlation coefficients. Boxed areas, with labels, highlight areas referred to in the text.

4.3 Conclusions

Incorporating a stringent *p*-value into both Pearson's and Spearman's tests for covariance reduced the chance of spurious correlations between urine and plasma data but because of the many thousands of data points some erroneous correlations have occurred. Deciphering true and false correlations was problematic though larger areas and greater intensity correlations indicated potentially biologically relevant correlations.

Both tests indicated glycerol in plasma from case samples was positively correlated to a urine species that could not be identified. Further, Spearman's analysis showed the urine area was correlated with the upfield section of lipid signals for which a speculative hypothesis has been reasoned involving triglyceride presence in lipoprotein species of different density.

Spearman's correlations identified a urine signal, tentatively identified as 3-hydroxyisovaleric acid, to be connected with plasma lipid signals in an inverse relationship. Relationships between 3-hydroxyisovaleric acid, biotin and lipids are established that corroborate the SHY findings.

SHY has provided data that could relate metabolites in urine and plasma for which it may have been difficult to discern connections. However the non-spurious nature of the correlations would need to be established before more definitive conclusions could be made. Spearman's analysis is an initial step towards this, here showing some correlations were due to unique or greatly elevated signal intensities and highlights the need for detailed medical records including consumption of non-prescription drugs and supplements.

Chapter 5. Gas Chromatography Analysis of Plasma

The concentrations of amino acids (AAs) in plasma, as determined by GC, will be described in the following chapter. Aliquots of plasma obtained from the same cohort as described in Chapter 2 were available, comprising of 33 case and 24 control samples.

MVA and univariate analysis of the GC data was performed in an attempt to identify possible AA biomarkers of breast cancer occurrence and progression. MVA and univariate analysis were performed as described in Section 8.4 and Section 8.5, respectively, with the exceptions that data were not normalised and UV scaling was performed rather than Pareto.

5.1 Determination of Amino Acid Concentrations

The commercially available system EZ:Faast (Phenomenex, Macclesfield, UK) has been developed for quantitative analysis of amino acids in a number of biological matrices, including plasma and urine. Briefly, the procedure involves a solid phase extraction step followed by derivatisation and a liquid/liquid extraction with the sample preparation procedure followed that was listed in the information booklet that accompanied the system. Further details of the methodology are in Section 8.7. Following earlier 'in-house' studies (E. Turner, unpublished work) it was decided to use this system to test whether AA biomarkers could be established for the sample set available.

Each sample was injected three times and any chromatogram that was deemed not to have sufficient signal to noise ratio, whereby the signal area was unduly affected by noise, was discarded. For two control samples and one case sample two chromatograms were considered too poor to use so without a replicate the data from the acceptable chromatogram was excluded from analysis. Resultantly, data from 32 case and 22 control samples were used in analysis. For five control samples

and six case samples data from one chromatogram was considered too poor to use so the average values for the concentrations of AAs were calculated using two runs of the samples.

Every sample contained an internal standard (IS) that allowed determination of AA concentrations relative to it. However, the detector response factor is dependent on the AA so a calibration curve for each of the 26 AAs was required (structures and retention times shown in Appendix 2, Table A2.1). Calibration curves were created using data obtained from calibration standard sets that consisted of three runs, each of a different concentration (50, 100 and 200 nmol ml^{-1}), containing all of the AAs. Multiple injections of the same standard set were performed. A gradient from the calibration curve was obtained for every run of each standard set. Upon optimisation of the gas chromatograph parameters three runs each from two standard sets were disregarded to account for variability in detector response resulting from a change of settings. A single run of a standard set was performed prior to the initial run of the first test sample. After every 10 test samples a standard set was run. Up to three runs per standard set were performed within a 24 hour time period (the maximum storage time at room temperature recommended by the kit manufacturers). In total, calibration curves were generated for 20 runs from seven calibration sets. A representative chromatogram from a standard is shown in Figure 5.1.

The calibration curve gradient was analysed across the 20 runs of calibration standard sets. The maximum value of the gradient divided by the minimum value (max/min) and residual standard deviation (RSD) determined the consistency of calibration curve gradient values. For six of the 26 AAs calibration curve gradients were not obtainable for all runs of standard sets. This was due to the concentration being below the lower limit of detection (LLOD) of the gas chromatograph for all three concentrations within a run of a standard set. When the max/min value was less than or equal to 1.5 the RSD range was 6.9-12.8% so the calibration curve gradient value was deemed acceptable; for metabolomics studies where GC-MS has been employed, use of an RSD of less than 20%[184] or 30%[185] as the determinant

for data retention has been applied. Seven of the max/min values were 1.6 or above (RSDs 13.0-31.8%) thus leaving concentrations of 13 AAs to be investigated in the biological samples (Table 5.1).

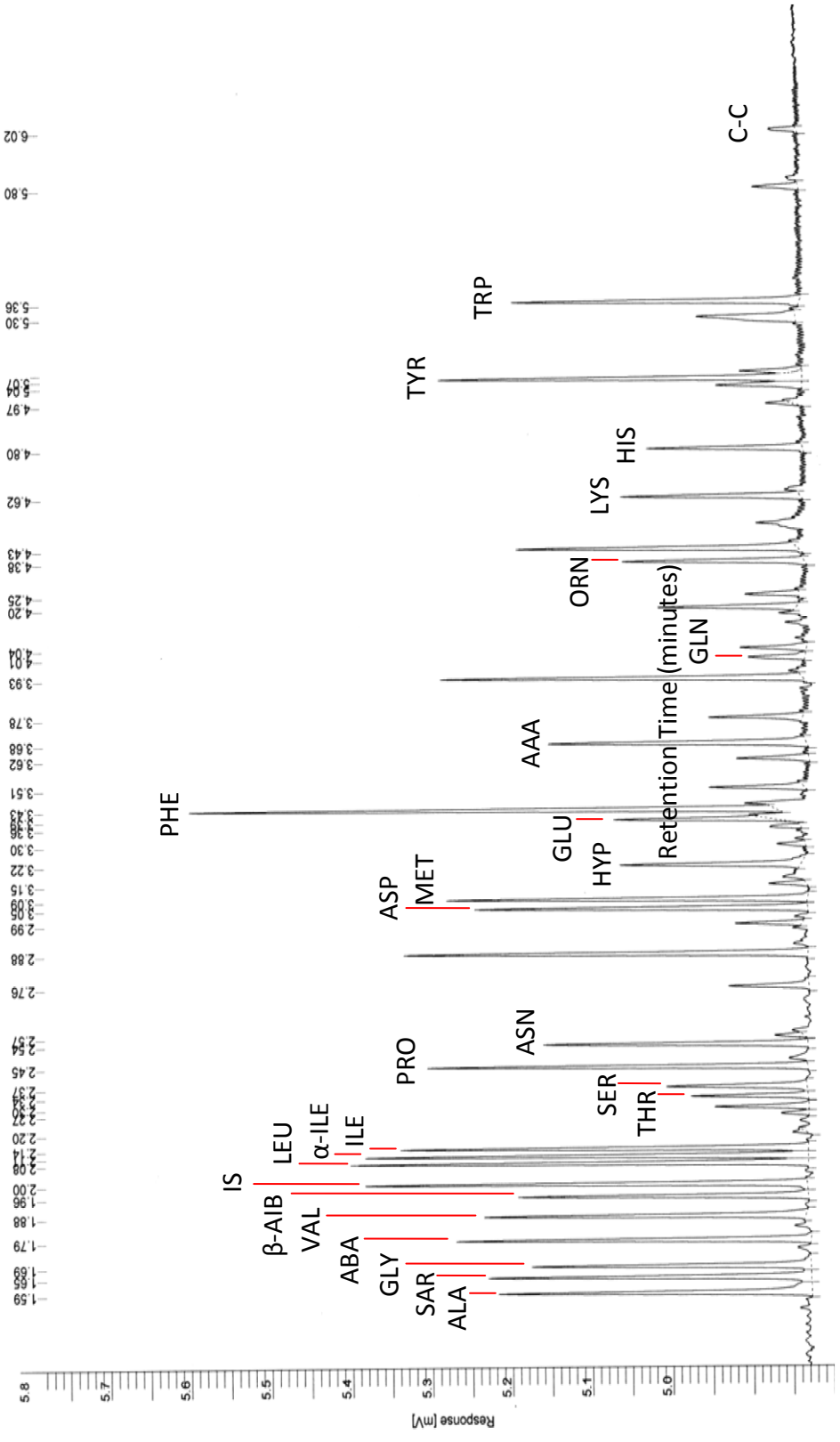


Figure 5.1 Chromatogram of a typical standard (200 nmol ml^{-1}) from a set used to generate a calibration curve for samples.

Numbers above the x-axis are retention times. AA abbreviations and full names are given in Table 5.1.

Table 5.1 Retention times and further investigation status of AAs.

AA		Retention Time (minutes)	Further Investigation
Full Name	**Abbreviation**		
Alanine	ALA	1.59	Yes
Sarcosine	SAR	1.65	Yes
Glycine	GLY	1.69	Yes
α-Aminobutyric acid	ABA	1.79	Yes
Valine	VAL	1.88	Yes
β-Aminoisobutyric acid	β-AIB	1.96	Yes
Norvaline (internal standard)	IS	2.00	N/A
Leucine	LEU	2.08	Yes
Allo-Isoleucine	α-ILE	2.11	Yes
Isoleucine	ILE	2.14	Yes
Threonine	THR	2.34	No
Serine	SER	2.38	No
Proline	PRO	2.45	Yes
Asparagine	ASN	2.54	No
Aspartic acid	ASP	3.05	Yes
Methionine	MET	3.09	Yes
Hydroxyproline	HYP	3.22	No
Glutamic acid	GLU	3.39	No
Phenylalanine	PHE	3.43	Yes
α-Aminoadipic acid	AAA	3.68	No
Glutamine	GLN	4.01	No
Ornithine	ORN	4.37	No
Lysine	LYS	4.62	No
Histidine	HIS	4.80	No
Tyrosine	TYR	5.07	No
Tryptophan	TRP	5.36	No
Cystine	C-C	6.01	No

In addition to the detector response factor differing inter-AA, variation could occur intra-AA with time due to column degradation. Running a set of standards regularly between samples reduced the impact of potential response factor deviations. For every AA that was investigated further the calibration curve gradient from every run of standards was plotted against run order. Despite all but one of the new gradients being negative, the r^2 values were low with values ranging between -0.03 for VAL

and -0.49 for both LEU and GLY (Appendix 2, Figure A2.1). For ASN, which had a positive gradient, the r^2 value was 0.04. For an intra-AA difference in response factor to be concluded much higher (modulus) r^2 values than those observed would be required.

Collectively, AA response factor variation was assessed through interpretation of IS area divided by total amino acid area (IS/AA) values. For each concentration the values were plotted against run order of calibration standard sets. Plots indicated weak positive association between IS/AA and run order (data not shown). Given the weak negative association shown between individual AA calibration curve gradient values and run order, the finding was not unexpected. Given that the maximum r^2 value of the three plots was 0.32 and the fourth set of standards consistently provided some of the highest IS/AA values there was no clear evidence of differing response factor for AAs collectively.

Additionally, for each concentration of every calibration standard set, IS/AA values of runs were plotted against run order. For some calibration standard sets there was strong negative association between IS/AA and run order. The first, second and third runs of Standard VII and Standard VIII at 100 nmol ml^{-1} and Standard XI at 50 nmol ml^{-1} showed almost perfect correlation with r^2 values all in excess of -0.99 (data not shown). All but one of the plots revealed a negative correlation (data not shown) but with just three points the gradient and r^2 value can be heavily influenced by a single point. Degradation of standards with time could also be assessed by evaluation of the plots. If degradation occurred, strongest negative association would be expected to be shown by plots for standards that had the longest time period between the first and last run. This time was 23 hours 55 minutes for both Standard X and Standard XI but stronger negative association was not observed than for standards with a shorter elapsed time such as Standard VII, which had an equivalent time of 6 hours 30 minutes (Appendix 2, Figure A2.2). There is no evidence to suggest degradation of standards and because the elapsed time between runs of test samples was no more than the maximum time between runs of standards it can be postulated no degradation of test samples occurred.

5.2 Results

Having performed the appropriate calibration procedures the concentrations of the 13 AAs were investigated in the plasma samples. No more than four samples exhibited detectable levels of SAR, β-AIB, α-ILE and ASP so these AAs were excluded from analysis. ABA and MET were detectable in the majority of samples so were retained in analyses; for some samples not all runs provided a signal for either AA or MET that was above the LLOD so the average was calculated using only the runs where the signal was quantifiable. A typical chromatogram of a sample is shown in Figure 5.2.

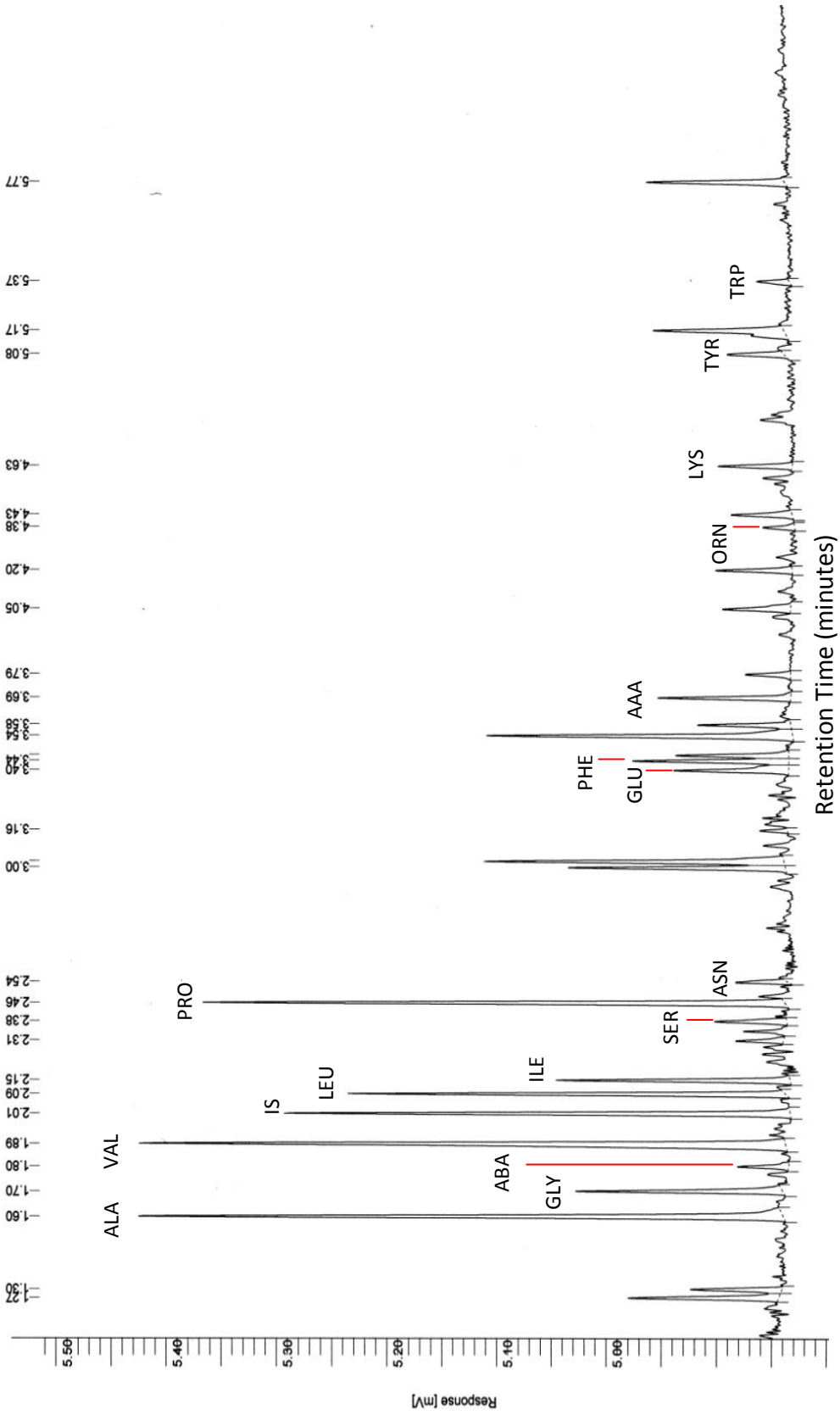


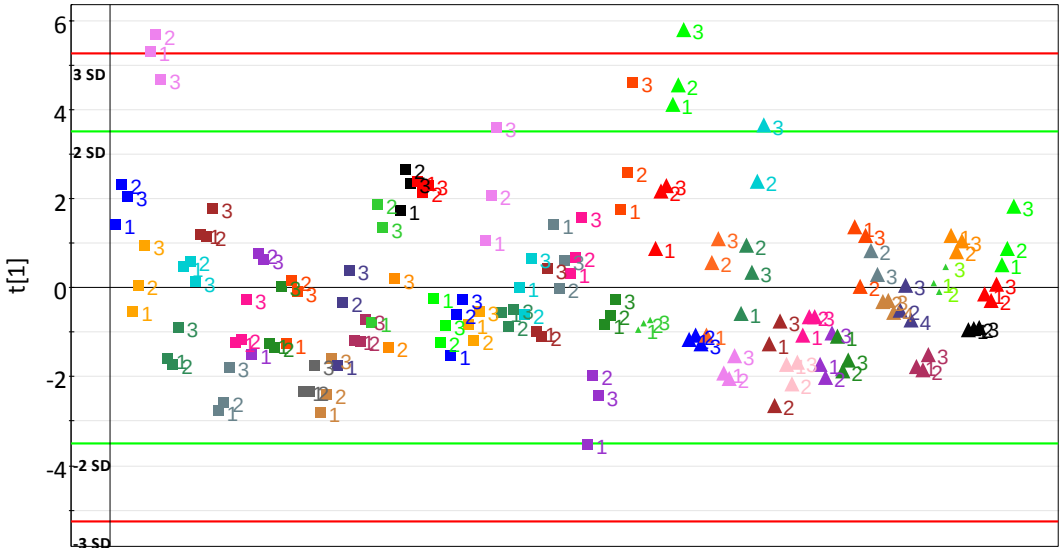
Figure 5.2 Chromatogram of a case sample (2001, run 3). Numbers above the x-axis are retention times.

5.2.1 Amino Acid Concentration as a Function of Case and Control Samples

5.2.1.1 Application of Multivariate Analysis

Usually MVA is used when there are more variables than observations. In the case of the GC analysis procedure used here this is not the situation. However, MVA has been employed to allow potential interconnectivities between AAs to be visualised. For NMR spectroscopy data pareto scaling is normally preferred to UV scaling to avoid the over-inflation of spectral noise regions (Section 1.4.1.3) but for GC data the difference in magnitude between the smallest and largest values of AA concentrations is substantially less so UV scaling is suitable.

Prior to averaging concentrations the reproducibility of GC data was assessed by PCA of individual runs. NMR spectroscopy produces data that is highly reproducible,[50] hence replicates are not required, but GC is less robust.[186] ABA and MET were excluded because presence was not detectable in all runs of samples. The scores plot (Figure 5.3) of the one component model ($R^2\boldsymbol{X}$ = 0.439 and $Q^2\boldsymbol{X}$ = 0.252) shows that multiple runs are needed because the positions of the runs from a sample in scores space are not the same. However, they are sufficiently similar to allow analysis of sample data that has been averaged from runs.



Key: shape - □ case and △ control

Figure 5.3 PCA plot of individual runs of case and control samples. Adjacent runs with the same colouring originate from the same sample. Label numbering refers to the order of runs.

AA concentrations from runs were averaged, including those for ABA and MET, and PCA was performed. A one component model was created with an $R^2\boldsymbol{X}$ value of 0.386 and a $Q^2\boldsymbol{X}$ value of 0.234 (data not shown). There was no separation between case and control samples. The data could not be modelled using PLS-DA.

The average concentration values for ABA and MET were not as accurate as for other AAs. Calculating the average signal area, and hence concentration, through division of the sum of detectable signal areas by the number of runs for which the signal was detectable produced a concentration value higher than the true value: if a signal was present but below the LLOD the area was not recorded and hence could not be included in concentration calculations. Averaging over all runs irrespective of whether the signal area was below the LLOD produced a lower value than the true one because any area less than that could be detected was set to zero. For ABA the average concentration of case (22.0 nmol ml^{-1}) and control samples (20.7 nmol ml^{-1}) combined was 21.5 nmol ml^{-1} (excluding samples for which the average was zero; range 9.3-38.2 nmol ml^{-1}) and the lowest concentration of any run was 8.2 nmol ml^{-1}, which is an indicator of the LLOD. Given this value is almost 40% of the overall average, values just below the LLOD could strongly influence the sample

average. For MET, the equivalent case, control and combined concentrations were 19.6, 19.9 and 19.7 nmol ml^{-1}, respectively, whilst the range was 9.4-33.2 nmol ml^{-1} and the lowest concentration of a run was 9.2 nmol ml^{-1}. Including zero sample averages the combined average was reduced to 17.9 and 11.2 nmol ml^{-1} for ABA and MET, respectively.

Repeating PCA with ABA and MET excluded had little effect on the appearance of scores and loadings plots (data not shown; one PC model, $R^2\boldsymbol{X}$ = 0.463 and $Q^2\boldsymbol{X}$ = 0.259).

5.2.1.2 Application of Univariate Analysis

The mean concentrations of AAs in case and control samples were evaluated to establish whether significant differences were present between sample types. All data were tested for normality with a null hypothesis that data distribution was normal (Section 1.4.2.1). A Shapiro-Wilk *p*-value ≥0.05 led to the conclusion that the data was normally distributed, hence the Student's *t*-test was used but if non-normal distribution of data was proven, as evidenced by *p*-value <0.05, the Mann Whitney U test was performed to determine whether potential differences were statistically significant. The null hypothesis was no difference between mean concentration values for the two sample types.

Box plots (Figure 5.4) summarised the data and allowed possible outliers to be identified. All chromatograms were reinvestigated if an AA concentration was observed beyond the whiskers of the plot. On 14 occasions (including plots for ABA and MET that excluded samples with zero average) for the nine AAs, case or control sample concentrations were more extreme than 1.5 times the respective interquartile range beyond the upper or lower quartile. Every sample was retained because all runs provided consistent signal areas therefore no reason could be identified to permit exclusion.

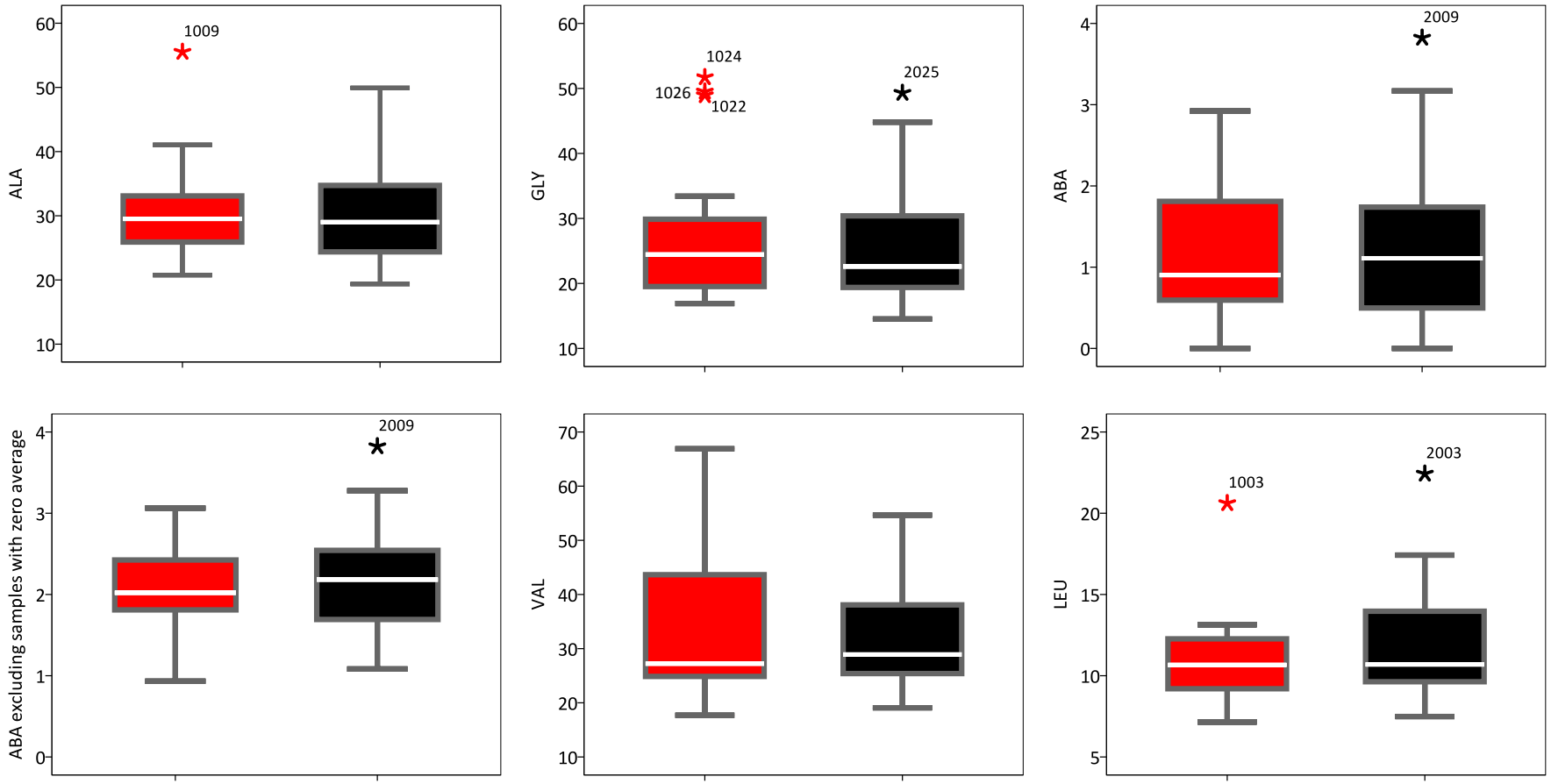


Figure 5.4 Box plots of nine AA concentrations for case and control samples. Except where specified the plots represent data from 54 samples. Red filled box = control samples, black filled box = case samples. * = samples more extreme than 1.5 times the interquartile range beyond the upper quartile, coloured as per filled box with sample label. Numbers on the *y*-axis relate to AA concentration (nmol ml^{-1} $x10^{1}$).

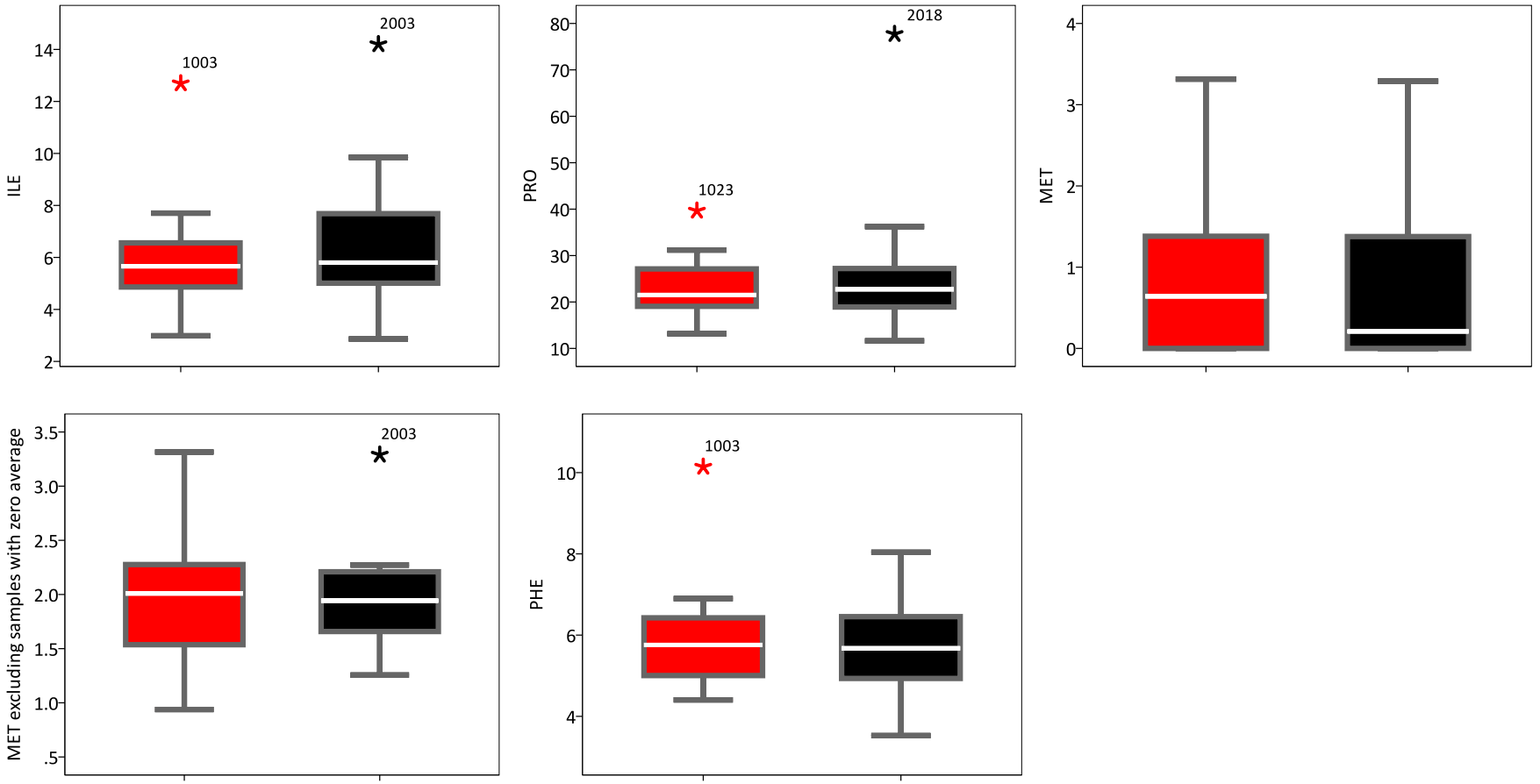


Figure 5.4 Continued.

The mean concentrations of case and control samples were calculated for each AA and a conclusion forwarded as to whether there was a significant difference between values (Table 5.2). For all tests performed a *p*-value >0.05 resulted so the null hypothesis was retained in each case: there was no difference between mean concentration values for case and control samples. Correction for multiple testing did not need to be performed. For ABA and MET, excluding samples that had an average concentration of zero did not change the conclusion that there were no differences between the concentration values of the two sample types.

Table 5.2 Summary of data regarding mean concentration differences of AAs between case and control samples.

AA	**Number of Samples**	**Mean Concentration (nmol ml^{-1})**		**Data Distribution**	**Uncorrected *p*-value**	**Mean Concentration Conclusion**
		Control	**Case**			
ALA	54	308.7	297.5	Non-normal	0.660	No difference
GLY	54	275.3	257.8	Non-normal	0.408	No difference
ABA	54	17.0	18.6	Normal	0.819	No difference
ABA (excluding zero sample averages)	45	20.7	22.0	Normal	0.499	No difference
VAL	54	342.8	320.6	Non-normal	0.958	No difference
LEU	54	109.0	117.4	Non-normal	0.379	No difference
ILE	54	59.4	62.8	Non-normal	0.561	No difference
PRO	54	230.4	243.2	Non-normal	0.916	No difference
MET	54	13.5	9.5	Non-normal	0.322	No difference
MET (excluding zero sample averages)	30	19.9	19.6	Normal	0.927	No difference
PHE	54	58.9	57.0	Non-normal	0.846	No difference

5.2.2 Amino Acid Concentration as a Function of Breast Cancer Grade

Two individuals included in this study were reported to have been diagnosed with more than one invasive tumour. Consequently, samples from these patients were excluded from analysis. The remaining 30 case samples were graded 0, 1, 2 or 3 with 7, 5, 9 and 9 samples, respectively.

5.2.2.1 Application of Multivariate Analysis

PCA was performed on the 30 samples resulting in a one component model ($R^2\boldsymbol{X}$ = 0.392 and $Q^2\boldsymbol{X}$ = 0.219; scores plot shown in Figure 5.5 but loadings plot not shown). No clear separation was observed between the four groups and a PLS-DA model was not able to be created. Excluding ABA and MET from PCA resulted in a one component model ($R^2\boldsymbol{X}$ = 0.478 and $Q^2\boldsymbol{X}$ = 0.282; data not shown) but, again, no clear separation was observed between the groups in the scores plot. A PLS-DA model was not able to be generated.

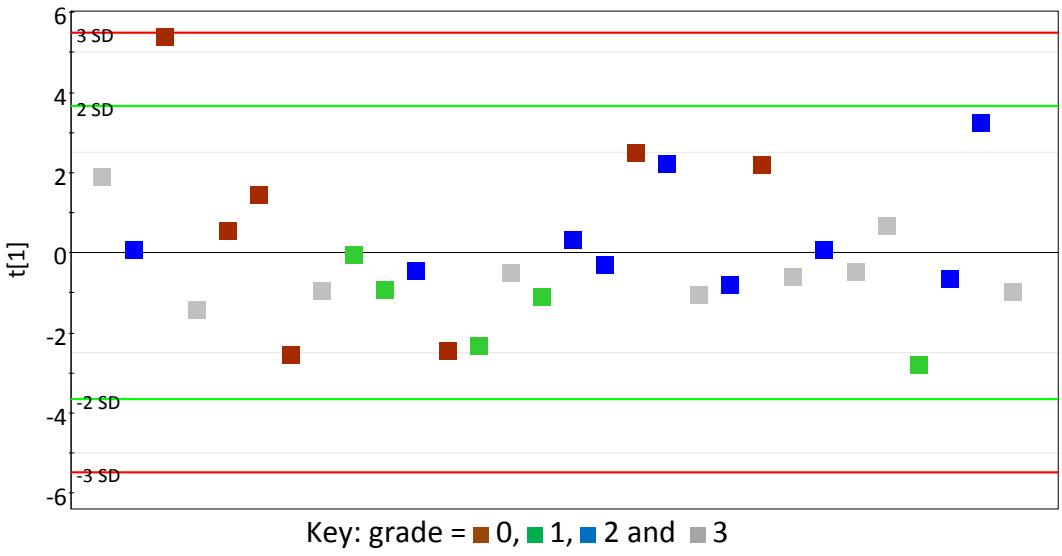


Figure 5.5 PCA scores plot for single tumour case samples coloured according to tumour grade.

It was proposed that assessing two groups might reveal separation of samples that was not apparent when the four groups were considered together. The model parameters, including and excluding ABA and MET, are summarised in Table 5.3 and

Table 5.4, respectively.

Table 5.3 Parameters for models that included ABA and MET.

Grades Compared	PCA			PLS-DA			
	PCs	R^2X(cum)	Q^2X(cum)	PCs	R^2X(cum)	R^2Y(cum)	Q^2Y(cum)
0 v. 1	2	0.812	0.453	4	0.915	0.917	0.658
0 v. 2	1	0.385	0.107	0	/	/	/
0 v. 3	2	0.707	0.269	0	/	/	/
1 v. 2	1	0.345	0.040	2	0.542	0.682	0.454
1 v. 3	0	/	/	1	0.268	0.54	0.203
2 v. 3	0	/	/	0	/	/	/

Table 5.4 Parameters for models that excluded ABA and MET.

Grades Compared	PCA			PLS-DA			
	PCs	R^2X(cum)	Q^2X(cum)	PCs	R^2X(cum)	R^2Y(cum)	Q^2Y(cum)
0 v. 1	2	0.841	0.425	1	0.647	0.352	0.212
0 v. 2	1	0.480	0.191	0	/	/	/
0 v. 3	2	0.787	0.388	0	/	/	/
1 v. 2	1	0.425	0.095	3	0.759	0.759	0.523
1 v. 3	2	0.784	0.495	1	0.375	0.437	0.202
2 v. 3	0	/	/	0	/	/	/

PCA did not show separation between any of the pairs of groups when incorporating all nine AAs (data not shown). Tentative separation could be viewed between the groups in the three PLS-DA models (data not shown). For each model all AA concentrations except ABA were indicated as being lower in grade 1 samples. Given the previously stated problems with ABA (and MET) concentration measurements and small number of samples in the groups, it is not possible to make reliable conclusions at present and further samples would be required to increase confidence in this observation.

Excluding ABA and MET data resulted in tentative separation between most grade 0 and 1 samples when PCA was employed (Figure 5.6); this model gave the best separation between any combinations of two groups. $R^2\boldsymbol{X}$(cum) and $Q^2\boldsymbol{X}$(cum) values (Table 5.4) were high though with only seven variables it would be expected

that the model represented the data to a high level. The loadings plot (data not shown) indicated all of the AA concentrations were elevated in samples located in positive PC 1 scores space compared to those in negative space. Similar observations, though not as pronounced, were made when grade 1 samples were compared to grade 2 or 3 samples. No improvement in separation was visible upon application of PLS-DA for grade 0 and 1 samples (data not shown); the scores plot generated using grade 1 and 3 samples displayed similar positioning of samples in scores space (data not shown), as did the scores plot for grade 1 and 2 samples (Figure 5.7).

Validation of PLS-DA results is always required; the 'leave one out' method was employed for the three models that were generated using data from seven AAs. Associated information for the grade 1 and 2 model is shown in Table 5.5.

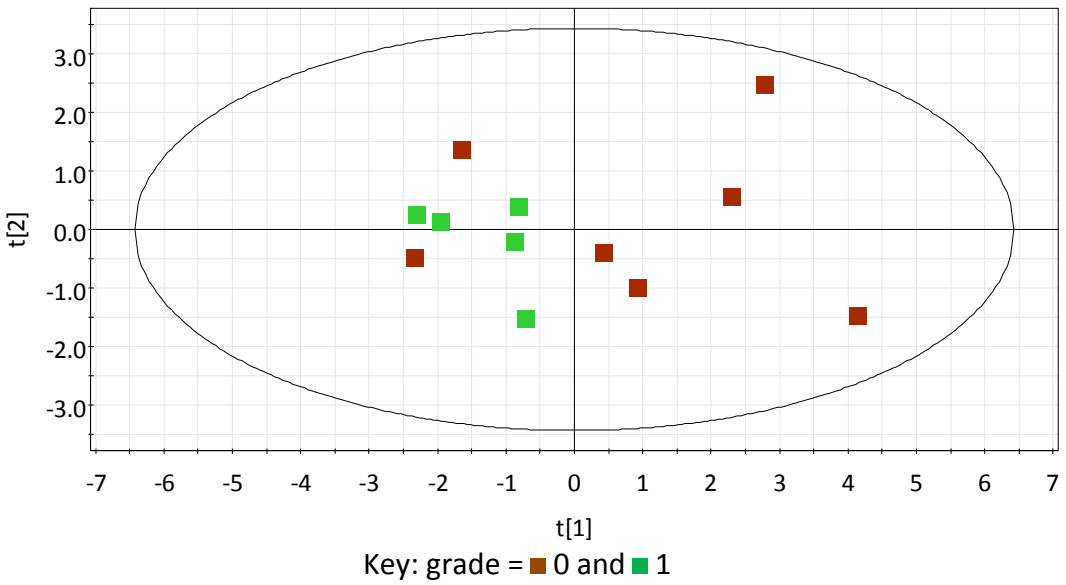


Figure 5.6 PCA scores plot for grade 0 and grade 1 samples excluding ABA and MET data.

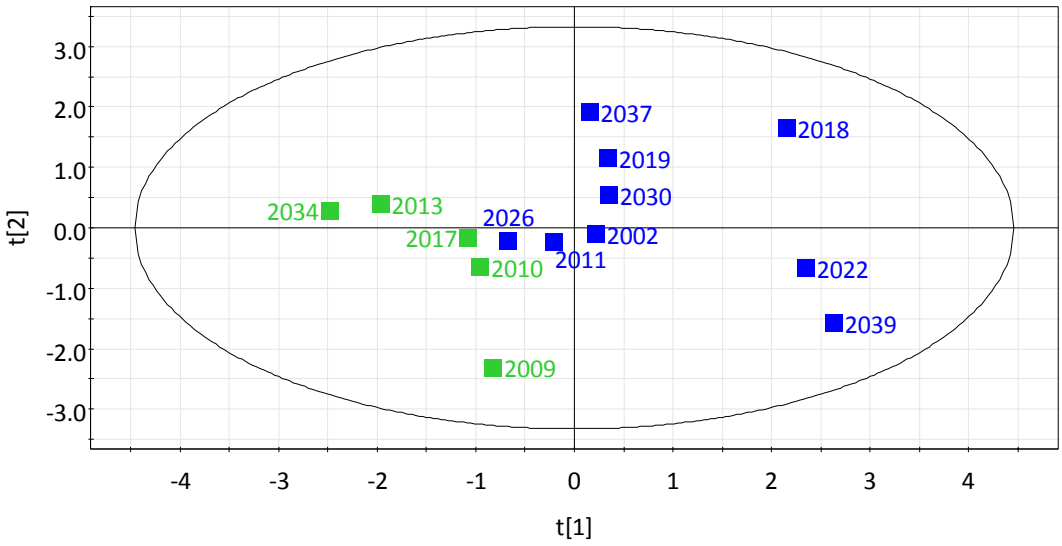


Figure 5.7 PLS-DA scores plot for grade 1 and grade 2 samples excluding ABA and MET data. R^2X = 0.375 and 0.266, R^2Y = 0.541 and 0.108 and Q^2Y = 0.393 and 0.118 for PC 1 and PC 2, respectively.

Table 5.5 'Leave one out' cross-validation parameters for the PLS-DA model shown in Figure 5.7.

Sample Excluded	Number of Components	R^2X (cum)	R^2Y (cum)	Q^2Y (cum)	Grade	*Y*-Predicted*	
						1	2
2002	3	0.774	0.781	0.521	2	0.396	0.604
2009	4	0.877	0.848	0.463	1	1.087	-0.087
2010	1	0.377	0.527	0.294	1	0.530	0.470
2011	3	0.760	0.804	0.465	2	0.550	0.450
2013	1	0.329	0.521	0.308	1	0.769	0.231
2017	1	0.370	0.539	0.349	1	0.556	0.444
2018	2	0.703	0.665	0.451	2	-1.307	2.307
2019	1	0.432	0.539	0.449	2	0.392	0.608
2022	1	0.349	0.575	0.486	2	-0.407	1.407
2026	5	0.957	0.883	0.697	2	0.334	0.666
2030	1	0.379	0.542	0.468	2	0.307	0.693
2034	1	0.314	0.509	0.415	1	0.925	0.075
2037	1	0.393	0.565	0.493	2	0.355	0.645
2039	1	0.326	0.581	0.419	2	-0.330	1.330

*A sample was regarded as belonging to a grade by having a ***Y***-predicted value >0.50. Incorrect classification is represented by red shading and correct by pink, orange or green corresponding to a ***Y***-predicted value of <0.60, 0.60-0.70 or >0.70, respectively.

All but one of the 14 samples was correctly predicted, giving a 93% predictive ability. The PLS-DA loadings plot (data not shown) indicated all AAs were elevated in grade 2 samples. The accuracy of class prediction is reduced for samples as the ***Y***-predicted value tends towards 0.50; predictions made based on values less than 0.60, of which there are two, as indicated by pink shading in Table 5.5, should be treated with caution whilst those made using values greater than 0.70 (shaded green) could be regarded as the most reliable. The extent of the ***Y***-predicted value beyond 0.5 is not regularly considered.[140]

Of the 12 grade 0 and 1 samples 50% were correctly predicted, the percentage expected by random chance, with two having a ***Y***-predicted value of over 0.70. For grade 1 and 3 samples, nine from 14 (64%) samples were correctly predicted. Only two predictions were made based on 0.60-0.70 ***Y***-predicted values and the rest from values greater than 0.70, though one value was generated using a one component PLS-DA model that had a negative $Q^2\boldsymbol{Y}$ value. Despite high ***Y***-predicted values for correctly predicted grade 1 and 3 samples, the combination of a predictive ability that is too close to 50% and high ***Y***-predicted values (up to 1.010) for incorrectly predicted samples does not allow the conclusion to be made that grade 1 samples have lower concentrations of seven AAs.

Although there is closeness in scores space between some of the grade 1 and 2 samples, the high predictive ability percentage could indicate that grade 1 samples have lower concentrations of ALA, GLY, VAL, LEU, ISO, PRO and PHE. Validation using a prediction set is a more rigorous method than 'leave one out' cross-validation because for the latter method the predicted samples would have been used to build the original model. Given the small number of samples in each grade a prediction set was not available. However, permutation testing did validate the model and supported 'leave one out' cross-validation findings. The associated plot for the grade 2 class (Figure 5.8) shows all but 15 of the 999 permuted models to have lower $R^2\boldsymbol{Y}$ values than the original model whilst the equivalent for $Q^2\boldsymbol{Y}$ is two permuted plots. The intercept values of the regression lines for $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ are 0.367 and -0.369, respectively. The permutation testing plot for the grade 1 class

(plot not shown) displayed $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ intercept values of 0.357 and -0.375, respectively, with six and one permuted models having higher $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ values, respectively. Although the original model should have higher $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ values than the permuted models as one of the criteria for validation, using 999 permutations, the maximum that SIMCA-P+ software can perform, it is not unexpected that a small proportion will be higher than the original model in some cases. The $R^2\boldsymbol{Y}$ intercept values are close to 0.4, the maximum value to allow the original model to be validated. The better the model, the lower the intercept values and the fewer permuted models that will have $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ values above those of the original model. This could explain why classes were correctly predicted using 'leave-one-out' cross-validation for seven of the 13 samples with ***Y***-predicted values between 0.50 and 0.70.

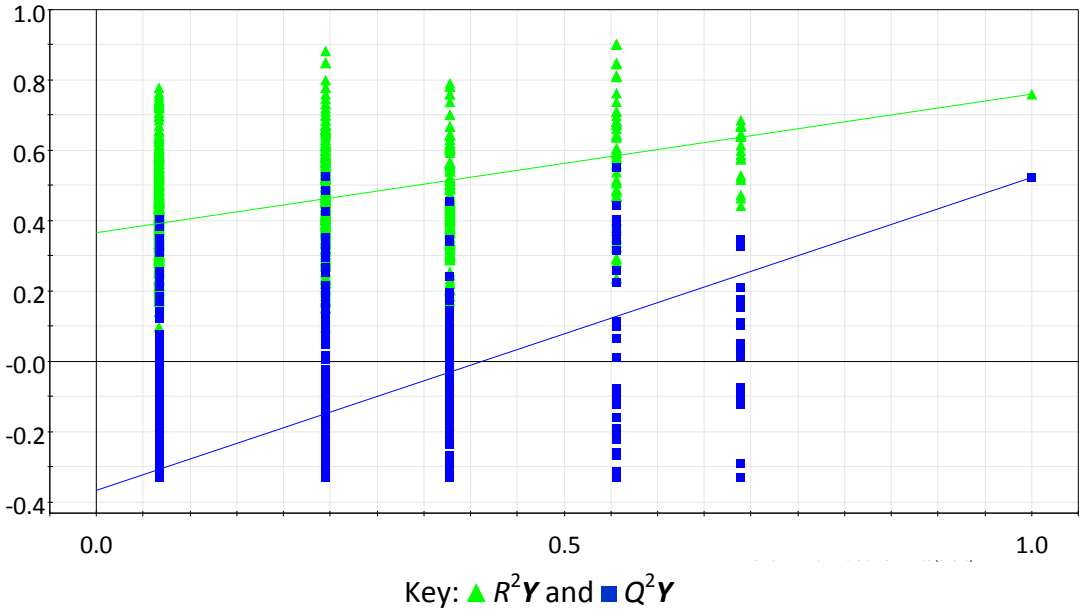


Figure 5.8 Permutation testing plots for the grade 2 class in the PLS-DA model shown in Figure 5.7. The R^2Y and Q^2Y intercept values of the regression lines are 0.367 and -0.369, respectively.

5.2.2.2 Application of Univariate Analysis

Mean concentrations of nine AAs were compared for every combination of two groups to determine whether there were significant differences between tumour grades. Table 5.6 shows the mean concentrations of AAs for the different grades

and Figure 5.9 the box plots. All chromatograms from a sample were reinvestigated if any AA concentration was more extreme than 1.5 times the interquartile range beyond the upper or lower quartile of the group; this was performed for 19 instances. Every sample was retained because all runs provided consistent signal areas therefore no reason could be identified to permit exclusion.

Table 5.6 Mean concentration of AAs with different breast cancer grades.

AA	Number of Samples	Sample Grade Mean Concentration (nmol ml^{-1})			
		0	1	2	3
ALA	30	321.8	233.6	321.6	288.4
GLY	30	247.1	205.2	266.7	255.8
ABA	25	25.0	23.0	22.7	19.2
VAL	30	348.2	282.8	359.9	286.6
LEU	30	136.3	101.9	117.7	109.8
ILE	30	77.3	51.7	59.8	60.9
PRO	30	2.419	165.7	301.1	232.3
MET	15	20.5	19.4	23.7	17.2
PHE	30	61.5	48.9	59.7	56.1

For the seven AAs present in all samples, as listed previously, all were normally distributed for all grades with the exception of GLY for grade 1, PRO for grade 2 and both VAL and ILE for grade 3. Significant differences in concentrations of these seven AAs between grades are summarised in Table 5.7. Corrections for multiple comparisons were performed using FDR. Due to different numbers of samples in groups that contained ABA and MET, these two AAs were analysed separately and their *p*-values were not included in the aforementioned multiple test correction calculations.

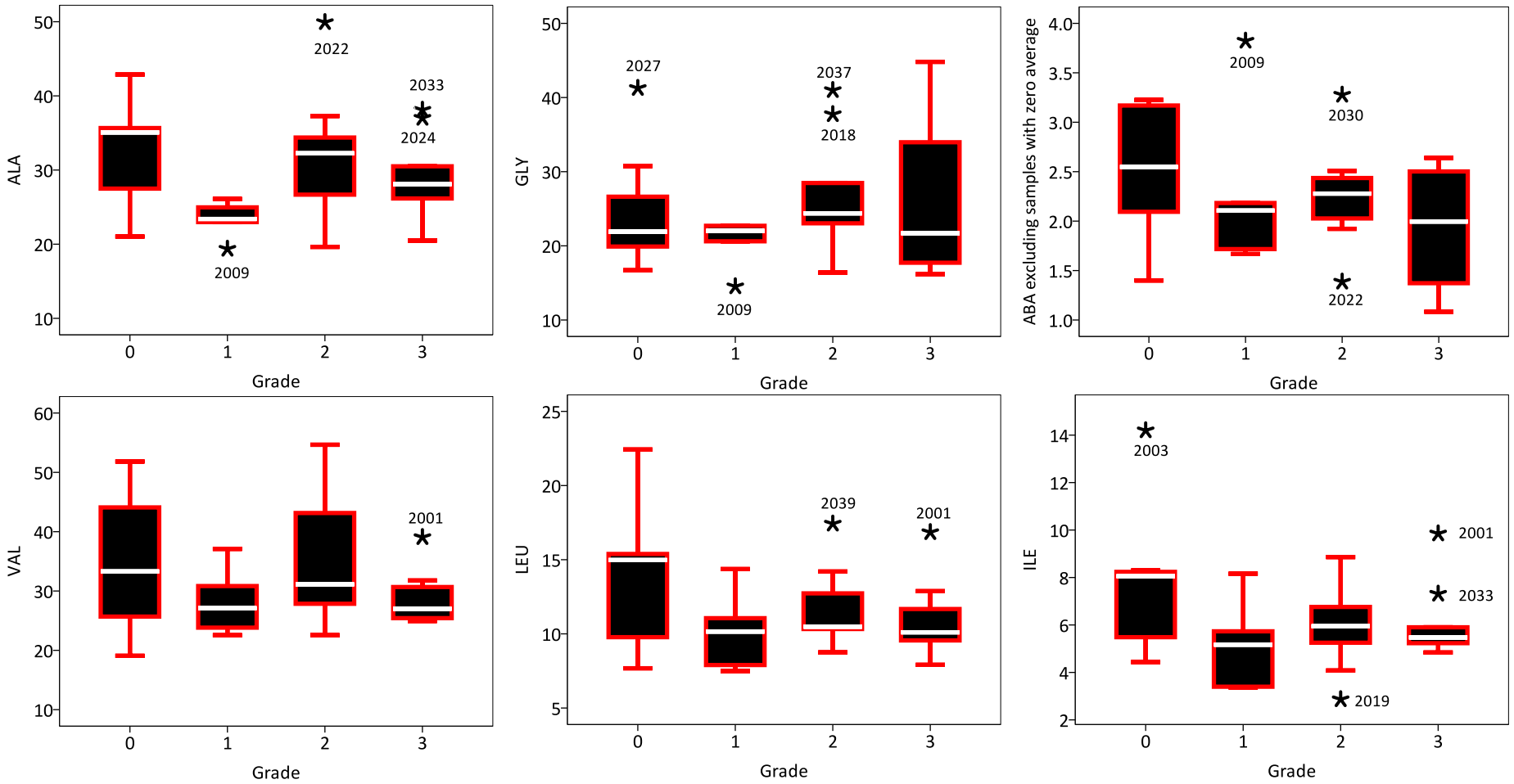


Figure 5.9 Box plots of concentrations of AAs for case samples of different grade. * = samples more extreme than 1.5 times the interquartile range beyond the upper or lower quartile. Numbers on the *y*-axis are AA concentration (nmol ml^{-1} $x10^{1}$).

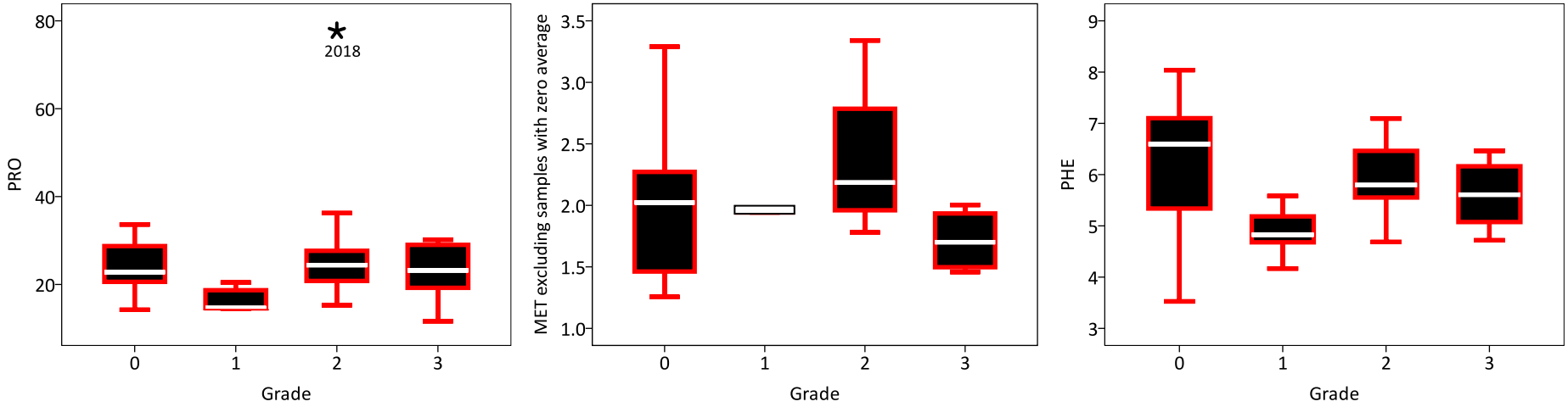


Figure 5.9 Continued.

Table 5.7 Summary of differences in the concentrations of AAs excluding ABA and MET between grades. ↑ or ↓ indicates whether the AA mean concentration is higher or lower (though not necessarily significantly) in the first listed grade compared to the second; only concentrations of AAs listed under the 'mean concentration conclusion' heading are significantly different.

Grades Compared	AA Identified	Data Distribution	*p*-value		Mean Concentration Conclusion
			Uncorrected	Corrected	
0 v. 1	↑ALA	Normal	0.033	0.154	No difference
	↑PRO	Normal	0.044	0.154	No difference
0 v. 2	None	/	/	/	No difference
0 v. 3	None	/	/	/	No difference
1 v. 2	↓ALA	Normal	0.047	0.122	No difference
	↓PRO	Non-normal	0.007	0.049	↓PRO
	↓PHE	Normal	0.018	0.063	No difference
1 v. 3	None	/	/	/	No difference
2 v. 3	None	/	/	/	No difference

MVA had indicated concentrations of all AAs except ABA were lower in grade 1 samples compared to other grades. This was substantiated by Table 5.6 but nearly all of the differences were not statistically significant (Table 5.7). A few AAs were identified as potentially significantly different between grades (uncorrected *p*-value <0.05) but after correction for multiple comparisons only PRO was significantly different between grades 1 and 2. Data verification proceeded to investigate the reliability of the result due to the influence of sample 2018. Although the sample's chromatograms had been concluded to be suitable for analysis the mean PRO concentration was much higher than for any other sample. The concentration of the three runs ranged from 722.1 to 804.7 nmol ml^{-1} with a mean of 777.0 nmol ml^{-1}. The mean concentration range was 144.7-204.5 nmol ml^{-1} for grade 1 and 152.7-362.4 nmol ml^{-1} for grade 2 excluding sample 2018. With a mean of more than double that of the sample with the next greatest concentration, sample 2018 had a large influence on the group mean especially given the small number of samples in the group (nine). Analysis was repeated that excluded sample 2018. The difference in mean concentrations between the groups was no longer significant (Table 5.8). Clear justification for removing the sample cannot be provided but due to the

sample's vastly greater concentration and the borderline difference at the 95% significance level when the sample is included, further samples would be required to verify a potential difference in the mean concentration of PRO between grades 1 and 2.

Table 5.8 Summary of differences in the concentration of PRO between grades 1 and 2. ↓ indicates (non-significantly) lower mean concentration in grade 1 compared to grade 2.

Grades Compared	AA Identified	Data Distribution	*p*-value		Mean Concentration Conclusion
			Uncorrected	Corrected	
1 v. 2 (excluding 2018)	↓PRO	Normal	0.027	0.095	No difference

Table 5.9 shows the sample distribution of ABA and MET. It was deemed there were sufficient numbers of ABA samples in each of the groups for mean concentrations to be compared for all combinations of two groups. For MET, only one sample belonged to the grade 1 group so comparison of concentrations excluded this group thus reducing the number of combinations of groups to three. For all tests no *p*-value was less than 0.05 so it was concluded there was no difference in the concentration of ABA or MET between different grades of tumour.

Table 5.9 Sample distribution of ABA and MET with breast cancer grade.

Grade / AA	0	1	2	3
ABA	6	5	7	7
MET	6	1	4	4

5.3 Conclusions

Due to max/min and RSD values obtained from calibration standards being unacceptably high for some AAs only 13 from 26 were able to be quantified in samples. For the 9 AAs present in the majority or all of the samples none of the

mean concentrations were revealed to be significantly different between case and control samples. MVA did not show clear separation between the different types of samples. In an analogous manner to Chapter 2 where NMR spectroscopy was used, there is no evidence to suggest an association between breast cancer and a change in AA concentrations that were determined by analysis of data acquired by GC. Furthermore, the conclusion can generally be extended to different grades of breast cancer though there is some evidence to suggest that grade 1 samples have lower concentrations of ALA, GLY, VAL, LEU, ISO, PRO and PHE. The numbers in each class were small and further samples would be necessary for rigorous validation of models. Univariate analysis generally did not confirm MVA observations regarding grade 1 samples to the extent that they were statistically significant thus indicating the differences observed were not due to individual AA concentration variance, with the possible exception of PRO. Further experimentation in the form of increased sample numbers would be needed to determine whether there is a significant difference in the mean concentration of PRO between grades 1 and 2 as identified when no potential outlying samples were excluded from analysis.

Chapter 6. NMR Analysis of Tissue Extracts

Analysis of aqueous and lipophilic tissue extract data acquired by ^{1}H-NMR spectroscopy will be detailed in the following chapter. Breast cancer tissue (Tumour) samples and adjacent non-cancerous tissue (Normal) samples were available from 15 post-menopausal women with differing severity of disease: 5 each with single occurrence grade 1, 2 or 3 tumours. The receptor status of the tumours was highly uniform with 12 samples having ER+, PR+ and HER- status, 11 of which are known to be DCIS positive (Table 6.1). The patient cohort is different to that of previous chapters.

Table 6.1 Receptor and DCIS statuses of Tumour samples.

Tumour Receptor Status (Subtype)	**DCIS Status**	
	Present	Unknown
ER+, PR+, HER- (Luminal A)	11	1
ER+, PR-, HER- (Luminal A)	0	1
ER+, PR+, HER unknown (Luminal A or B)	0	1
ER-, PR-, HER- (Basal)	0	1

Samples were prepared as per Section 8.1.1.3 and data collected as detailed in Section 8.2.3. Section 8.3 applies for spectral processing with dark regions listed in Table 8.5 and Table 8.6 for aqueous and lipophilic extracts, respectively. Constant sum normalised data was used initially. MVA and univariate analysis were performed as detailed in Section 8.4 and Section 8.5, respectively, in an attempt to identify possible biomarkers of breast cancer occurrence and progression in both extract types.

6.1 Results

Each tissue sample was given an identifier during collection, the format dependent on the source site and generalised for this work as a three or four digit number to

establish the 'non-identifiable' patient provider combined with N or T to describe Normal or Tumour tissue sample types, respectively.

Samples from nine patients, thus 18 samples in total, were sourced from the Breast Cancer Campaign Tissue Bank (BCCTB). For this chapter a paired sample is described as a single sample, either N or T, for which the other tissue sample type is available from the same patient. Of the 18 paired samples, 12 were collected at Site A and six at Site B. The remaining 12 paired samples were sourced directly from Site A.

The frozen weight and appearance of samples was recorded. In addition the appearance of the non-homogenised part of the sample was noted along with both wet and air-dried non-homogenised weights. After addition of extraction liquids the colour was recorded as for cellular debris following centrifugation. Due to the small size of some samples, volumes of extraction liquids had to be scaled up from the values calculated using the ratios detailed in Section 8.1.1.3.1 to achieve the minimum required for homogeniser use (200 µl).

One tissue sample (T6876; ER+, PR+, HER-, DCIS present) was destroyed during homogenisation and one lipophilic extract sample (T6758; ER+, PR+, HER-, DCIS present) contaminated during processing. An aqueous sample (N116; paired Tumour sample ER+, PR+, HER-, DCIS present) spectrum was not suitable for analysis. Numbers of samples available for data analysis are summarised in Table 6.2.

Table 6.2 Summary of sample numbers available for data analysis. Paired relates to number of samples, either Normal or Tumour type, with opposite sample type from the same patient. Values in parentheses refer to breast cancer grades 1, 2 and 3, respectively.

Sample / Extract Type	Normal	Tumour	Paired
Aqueous	14 (4, 5, 5)	14 (5, 4, 5)	26 (8, 8, 10)
Lipophilic	15 (5, 5, 5)	13 (5, 4, 4)	26 (10, 8, 8)

6.1.1 Aqueous Extracts Analysis

A high level of citrate was present in the aqueous extract spectrum of two samples. This unnatural level is undesirable for metabolomics studies and is known to source from sodium citrate that is present in certain collection tubes[185] so for the two samples it is possible that the collection tubes were the source of citrate. The region 2.517-2.750 ppm that contained citrate signals was removed from all samples in further analysis. The region 3.341-3.371 ppm was excluded because it contained methanol from the extraction process. A typical spectrum that did not contain citrate is shown in Figure 6.1. A variety of sources were used to identify metabolite signals.[103,118,135,136,138,141,142,187-191]

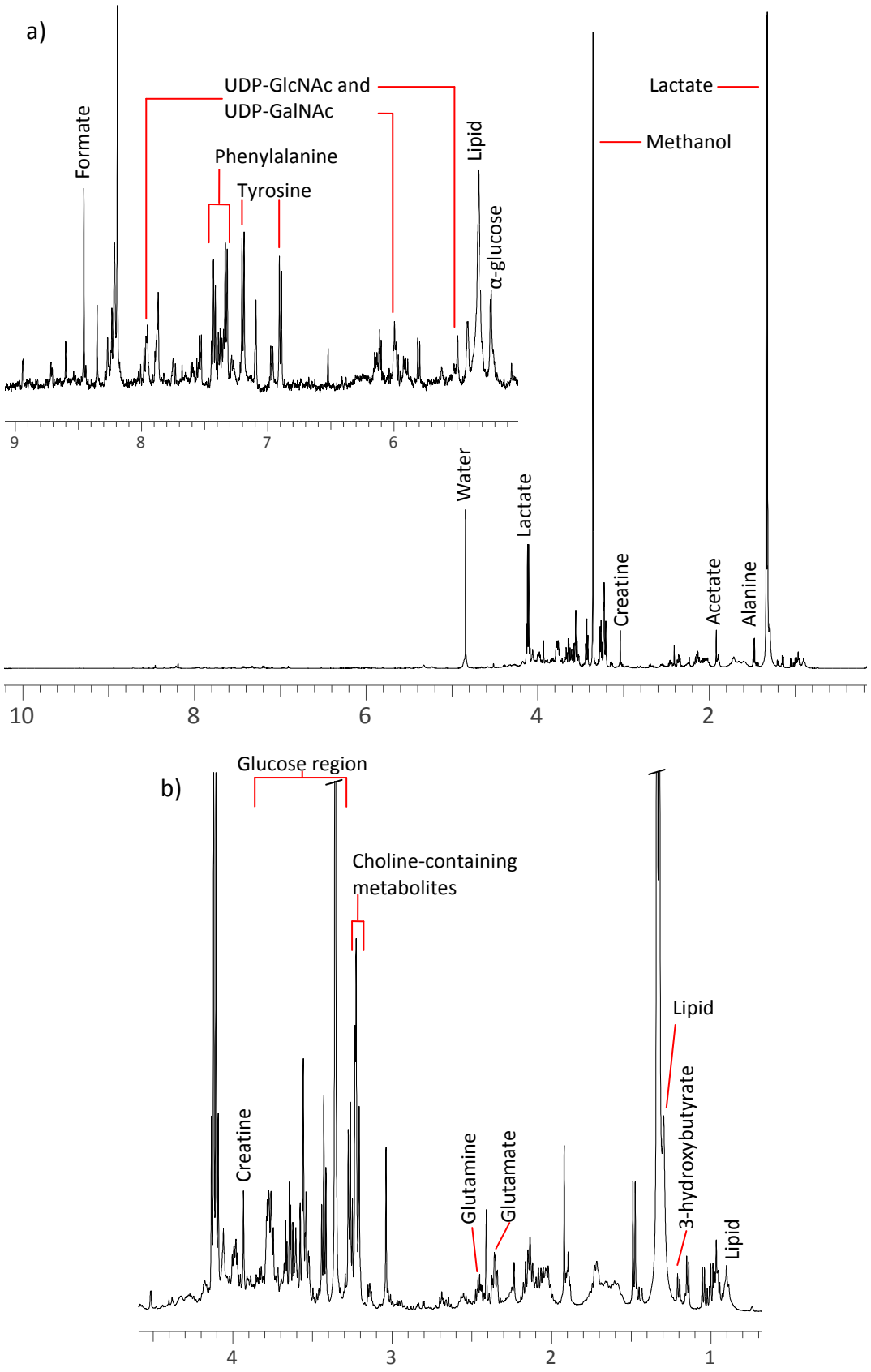


Figure 6.1 ^{1}H-NMR spectrum of tissue aqueous extracts from a Tumour sample. a) main: whole spectrum; inset: aromatic region. b) aliphatic region. The *x*-axis is chemical shift in ppm.

6.1.1.1 Evaluation of Breast Cancer Occurrence

6.1.1.1.1 Analysis Implementing Sum Normalisation

PCA was performed initially for 28 samples. A model could not be generated so components were forced to investigate the samples. The first two components showed T132 to be greatly separated in scores space (Figure 6.2). The sample was noted as unusual during data acquisition because spectrometer tuning was not able to be performed to the usual standard and presence of signals (large, broad singlets at 3.619 and 3.882 ppm) that were non-standard compared to other samples (excluding citrate signals). The loadings plot (Figure 6.3) confirmed the signals unique to the sample were causing scores space isolation for the sample.

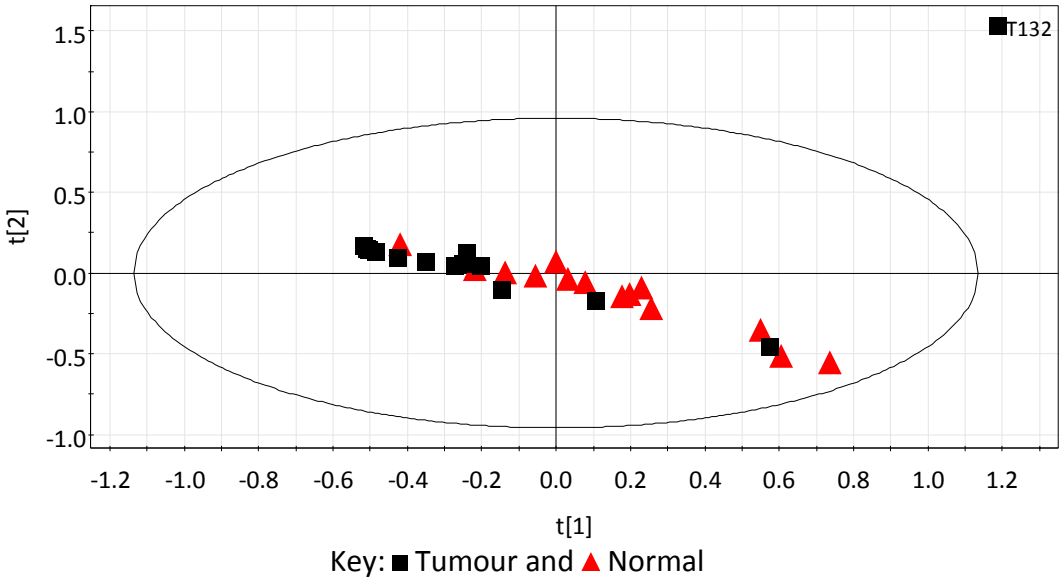


Figure 6.2 Forced PCA scores plot coloured according to tissue type for all 28 aqueous extract samples.

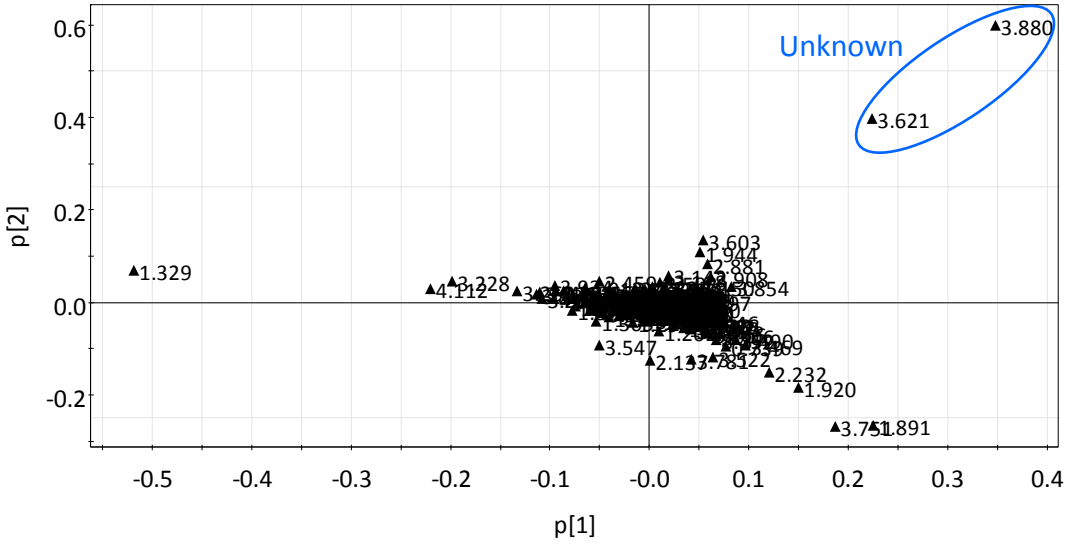


Figure 6.3 PCA loadings plot corresponding to the model displayed in Figure 6.2. The blue ellipse highlights bins containing unknown signals unique to T132 areas.

Unlike citrate, the unidentified signals could not be attributed to a source, either natural or 'contaminant'. Given this, the influence of the signals in PCA and importantly irregularities during data acquisition, sample T132 was removed from all further analysis. The signals were not present in the Normal sample from the same patient.

PCA was repeated and a three PC model was generated with $R^2\boldsymbol{X}$(cum) and $Q^2\boldsymbol{X}$(cum) values of 0.594 and 0.337, respectively. The scores plot is shown in Figure 6.4 and loadings plot in Figure 6.5.

The majority of Normal and Tumour samples were separated diagonally along PC 1 and PC 2 (Figure 6.4). The loadings plot indicated bins centred at 1.329, 3.228 and 4.112 ppm strongly contributed to positive $\boldsymbol{p}$[1] and negative $\boldsymbol{p}$[2] loadings space in Figure 6.5, which discriminated Tumour from Normal samples and indicated increased levels in Tumour samples. The first and last bin contained lactate signals whilst for the other both GPC and PCho were present.

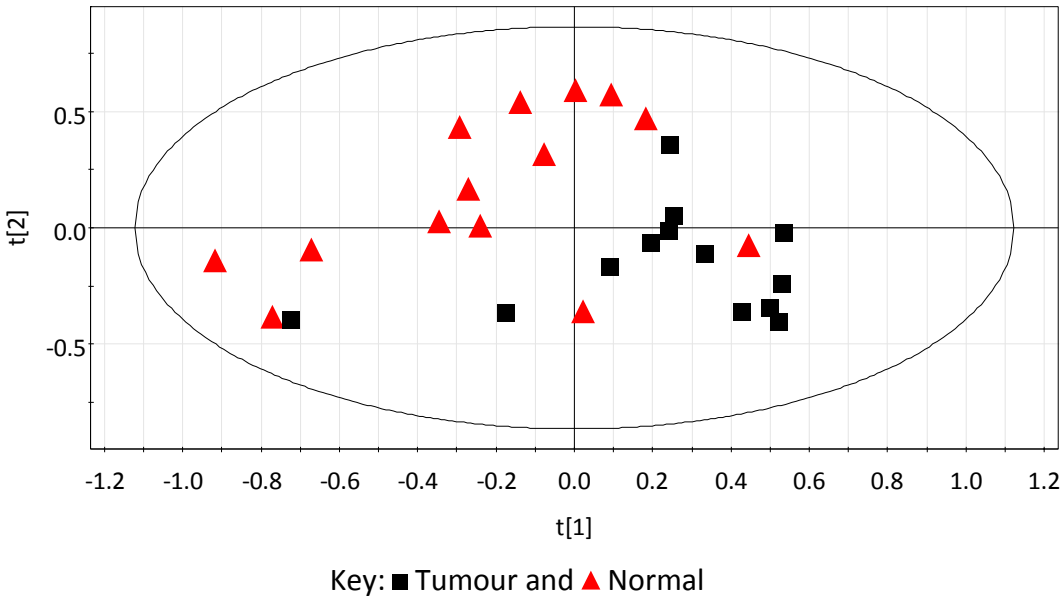


Figure 6.4 PCA scores plot coloured according to tissue type for aqueous extract samples excluding T132 showing the first two model components. R^2X = 0.311 and 0.184, and Q^2X = 0.171 and 0.151 for PC 1 and PC 2, respectively.

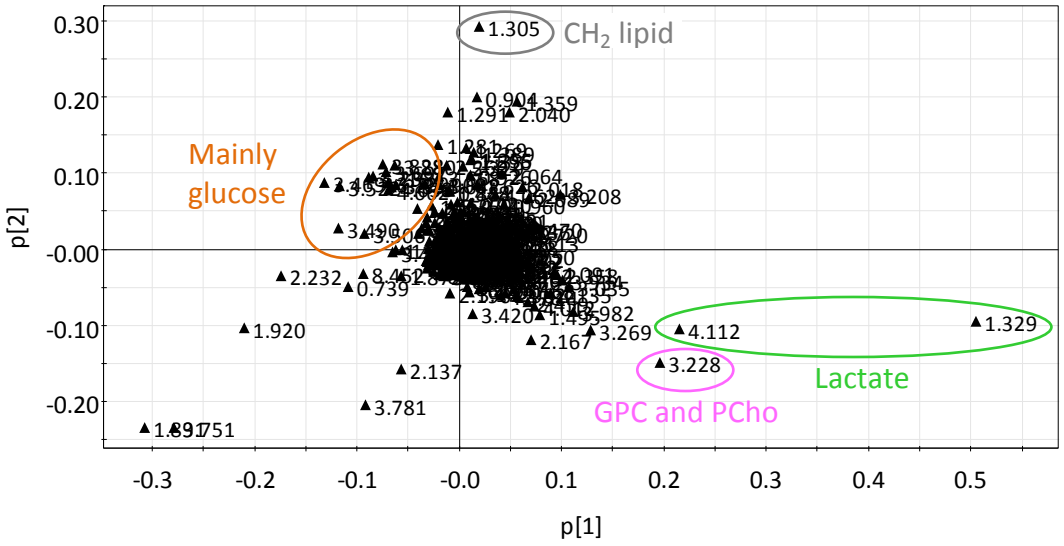


Figure 6.5 PCA loadings plot corresponding to the model displayed in Figure 6.4.

Other bins were present in the same loadings quadrant, though less influential, as well as in the opposite quadrant, the latter indicating lower signal levels in Tumour samples. A cut-off loadings value was used to determine whether the bin would be investigated further in determining significant differences between mean levels of metabolites in this study for the two sample types. One of the lesser intensity peaks of the lactate quartet (bin centred at 4.136 ppm) had a loadings value between 0.08

and 0.09 in the first component; due to the diagonal nature of separation the modulus loadings values of the first and second components were combined and a cut-off value of 0.08 employed for bins located in the two quadrants previously referred to. This allowed signals of lesser intensity to be considered without incorporating too many bins otherwise the purpose of using MVA would not be able to be upheld. Some discretion was used regarding bins that had a high loadings value in one component but were just outside of either quadrant in question, such as the bin centred at 1.305 ppm (Figure 6.5) For instances where a metabolite signal was present in an adjacent bin(s) to those identified in the loadings plot the normalised integrals were summed across all of the bins.

In total, 24 integral regions were evaluated (Figure 6.6) though other bins were also identified. The bin centred at 3.469 ppm had the largest ***p***[1] scores value in the top left (negative ***p***[1] and positive ***p***[2]) quadrant; the signal present was attributed to glucose. Numerous other glucose containing bins that fulfilled the criteria for further investigation were located in the same quadrant including those centred at 3.491, 3.507, 3.522, 3.669, 3.697, 3.716, 3.729, 3.829, 3.855, 3.880, 3.909, 4.638, 4.663 and 5.239 ppm. Due to α- and β-glucose signals centred at 5.233 and 4.648 ppm, respectively, being readily resolved and no signals from other metabolites contained in the same bins, the integrals of the bins for these resonances were analysed to determine whether there was a difference between mean integrals of glucose for Tumour and Normal samples. α- and β-glucose are the same molecule but were evaluated separately to enable a relative overview of the ratios between the two sample types.

Due to a Normal and Tumour sample emanating from the same patient the samples could not be treated as independent and metabolite level differences between the sample types from a patient, *i.e.* paired samples, had to be considered. As a result, the unpaired samples were excluded, thus leaving 24 samples for univariate analysis. Box plots in Figure 6.6 summarised the paired sample data.

All data were tested for normality with a null hypothesis that data distribution was normal. A Shapiro-Wilk *p*-value <0.05 led to the conclusion that the data was not normally distributed, hence the Wilcoxon signed rank test was used but if normal distribution of data was proven, as evidenced by *p*-value ≥0.05, a paired Student's *t*-test was performed to determine whether there was a difference between the mean levels of metabolites in Normal and Tumour samples. The null hypothesis was no difference between mean integral values for the two sample types. FDR was used to correct for multiple comparisons. Table 6.3 displays the bins identified as having potentially different mean integrals and conclusions regarding whether the difference was statistically significant.

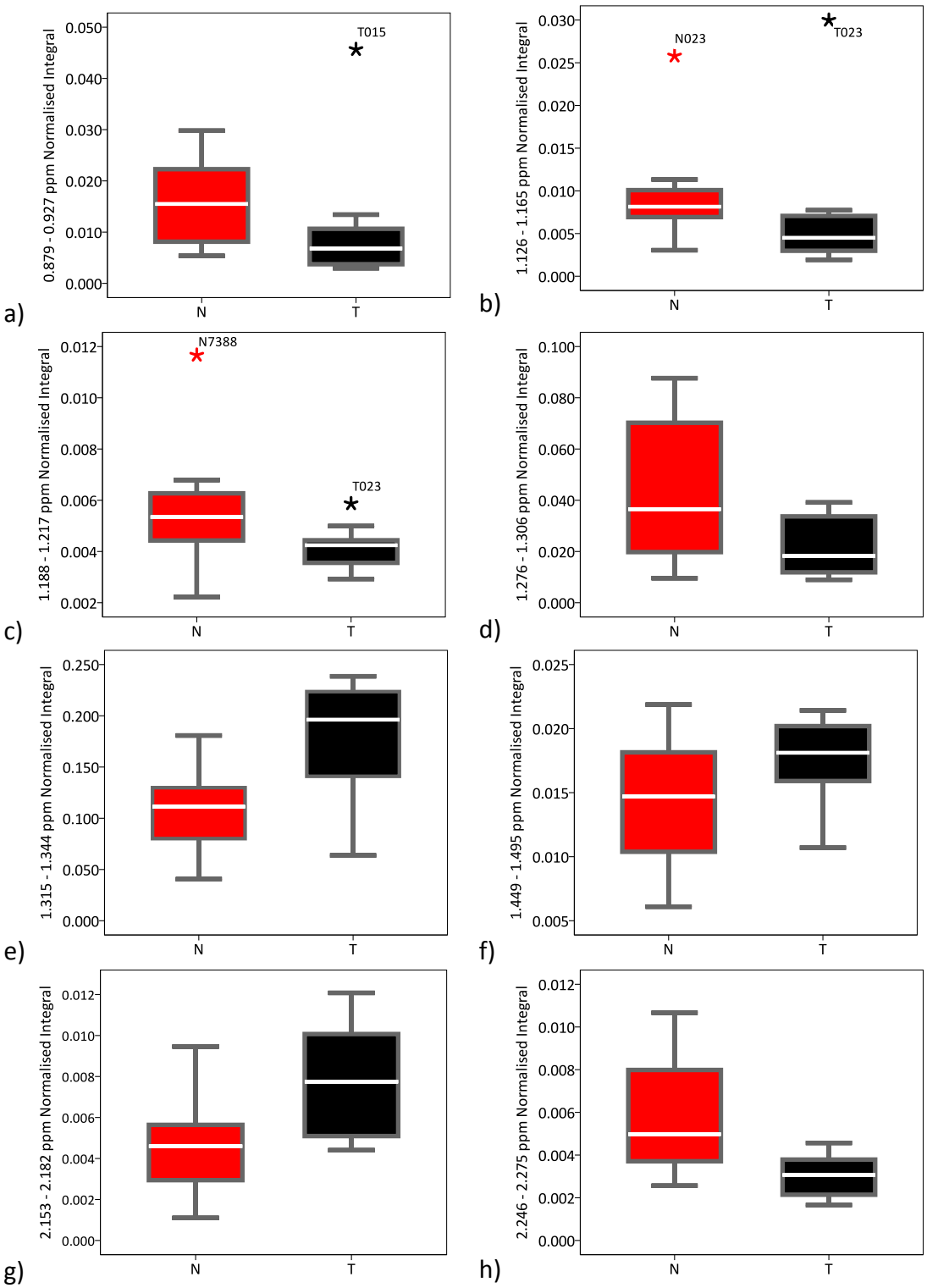


Figure 6.6 Box plots of integrals for Normal (N, red filled box) and Tumour (T, black filled box) paired samples. * = samples more extreme than 1.5 times the interquartile range beyond the upper or lower quartile, coloured as per filled box with sample label. Refer to Table 6.3 for assignments of spectral regions.

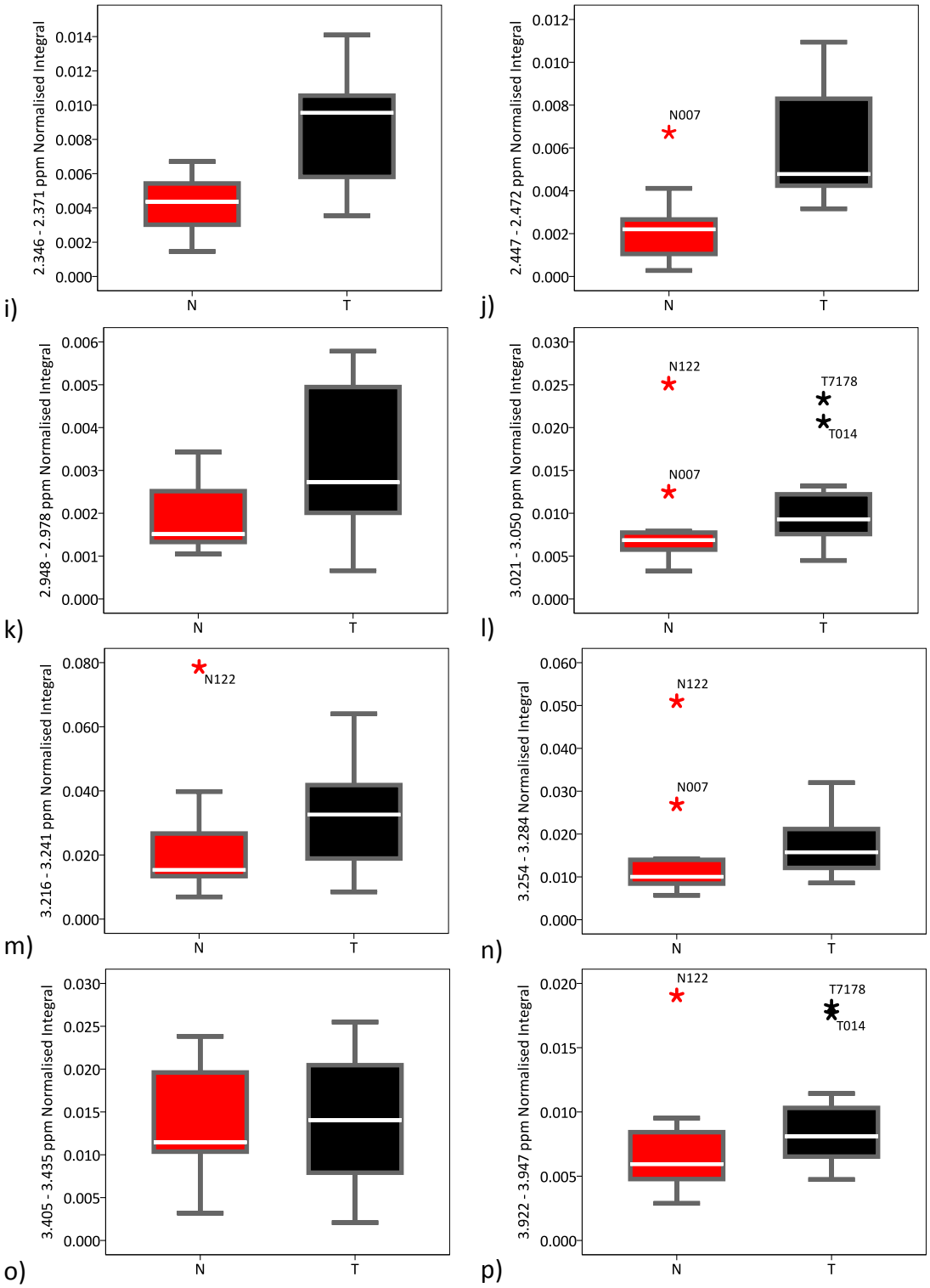


Figure 6.6 Continued.

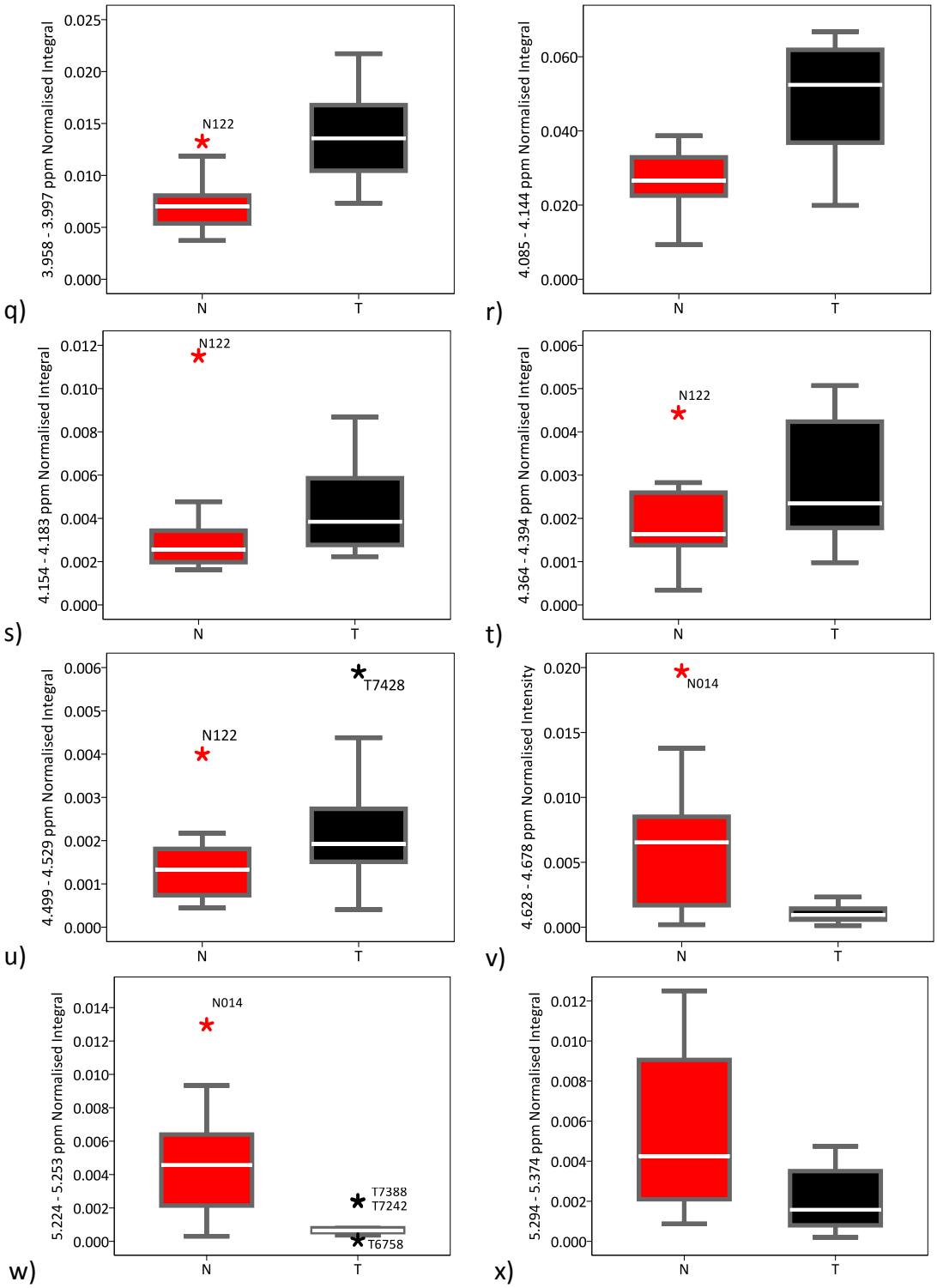


Figure 6.6 Continued.

Table 6.3 Summary of data regarding mean integral differences between tissue types from bins identified in Figure 6.5.

Identified Bin [centre] (ppm)	Bin Range Tested (ppm)	Major Metabolite	Shapiro-Wilk *p*-value (1st = N, 2nd = T)	Data Distribution	*p*-value		Mean Integral Conclusion
					Uncorrected	Corrected	
0.879-0.890 [0.885]	0.879-0.934	CH_3 of fatty acid chain	0.400, 0.000	Non-normal	0.084	0.119	No difference
0.890-0.920 [0.905]							
1.136-1.165 [1.151]	1.126-1.165	Unidentified	0.056, 0.647	Normal	0.044	0.075	No difference
1.188-1.217 [1.203]	1.188-1.217	3-hydroxybutyrate	0.058, 0.082	Normal	0.077	0.116	No difference
1.276-1.286 [1.281]	1.276-1.315	CH_2 of fatty acid chain	0.001, 0.000	Non-normal	0.034	0.063	No difference
1.286-1.297 [1.292]							
1.297-1.315 [1.306]							
1.315-1.344 [1.329]	1.315-1.344	Lactate	0.826, 0.213	Normal	0.003	0.018	↑T
1.483-1.507 [1.495]	1.439-1.507	Alanine	0.504, 0.382	Normal	0.026	0.052	No difference
2.153-2.182 [2.168]	2.153-2.182	Glutamate	0.849, 0.202	Normal	0.013	0.035	↑T
2.246-2.275 [2.261]	2.246-2.275	Lipid	0.061, 0.426	Normal	0.007	0.024	↑N
2.346-2.371 [2.359]	2.346-2.371	Glutamate	0.874, 0.452	Normal	0.002	0.016	↑T
2.447-2.472 [2.459]	2.447-2.472	Glutamine	0.041, 0.023	Non-normal	0.012	0.035	↑T
2.948-2.978 [2.963]	2.948-2.978	Unidentified	0.042, 0.443	Non-normal	0.060	0.096	No difference
3.021-3.050 [3.035]	3.021-3.050	Creatine	0.000, 0.030	Non-normal	0.182	0.190	No difference
3.216-3.241 [3.229]	3.216-3.241	GPC and PC	0.010, 0.887	Non-normal	0.158	0.181	No difference
3.254-3.284 [3.269]	3.254-3.284	Taurine	0.000, 0.400	Non-normal	0.099	0.132	No difference
3.405-3.435 [3.420]	3.405-3.435	Taurine	0.370, 0.469	Normal	0.782	0.782	No difference
3.922-3.947 [3.935]	3.922-3.947	Creatine	0.004, 0.020	Non-normal	0.182	0.190	No difference

Table 6.3 Continued.

Identified Bin [centre] (ppm)	Bin Range Tested (ppm)	Major Metabolite	Shapiro-Wilk *p*-value (1st = N, 2nd = T)	Data Distribution	*p*-value		Mean Integral Conclusion
					Uncorrected	Corrected	
3.958-3.968 [3.963]	3.958-3.997	Unidentified	0.127, 0.883	Normal	0.002	0.016	↑T
3.968-3.997 [3.983]							
4.085-4.098 [4.092]	4.085-4.144	Lactate	0.759, 0.419	Normal	0.000	0.000	↑T
4.098-4.128 [4.113]							
4.128-4.144 [4.136]							
4.154-4.183 [4.169]	4.154-4.183	Glycerophospholipid	0.000, 0.127	Non-normal	0.136	0.172	No difference
4.364-4.394 [4.379]	4.364-4.394	Glycerophospholipid	0.302, 0.180	Normal	0.149	0.179	No difference
4.499-4.529 [4.514]	4.499-4.529	Unidentified	0.040, 0.090	Non-normal	0.019	0.044	↑T
4.628-4.648 [4.638]	4.628-4.678	β-glucose	0.182, 0.554	Normal	0.007	0.024	↑N
5.224-5.253 [5.239]	5.224-5.253	α-glucose	0.452, 0.001	Non-normal	0.005	0.024	↑N
5.315-5.344 [5.330]	5.294-5.374	C*H*=C*H* of fatty acid chain	0.070, 0.135	Normal	0.020	0.044	↑N

Discussion will ensue on conclusions of mean integral analysis for those integrals where a significant difference was identified. Many of the species are connected and their relationships are summarised in Figure 6.7.

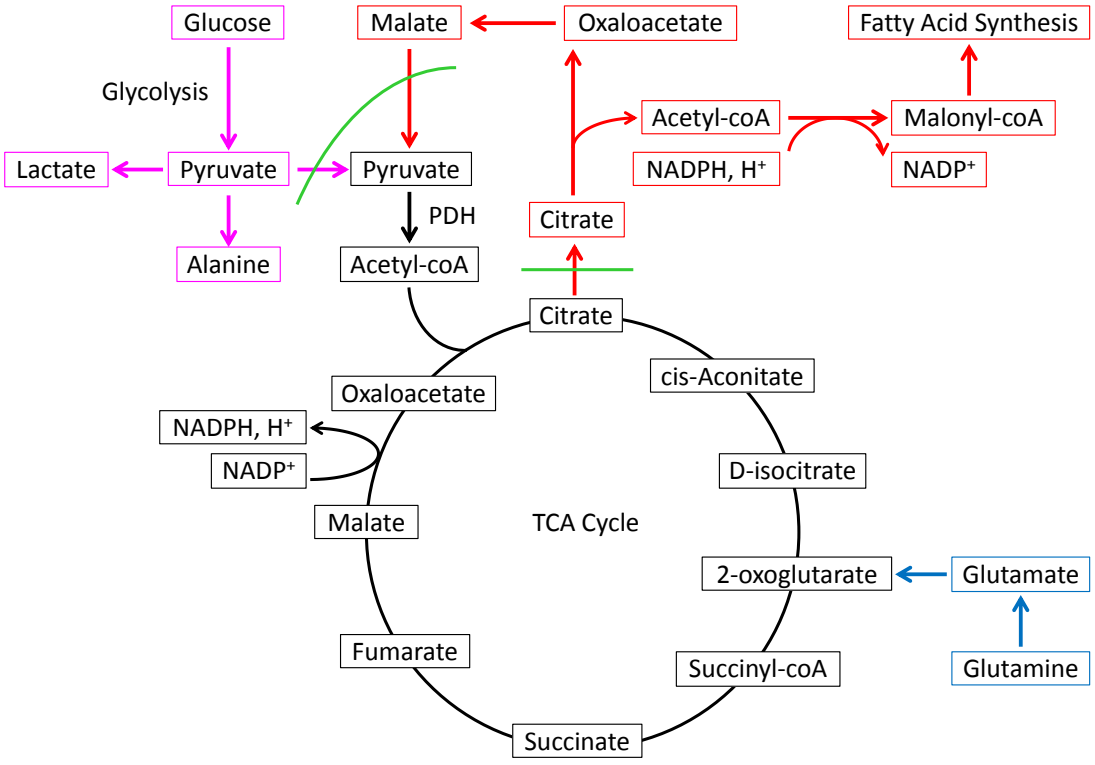


Figure 6.7 Simplified summary of relationships that include many metabolites whose levels were significantly different between Tumour and Normal samples. Green lines indicate the border of mitochondria and cytosol; pink and red boxed species are present in the cytosol. $NADP^+$ = nicotinamide adenine dinucleotide phosphate; NADPH = reduced form of $NADP^+$; PDH = pyruvate dehydrogenase; TCA = tricarboxylic acid and coA = coenzyme A.

Lactate levels were concluded to be significantly higher in Tumour samples compared to Normal samples. This is in agreement with the Warburg effect whereby glucose uptake and lactate formation are both increased.[25,147] Glucose levels were higher in Normal tissue, again supporting the Warburg effect. Additionally, the same *p*-values indicated a similar ratio of α- and β-glucose for Normal and Tissue samples. Production of lactate and alanine has been attributed as accounting for over 90% of total glucose metabolism in cancer cells.[28]

Increased alanine has been associated with cancerous tissue[135,192] though for this study after adjustment for multiple testing the *p*-value was just above the threshold for which the null hypothesis could be rejected. When cells undergo glycolysis greater levels of pyruvate result, a precursor of both alanine and lactate, but under these conditions activity of pyruvate dehydrogenase (PDH) is impaired.[135] PDH converts pyruvate to acetyl-CoA, which is part of the TCA cycle, so with reduced levels of conversion more lactate is produced[193] and more pyruvate could be converted into alanine through a transaminase reaction with glutamate that also produces 2-oxoglutarate.[118,135] Due to reduced PDH activity acetyl-CoA also derives from glutaminolysis (oxidation of glutamine).[194]

Glutamine is readily consumed by cancerous cells in addition to glucose.[28,195] A high proportion of glutamate is initially synthesised from glutamine[135] with lactate and alanine subsequently produced.[28,196] Glutamine can be converted into 2-oxoglutarate through a transaminase reaction as described previously, along with production of alanine, or by glutamate dehydrogenase (GDH). Upon implementation of the latter, malate is formed from 2-oxoglutarate after a number of intermediate metabolites and can subsequently undergo reaction with $NADP^+$ to form NADPH, the reduced version of $NADP^+$, and pyruvate, and hence lactate.[194] Production of lactate and alanine has accounted for 60% of total glutamine utilisation in cancer cells.[28] From this evidence it would be expected levels of the first two metabolites to be elevated in Tumour tissue and the opposite for glutamine and glutamate. Levels of lactate and alanine have been discussed; glutamine and glutamate, however, were also elevated in Tissue samples. An increase in glutamine and glutamate has previously been associated with breast cancer[135] and severity of meningioma,[197] a common form of brain tumour; an elevated glutamate level has been associated with tumour recurrence[198] and tumour versus normal cell lines.[199] Kung *et al.*[200] note that whilst the importance of glutamine has been shown in many cancer types, for breast cancers its prominence is not well defined. The authors provide evidence that breast tumour subtypes vary in their glutamine dependence observing luminal-type breast cancer cells were much more glutamine independent than basal-type breast cancer cells

with regards to survival. In this study 14 out of 15 Tumour samples were luminal type (13 luminal A and one unknown luminal A or B) with just one basal-type thus greatly reducing metabolic effects due to subtype variation.

Glutathione can be synthesised from glutamate,[201] and *vice-versa*,[199] and a similar trend between levels of the two metabolites have been observed with regards to tumour severity in MAS of brain tissue: increased amounts with higher cancer grade.[197] Glutathione may act both as a protective and pathogenic factor in that it is an antioxidant in cells but increased levels in tumour cells, including breast,[202] can interfere with the cytotoxic action of a number of anticancer drugs.[203] Although Monleon *et al*.[197] identified glutathione in tissue they have stated it is relatively difficult to find in tissue extracts and signals cannot be attributed to the metabolite in this study.

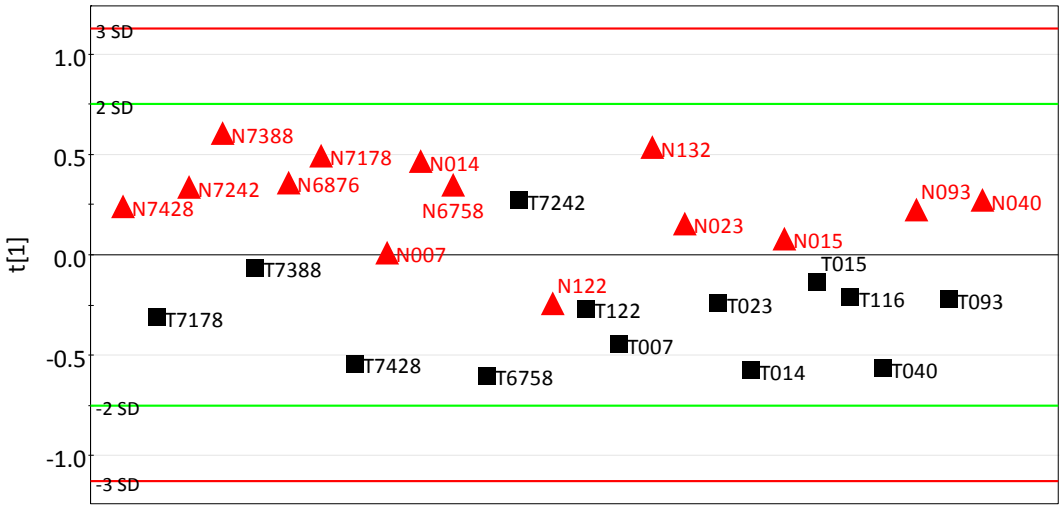
Citrate is formed in the TCA cycle from the reaction of acetyl-CoA and oxaloacetate (OAA). Much citrate reforms OAA and acetyl-CoA in the cytosol.[194] Fatty acids are synthesised from acetyl-CoA and malonyl-CoA in the presence of NADPH.[204] Due to the high level of lactate produced from glutamate sufficient NADPH is generated to enable fatty acid synthesis.[28] OAA can lead to the formation of pyruvate *via* malate, which generates further NADPH and allows the cycle to continue.[194] Due to much citrate leaving the mitochondria the TCA cycle would be depleted but glutamate can contribute to formation of metabolites within the cycle further to previously described. Instead of conversion to pyruvate from glutamate derived malate, OAA can be produced thus providing an essential metabolite for the synthesis of fatty acids.

Lipid levels at 2.261 ppm and 5.330 ppm were adjudged to be higher in Normal samples. Healthy tissue samples have a higher level of adipocytes than Tumour tissue samples, which are mainly composed of epithelial tissue.[135] It could be expected to observe increased levels of some lipids in samples that contain a higher amount of fat. Changes in metabolite levels have been observed between healthy samples that had different amounts of fatty tissue and fibrotic tissue.[135] Histology

records were not available for the samples but notes made related to appearance during the homogenisation process were assessed relative to lipid level in the aforementioned regions. The majority of samples were pink or light orange with some primarily dark (black/brown/grey) and others mainly white. No correlation was observed between lipid amounts and any of the groups.

Alanine, CH_2 of fatty acid chain and an unassigned resonance in the 1.126-1.165 ppm bin displayed significantly different levels before multiple test correction but not afterwards. With an increased number of tests there is an increased chance of false outcomes; if fewer tests were performed the multiple test correction applied would be less severe.[92]

PLS-DA was also applied, the scores and loadings plots are shown in Figure 6.8 and Figure 6.9, respectively.



Key: ■ Tumour and ▲ Normal

Figure 6.8 PLS-DA scores for aqueous extract samples excluding T132. Classed according to tumour type.

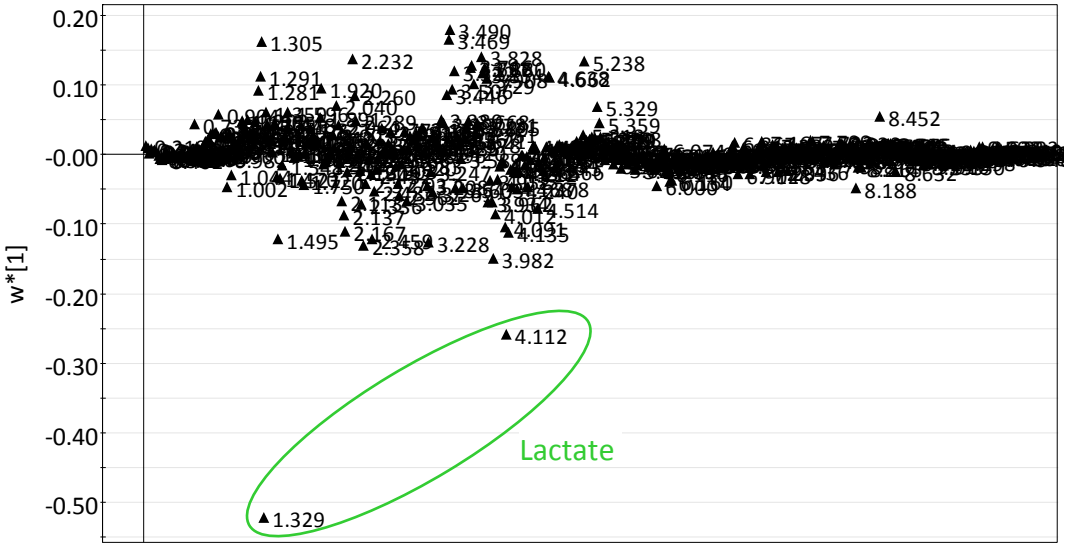


Figure 6.9 PLS-DA loadings plot corresponding to the model displayed in Figure 6.8.

The one component model ($R^2\boldsymbol{X}$ = 0.275, $R^2\boldsymbol{Y}$ = 0.613 and $Q^2\boldsymbol{Y}$ = 0.487) separated the majority of Tumour and Normal samples from each other. The bins that contained lactate, centred at 1.329 and 4.112 ppm, were the two most influential in the loadings plot (Figure 6.9) with the bin centred at 3.228 ppm also contributing to the separation, all three of which were indicated to be raised in Tumour samples. PLS-DA can overfit data so validation is required.[79] 'Leave-one-out' cross-validation was performed using 27 samples with unpaired samples predicted singly and paired samples predicted together. Table 6.4 summarises the results.

Table 6.4 'Leave-one-out' cross-validation parameters of the PLS-DA model shown in Figure 6.8.

Sample Excluded	Number of Components	R^2X (cum)	R^2Y (cum)	Q^2Y (cum)	*Y*-Predicted*	
					N	T
N007	1	0.261	0.647	0.456	0.415	0.585
T007	1				0.104	0.896
N014	1	0.256	0.594	0.383	1.000	0.000
T014	1				-0.017	1.017
N023	1	0.254	0.648	0.470	0.693	0.307
T023	1				0.384	0.616
N040	1	0.244	0.624	0.420	0.768	0.232
T040	1				-0.164	1.164
N093	1	0.266	0.620	0.457	0.681	0.319
T093	1				0.298	0.702
N122	1	0.271	0.668	0.523	0.137	0.863
T122	1				0.288	0.712
N6758	2	0.417	0.779	0.530	0.307	0.693
T6758	2				-0.161	1.161
N7178	1	0.238	0.635	0.460	0.996	0.004
T7178	1				0.174	0.826
N7242	1	0.274	0.672	0.568	0.791	0.209
T7242	1				0.983	0.017
N7388	1	0.246	0.648	0.502	1.146	-0.146
T7388	1				0.453	0.547
N7428	1	0.251	0.613	0.426	0.689	0.311
T7428	1				-0.028	1.028
N015	1	0.273	0.663	0.533	0.556	0.444
T015	1				0.511	0.489
T116	1	0.251	0.630	0.460	0.298	0.702
N132	1	0.260	0.603	0.379	1.060	-0.060
N6876	1	0.449	0.744	0.405	0.935	0.065

*A sample was regarded as belonging to a grade by having a ***Y***-predicted value >0.50. Incorrect classification is represented by red shading and correct by pink, orange or green, corresponding to a ***Y***-predicted value of <0.60, 0.60-0.70 or >0.70, respectively.

Of the 14 Normal samples, 11 (79%) were predicted correctly whilst the predictive ability of Tumour samples was 85% due to having 11 of 13 samples correctly predicted. Both percentages were high, which indicated the classes were modelled

well and the data was not overfitted. Of those samples correctly predicted, 64% of Normal samples had a ***Y***-predicted value greater than 0.70 and for Tumour samples 82% achieved this figure. The accuracy of class prediction is reduced for samples as the ***Y***-predicted value tends towards 0.50 so classifications based on values less than 0.60, coloured pink in Table 6.4, should be treated with caution: both N015 and T7388 samples had low modulus ***t***[1] values. Given the positioning of N122 and T7242, both located in scores space that was predominantly occupied by the opposite sample type, incorrect class prediction of the two samples was not unexpected.

Permutation testing was also performed for further validation of the model. Figure 6.10 shows all but two of the 999 permuted models for the Tumour class to have lower $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ values than the original model. The intercept values of the regression lines for $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ are 0.239 and -0.148, respectively. The permutation testing plot for the Normal class (plot not shown) displayed $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ intercept values of 0.240 and -0.146, respectively, with no permuted models having higher values. Both $R^2\boldsymbol{Y}$ values are less than 0.4 and both $Q^2\boldsymbol{Y}$ values are below 0.05 so it is indicated that the model is validated thus furthering the conclusion from 'leave-one-out' cross validation.

Validation of the PLS-DA model supported the observation that Tumour samples can be distinguished from Normal tissue using MVA.

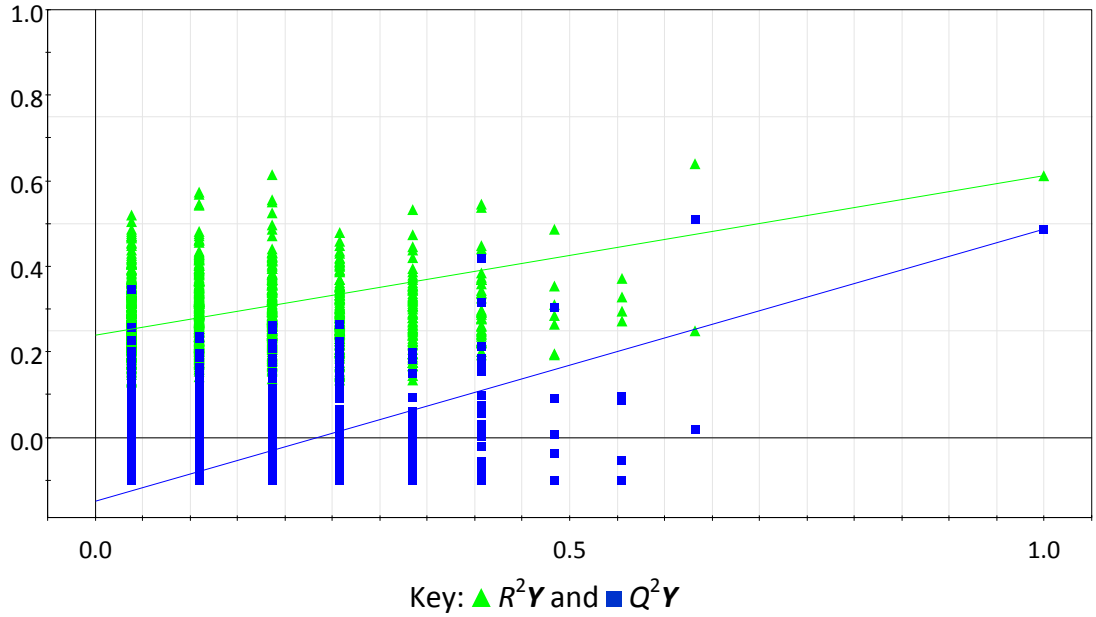


Figure 6.10 Permutation testing plots for the Tumour class in the PLS-DA model shown in Figure 6.8. The R^2Y and Q^2Y intercept values of the regression lines are 0.239 and -0.148, respectively.

6.1.1.1.2 Analysis Implementing Probabilistic Quotient Normalisation and Removal of Sample N122

Using sum normalisation N122 was observed on eight occasions as more extreme than 1.5 times the interquartile range beyond the upper quartile of Normal samples in Figure 6.6**.** Sample preparation and data acquisition records were investigated but no anomalous factors were apparent. Visual inspection of spectra showed N122 to contain very little glucose, which was surprising since glucose levels were generally higher in Normal compared to Tumour samples (Table 6.3). The next lowest Normal sample glucose level, as measured by β-glucose signal integral, was nearly five times that of N122. Glucose is prevalent in spectra so for samples with a low level, normalisation is affected resulting in increased integrals for all other signals. Ascertaining whether differences for this sample were due to genuine variation of metabolite levels or the effect of normalisation was difficult. Using the bin containing GPC and PCho as an example, N122's normalised integral value was more than 20% larger than the next greatest irrespective of sample type and almost twice the value of the second largest Normal sample integral. N122's integral was

more than nine-fold larger than that of the paired Tumour sample. Given the potential problem using sum normalisation, PQN was used.

PQN has been advocated as superior to constant sum normalisation[69] but the method has been shown to be not always the optimum choice when many normalisation methods were compared,[158] hence why it was not used from the beginning of analysis. Section 8.3.1.2 describes the PQN methodology and Table 8.7 lists regions included in the calculation of the median quotient. The scores and loadings plots were very similar using both normalisation methods as shown by PC 1 in Figure 6.11 and Figure 6.12, respectively. A three component model was produced in each case with $R^2\boldsymbol{X}$(cum) values of 0.594 and 0.602 for sum normalisation and PQN, respectively, and related $Q^2\boldsymbol{X}$(cum) values of 0.337 and 0.307.

Although the scores and loadings plots generated using the two different normalisation methods were of similar appearance PQN was employed due to potential weakness of sum normalisation: the inability to account for presence or absence of a large signal(s) in certain samples without affecting normalised intensities of other signals of that sample, in this case the much reduced presence of glucose signals in N122.

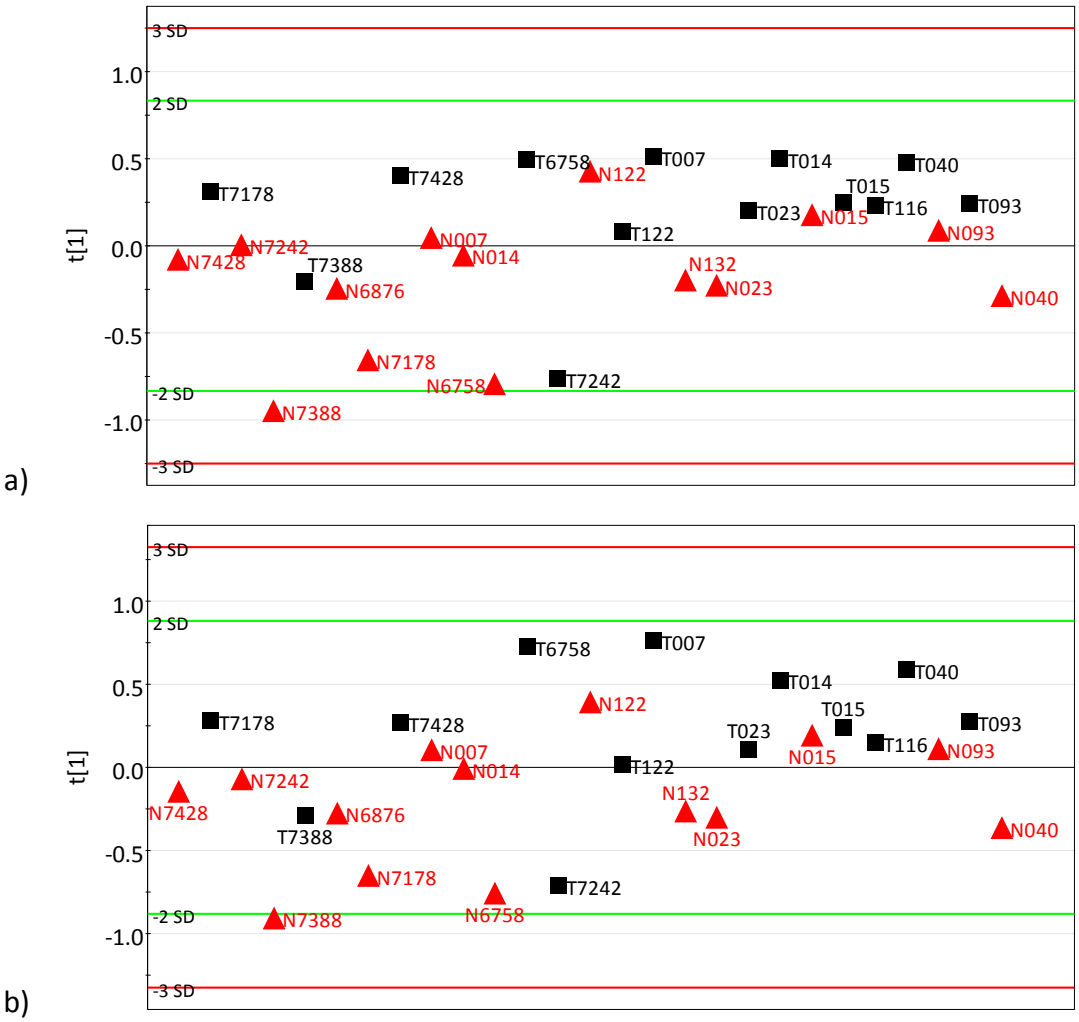


Figure 6.11 PCA scores plot of PC 1 coloured according to tissue type for aqueous extract samples using a) constant sum normalisation (0.311, 0.171) and b) PQN (0.311, 0.174); R^2X and Q^2X values, respectively, in parentheses.

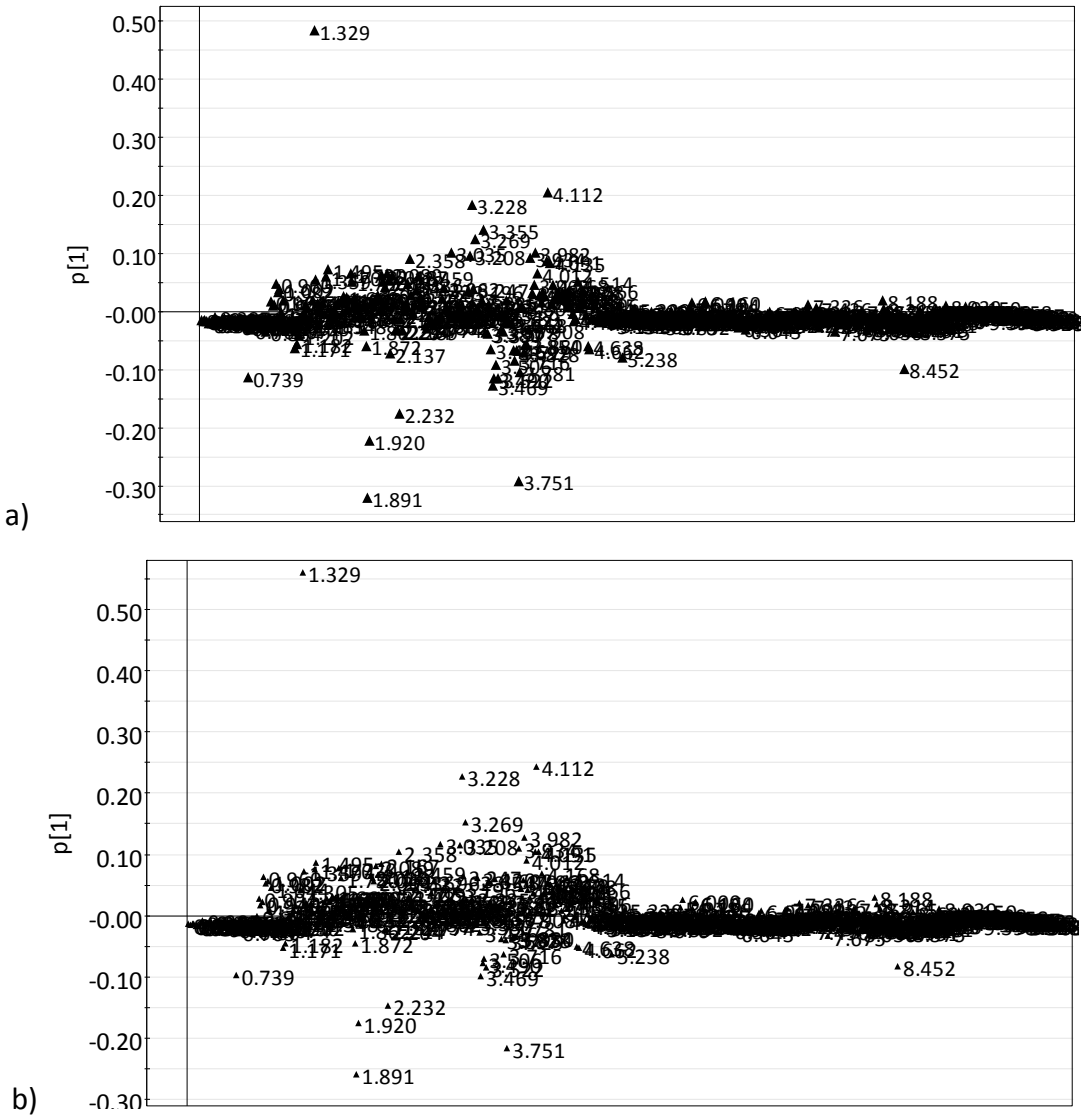


Figure 6.12 Loadings plot of PC 1 for aqueous extract samples using a) constant sum normalisation and b) PQN.

Quotients of spectra relative to the median (reference) spectrum varied between 0.74 and 1.25 with a value of 0.97 for N122. After multiple test correction the same conclusion regarding difference between mean integral values of sample types applied to 22 of the 24 bin regions tested including 10 of the 11 bin regions previously identified as significantly different (Table 6.5). Alanine had a *p*-value >0.05 when constant sum normalisation was used but upon implementation of PQN it could be concluded a significantly greater concentration was present in Tumour samples, however, the reverse was applicable to C*H*=C*H* of fatty acid chains (5.294-5.374 ppm) with the level no longer being significantly greater in Tumour samples.

Due to the small sample set of 12 pairs, N122 and T122 were removed from analysis because one paired sample could greatly influence univariate analysis results. PQ normalised data was retested for normality of distribution and pair-wise testing applied as previously. Ten bin region mean integral values that were significantly different between sample types when samples N and T 122 were included remained so with only 2.246-2.275 ppm, the signal assigned to lipids, no longer significantly different (Table 6.5). Seven additional bin regions were concluded to be significantly different and contained an unidentified resonance (2.948-2.978 ppm), creatine (3.021-3.050 ppm and 3.922-3.947 ppm), GPC and PCho (3.216-3.241 ppm), taurine (3.254-3.284 ppm) and glycerophospholipid (4.154-4.169 ppm and 4.364-4.394 ppm). Levels were increased in Tumour compared to Normal samples for all bins. Metabolic reasoning for the findings will be discussed.

Table 6.5 Summary of data regarding mean integral difference between tissue types for different normalisation methods and after exclusion of a sample using PQN; sum normalised data, also shown in Table 6.3, has been included to aid comparison. Non = non-normal, ND = no difference, N = Normal, T = Tumour and ↑ = significant increase.

Identified Bin [centre] (ppm)	Bin Range Tested (ppm)	Major Metabolite	Data Distribution			Corrected *p*-value			Mean Integral Conclusion		
			Sum	PQN	Exc 122	Sum	PQN	Exc 122	Sum	PQN	Exc 122
0.879-0.890 [0.885]	0.879-0.934	CH_3 of fatty acid chain	Non	Non	Non	0.119	0.284	0.328	ND	ND	ND
0.890-0.920 [0.905]											
1.136-1.165 [1.151]	1.126-1.165	Unidentified	Normal	Non	Non	0.075	0.176	0.052	ND	ND	ND
1.188-1.217 [1.203]	1.188-1.217	3-hydroxybutyrate	Normal	Normal	Normal	0.116	0.214	0.219	ND	ND	ND
1.276-1.286 [1.281]	1.276-1.315	CH_2 of fatty acid chain	Non	Non	Non	0.063	0.181	0.208	ND	ND	ND
1.286-1.297 [1.292]											
1.297-1.315 [1.306]											
1.315-1.344 [1.329]	1.315-1.344	Lactate	Normal	Normal	Normal	0.018	0.032	0.032	↑T	↑T	↑T
1.483-1.507 [1.495]	1.439-1.507	Alanine	Normal	Normal	Normal	0.052	0.033	0.032	ND	↑T	↑T
2.153-2.182 [2.168]	2.153-2.182	Glutamate	Normal	Normal	Non	0.035	0.043	0.032	↑T	↑T	↑T
2.246-2.275 [2.261]	2.246-2.275	Lipid	Normal	Normal	Non	0.024	0.043	0.052	↑N	↑N	ND
2.346-2.371 [2.359]	2.346-2.371	Glutamate	Normal	Normal	Normal	0.016	0.030	0.012	↑T	↑T	↑T
2.447-2.472 [2.459]	2.447-2.472	Glutamine	Non	Non	Non	0.035	0.033	0.032	↑T	↑T	↑T
2.948-2.978 [2.963]	2.948-2.978	Unidentified	Non	Normal	Non	0.096	0.092	0.032	ND	ND	↑T
3.021-3.050 [3.035]	3.021-3.050	Creatine	Non	Non	Normal	0.190	0.181	0.049	ND	ND	↑T
3.216-3.241 [3.229]	3.216-3.241	GPC and PCho	Non	Non	Non	0.181	0.144	0.032	ND	ND	↑T
3.254-3.284 [3.269]	3.254-3.284	Taurine	Non	Non	Non	0.132	0.181	0.049	ND	ND	↑T

Table 6.5 Continued.

Identified Bin [centre] (ppm)	Bin Range Tested (ppm)	Major Metabolite	Data Distribution			Corrected *p*-value			Mean Integral Conclusion		
			Sum	PQN	Exc 122	Sum	PQN	Exc 122	Sum	PQN	Exc 122
3.405-3.435 [3.420]	3.405-3.435	Taurine	Normal	Normal	Non	0.782	0.730	0.298	ND	ND	ND
3.922-3.947 [3.935]	3.922-3.947	Creatine	Non	Non	Normal	0.190	0.176	0.049	ND	ND	↑T
3.958-3.968 [3.963]	3.958-3.997	Unidentified	Normal	Normal	Normal	0.016	0.030	0.012	↑T	↑T	↑T
3.968-3.997 [3.983]											
4.085-4.098 [4.092]	4.085-4.144	Lactate	Normal	Normal	Normal	0.000	0.024	0.024	↑T	↑T	↑T
4.098-4.128 [4.113]											
4.128-4.144 [4.136]											
4.154-4.183 [4.169]	4.154-4.183	Glycerophospholipid	Non	Non	Non	0.172	0.181	0.049	ND	ND	↑T
4.364-4.394 [4.379]	4.364-4.394	Glycerophospholipid	Normal	Normal	Normal	0.179	0.181	0.049	ND	ND	↑T
4.499-4.529 [4.514]	4.499-4.529	Unidentified	Non	Non	Normal	0.044	0.050*	0.024	↑T	↑T	↑T
4.628-4.648 [4.638]	4.628-4.678	β-glucose	Normal	Non	Non	0.024	0.032	0.024	↑N	↑N	↑N
5.224-5.253 [5.239]	5.224-5.253	α-glucose	Non	Non	Non	0.024	0.030	0.024	↑N	↑N	↑N
5.315-5.344 [5.330]	5.294-5.374	C*H*=C*H* of fatty acid chain	Normal	Non	Non	0.044	0.092	0.074	↑N	ND	ND

*p-value <0.05 when more than three decimal places retained.

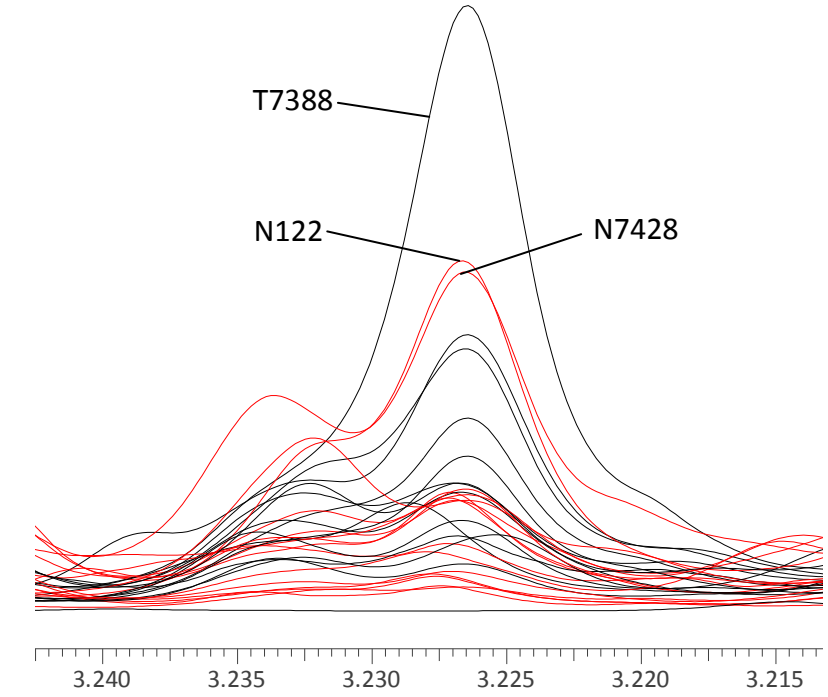
The creatine level was significantly elevated in Tumour samples when samples N and T 122 were not included in analysis (Table 6.5). HR-MAS of tissue showed raised creatine levels in head and neck squamous cell carcinoma compared to normal adjacent tissue.[205] This observation was supported by reduction of creatine kinase in oral squamous cell carcinoma.[206] The enzyme catalyses the reversible process of phosphorylation of creatine to creatine phosphate so a reduction in the enzyme level would inhibit the process in this direction thus creatine, as the reverse direction product, would be expected to be present in a greater quantity.[207]

The same conclusion as for creatine was made regarding the bin at 3.254-3.284 ppm that contained taurine. Taurine has been shown to be increased in breast tumour samples compared to non-tumour samples.[135,141] Taurine has been postulated as being an antioxidant[208,209] and marker of increased cell proliferation.[135] Glucose was also present in the downfield bin (3.405-3.435 ppm) that contained taurine. This observation could explain why a significant difference in mean integrals was not observed for this bin; increased glucose in Normal samples could be counteracted by up regulation of taurine in Tumour samples.

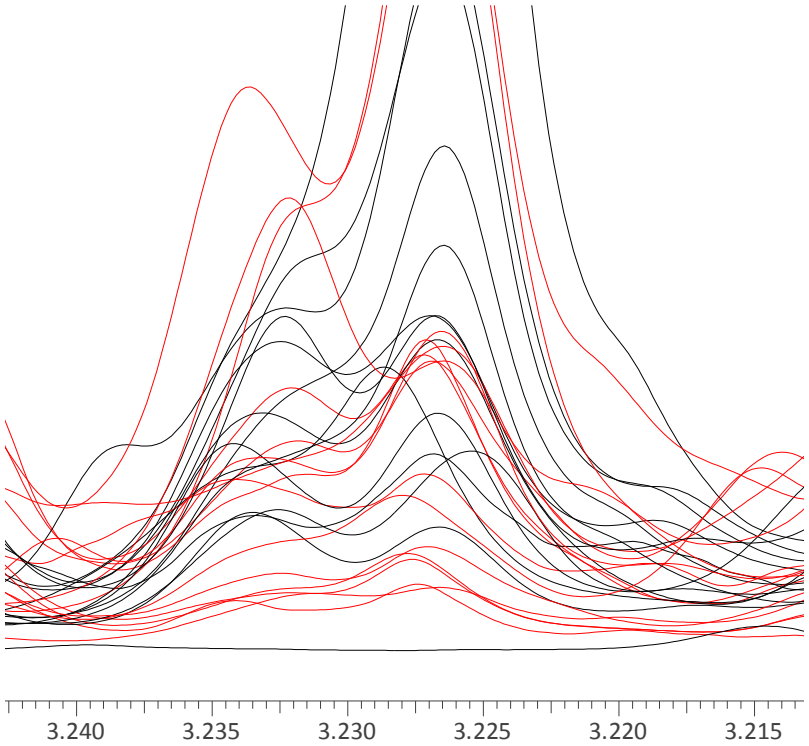
Altered choline metabolism is indicative of cancer but the role of choline-containing metabolites is poorly understood.[119,210] In this study the mean integral level for the bin 3.216-3.241 ppm containing GPC and PCho (Table 6.5) was significantly increased in Tumour samples when samples N and T 122 were not included in the analysis. GPC has a higher chemical shift than PCho.[118,188,211] Elevations in total choline metabolites[210,212] and PCho[134,210,212] have been observed though both an increase[141] and decrease[213] in GPC have been reported and consensus cannot be achieved regarding the GPC to PCho ratio. Breast cancer tissue and cells have exhibited a decrease in the ratio[141,213] and Stewart *et al*.[210] cite this ratio change as indicative of cancer whereas Moestue *et al*.[214] argue the opposite. It has been postulated that the micro-environment of the tumour affects choline metabolism[212] such that *in vivo* tumours could evolve further mechanisms compared to *in vitro* tumours.[215] Due to the aforementioned region containing two overlapping signals assessment of the mean values between the tissue sample types

for GPC and PCho cannot be performed. Figure 6.13a shows N122 and N7428 to have very similar levels of the greater intensity signal (GPC) but different amounts of the lower intensity signal (PCho) whilst for T7388 it is difficult to establish the two separate signals. Figure 6.13b shows more clearly the composition of the signals within the region for the majority of samples but it cannot be ascertained whether the ratio trend of the two signals is different between Normal and Tumour samples nor can it be established whether the increased Tumour sample mean integral of the region was due to an integral value change in both signals or just one.

Betaine production and phosphatidylcholine (PDC) synthesis are the two major pathways of intracellular choline metabolism.[212] Betaine is further synthesised to glycine[119] and subsequently, in order, to glutathione, glutamate and glutamine,[199] the last two of which were identified in spectra and raised in Tumour samples. PCho is an intermediary of PDC production and GPC and free fatty acids are degradation products of PDC.[119] PDC, a glycerophospholipid, is the most abundant membrane-forming phospholipid in cells.[119,216]



a)



b)

Figure 6.13 ACD normalised view of ^{1}H-NMR spectrum section of aqueous extracts including 3.216-3.241 ppm region, which contains GPC (most downfield signal) and PCho signals. a) Vertical integral range to include all samples and b) lower vertical integral range to emphasise the majority of samples. Spectral line of sample: red = Normal, black = Tumour.

The signals tentatively assigned as glycerophospholipids[187] in bins 4.154-4.183 ppm and 4.364-4.394 ppm were elevated for Tumour samples when samples N and T 122 were excluded. Given increased choline metabolites and GPC in tumour tissue and the relationship with glycerophospholipids an explanation is provided for the aforementioned observation.

6.1.1.1.3 Analysis Excluding Lactate

Lactate is prevalent in Tumour samples (Figure 6.1) and can strongly influence positions of samples in scores space. Scaling can be used to reduce the disparity in influence signals of different intensities have. Pareto has been used in these analyses as the *de facto* scaling method but MVA is still strongly influenced by large intensity bins. Non-normalised data from lactate containing bins (1.314-1.374 ppm and 4.085-4.144 ppm) was removed before PQN was employed.

The scores (Figure 6.14) and loadings plots (Figure 6.15) for PC 2 versus PC 1 of the three component model ($R^2\boldsymbol{X}$(cum) = 0.584 and $Q^2\boldsymbol{X}$(cum) = 0.251) were very similar in appearance (vertical reflection excluded) to those for which lactate was included (Figure 6.4 and Figure 6.5, respectively).

Scores and loadings plots (data not shown) associated with the one component PLS-DA model ($R^2\boldsymbol{X}$ = 0.217, $R^2\boldsymbol{Y}$ = 0.541 and $Q^2\boldsymbol{Y}$ = 0.362) were very similar to the equivalent that included lactate (Figure 6.8 and Figure 6.9, respectively).

Further information in relation to metabolites differing between Tumour and Normal samples was not gained upon exclusion of lactate from MVA.

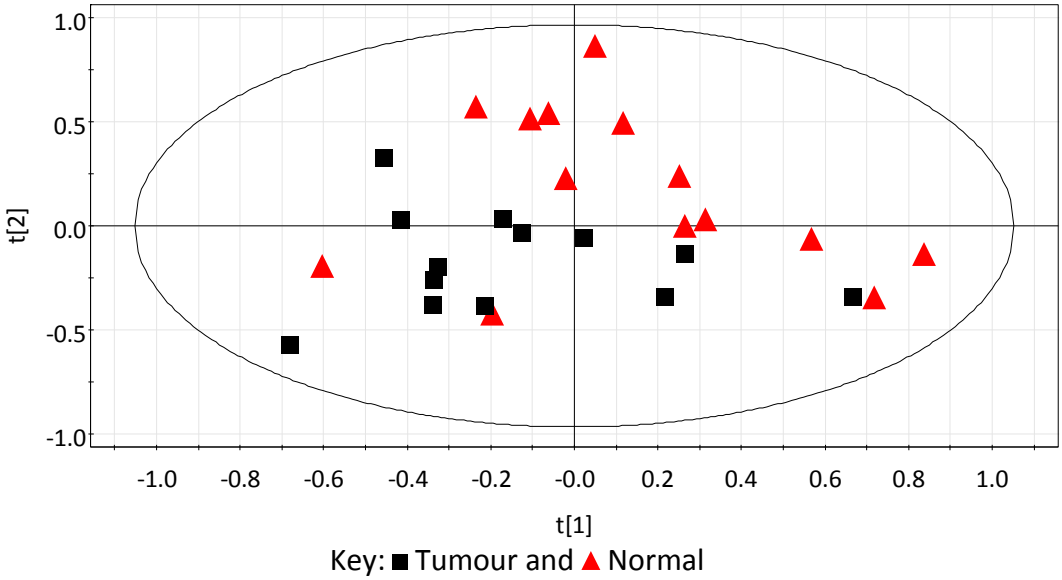


Figure 6.14 PCA scores plot coloured according to tissue type for aqueous extract samples excluding lactate regions showing the first two model components. R^2X = 0.251 and 0.212, Q^2X = 0.039 and 0.041 for PC 1 and PC 2, respectively.

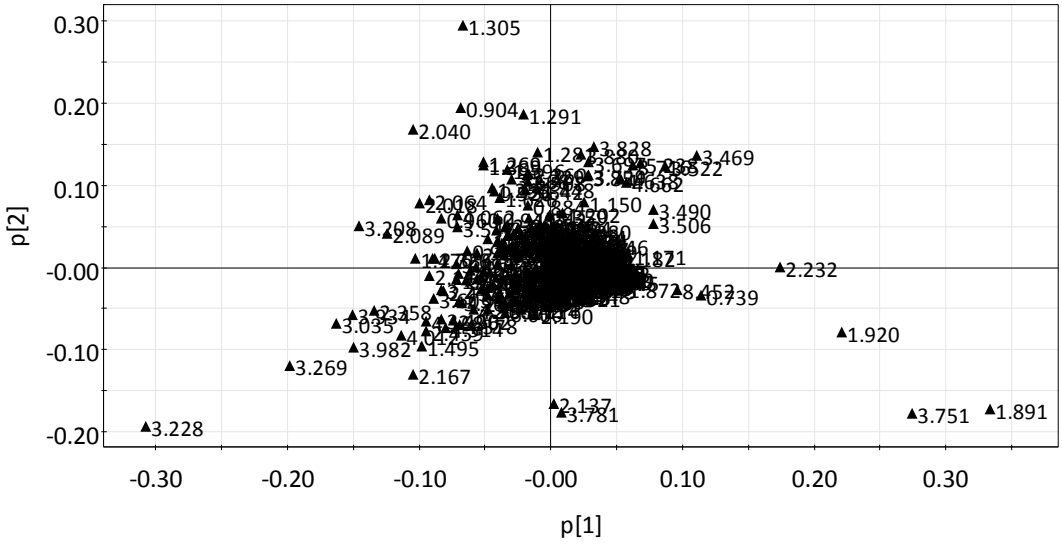


Figure 6.15 PCA loadings plot corresponding to the model displayed in Figure 6.14.

6.1.1.1.4 Analysis of Aromatic Region Only

Figure 6.1a shows the spectrum of aqueous extracts is dominated by the aliphatic region and signals in the aromatic region are of much lower intensity. Beckonert *et al.*[135] reported differences between Tumour and Normal samples in the aromatic region of aqueous extracts of breast cancer tissue. In this study visual differences were observed but no bins were identified by MVA when considering the whole

spectrum. To circumvent the scaling issue referred to in Section 6.1.1.1.3 only the aromatic region (5.446-9.362 ppm) was analysed.

The majority of Tumour and Normal samples were separated along PC 1 (Figure 6.16) in the two PC model ($R^2\boldsymbol{X}$(cum) = 0.331 and $Q^2\boldsymbol{X}$(cum) = 0.007) with the same two Normal samples, N007 and N122, located amongst the majority of Tumour samples as per inclusion of the aliphatic region (Figure 6.4). The loadings plot (Figure 6.17) indicated levels of many metabolites were raised in the aforementioned grouping of samples with a corresponding decrease in formate (8.452 ppm) and a singlet at 7.675 ppm tentatively assigned as pyridoxine.[191] Four Tumour samples were isolated from the majority of Tumour samples, occupying similar PC 1 scores space to Normal samples. The reverse was true regarding levels of metabolites for this grouping.

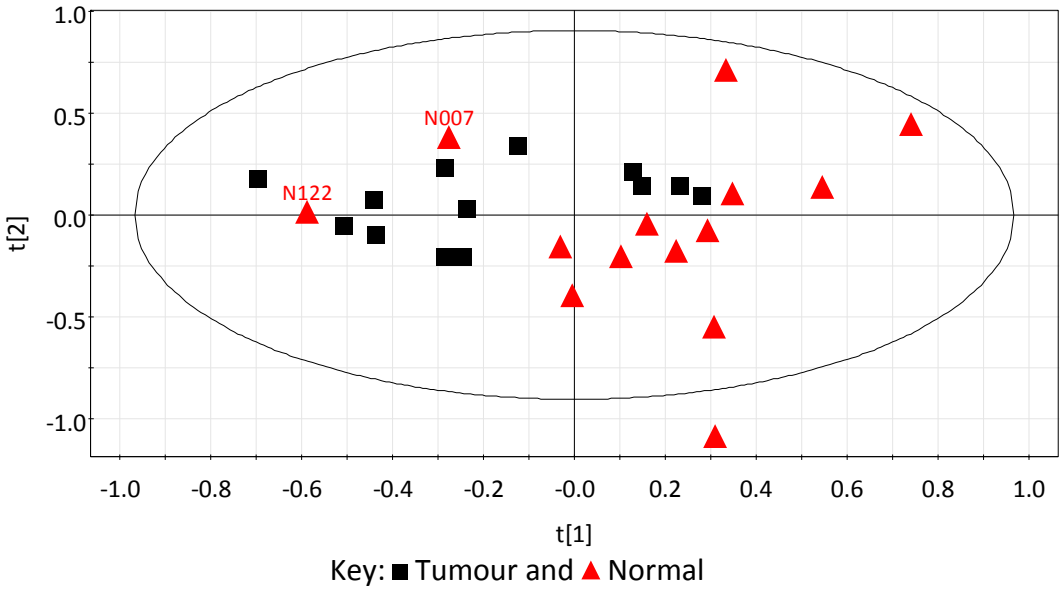


Figure 6.16 PCA scores plot coloured according to tissue type for the sum normalised aromatic region of aqueous extracts.

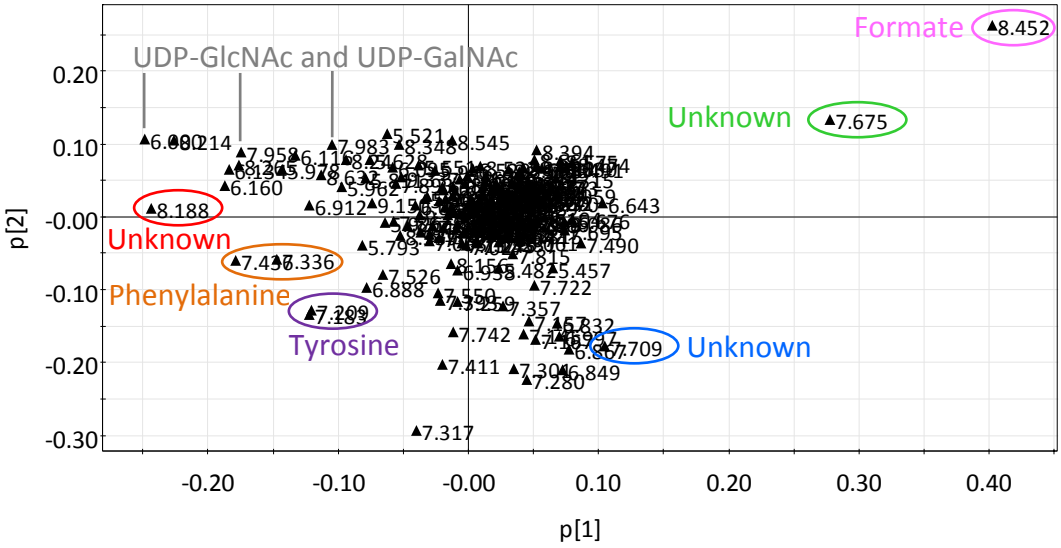


Figure 6.17 PCA loadings plot corresponding to the model displayed in Figure 6.16.

Metabolite levels were investigated within bins that possessed the greatest modulus PC 1 loadings values in Figure 6.17. Bins with a modulus ***p***[1] value equal to or greater than 0.105 were analysed: inspection of spectra showed the bin centred at 7.709 ppm contained distinguishable signals for only four samples, this was deemed an insufficient number upon which conclusions could be based. Bins with a ***p***[1] value at least equal to that of the aforementioned bin were investigated to ensure a signal(s) was present for numerous samples before univariate analysis was performed and all bins identified conformed. In total 14 bin regions were used in analysis. Box plots in Figure 6.18 provide an overview of paired sample data whilst Table 6.6 displays the bins identified as having potentially different mean integrals and conclusion regarding whether the difference was statistically significant. Adenosine 3',5'-diphosphate,[103] pyridoxine[191] and UDP-GlcNAc and UDP-GalNAc[189,190] have been tentatively assigned.

When all 24 paired samples were included no bin mean integral values were significantly different between Tumour and Normal samples once adjustment for multiple testing was performed. The box plots showed N122 to be more extreme than 1.5 times the interquartile range beyond the upper quartile on four occasions. Previously, when the whole spectrum was considered analysis was performed that excluded the sample, and the paired Tumour sample, because the same observation

was noted on multiple occasions (Figure 6.6) and was attributed to the sample containing a very low level of glucose. The paired 122 samples were also excluded in aromatic region only analysis and results incorporated into Table 6.6. Resultantly, UDP-GlcNAc and UDP-GalNAc in the regions 5.973-6.015 ppm and 6.108-6.124 ppm and unidentified signals in 8.178-8.200 ppm and 8.200-8.229 ppm regions were significantly increased in Tumour samples.

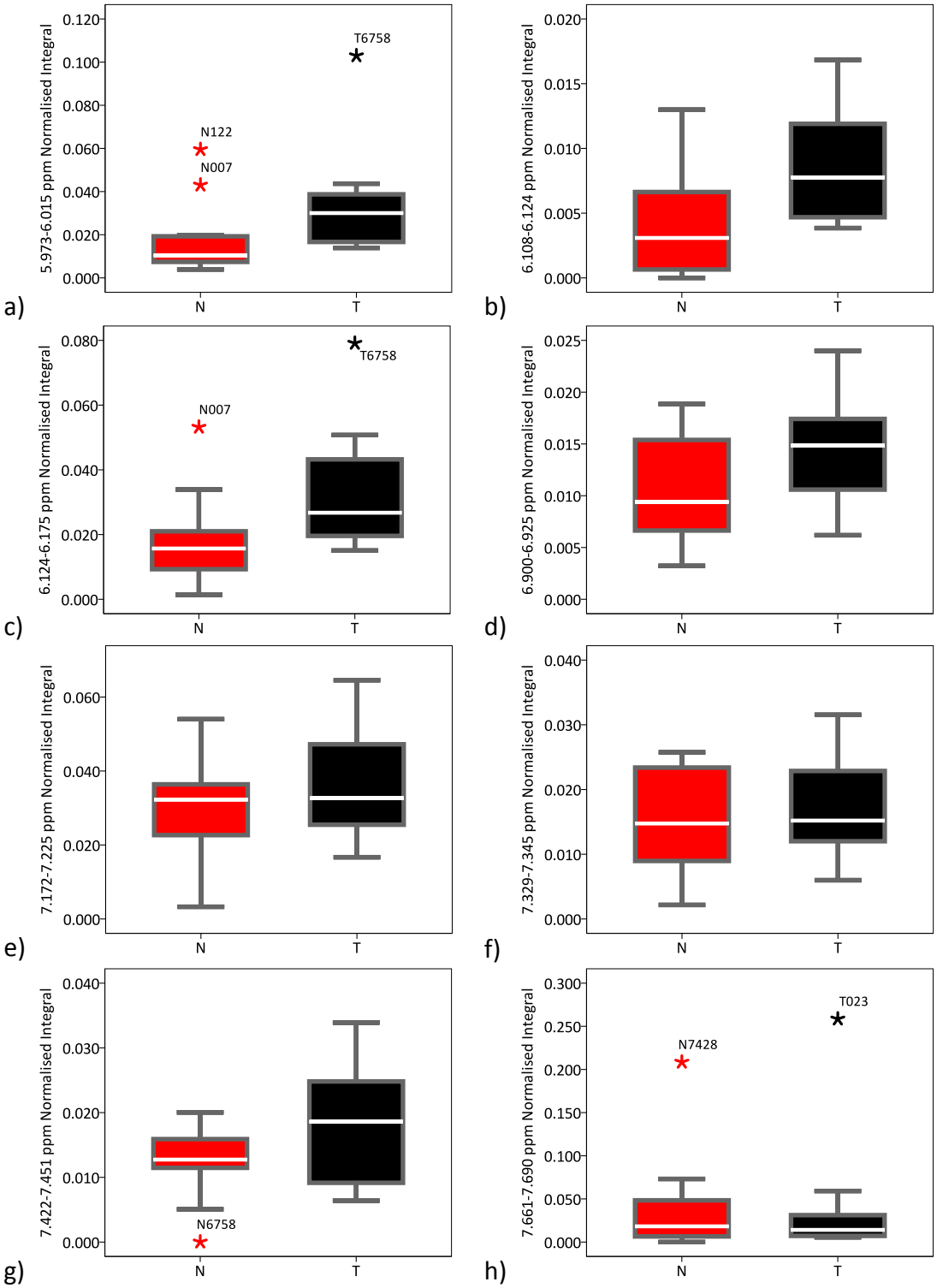


Figure 6.18 Box plots of aromatic region integrals for Normal (N, red filled box) and Tumour (T, black filled box) paired samples. * = samples more extreme than 1.5 times the interquartile range beyond the upper or lower quartile, coloured as per filled box with sample label. Refer to Table 6.6 for assignments of spectral regions.

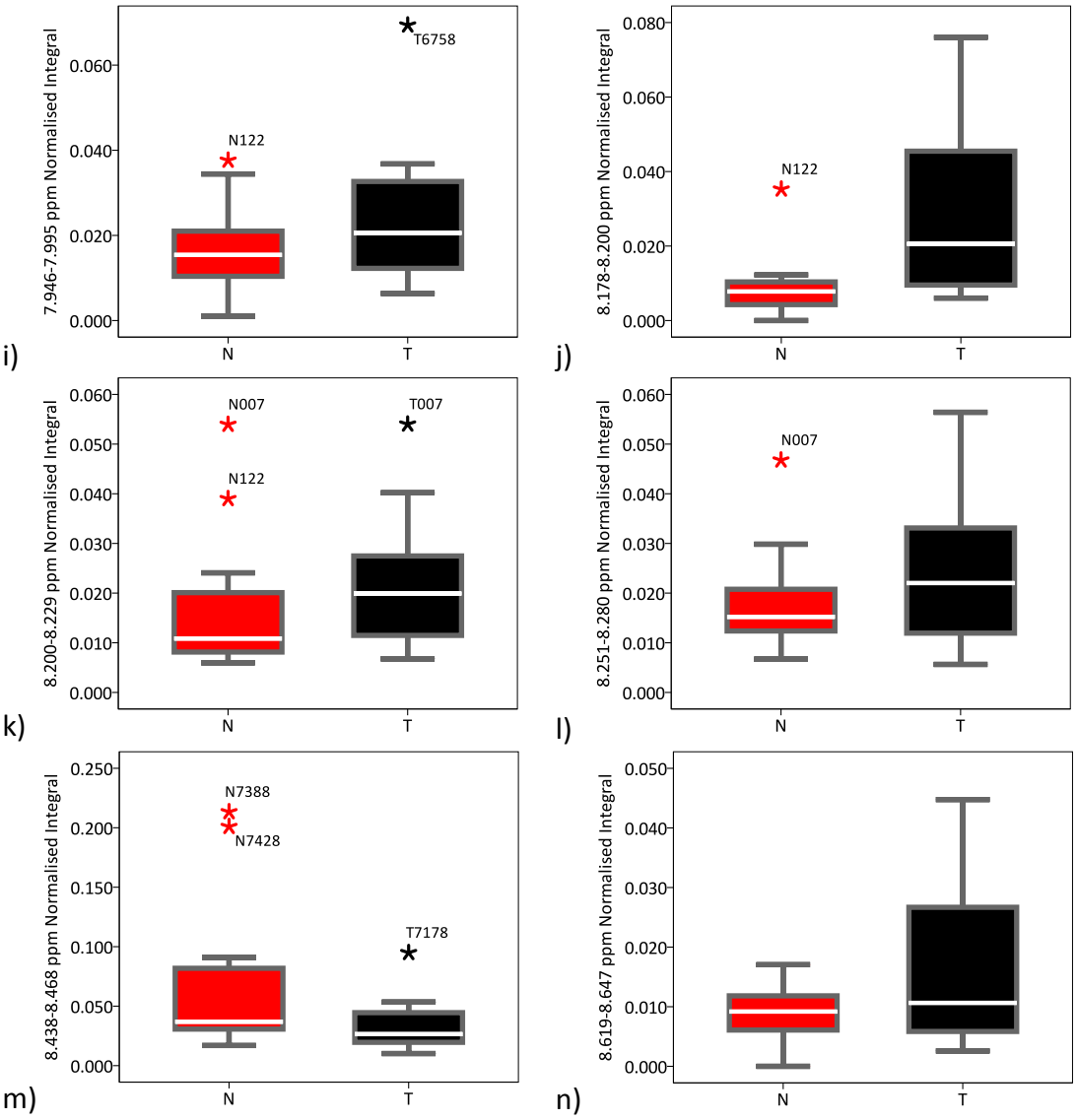


Figure 6.18 Continued.

Table 6.6 Summary of data regarding mean integral difference between tissue types before and after exclusion of samples from aromatic region bins identified in Figure 6.17. Non = non-normal; ND = no difference; N = Normal; T = Tumour; ↑ = significant increase.

Identified Bin [centre] (ppm)	Bin Range Tested (ppm)	Major Metabolite	Data Distribution		*p*-value				Mean Integral Conclusion	
					Uncorrected		Corrected			
			All	Exc 122	All	Exc 122	All	Exc 122	All	Exc 122
5.973-5.985 [5.978]	5.973-6.015	UDP-GlcNAc and UDP-GalNAc	Non	Non	0.071	0.013	0.168	0.046	ND	↑T
5.985-6.015 [6.000]										
6.108-6.124 [6.116]	6.108-6.124	UDP-GlcNAc and UDP-GalNAc	Non	Non	0.012	0.003	0.168	0.042	ND	↑T
6.124-6.146 [6.134]	6.124-6.175	Adenosine 3',5'-diphosphate	Non	Non	0.071	0.033	0.168	0.077	ND	ND
6.146-6.175 [6.160]										
6.900-6.925 [6.912]	6.900-6.925	Tyrosine	Normal	Normal	0.146	0.081	0.252	0.126	ND	ND
7.172-7.195 [7.183]	7.172-7.225	Tyrosine	Normal	Normal	0.224	0.090	0.285	0.126	ND	ND
7.195-7.225 [7.209]										
7.329-7.345 [7.336]	7.329-7.345	Phenylalanine	Normal	Normal	0.586	0.381	0.631	0.410	ND	ND
7.422-7.451 [7.436]	7.422-7.451	Phenylalanine	Normal	Normal	0.077	0.026	0.168	0.073	ND	ND
7.661-7.690 [7.675]	7.661-7.690	Pyridoxine	Non	Non	0.638	0.424	0.638	0.424	ND	ND
7.946-7.972 [7.958]	7.946-7.995	UDP-GlcNAc and UDP-GalNAc	Non	Non	0.182	0.050	0.255	0.100	ND	ND
7.972-7.995 [7.983]										
8.178-8.200 [8.188]	8.178-8.200	Unknown	Non	Normal	0.028	0.006	0.168	0.042	ND	↑T
8.200-8.229 [8.214]	8.200-8.229	Unknown	Non	Non	0.071	0.010	0.168	0.046	ND	↑T
8.251-8.280 [8.265]	8.251-8.280	Adenosine 3',5'-diphosphate	Non	Non	0.347	0.182	0.405	0.212	ND	ND
8.438-8.468 [8.452]	8.438-8.468	Formate	Non	Non	0.084	0.062	0.168	0.109	ND	ND
8.619-8.647 [8.632]	8.619-8.647	Adenosine 3',5'-diphosphate	Normal	Normal	0.162	0.105	0.252	0.133	ND	ND

From glucose, fructose 6-phosphate can be produced *via* glycolysis, which in turn can react with glutamine to produce glutamate and glucosamine 6-phosphate.[217] The former product has been identified as raised in Tumour samples previously in this study whilst the latter is a constituent of UDP-GalNAc. UDP-GalNAc can be reversibly catalysed to UDP-GlcNAc. Another UDP-GalNAc moiety, acetyl CoA, can be made from pyruvate or by fatty acid oxidation.[189] Reduced glucose and increased glycolysis product levels, such as lactate, in Tumour samples was concluded earlier in this work therefore increased levels of UDP-GlcNAc and UDP-GalNAc, derived from glycolysis, in Tumour samples complements the previous conclusion.

O-linked β-N-acetylglucosamine glycosylation (O-GlcNAcylation) is the glycosylation of proteins with O-linked β-N-acetylglucosamine (O-GlcNAc).[190] Several tumour-associated proteins have been identified as glycosylated O-GlcNAc proteins and levels in breast tumour tissue are increased compared to corresponding adjacent tissue.[218] Gu *et al*.[218] have reasoned that a raised level of UDPGlcNAc and subsequently GlcNAcylation might be one of the reasons why the Warburg effect is present for tumours.

It can be speculated as to why the UDP-GlcNAc and UDP-GalNAc signals between 7.946 ppm and 7.995 ppm were not significantly raised in Tumour samples: underlying signals could be present that do not follow the same trend or due to the low signal intensity noise could reduce the magnitude of difference between sample types.

Box plots (Figure 6.18) for 8.178-8.200 ppm and 8.200-8.229 ppm bins show sample N122 was more extreme than 1.5 times the interquartile range beyond the upper quartile as it was for two of the three UDP-GlcNAc and UDP-GalNAc containing regions. It can be tentatively speculated that signals could be from metabolites related to glycolysis.

6.1.1.2 Evaluation of Breast Cancer Severity

6.1.1.2.1 Analysis of Whole Spectrum

Further to discriminating between Tumour and Normal samples, breast cancer grade was investigated in an attempt to reveal metabolic markers of disease severity.

PCA of PQ normalised data produced a one component model ($R^2\boldsymbol{X}$ = 0.387 and $Q^2\boldsymbol{X}$= 0.174) when Tumour samples alone were included. Scores and loadings plots are shown in Figure 6.19 and Figure 6.20, respectively.

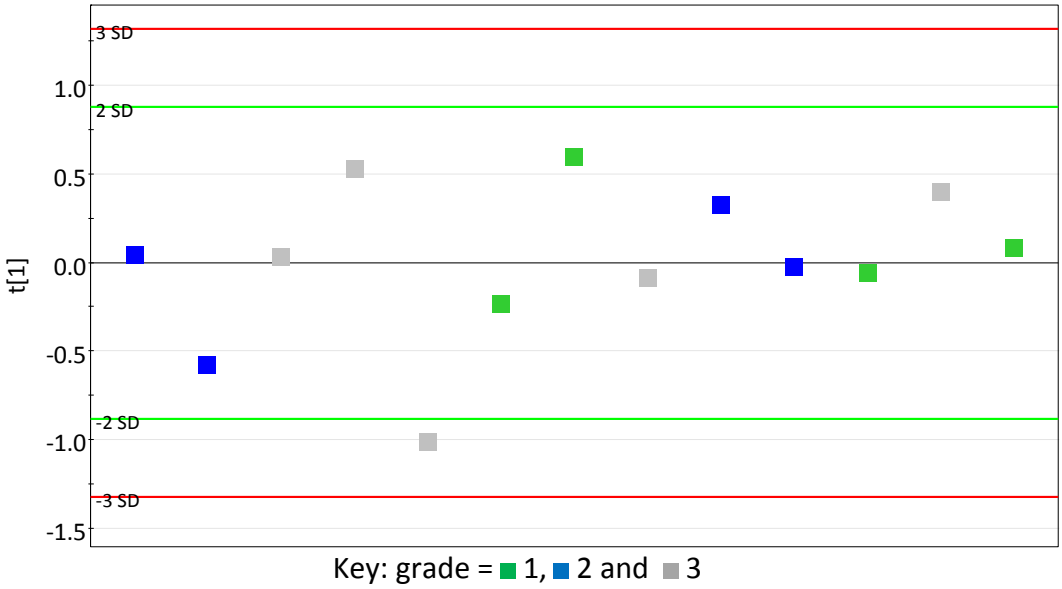


Figure 6.19 PCA scores plot for aqueous extract Tumour samples. Samples coloured according to tumour grade.

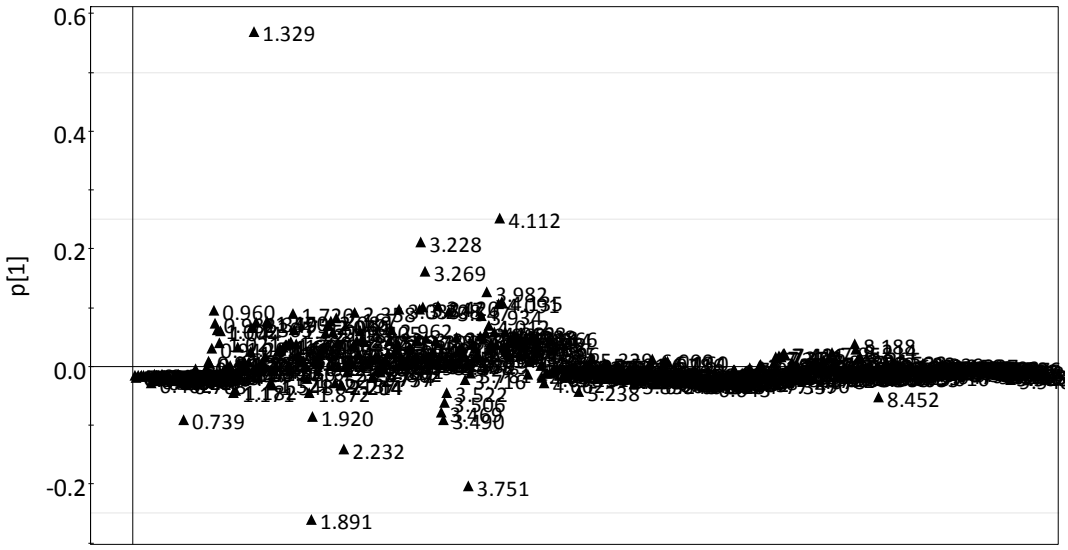
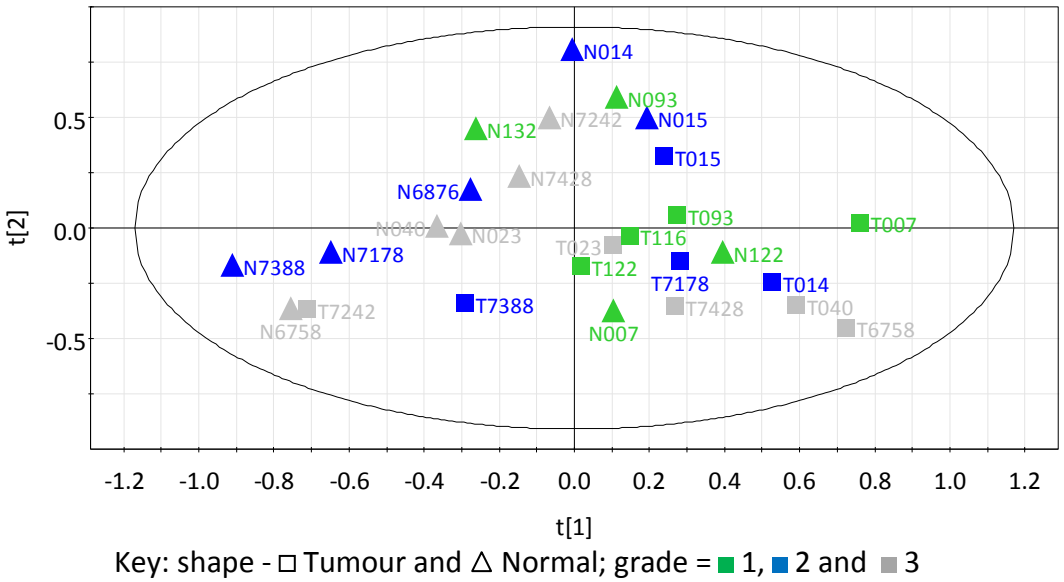


Figure 6.20 PCA loadings plot corresponding to the model displayed in Figure 6.19.

No distinction in scores space based on tumour grade was observed. Although many studies have identified metabolic differences between tumour and non-tumour breast tissue samples very few have reported differences between tumour grades. This could be due to composition of the tissue because the composition is linked with levels of metabolites.[136,141] For tumours of the same grade the cancer cell percentage has been reported to vary from 0% to over 50%. Histopathological analysis was not possible for this study but if a similar observation was present for all grades this could explain why no separation between samples based on grade was present. Hilvo *et al*.,[219] using LC-MS, showed phospholipids were increased in grade 3 samples compared to those in the combined grade 1 and 2 group but triglycerides were unchanged and no change was observed for either lipid species between grade 1 and 2 samples. In one of the first tissue MAS studies Cheng *et al*.[137] indicated lactate and PCho increased in grade 3 compared to grade 2 samples through an elevation in the lactate:choline and PCho:choline ratios. For this study, although lactate (1.329 and 4.112 ppm) and PCho and GPC (3.228 ppm) containing bins were shown to be highly discriminatory between samples with levels significantly higher in Tumour samples, correlation was not present with tumour grade.

Although metabolites were not identified that were reflective of tumour grade investigation ensued to determine whether there was a connection between Tumour sample grade and the paired Normal sample. For example, whether both paired samples were in the same scores space relative to other samples of the same type or whether positioning in scores space of Normal samples was dependent on the grade of the Tumour counterpart. Figure 6.21 shows the scores plot whilst the loadings plot (data not shown) was the same as Figure 6.5 (model statistics related to this figure apply). Samples T116, N6876 and N132 were included to show positions relative to other samples despite not having a paired sample.



Key: shape - □ Tumour and △ Normal; grade = ■ 1, ■ 2 and ■ 3

Figure 6.21 PCA scores plot for aqueous extract samples. Tumour samples coloured according to tumour grade and Normal samples coloured according to the grade of Tumour counterpart. R^2X = 0.311 and 0.184, Q^2X = 0.171 and 0.151 for PC 1 and PC 2, respectively.

No pattern is apparent in scores space between paired Tumour and Normal samples. Normal samples do not show a metabolite profile that is related to Tumour sample grade and through non-clustering of Normal samples coloured the same in Figure 6.21, grade of the paired Tumour sample cannot be estimated, *i.e.* a raised or reduced metabolite(s) level in Normal samples does not indicate corresponding Tumour sample grade. Additionally, using PLS-DA, models were not able to be generated for the data used in the above two PCA models.

6.1.1.2.2 Analysis Excluding Lactate

Lactate has been shown to be significantly increased in Tumour samples compared to Normal samples but different grades of Tumour samples have not been distinguished when lactate signals have been included. It was hypothesised that exclusion of lactate (1.314-1.374 ppm and 4.085-4.144 ppm) could reveal influential metabolites of smaller intensities.

PCA generated a three component model ($R^2\boldsymbol{X}$(cum) = 0.683 and $Q^2\boldsymbol{X}$(cum) = 0.215) but separation based on grade was not observed (data not shown; PC 1 similar to Figure 6.19 and Figure 6.20 for scores and loadings plots, respectively). A PLS-DA model was not able to be created nor a PCA model incorporating Tumour and Normal samples.

6.1.1.2.3 Analysis of Aromatic Region Only

For reasons previously stated the aromatic region alone was analysed. PCA produced a one component model ($R^2\boldsymbol{X}$ = 0.274 and $Q^2\boldsymbol{X}$= 0.071) when only Tumour samples were included. No separation based on tumour grade was observable (Figure 6.22). Nearly all of the bins identified when Tumour and Normal samples were analysed displayed the greatest modulus $\boldsymbol{p}$[1] values in the loadings plot (data not shown). Two of the most prominent bins were centred at 6.000 ppm and 7.958 ppm; both contained UDP-GlcNAc and UDP-GalNAc. The two metabolites have been associated with breast cancer progression.[135,218] Beckonert *et al*.[135] used self organising map (SOM) plots to visualise tissue extract data and these indicated UDP-GlcNAc and UDP-GalNAc were highest in grade 3 samples and lowest in control with grade 2 having an intermediate level though the observation was much closer to that of the former. Quantification and hence statistical significance of differences was not performed by the authors.

Application of PLS-DA was made and a one PC model ($R^2\boldsymbol{X}$ = 0.148, $R^2\boldsymbol{Y}$ = 0.378 and $Q^2\boldsymbol{Y}$ = 0.070) was built. Group 1 samples were separated from group 2 and 3

samples; there was no separation between groups 2 and 3 (Figure 6.23). The loadings plot (data not shown) indicated the unidentified signal at 8.188 ppm was elevated in grade 1 samples and UDP-GlcNAc and UDP-GalNAc (5.978 ppm, 6.000 ppm and 7.598 ppm), adenosine 3',5'-diphosphate (6.134 ppm, 6.160 ppm 8.265 ppm and 8.632 ppm), pyridoxine (7.675 ppm) and formate (8.452 ppm) were raised in grade 2 and 3 samples. The six metabolites were identified in the loadings plot (Figure 6.17) as strongly contributing to Tumour and Normal samples' scores positions.

PLS-DA was repeated using two classes: grade 1 and grades 2 and 3 combined. The scores and loadings plots of the one PC model ($R^2\boldsymbol{X}$ = 0.152, $R^2\boldsymbol{Y}$ = 0.727 and $Q^2\boldsymbol{Y}$ = 0.264) were not shown due to the similarity when three classes were used.

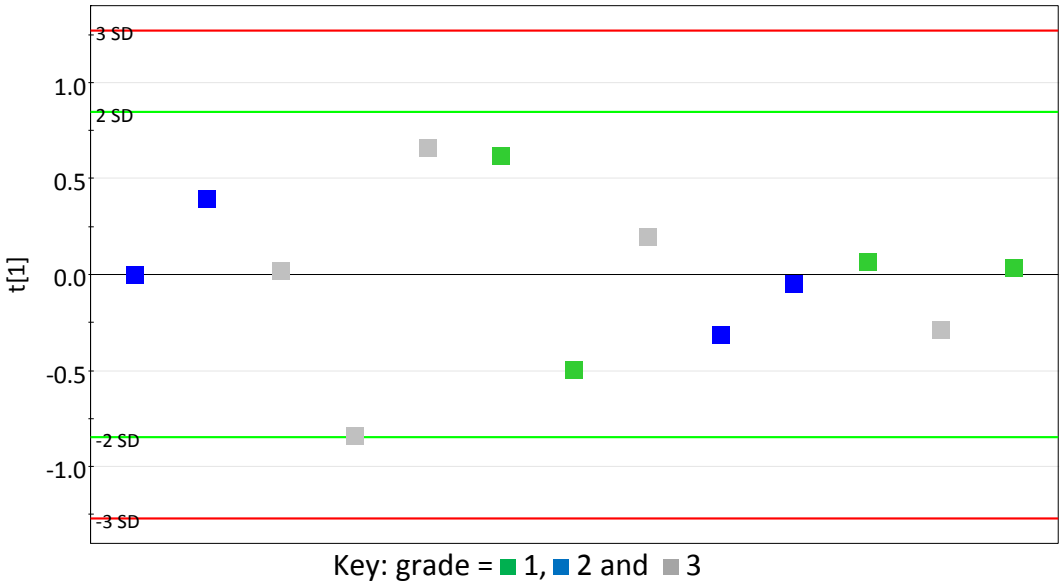


Figure 6.22 PCA scores plot of the aromatic region for aqueous extract samples of Tumour samples. Coloured according to tumour grade.

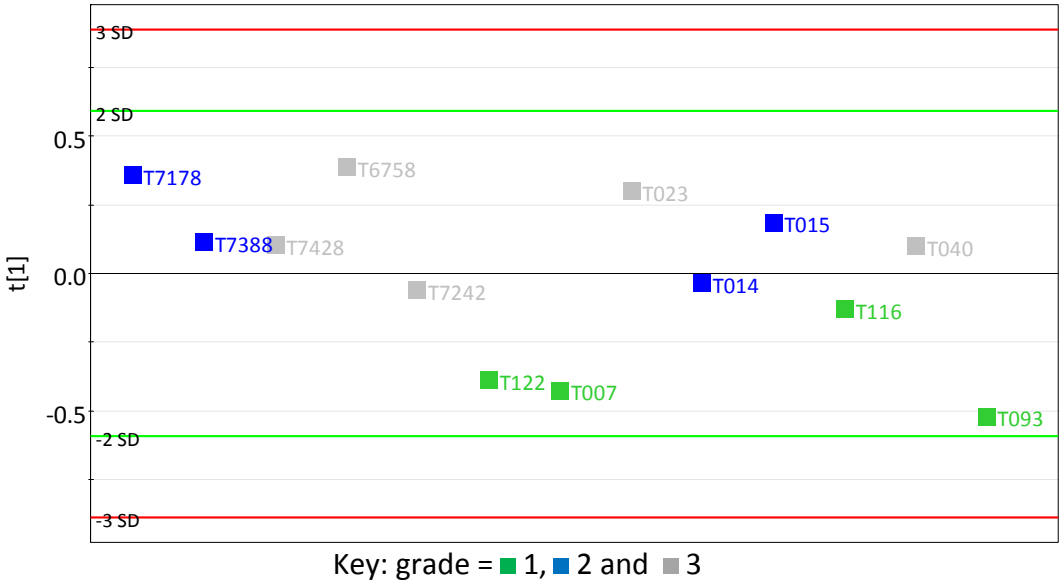


Figure 6.23 PLS-DA scores plot of the aromatic region for aqueous extract samples of Tumour samples. Classed according to tumour grade.

Univariate analysis was performed to determine whether levels of metabolites were significantly different between grade 1 and grades 2 and 3 combined. Box plots (Figure 6.24) provided an overview of data from the two classes and Table 6.7 concluded whether any difference was statistically significant. Before correction for multiple testing adenosine 3',5'-diphosphate (8.619-8.647 ppm) was significantly increased in grade 2 and 3 samples compared to grade 1 samples but after correction the difference was no longer statistically significant. There was no difference between classes for all other bin integral values tested.

'Leave-one-out' cross-validation was performed to test whether PLS-DA overfitted the data (Table 6.8). Although seven of nine (78%) grade 2 or 3 samples were correctly predicted with five of those having a ***Y***-predicted value of over 0.70, only one of four (25%) grade one samples were correctly predicted and the *Y*-predicted value for that sample was only 0.557. Figure 6.25 shows many permuted models for the 'grades 2 and 3 combined' class to have higher $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ values than the original model. The intercept values of the regression lines for $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ are 0.586 and -0.058, respectively. The $R^2\boldsymbol{Y}$ value is much greater than 0.4 so it is indicated that the model has overfitted the data thus furthering the implication from 'leave-one-out' cross validation. The permutation testing plot for the control

class is not shown but with $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ intercept values of 0.582 and -0.053, respectively, the same conclusion was made. Being unable to validate the model would explain why no significant differences between classes were observed. Using PQN a PCA model was not able to be built when all samples were included.

It is concluded that there is no evidence to suggest a difference between metabolite levels of samples with different tumour grades.

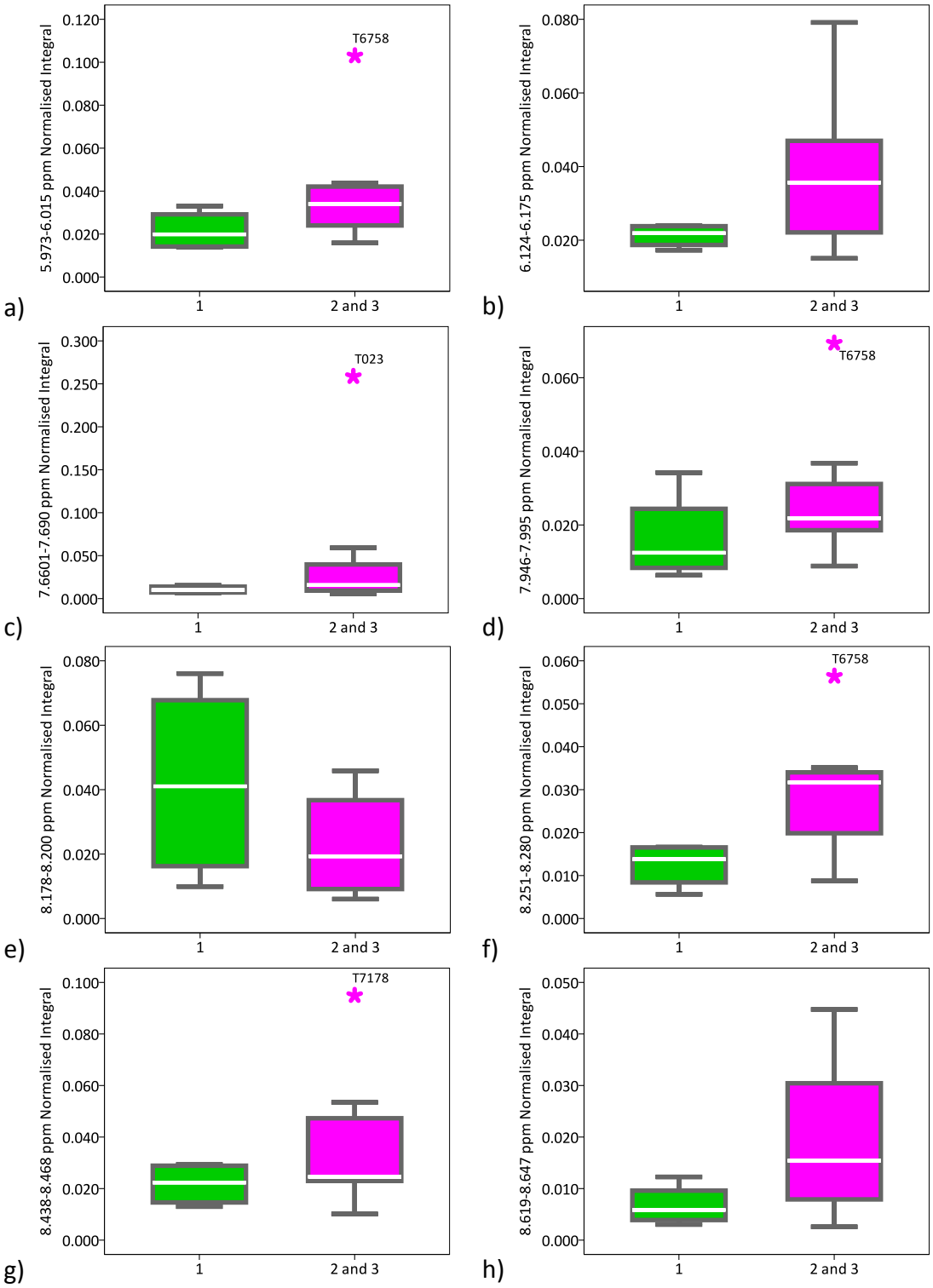


Figure 6.24 Box plots of aromatic region integrals for grade 1 (green filled box) and grade 2 and 3 combined (pink filled box) Tumour samples. * = samples more extreme than 1.5 times the interquartile range beyond the upper quartile, coloured as per filled box, with sample label.

Table 6.7 Summary of Tumour sample data regarding mean integral differences of aromatic region bins between grade 1 and grades 2 and 3 combined.

Identified Bin [centre] (ppm)	Bin Range Tested (ppm)	Major Metabolite	Data Distribution	*p*-value		Mean Integral Conclusion
				Uncorrected	Corrected	
5.973-5.985 [5.978]	5.973-6.015	UDP-GlcNAc and UDP-GalNAc	Non-normal	0.106	0.243	No difference
5.985-6.015 [6.000]						
6.124-6.146 [6.134]	6.124-6.175	Adenosine 3',5'-diphosphate	Normal	0.139	0.243	No difference
6.146-6.175 [6.160]						
7.661-7.690 [7.675]	7.661-7.690	Pyridoxine	Non-normal	0.260	0.320	No difference
8.178-8.200 [8.188]	8.178-8.200	Unknown	Normal	0.320	0.320	No difference
8.251-8.280 [8.265]	8.251-8.280	Adenosine 3',5'-diphosphate	Normal	0.057	0.200	No difference
8.438-8.468 [8.452]	8.438-8.468	Formate	Normal	0.284	0.320	No difference
8.619-8.647 [8.632]	8.619-8.647	Adenosine 3',5'-diphosphate	Normal	0.028	0.196	No difference

Table 6.8 'Leave one out' cross-validation of the PLS-DA model in Figure 6.23.

Sample Excluded	Number of Components	R^2X(cum)	R^2Y(cum)	Q^2Y(cum)	*Y*-Predicted*	
					1	2 or 3
T7178	3	0.506	0.918	0.338	-0.107	1.107
T7388	3	0.493	0.926	0.293	0.292	0.708
T7428	1	0.157	0.739	0.260	0.274	0.726
T6758	1	0.165	0.732	0.271	-0.085	1.085
T7242	1	0.172	0.768	0.411	0.728	0.272
T122	1	0.175	0.673	0.325	0.304	0.696
T007	3	0.499	0.899	0.118	0.296	0.704
T023	1	0.179	0.701	0.208	-0.089	1.089
T014	1	0.153	0.784	0.272	0.505	0.495
T015	1	0.169	0.716	0.228	0.346	0.654
T116	1	0.153	0.812	0.211	0.323	0.677
T040	1	0.157	0.750	0.249	0.400	0.600
T093	3	0.506	0.917	0.193	0.557	0.443

*A sample was regarded as belonging to a grade by having a ***Y***-predicted value >0.50. Incorrect classification is represented by red shading and correct by pink, orange or green, corresponding to a ***Y***-predicted value of <0.60, 0.60-0.70 or >0.70, respectively.

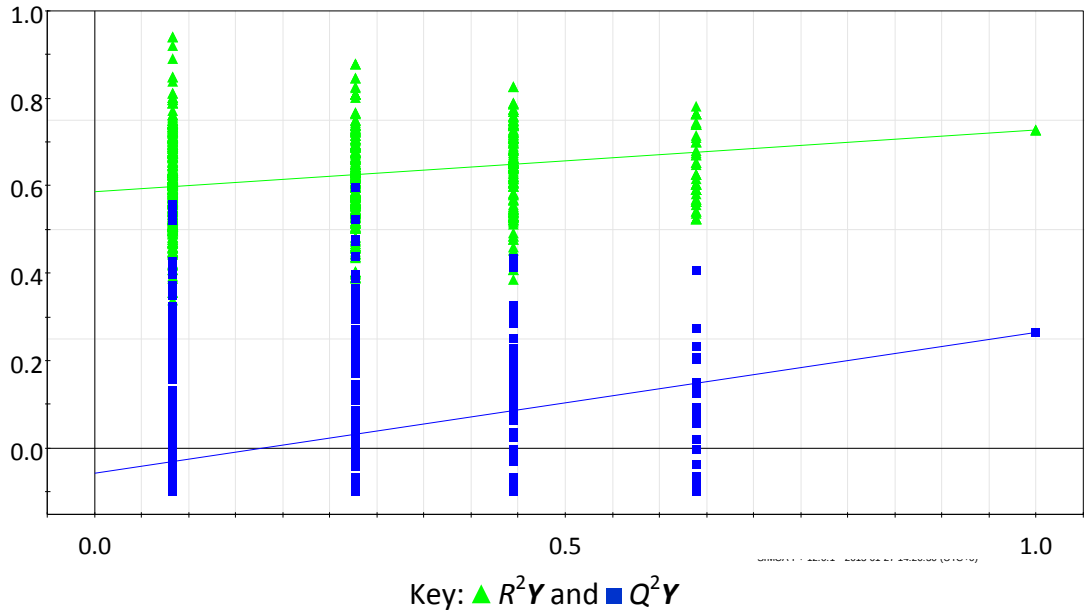


Figure 6.25 Permutation testing plots for the 'grades 2 and 3 combined' class in the PLS-DA model shown in Figure 6.23. The R^2Y and Q^2Y intercept values of the regression lines are 0.586 and -0.058, respectively.

6.1.1.3 Investigation of Processing Time

The time that tissue samples are not in the frozen state must be kept to a minimum both at the time of collection and processing.[51] Regarding the latter, a number of steps were taken to ensure the extraction procedure adhered to this and was consistent. The samples were still frozen upon first contact with organic extraction solvents as stipulated by Beckonert *et al*.[51] but one variable that could not be controlled was the time required to ensure as much as possible of the sample was homogenised in a suitable timeframe due to variation in size and composition. Enzymatic changes can occur rapidly and levels of certain metabolites can alter on a millisecond timescale.[51]

To test the effect of variable homogenisation time a large sample (T6758) was processed as normal. Immediately after the sample was placed on ice following all solvent addition steps non-homogenised tissue was removed and the extraction procedure repeated. The processing steps between the end of homogenising and tissue removal was approximately two minutes.

Both samples were included in PCA. The samples were positioned very closely in scores space both in PC 2 versus PC 1 plot (data not shown; very similar to Figure 6.4) and PC 3 versus PC 2 plot (Figure 6.26). For the three PC model $R^2\boldsymbol{X}$(cum) was 0.604 and $Q^2\boldsymbol{X}$(cum) equalled 0.324. T6758L was not used in any other analysis.

The closeness in scores space indicates the time taken for homogenisation due to variance in sample size and consistency does not impact strongly on the aqueous extraction profile and consistency of homogenisation is a more important factor.

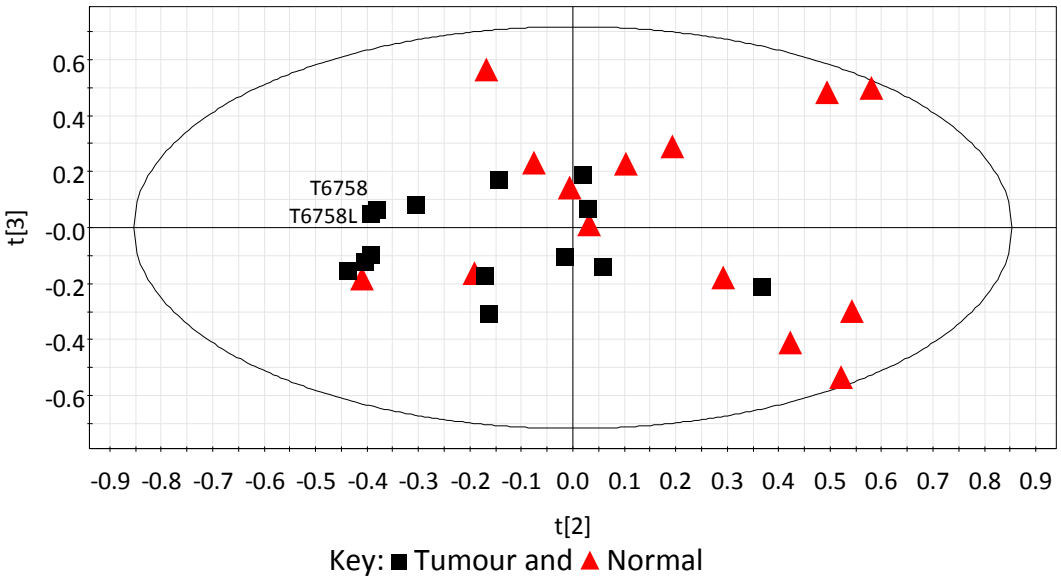


Figure 6.26 PCA scores plot coloured according to tissue type for aqueous extract samples showing the second and third model components. R^2X = 0.179 and 0.127, and Q^2X = 0.172 and 0.084 for PC 2 and PC 3, respectively. 'L' at the end of the tissue sample code denotes the sample with the longer processing time.

6.1.2 Lipophilic Extract Analysis

Regions pertaining to the solvent signals of methanol (3.321-3.450 ppm) and chloroform (7.403-7.594 ppm) were removed and PCA of sum normalised data performed initially. These signals are shown on a typical spectrum in Figure 6.27. A variety of sources were used to identify metabolite signals.[103,135,142,220-223]

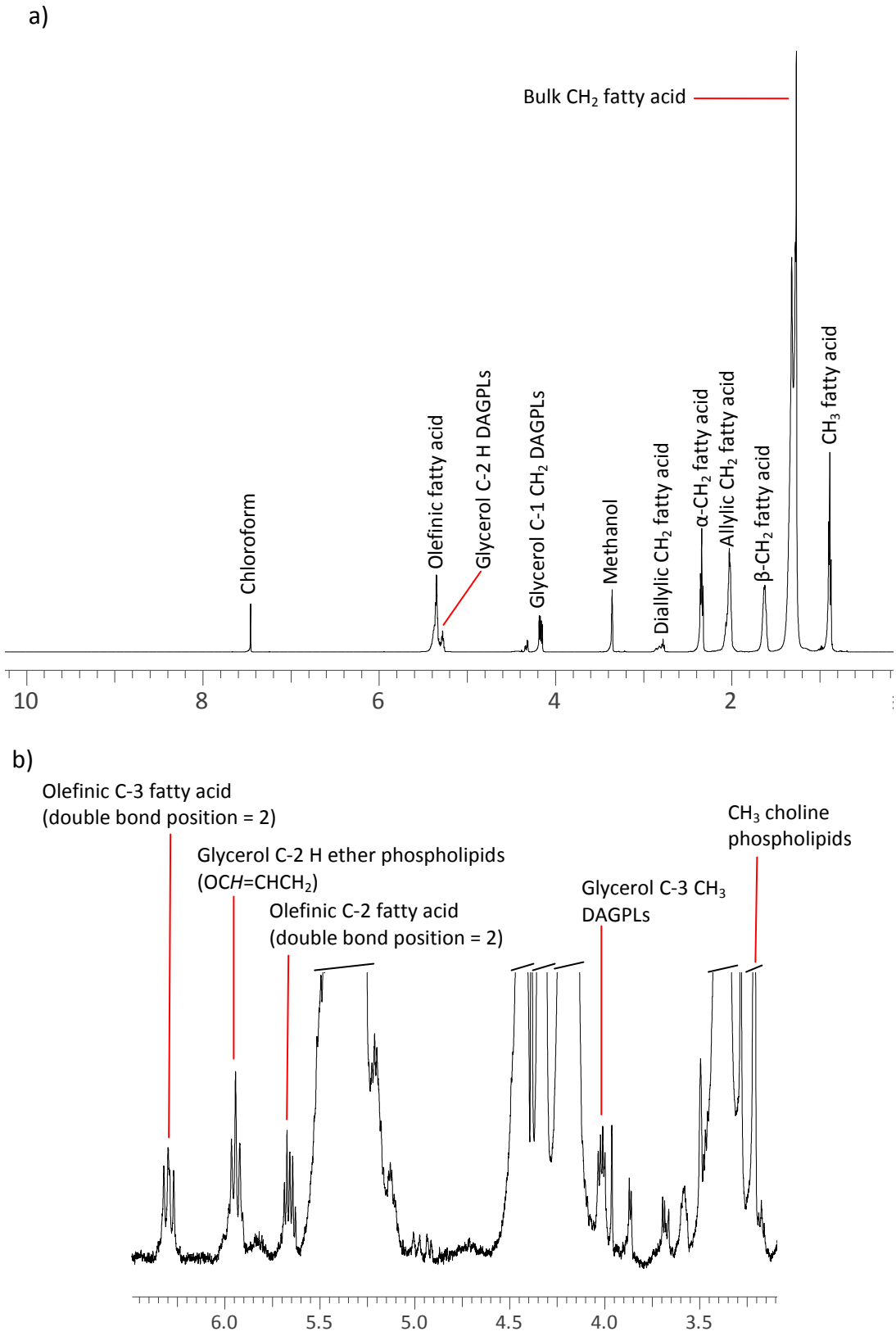


Figure 6.27 ^{1}H-NMR spectrum of tissue lipophilic extracts from a Normal sample. a) Whole spectrum, b) expansion of 3.1-6.5 ppm region. The *x*-axis is chemical shift in ppm. DAGPLs = diacylglycerophospholipids

6.1.2.1 Evaluation of Breast Cancer Occurrence

PCA was performed and a two PC model ($R^2\boldsymbol{X}$(cum) = 0.821 and $Q^2\boldsymbol{X}$(cum) = 0.735) was generated with scores and loadings plots shown in Figure 6.28 and Figure 6.29, respectively.

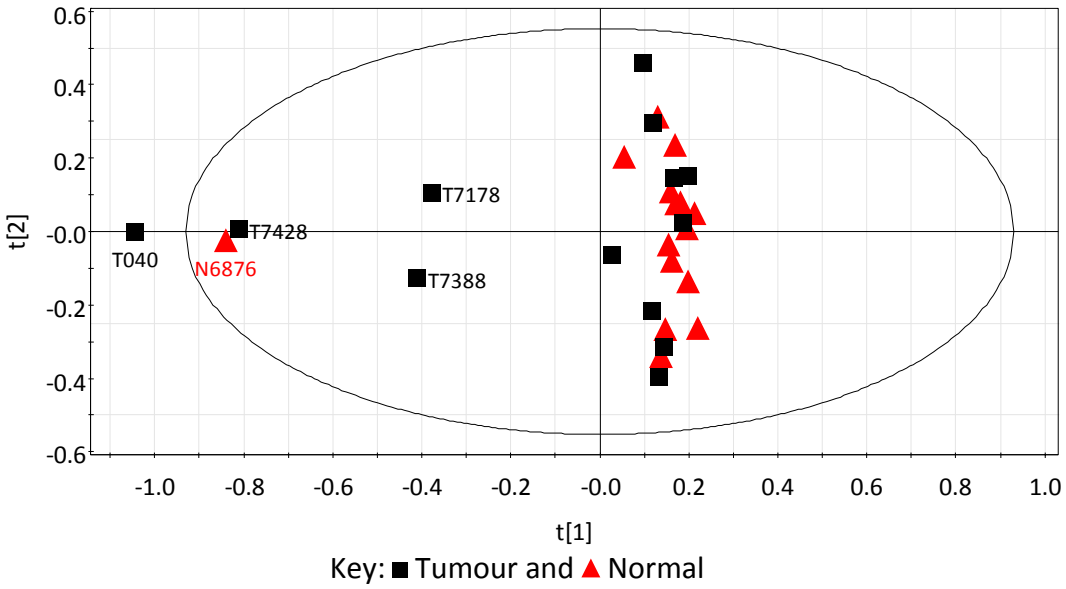


Figure 6.28 PCA scores plot for lipophilic extract samples using 0.02 ppm variable bins coloured according to tissue type. Samples most remote from the majority are labelled.

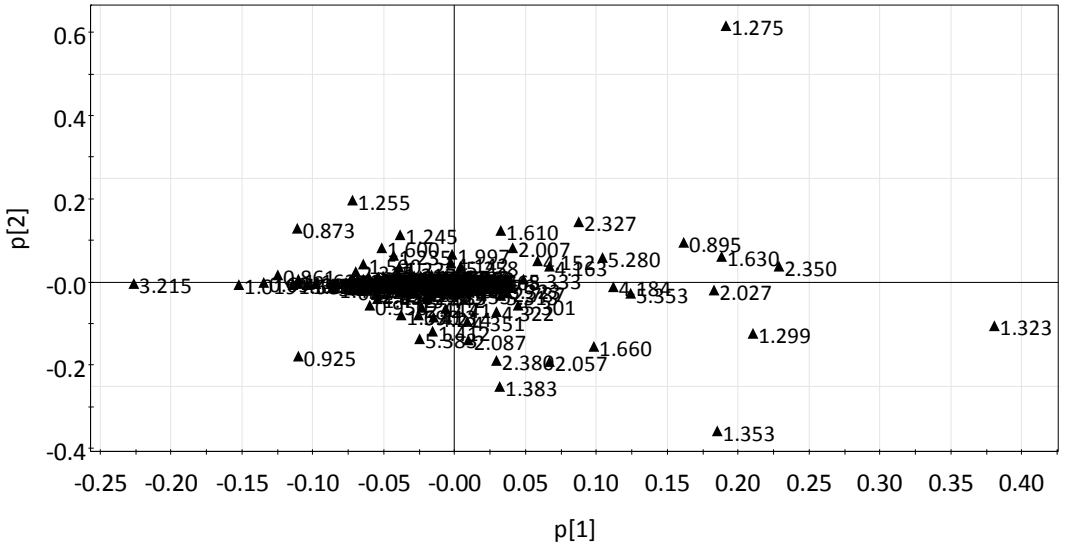


Figure 6.29 PCA loadings plot corresponding to the model displayed in Figure 6.28.

Table 6.9 shows the bins that had a positive or negative ***p***[1] value greater than 0.15 and were the most influential regarding separation of samples along PC1. One Normal and four Tumour samples were clearly separated from the other 23 samples in this component. The high $R^2\boldsymbol{X}$ value of 0.607 indicated the PC captured a large amount of the total variance of the data.

Table 6.9 Identity of lipids causing separation of samples in Figure 6.28.

Bin Centre (ppm)	**Lipid**	**Lipid Level in Five Most Separated Samples in PC 1 Scores Space**
0.895	CH_3 fatty acid	↓
1.019	C-19 CH_3 cholesterol	↑
1.275	Bulk CH_2 fatty acid	↓
1.299		
1.323		
1.353		
1.630	β-CH_2 fatty acid	↓
2.027	Allylic CH_2 fatty acid	↓
2.350	α-CH_2 fatty acid	↓
3.215	Choline phospholipids	↑

With only four Tumour samples separated from the majority of Normal samples little information could be gained about potential metabolic differences between cancerous and non-cancerous tissue. A number of strategies to resolve this were employed including PQN, removal of select samples, removal of heavily influential bins and use of smaller bin size.

Employing PQN rather than sum normalisation on the same pre-normalised data generated a two PC model with $R^2\boldsymbol{X}$(cum) = 0.938 and $Q^2\boldsymbol{X}$(cum) = 0.914. The scores plot is shown in Figure 6.30 and loadings plot in Figure 6.31.

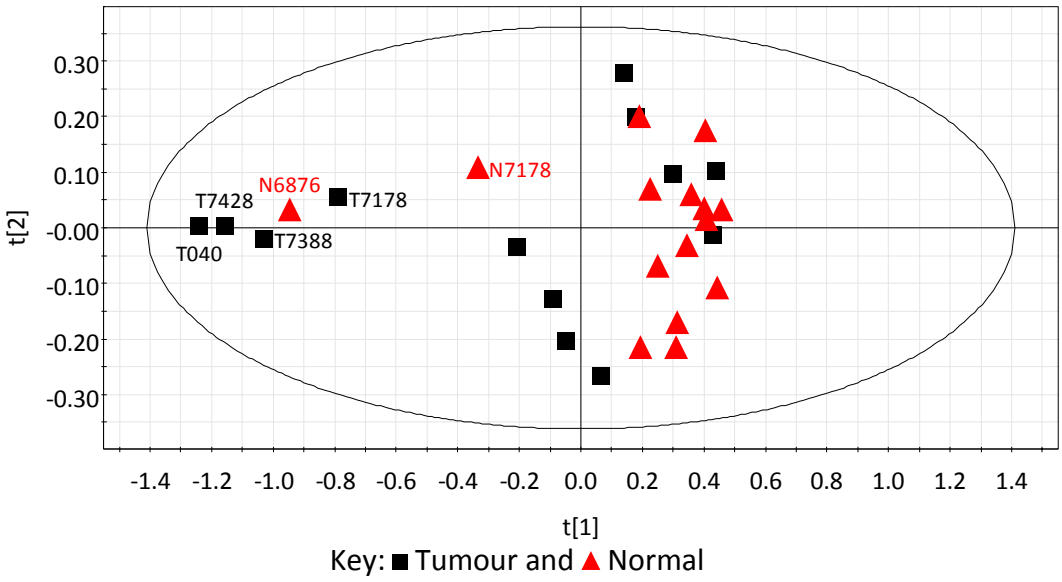


Figure 6.30 PCA scores plot for lipophilic extract samples using PQN coloured according to tissue type. Samples most remote from the majority are labelled.

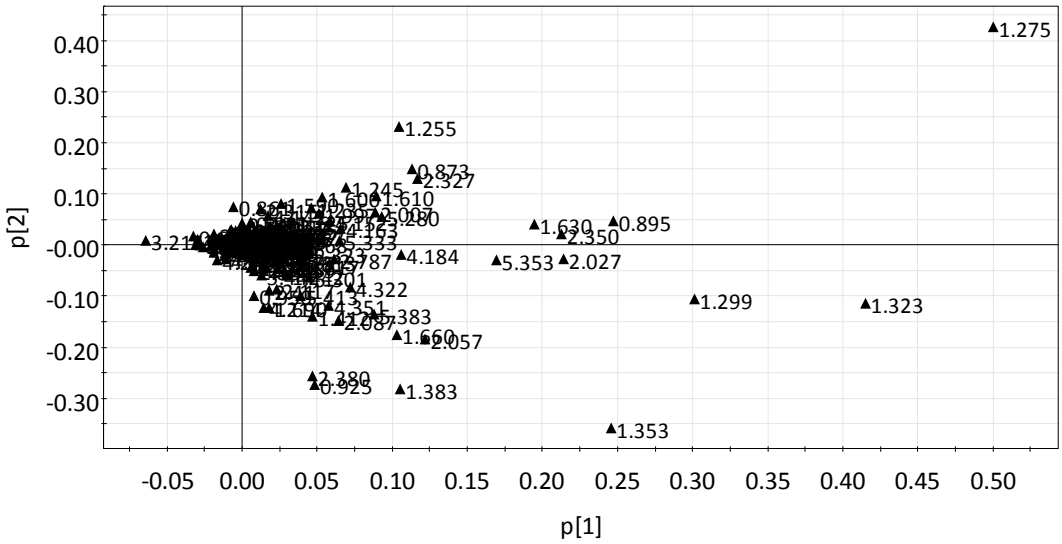


Figure 6.31 PCA loadings plot corresponding to the model displayed in Figure 6.30.

Some minor differences could be observed using PQN compared to sum normalisation, for example a few samples, such as N7178, were not positioned as closely to the majority of samples in PC 1 scores space using the former. However, the same five samples still had the greatest PC 1 modulus scores values and were separated from the other samples with the same down-regulated lipid species contributing to the separation. C-19 CH_3 cholesterol and choline-containing compounds had lower ***p***[1] modulus loadings values indicating increased levels in

the five separated samples were less prominent using PQN. In conclusion, PQN alone did not enhance knowledge of potential metabolic changes between tissue types compared to sum normalisation but unless specified otherwise PQN data was used in further work due to its potential advantages.[69]

The next approach was to remove Tumour samples 040, 7178, 7388 and 7428 and Normal sample 6876. The first two PCs of the six component model ($R^2\boldsymbol{X}$(cum) = 0.968, $Q^2\boldsymbol{X}$(cum) = 0.877) are exhibited in the scores plot shown in Figure 6.32. The complementary loadings plot was very similar to Figure 6.29, for which no samples were removed, so is not shown. Approximately, an amplification of the centre of the scores plot shown in Figure 6.30, *i.e.* where 23 of the 28 samples were located before any were excluded, is exhibited in Figure 6.32. Removal of five samples still resulted in the same bins influencing the position of samples in scores space and similar relative positions of the remaining samples.

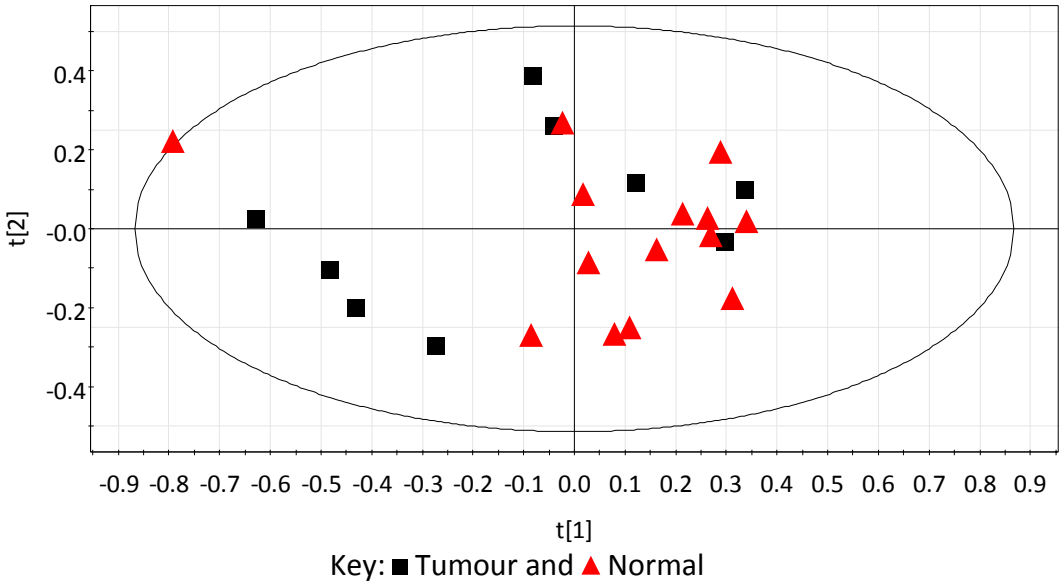


Figure 6.32 PCA scores plot showing the first two model components for lipophilic extract samples using PQN excluding the five samples with largest scores values in Figure 6.28. Coloured according to tissue type. R^2X = 0.649 and 0.227, and Q^2X = 0.528 and 0.270 for PC 1 and PC 2, respectively.

The spectrum, as shown in Figure 6.27, contains a large resonance from bulk methylene protons at 1.30 ppm and also a number of very small intensity signals.

Excluding the aforementioned large resonance (1.200-1.458 ppm), which accounted for 59-71% of the total spectral integral, would increase the influence of all other signals when repeating analysis. The loadings plot of the two PC model ($R^2\boldsymbol{X}$(cum) = 0.886 and $Q^2\boldsymbol{X}$(cum) = 0.855) in Figure 6.33 shows an amplification of the ***p***[1] values of bins compared to the same bins in Figure 6.31, for which the 1.200-1.458 ppm region was not excluded. The complementary scores plots were similar: data not shown for exclusion of bulk methylene resonance.

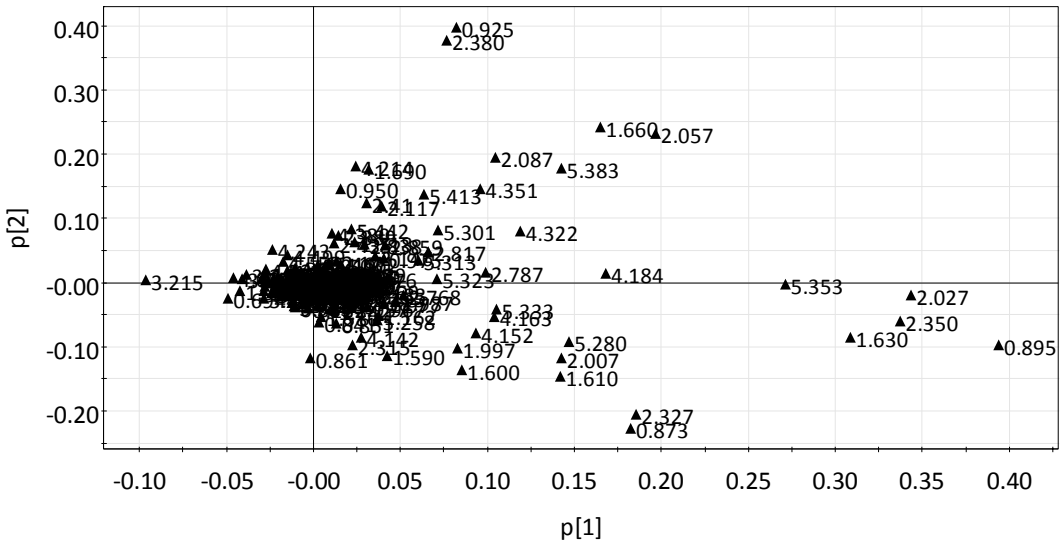


Figure 6.33 PCA loadings plot for lipophilic extract samples using PQN excluding 1.200-1.458 ppm.

Removal of the bulk methylene region provided little extra information on influential bins. To allow assessment of small intensity signals all of the large signals were excluded. Large signals were defined as being visible in the spectrum in Figure 6.27a and accounted for 89.7-99.3% of normalised integrals of samples. Over 99% of the lipophilic extracts spectrum has been attributed to triglycerides.[135] The five labelled samples in Figure 6.28 possessed the five greatest amounts of small signals (4.9-10.3%) thus providing numerical evidence as to why scores plots indicated the five samples contained lesser amounts of triglyceride signals.

The same five samples were still separate from the majority in PC 1 though T014 was additionally present in similar scores space (Figure 6.34). The complementary

loadings plot of the two PC model ($R^2\boldsymbol{X}$(cum) = 0.637 and $Q^2\boldsymbol{X}$(cum) = 0.402) is displayed in Figure 6.35.

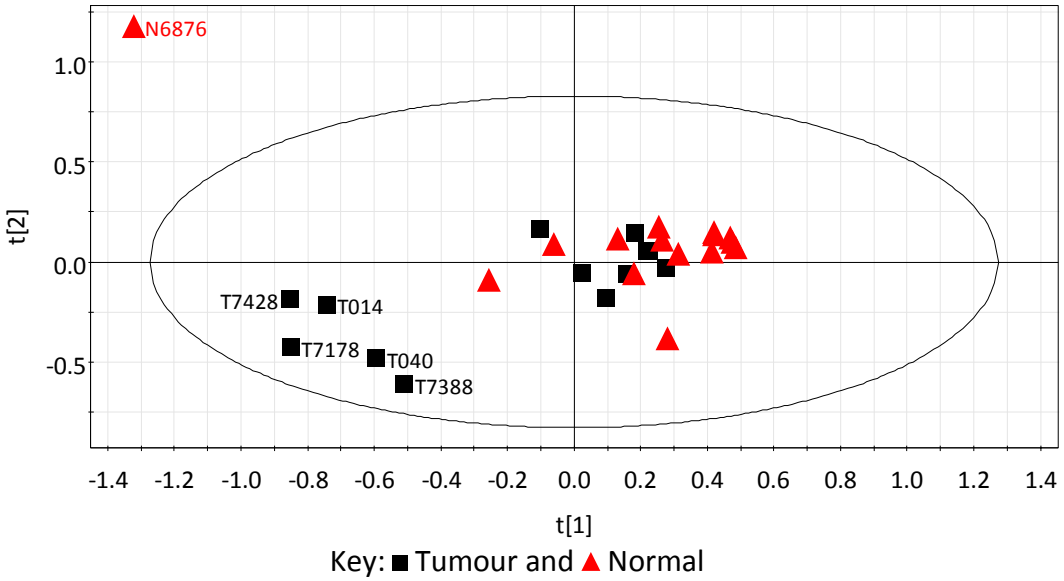


Figure 6.34 PCA scores plot for lipophilic extract samples excluding large signals. Coloured according to tissue type. Samples most remote from the majority are labelled.

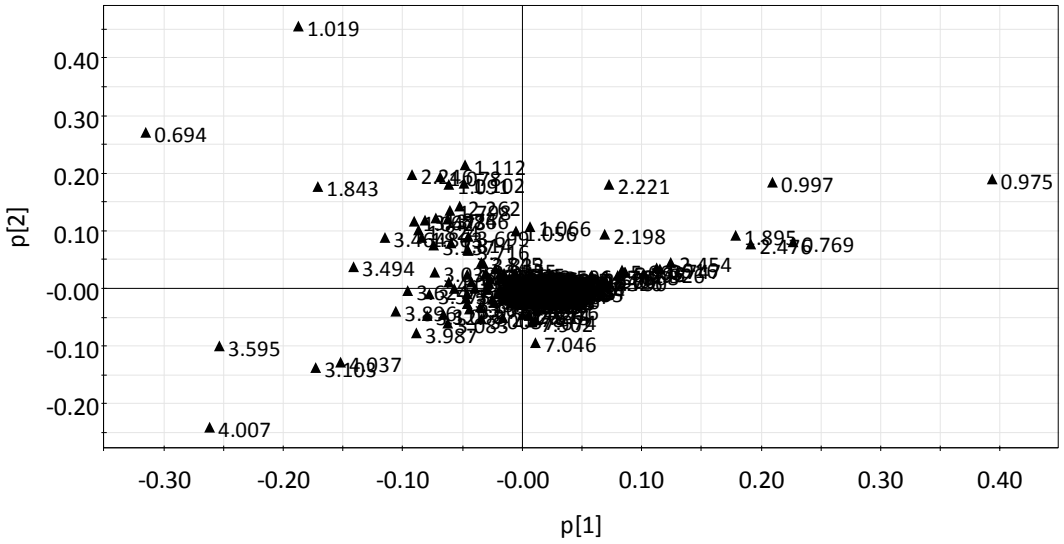


Figure 6.35 PCA loadings plot corresponding to the model displayed in Figure 6.34.

Tumour samples 014, 040, 7178, 7388 and 7428 exhibited raised levels of signals in bins centred at 3.103, 3.595, 4.007 and 4.037 ppm. Signals in the first bin were attributed to –CH_2NH_2 head group of ethanolamine phospholipids, the second bin to –CH_2N– head group of choline phospholipids and the third and fourth bins to both

glycerol C-3 methylene proton of diacylglycerophospholipids (DAGPLs) and $-OCH_2$ head group of ethanolamine phospholipids.[220] N6876 contained greater levels of C-18 and C-19 CH_3 cholesterol at 0.694 and 1.019 ppm,[220] respectively, and an unidentified metabolite at 1.843 ppm. C-18 CH_3 cholesterol has been shown to be raised in grade 3 versus non-tumour breast cancer tissue.[135] Assessment concerning sample grade will be discussed in section 6.1.2.2.2. Assignment of signals that were decreased in the aforementioned six samples was unable to be made. PLS-DA scores plot (Figure 6.36; one PC model, $R^2\boldsymbol{X}$ = 0.395, $R^2\boldsymbol{Y}$ = 0.324 and $Q^2\boldsymbol{Y}$ = 0.196) identified the same five Tumour samples as separate from the majority of samples but N6876 was no longer distinctly positioned in scores space, being closer to the majority of samples rather than the five separated Tumour samples. The four bins previously discussed as increased in the five Tumour samples were again highlighted in the loadings plot (not shown) and likewise those bins associated with a decreased level.

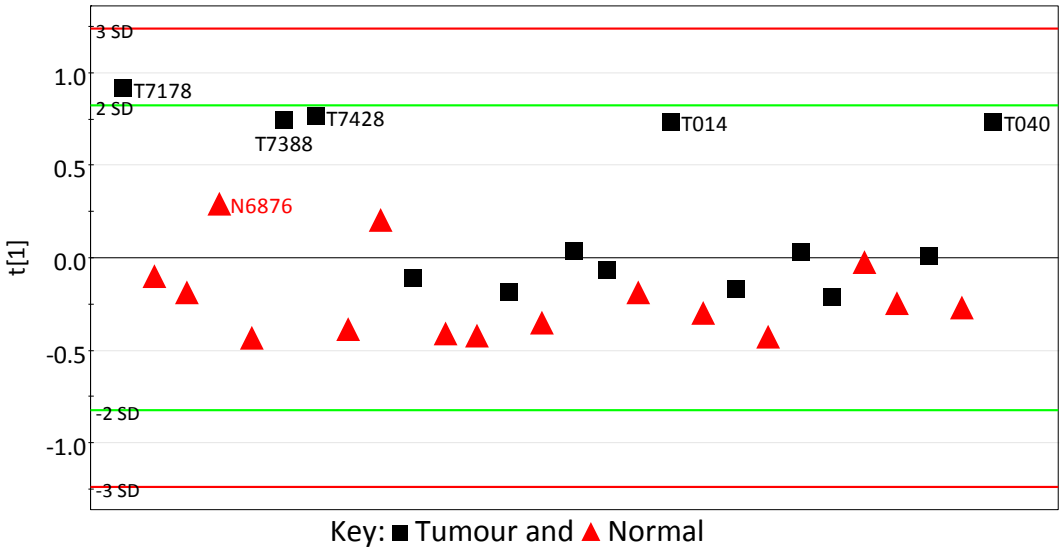


Figure 6.36 PLS-DA scores plot for lipophilic extract samples excluding large signals. Classed according to tissue type. Samples labelled as per Figure 6.34.

Similar results for the majority of samples in terms of PCA scores space separation were observed when including and excluding large signals. A previous study[135] concluded self-organising map (SOM) plots showed similar results when sum normalisation was performed for the whole spectrum and non-triglyceride regions, which accounted for less than 1% of the total spectral integral. Although N6876

contained a greater amount of low intensity signals along with five Tumour samples than the majority, the composition of these signals was shown to be different between the samples in question.

The six labelled samples in Figure 6.34 were excluded from PCA to further investigate potential separation between the other 22 samples. A one component model ($R^2\boldsymbol{X}$ = 0.262 and $Q^2\boldsymbol{X}$ = 0.134) was generated but the scores plot (Figure 6.37) did not show separation between Tumour and Normal samples. However, it is interesting to note that nine of the 14 Normal samples had the greatest $\boldsymbol{t}$[1] values with six clearly separated from Tumour samples. The loadings plot (Figure 6.38) indicated levels of C-18 CH_3 cholesterol (0.694 ppm), $–CH_2N–$ head group of choline phospholipids (3.595 ppm) and glycerol C-3 methylene proton of DAGPLs and $–OCH_2$ head group of ethanolamine phospholipids in the bin centred at 4.000 ppm were reduced in the nine aforementioned Normal samples. This observation will be investigated in context of breast cancer grade in Section 6.1.2.2.2. A PLS-DA model was not able to be generated.

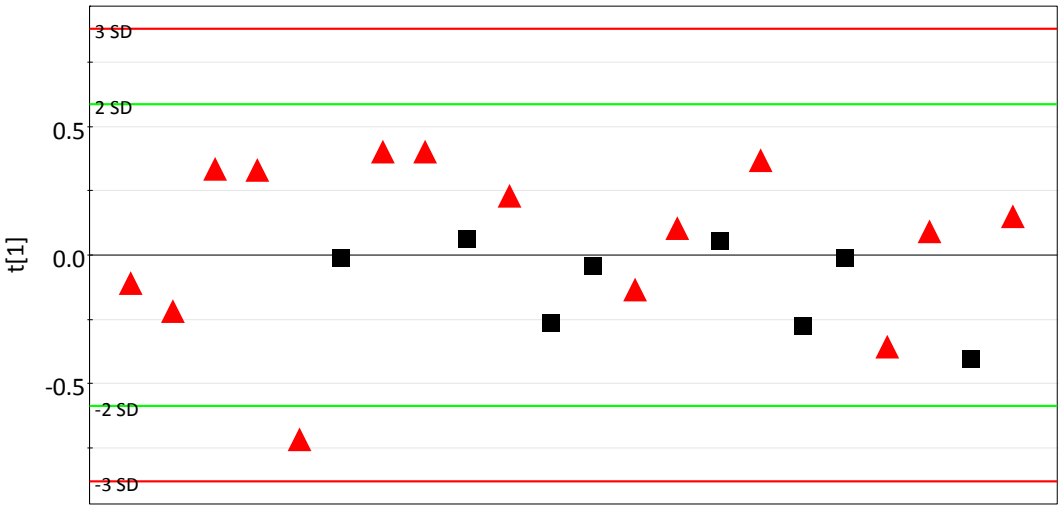


Figure 6.37 PCA scores plot for lipophilic extract samples excluding large signals and the (labelled) six samples most remote from the majority in Figure 6.34. Coloured according to tissue type.

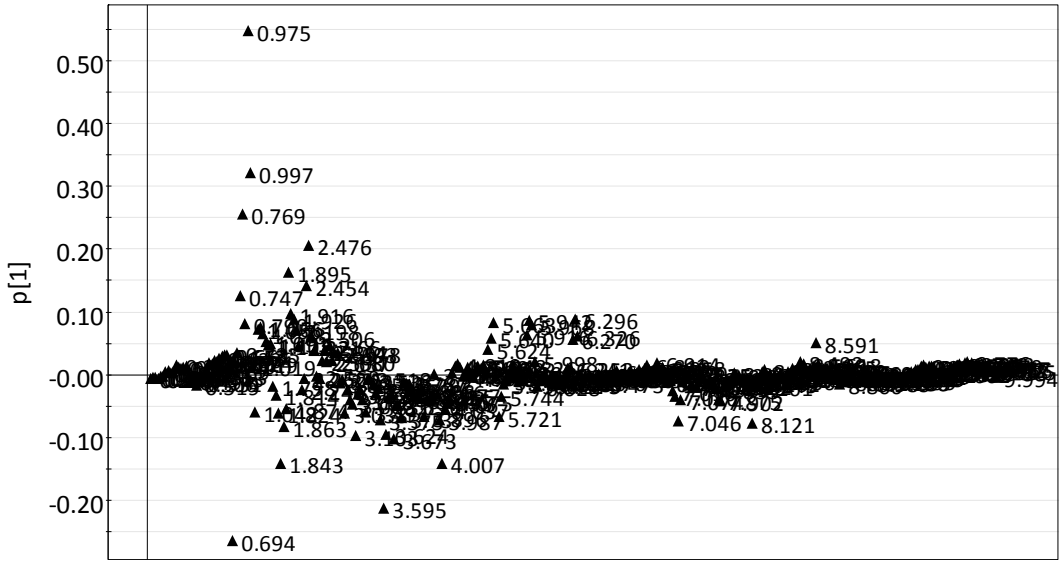


Figure 6.38 PCA loadings plot corresponding to the model displayed in Figure 6.34.

Composition of the same lipid species can affect signal positions in spectra, for example bulk CH_2 of fatty acid resonance appears further upfield with increased HDL lipoproteins, the reverse is applicable to increased VLDL lipoproteins.[128,224] It was postulated that use of small bins may provide information into lipid composition. 0.001 ppm fixed bins were employed but a similar outcome ensued in terms of position of samples in the scores plot (Figure 6.39). Smaller bins allowed further insight into which part of the CH_2 resonance discriminated samples primarily in PC 2 (Figure 6.40).

The bin centred at 1.275 ppm, range 1.261-1.291 ppm, discriminated between samples in PC 2 when 0.02 ppm variable bins were used but use of 0.001 ppm fixed bins showed many bins in the 1.26 ppm region caused the trend and bins in the 1.27-1.28 ppm region aided separation along PC 1. Spectral investigation showed samples with a high PC 2 scores value generally displayed the maximum intensity part of the resonance at the highest upfield chemical shift. Tumour and Normal samples are not separated based on this observation. As previously, part of the CH_2 resonance around 1.32 ppm and choline phospholipids had high ***p***[1] modulus values. There is potential to indicate predominant density, either high or low, of lipoproteins relative to other samples in the data set by using small bins though peak shifting, for instance due to difference in pH and small variation in

temperature between samples,[225] needs to be fully accounted for and deconvolution of the signal is advocated to accurately determine the amount of different density lipoproteins.[224]

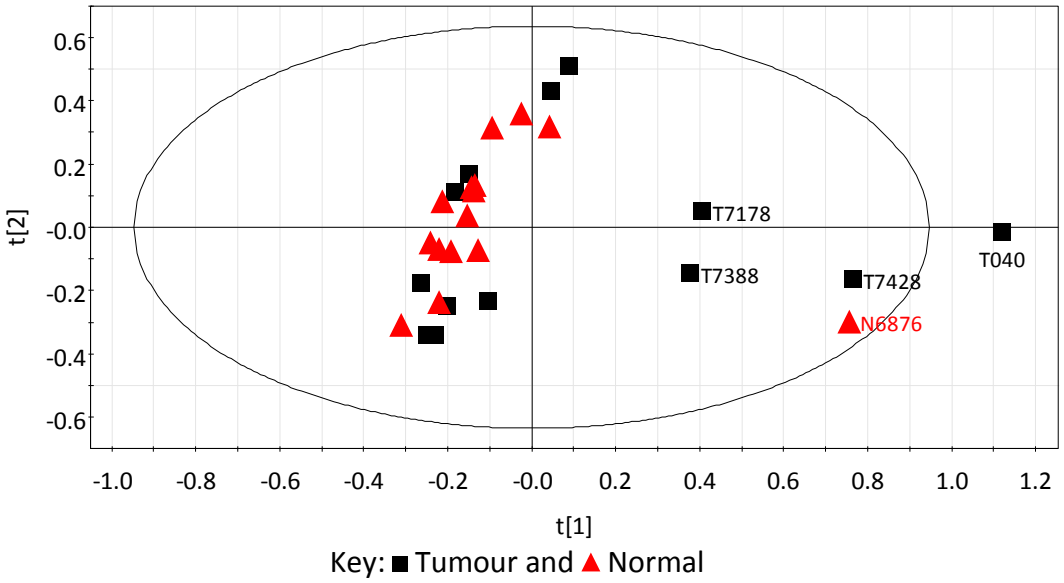


Figure 6.39 PCA scores plot for lipophilic extract samples using 0.001 ppm fixed bins. Coloured according to tissue type. Samples most remote from the majority are labelled.

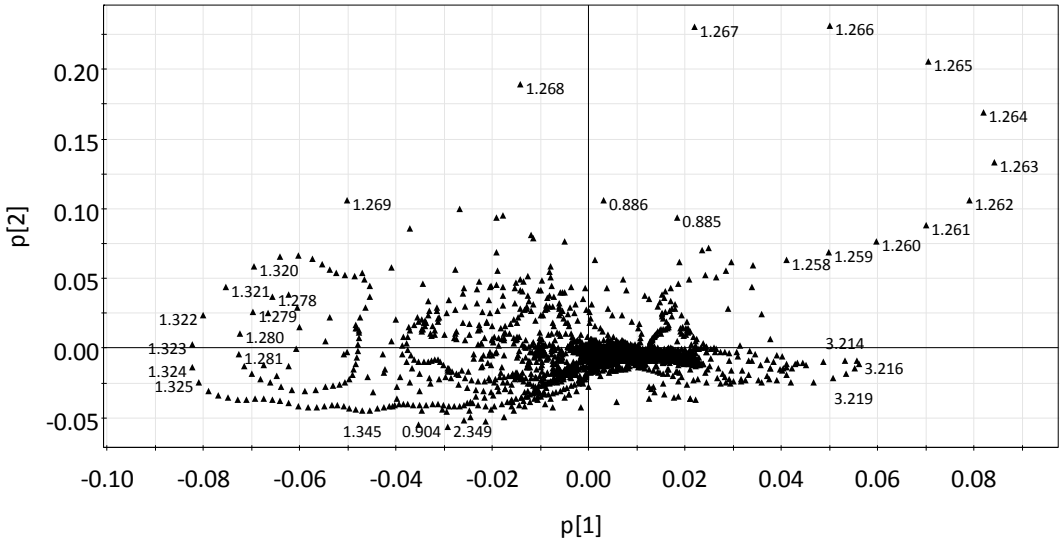


Figure 6.40 PCA loadings plot corresponding to the model displayed in Figure 6.39. Selected bins are labelled.

6.1.2.2 Evaluation of Breast Cancer Severity

6.1.2.2.1 Analysis of Whole Spectrum

In an analogous manner to aqueous extracts, breast cancer grade was investigated in an attempt to reveal metabolic markers of disease severity.

PCA of PQ normalised data produced a two component model ($R^2\boldsymbol{X}$(cum) = 0.964 and $Q^2\boldsymbol{X}$(cum) = 0.945) when Tumour samples alone were included: scores plot shown in Figure 6.41 but loadings plot not shown due to similarity with Figure 6.31, for which Normal samples were included.

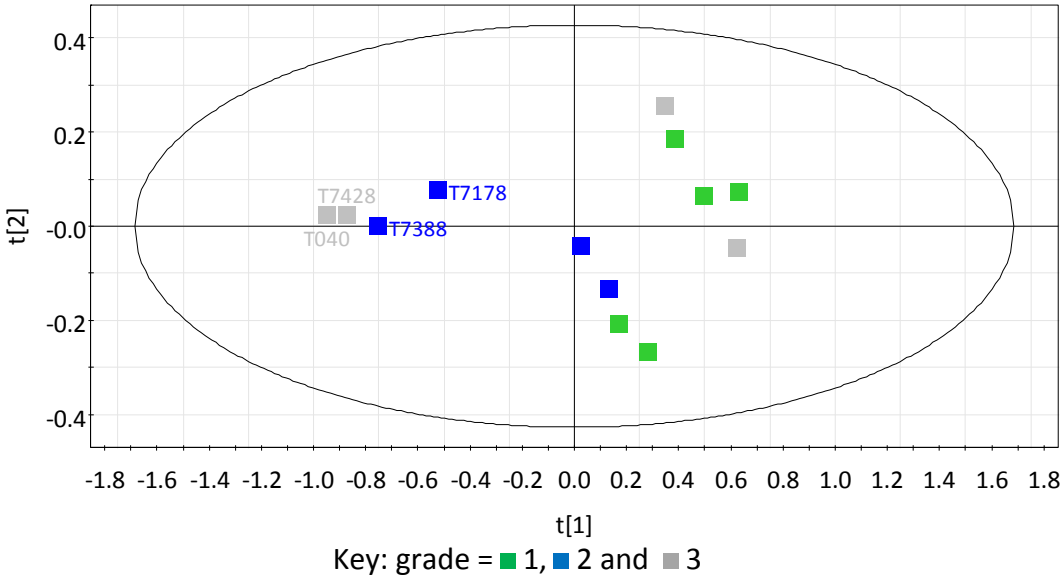


Figure 6.41 PCA scores plot for lipophilic extract Tumour samples. Samples coloured according to tumour grade. Samples most remote from the majority are labelled.

The four Tumour samples that were most remote from the majority along PC 1 when Normal samples were included in analysis were still separated. That two grade 3 samples (T023 and T7242) had high positive scores values indicated no separation based on tumour grade. One sample was ER, PR and DCIS positive and HER negative, which follows the status of descriptors for almost all samples for which data was recorded. The other sample was also ER and PR positive but status of DCIS and HER

was not recorded. Based on these four tumour descriptors there is no reason to explain the positioning of the two samples. Exclusive of these two samples tentative separation could be viewed, with the progressive nature of tumour severity being a possible reason for distinction within grade 2 samples, but grounds for exclusion of the two grade 3 samples in question do not exist.

As per aqueous extracts, investigation ensued to determine whether there was a connection between Tumour sample grade and the paired Normal sample. Figure 6.42 shows the scores plot; model statistics apply from previous work related to Figure 6.30 as does the loadings plot shown in Figure 6.31. Samples N6758 and N6876 were included to show positions relative to other samples despite not having a paired sample.

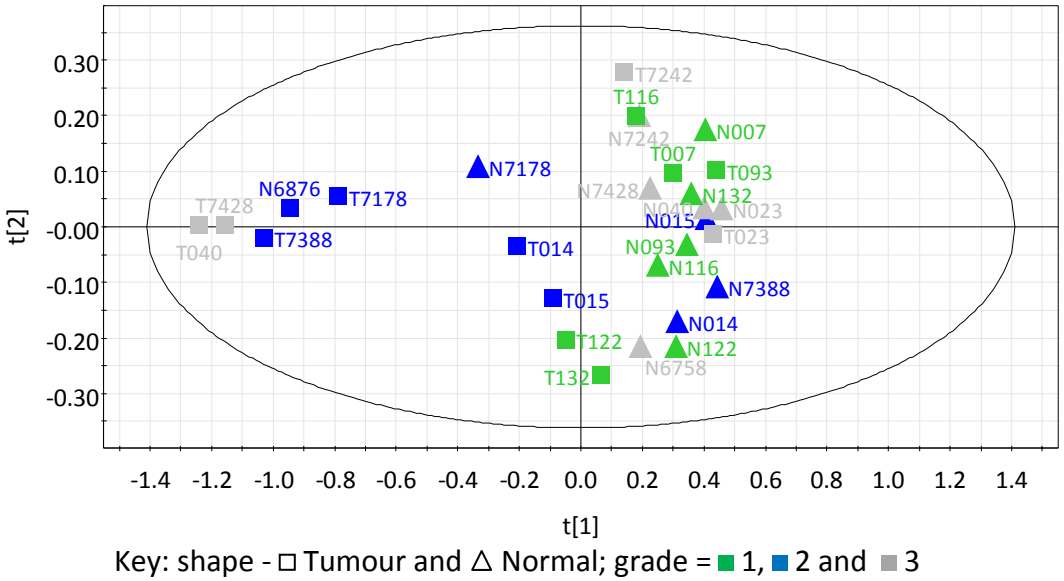


Figure 6.42 PCA scores plot for lipophilic extract samples. Tumour samples coloured according to tumour grade and Normal samples coloured according to the Tumour counterpart.

No pattern is apparent in scores space between paired Tumour and Normal samples though for each of 007, 023 and 7242 the paired samples are located in close proximity to each other; positioning of Tumour samples for the last two has recently been discussed. As also concluded from aqueous extracts analysis, adjacent tumour

free (Normal) tissue does not show a metabolite profile that is related to Tumour sample grade.

6.1.2.2.2 Analysis of Spectra Excluding Large Signals

Using PCA a one PC model ($R^2\boldsymbol{X}$ = 0.499 and $Q^2\boldsymbol{X}$ = 0.367) was built for the 13 Tumour samples. Figure 6.43 shows the scores plot whilst the loadings plot has been omitted due to similarity with Figure 6.38. T014 belongs to grade 2 class and is more distant from grade 1 samples than when the whole spectrum was used (Figure 6.41) but the two grade 3 samples are still positioned in scores space that is populated by grade 1 samples.

If a relationship between Normal samples and grade of Tumour samples was present it would be expected that groups of Normal samples in scores space would be present dependent on the grade of the paired Tumour sample irrespective of the positioning of the Tumour sample. Figure 6.44 does not show this observation indicating the relationship is not present. Samples N6758 and N6876 were included to show positions relative to other samples despite not having a paired sample. The scores plot in Figure 6.35 applies and associated model statistics.

As shown by Figure 6.37, the nine samples with greatest positive $\boldsymbol{t}$[1] values when the six samples most remote from the majority were removed were Normal samples but grade of the corresponding Tumour sample (Figure 6.45) is not related to this observation. Four, two and three samples are associated with grades 1, 2 and 3, respectively. The scores plot also shows more clearly than Figure 6.44 that positioning of grade 1 Tumour samples and counterpart Normal samples does not appear to be related.

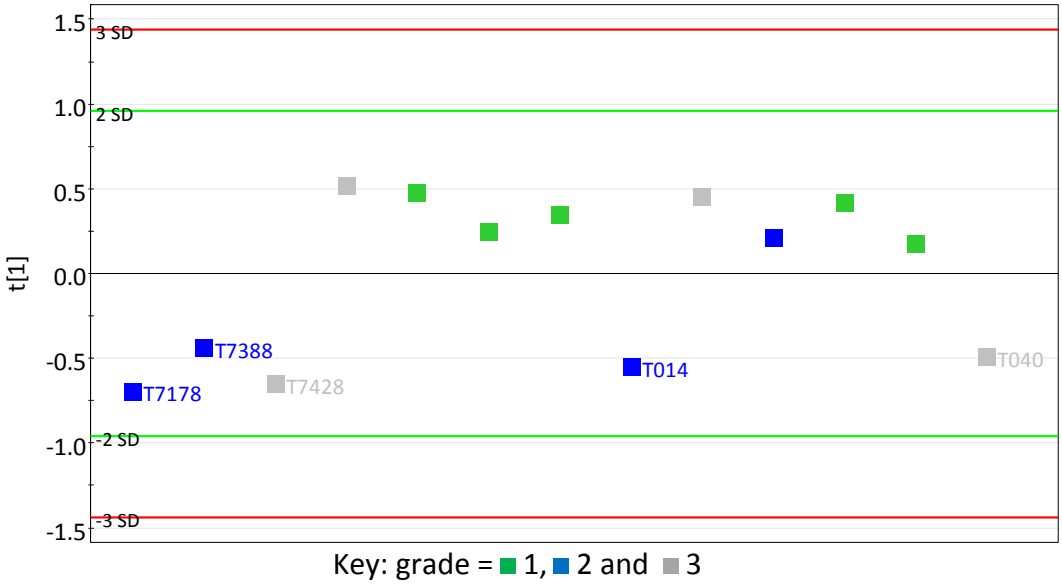


Figure 6.43 PCA scores plot for lipophilic extract Tumour samples excluding large signals. Samples coloured according to tumour grade.

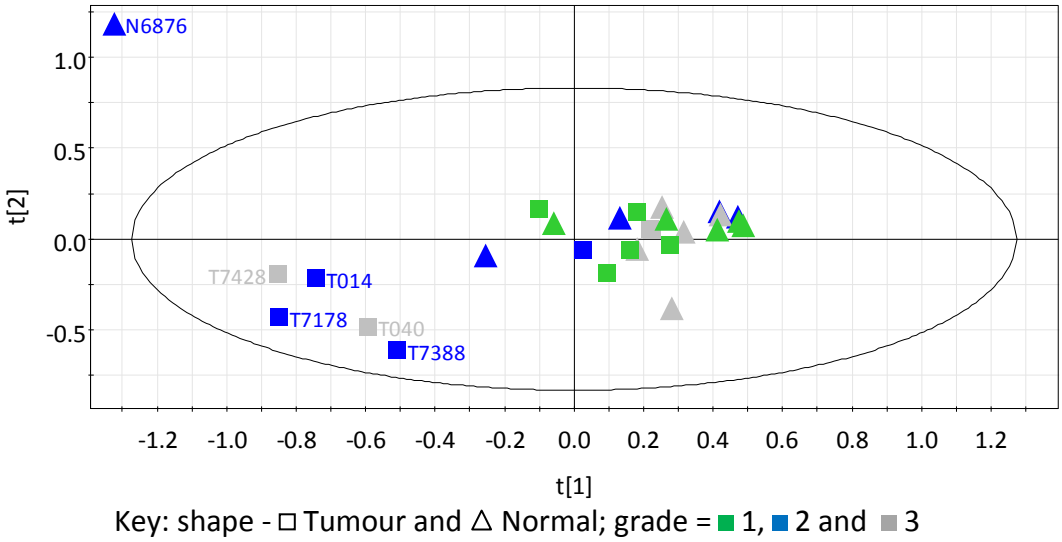
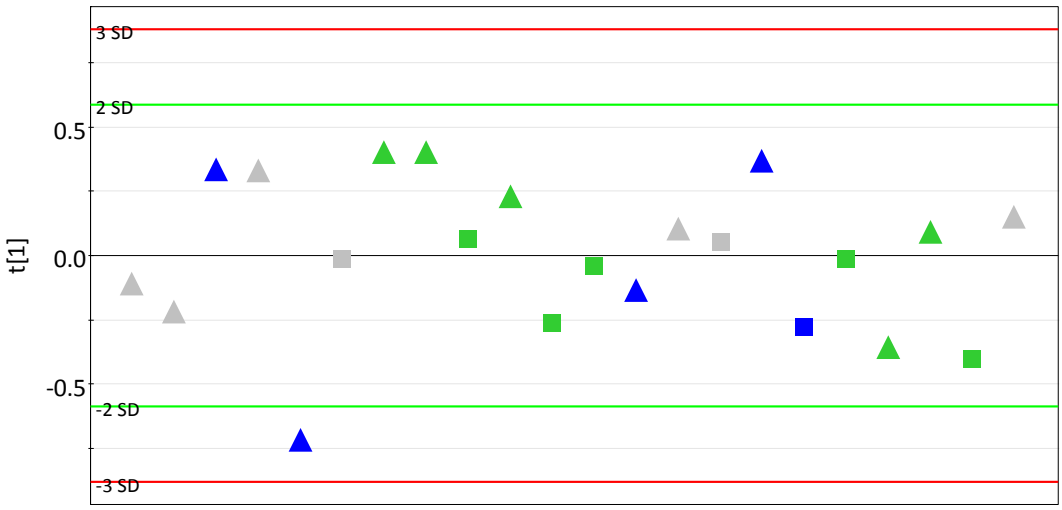


Figure 6.44 PCA scores plot for lipophilic extract samples excluding large signals. Tumour samples coloured according to tumour grade and Normal samples coloured according to the Tumour counterpart.



Key: shape - □ Tumour and △ Normal; grade = ■ 1, ■ 2 and ■ 3

Figure 6.45 PCA scores plot for lipophilic extract samples excluding large signals and the (labelled) six samples most remote from the majority in Figure 6.44. Tumour samples coloured according to tumour grade and Normal samples coloured according to the Tumour counterpart.

PLS-DA models were not able to be built for data used in the above three analyses. PCA of either whole spectra or small intensity signals did not provide information on metabolite changes that were related to breast cancer severity as assessed through tumour grade.

6.2 Conclusions

For proliferation cancer cells require large quantities of lipids and macromolecules, which are mostly synthesised from intermediates of glycolysis and glutaminolysis. Analysis of aqueous extracts from breast tissue biopsies revealed a number of metabolites were significantly up- or down-regulated between Normal and Tumour samples. In accordance with the Warburg effect, glucose was decreased and lactate and alanine increased in Tumour samples. Glutamine, glutamate, GPC, PCho, glycerophospholipids, taurine, creatine and as yet unidentified signals were raised in Tumour samples. Glycolysis derived UDP-GlcNAc and UDP-GalNAc were identified as raised in Tumour samples through investigation of the aromatic region only. PCA of lipophilic extracts revealed a small number of samples, both Normal and Tumour, contained lower amounts of triglycerides and upon removal of these prominent

signals increased levels predominantly of head groups of various phospholipids were observed in Tumour samples in the aforementioned sub-group. Normal and Tumour samples, however, could not be discriminated either inclusive or exclusive of the sub-group of samples.

Breast cancer severity as assessed through tumour grade did not correlate with differences in levels of metabolites in analyses of either aqueous or lipophilic extracts. Normal samples did not reveal the propensity of complementary Tumour samples in terms of grade nor was there correlation between metabolite levels of paired samples.

Results from both aqueous and lipophilic extracts showed the importance of normalisation and scaling methods.

The hospital site of sample collection, as inferred throughout by no separation of samples based on length of code identifiers (which was different for the two hospital sites) was not relevant to the findings of this study. This also applied to the source of samples, either BCCTB or direct, collected from the same site.

This is the first study, to the author's knowledge, that has used *de facto* pattern recognition techniques, *i.e.* PCA and PLS-DA, for analysis of both aqueous and lipophilic aqueous extracts of breast tissue. The sample cohort is very well defined in terms of menopausal status and subtype, grade and DCIS status of the tumour; due to heterogeneity of the disease reducing factors that can affect the outcome of results is very important especially in small sample number metabolomics studies. Conclusions made here with regards to relatively homogeneous breast cancer could be supposed to other statuses of the disease, due to certain common features, as hypotheses to be verified or disproved.

Chapter 7. Conclusions

^{1}H-NMR based metabolomics has successfully been applied to the field of breast cancer. GC analysis of amino acid concentrations has also been implemented.

7.1 Identification of Breast Cancer Status using Plasma and Urine

Metabolomics analysis of plasma or urine was unable to identify differences between patients with breast cancer and those who exhibited non-cancerous breast abnormalities.

Glucose, lactate and lipids were the most abundant metabolites in plasma. Exclusion of glucose due to levels being highly affected by diet did not result in improved separation between samples based on a number of descriptors.

Known potential confounding factors reported in the patients' medical records were accounted for as much as possible. Individually, none were shown to be the major contributor to positioning of samples in multivariate space. When differences in BMI, age and smoking status were minimised together (for the whole spectrum) tentative separation was observed between case and control samples based on increased levels of lipids and lactate and a decreased level of glucose for the former. A larger number of samples that conformed to strict criteria for all three of the variables would be needed to validate the observation.

Analysis of urine from the same cohort revealed that many patients consumed acetaminophen that was not recorded in their medical records. Neither exclusion of the signals pertaining to acetaminophen and acetaminophen by-products nor the exclusion of samples containing the signals revealed information to determine case from control samples.

Unlike plasma, minimising BMI, age and smoking status did not reveal metabolites that potentially distinguished between those afflicted with breast cancer and non-afflicted patients.

Amino acid (AA) concentrations were quantified using GC for plasma aliquots obtained from the same cohort as for the NMR analysis. Of the 26 AAs extracted using a commercially available system only nine were used in analysis due to presence at insufficient concentration for quantification or unsatisfactory reproducibility of the other AAs. Significant differences in the AA levels were not present between case and control samples.

7.2 Identification of Breast Cancer Tumour Descriptors using Plasma and Urine

Amongst those afflicted by breast cancer, the grade of tumour was not able to be distinguished using either plasma or urine. This was true of all tumour types and the preponderant type, invasive ductal carcinomas. Samples were not differentiated based on ER or PR scores or HER2 status. For urine, whereby all three descriptors were combined, tumour subtype was not responsible for the positioning of samples in scores space.

Using GC for quantification of AAs, tentative separation was observed between certain tumour grades. A validated PLS-DA model indicated concentrations of the seven AAs present in all samples (ALA, GLY, VAL, LEU, ISO, PRO and PHE) were greater in grade 2 samples compared to grade 1 samples. Univariate analysis revealed that PRO was decreased in the latter grade. However, a grade two sample had much a greater concentration of PRO and upon removal the difference was no longer significant. It must be stressed that there was no justification for this action because the concentrations in all three replicates were comparable but highlights the influence that one sample can have on small data sets.

7.3 SHY Analysis of Plasma and Urine Data Related to Breast Cancer

Tentative correlations were revealed between full resolution plasma and urine data when SHY was applied that otherwise would not have been identified.

Pearson's analysis was susceptible to outlying samples. Broad, unique singlets in the urine range 8.123-8.579 ppm for a case sample caused positive correlation with two plasma regions. The origin of the urine signals was not able to be established. For control samples, many areas of positive correlation were present that were related to the signal at 3.15 ppm in plasma. The signal was attributed to $DMSO_2$, a common dietary supplement, and one sample contained a substantially greater amount than any other. However, it was revealed that for case samples glycerol in plasma was positively correlated to (an) unidentified urine species within the range 4.259-4.296 ppm and was confirmed by the more rigorous Spearman's analysis.

Spearman's analysis also revealed that the aforementioned urine area was negatively correlated with the upfield section of lipid signals. This indicates a negative correlation between glycerol and lipids. It was speculatively hypothesised that levels of HDL, which exhibits the greatest upfield shift and contains lesser amounts of triglycerides compared to LDL and VLDL, were related to glycerol due to the connectivity between glycerol and glycerides and hence triglycerides. Negative association has been observed between triglycerides and HDL-cholesterol in CVD but patients' medical records do not contain information related to this affliction.

7.4 Identification of Breast Cancer Status and Tumour Descriptors using Tissue Extracts

Due to the heterogeneity of breast cancer reducing factors that can affect the outcome of results is very important especially in small sample number metabolomics studies. The sample cohort is very well defined in terms of menopausal status, DCIS presence and subtype and grade of the tumour.

7.4.1 Aqueous Extracts

Analysis of aqueous extracts from breast tissue biopsies revealed a number of metabolites that were significantly up-regulated or down-regulated in Tumour samples compared to Normal samples. Decreased glucose and increased lactate was present in Tumour samples, which is in accordance with the Warburg effect. The different levels were clearly visible from inspection of spectra. Alanine, glutamine, glutamate, GPC, PCho, glycerophospholipids, taurine, creatine and as yet unidentified signals were raised in Tumour samples.

For proliferation, cancer cells require large quantities of lipids and macromolecules, which are mostly synthesised from intermediates of glycolysis and glutaminolysis; many metabolites are connected with the TCA cycle.

Due to the small size of signals in the aromatic region compared to those in the aliphatic region, changes in signal intensity were not apparent in loadings space. Upon MVA of the aromatic region alone UDP-GlcNAc, UDP-GalNAc and unknown signals in the range 8.178-8.229 ppm were identified as elevated in Tumour samples. Glucose can undergo a series of reactions to form a moiety of UDP-GalNAc, which can be reversibly catalysed to UDP-GlcNAc. Acetyl CoA, another moiety, can be made from pyruvate, which is abundant due to increased glycolysis in Tumour cells.

Grade of tumour could not be distinguished by levels of metabolites and there was no indicator of prediction from Normal samples of the grade of the paired Tumour sample.

7.4.2 Lipophilic Extracts

Five samples, four Tumour and one Normal, were distinct from other samples and dominated scores space. This was attributed to containing lower levels of α-, β-, allylic and bulk CH_2 fatty acid and bulk methylene fatty acid whilst levels of choline

phospholipids were increased. Removing the five samples did not discriminate between Normal and Tumour samples. Neither removal of the bulk CH_2 fatty acid region, which contributed to over half of the spectrum integral, or all of the large signals that accounted for approximately 90% of the spectrum integral separated Tumour from Normal samples. The same samples were still isolated and the Tumour ones were shown to contain increased levels of head groups of various phospholipids.

Grade of tumour could not be distinguished by levels of metabolites and the grade of Tumour samples could not be indicated from the paired Normal samples.

7.5 Further Work

Verification of the findings that indicated a difference between groups of samples is required in a larger cohort, which is more representative of the patient population and reduces the impact of individual samples on the statistical significance of results.

Confounding factors can obscure potential differences between classes of samples. A larger cohort could provide greater numbers of samples that are matched according to various parameters known to influence metabolomics studies or maintain similar numbers that are sourced from a more defined cohort.

Breast cancer is a heterogeneous disease so determining whether the findings from the tissue extracts analysis, which used a well defined cohort in terms of tumour status, applied to other types of tumour would provide information on the extent to which the various types are different.

In addition to collecting tumour and adjacent normal tissue samples, if less-invasive samples were obtained from the same patients, such as plasma and urine, SHY could be performed on multiple sample types. If covariances were identified

between extracts of tissue samples and biofluid samples the need to collect the tissue sample would be reduced. In addition to benefits to the patient, sample integrity would be easier to maintain because tissue samples require immediate snap-freezing and the processing steps for metabolomics studies are more involved. In this study, SHY revealed covariances between plasma and urine that included lipids and non-lipids, which indicated that there could be increased value for investigation into covariances between different biofluids and both lipophilic and aqueous tissue extracts. For SHY analysis of control samples, identification of the unknown urine metabolite that was shown to correlate with glycerol and lipids in plasma could aid interpretation of the metabolic pathways connecting the species.

MS has greater sensitivity than NMR spectroscopy[45,52] but metabolite identification is not as universal.[52,53] Use of the former technique could reveal further differences in metabolite levels with SHY being employed to determine whether correlations were present between the two platforms, which could reveal information related to metabolic pathways.

Due to SHY being in its infancy a reduction in the number of spurious covariances and the processing power required whilst retaining non-spurious covariances could result from establishing the optimum number of data points to be used.

Chapter 8. Experimental Methods

8.1 NMR Sample Preparation

Chemicals were purchased from Sigma-Aldrich Company Ltd. (Poole, Dorset, UK), unless otherwise stated. NMR tubes (S-5-500-7, Norell) were purchased from GPE Scientific Ltd. (Leighton Buzzard, Bedfordshire, UK).

8.1.1 Samples for ^{1}H-NMR Analysis

8.1.1.1 Plasma Samples (Chapter 2)

Samples were stored at -80^{o}C before defrosting at room temperature. Samples were centrifuged (Hettich Mikro 120 (C1204) Centrifuge, angle rotor A1242) at 11,992 *g* for 5 minutes. 350 µl of plasma supernatant was added to 408 µl of deuterium oxide (D_2O). The mixture was vortexed for 8 s before transferring 600 µl to a 5 mm NMR tube. Samples not analysed immediately were stored at 4^{o}C for a maximum of 1.5 hours.

8.1.1.2 Urine Samples (Chapter 3)

Samples were stored at -80^{o}C before defrosting at room temperature. Samples were centrifuged (Hettich Mikro 120 (C1204) Centrifuge, angle rotor A1242) at 11,992 *g* for 5 minutes. 460 µl of urine supernatant was added to 230 µl of phosphate buffer. 100 ml of stock phosphate buffer solution (pH 7.43) contained 2.885 g sodium phosphate monobasic (Na_2HPO_4), 0.525 g sodium phosphate dibasic (NaH_2PO_4), 0.0172 g (1 mM) trimethylsilyl propanoic acid (TSP) and 0.0195 g (3 mM) sodium azide (NaN_3) in 20 ml of D_2O and 80 cm^3 of ribonuclease (RNase) free water. The phosphate buffer was shaken thoroughly and placed in a sonicator, interspersed by shaking, until salts dissolved. The urine/phosphate buffer mixture

was vortexed for 8 s before transferring 600 µl to a 5 mm NMR tube. Samples not analysed immediately were stored at 4^{o}C for a maximum of 1.5 hours.

8.1.1.3 Tissue Extracts (Chapter 6)

8.1.1.3.1 Separating Aqueous and Lipophilic Components

Samples were stored at -80^{o}C and surrounded by dry-ice during transferral (<2 minutes) to the processing laboratory. The frozen weight was recorded and the sample returned to the dry-ice. For samples >41 mg, 4 ml g^{-1} of methanol and 0.85 ml g^{-1} of water were added. The volume to weight ratio was increased for samples $\leq$41 mg to provide a minimum of 200 µl and the same scaling factor applied to subsequent volumes of extraction solvents. Following homogenisation (Omni Tissue Homogeniser; Camlab, Over, Cambridgeshire, UK) at 35,000 rpm using a stainless steel probe and glass vial, the sample was vortexed for 5 s. 2 ml g^{-1} of chloroform was added and vortexed for 5 s. A further 2 ml g^{-1} of chloroform and 2 ml g^{-1} of water were added and vortexed for 5 s. The sample was left on ice for 15 minutes before centrifugation at 1,023 *g* for 15 minutes. The solutions separated into an upper methanol/water phase (with polar metabolites) and a lower chloroform phase (with lipophilic compounds) with protein and cellular debris between the two layers. The upper layer was removed and was snap frozen using N_2. The lower layer was removed and transferred to a glass vial.

8.1.1.3.2 Processing of the Aqueous Component

Methanol and water were removed from the upper layer using a speed vacuum concentrator and the extracts stored at -80^{o}C. The extracts were resuspended in 680 µl of a 0.17% weight to volume solution of the sodium salt of TSP in D_2O, vortexed for 5 sec and centrifuged at 11,992 *g* for 5 minutes. 600 µl was transferred to a 5 mm NMR tube. All samples were analysed immediately.

8.1.1.3.3 Processing of the Lipophilic Component

Chloroform was removed from the lower layer under a stream of N_2 and the extracts resuspended in 430 µl of deuterated chloroform ($CDCl_3$; 99.8%), containing 0.03% volume to volume of tetramethylsilane (TMS), and 215 µl of deuterated methanol (CD_3OD; 99.8%; Eurisotop, Gif-sur-Yvette, Paris, France). The solution was stored at 4^{o}C for a maximum of three days before centrifugation at 1,023 *g* for 5 minutes and transferral of 600 µl to a 5 mm NMR tube. All samples were analysed immediately.

8.2 NMR Data Collection

All ^{1}H-NMR spectra were acquired on a Varian Unity Inova 500 spectrometer (Varian Inc., Palo Alto, California, USA) operating at 499.97 MHz proton frequency, at 20^{o}C. Samples were loaded into the probe and left for 5 minutes to allow temperature equilibration.

8.2.1 CPMG Experiment (Chapter 2)

The CPMG pulse sequence [RD – 90^{o} – (τ – 180^{o} – τ)$_n$ – acq] was used to obtain metabolic profiles for all plasma samples. The relaxation delay (RD) was 2 s, during which the water resonance was selectively saturated, τ was 1.5 ms and *n* was 150. For each spectrum 512 transients were collected into 16,384 pairs of data points with a spectral width of 8,000.00 Hz.

8.2.2 1D NOESY Experiment (Chapter 3)

The 1D NOESY pulse sequence [RD – 90^{o} – t_1 – 90^{o} – t_m – 90^{o} – acq] was used to obtain metabolic profiles for all urine samples. The RD was 2 s, t_m was 1.5 ms and t_1 was 3 µs. For each spectrum 512 transients were collected into 16,384 pairs of data points with a spectral width of 8,000.00 Hz.

8.2.3 Presaturation Experiment (Chapter 6)

The presaturation (PRESAT) pulse sequence was used to obtain metabolic profiles for all tissue extracts. For each aqueous and lipophilic spectrum 512 and 256 transients, respectively, were collected into 16,384 pairs of data points with a spectral width of 6,000.15 Hz. The RD was 0.5 s.

8.3 NMR Spectral Processing

All spectral data were processed using ACD Labs software 12.01 (Advanced Chemistry Development, Inc., (ACD/Labs), Toronto, Canada). An exponential line broadening was applied to each free induction decay (FID) prior to zero filling to 65,536 points and Fourier transformation. The resulting spectra were phased, baseline corrected and referenced. Details of parameters are listed in Table 8.1

Table 8.1 NMR spectral processing parameters

Samples	**Pulse Sequence**	**Line Broadening (Hz)**	**Reference**	
			Compound	**Chemical Shift (ppm)**
Plasma	CPMG	0.5	α-glucose	5.23[142]
Urine	1D NOESY	0.5	TSP	0.000[142]
Aqueous tissue extracts	PRESAT	0.5	TSP	0.000[142]
Lipophilic tissue extracts	PRESAT	0.5	TMS	0.000[226]

8.3.1 Additional NMR Data Processing for Multivariate Statistical Analysis

8.3.1.1 Binning and Dark Regions

Normalisation to constant sum and spectral binning was performed using ACD Labs software 12.01. All datasets were integrated into bins of 0.02 ppm width implementing 50% ‘looseness’, which allowed bin width to vary between 0.01 and

0.03 ppm. The software used an algorithm that simultaneously adjusted starting and final points of two adjacent bins to minimise the overall height of their borders resulting in bin divisions being more likely to occur at the edges of peaks. The lipophilic extract dataset was also binned using 0.001 ppm fixed widths.

Prior to binning over a spectral range (Table 8.2), several dark regions were created for spectra from plasma (Table 8.3), urine (Table 8.4), aqueous tissue extracts (Table 8.5) and lipophilic tissue extracts (Table 8.6) to exclude signals, such as from the water region, contaminants and certain metabolites, from subsequent MVA.

Table 8.2 Spectral range for which binning was performed.

Sample	**Range (ppm)**
Plasma	0.000-10.000
Urine	0.150-10.000
Aqueous tissue extracts	0.210-10.000
	5.446-9.362
Lipophilic tissue extracts	0.230-10.000

Table 8.3 Dark regions used for the plasma spectra (Chapter 2).

Dark Region (ppm)	**Excluded**	**Comment**
4.20-5.70	Water region	Variable water suppression efficiency
3.18-3.94	Glucose	Excluded later to remove influence from statistical models

Table 8.4 Dark regions used for the breast cancer urine spectra (Chapter 3).

Dark Region (ppm)	Excluded	Comment
4.500-6.200	Water region and urea	Variable water suppression efficiency
2.143-2.200 3.588-3.664 3.883-3.926 7.120-7.166 7.300-7.386 7.434-7.481	Acetaminophen and acetaminophen by-products	Excluded later to remove influence from statistical models
3.034-3.064 4.043-4.073	Creatinine	Excluded later to remove influence from statistical models
3.954-3.984 7.530-7.590 7.611-7.662 7.815-7.867	Hippurate	Excluded later to remove influence from statistical models

Table 8.5 Dark regions used for the breast cancer aqueous tissue extracts spectra (Chapter 6).

Dark Region (ppm)	Excluded	Comment
2.520-2.750	Citrate	Possible contaminant from collection tube
3.341-3.371	Methanol	Extraction solvent
4.750-5.000	Water region	Variable water suppression efficiency
1.314-1.374 4.085-4.144	Lactate	Excluded later to remove influence from statistical models

Table 8.6 Dark regions used for the breast cancer lipophilic tissue extracts spectra (Chapter 6).

Dark Region (ppm)	Excluded	Comment
3.321-3.450	Methanol	Extraction solvent
7.403-7.594	Chloroform	Extraction solvent
1.200-1.458	Bulk CH_2 fatty acid	Excluded later to remove influence from statistical models
0.796-0.961	CH_3 fatty acid	Excluded later to remove influence from statistical models
1.117-1.787	Bulk CH_2 and β-CH_2 fatty acids	
1.932-2.191	Allylic CH_2 fatty acid	
2.268-2.446	α-CH_2 fatty acid	
2.701-2.964	Diallylic CH_2 fatty acid	
3.147-3.450	Methanol and choline phospholipids	
4.083-4.758	glycerol C-1 methylene protons of diacylglycerophospholipids	
5.056-5.606	Olefinic CH fatty acid and glycerol C-2 proton of diacylglycerophospholipids	
7.210-7.725	Chloroform containing region	

8.3.1.2 Probabilistic Quotient Normalisation

Urine (Chapter 3), full resolution plasma and urine (Chapter 4), and aqueous and lipophilic tissue extracts (Chapter 6) data sets were PQ normalised.

Constant sum normalised data was exported from ACD Labs software 12.01 into Excel 2007 (Microsoft, Redmond, Washington, USA). For all "control" spectra the median of every variable (bin/data point) was calculated; this acted as the reference spectrum.[69] For each test spectrum quotients were calculated for every variable relative to the reference spectrum. The median of these quotients, excluding noise regions (regions included in the calculations are shown in Table 8.7), was calculated per test spectrum and all variables of the test spectrum were divided by this median. If regions were subsequently excluded from MVA, new constant sum normalised data was obtained and the PQN steps repeated.

Table 8.7 Regions included for the calculation of the median quotient of test spectra.

Samples	Urine	Plasma (full resolution)	Urine (full resolution)	Aqueous tissue extracts	Lipophilic tissue extracts
Region (ppm)	0.494-4.500	0.807-1.053	0.493-3.588	0.725-2.520	0.559-3.321
	6.200-9.162	1.173-1.385	3.664-3.883	2.750-3.341	3.450-3.752
	9.261-9.301	1.452-1.495	3.926-4.500	3.371-4.750	3.828-5.766
	9.344-9.374	1.851-1.921	6.200-7.120	5.164-6.175	5.892-6.043
	9.668-9.726	1.965-2.165	7.166-7.300	6.507-6.544	6.249-6.371
	9.957-10.000	2.188-2.295	7.386-7.434	6.764-8.045	7.011-7.403
		2.317-2.480	7.481-9.000	8.098-8.730	7.594-7.748
		3.022-3.978	9.022-9.162	8.812-8.872	8.107-8.137
		4.035-4.149	9.261-9.301	8.926-8.953	
		6.861-6.910	9.344-9.375	9.136-9.165	
		7.021-7.051	9.666-9.726	9.323-9.352	
		7.162-7.206			
		7.721-7.748			
		8.440-8.459			

8.4 Multivariate Statistical Analysis

All multivariate statistical analyses were performed using SIMCA-P+ software, version 12.0.1.0 (Umetrics, Umeå, Sweden). This software was used to scale (Pareto unless otherwise stated) and mean centre the data.

PCA was performed to view any intrinsic clustering in the samples, which can be seen in the scores (***t***) plots. The loadings (***p***) plots were used to identify regions of the spectra (bins) responsible for any clustering or outliers in the scores plot.

PLS-DA was performed to improve distinction of separation between groups of interest and to produce models from which classes of samples not used in that model could be predicted. Scores plots allowed observation of class discrimination and ***w**** plots were used to identify the bins accounting for the discrimination.

The quality of each model was assessed by the goodness of fit ($R^2\boldsymbol{X}$) and the ability to predict the class membership of new sample ($Q^2\boldsymbol{X}$ for PCA and $Q^2\boldsymbol{Y}$ for PLS-DA) . In addition, for PLS-DA the proportion of the classification data ($\boldsymbol{Y}$) accounted for ($R^2\boldsymbol{Y}$) was also considered. The predictive ability of the PLS-DA models was assessed by 'leave-one-out' cross-validation. Further validation was performed using permutation testing whereby the classes of samples were randomised and a PLS-DA model built. The $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ values should be less than those generated for the original model, which were based on the real classifications. The maximum number of permutations permissible by the software was performed, 999.

8.5 Univariate Statistical Analysis

Univariate tests of statistical significance were performed on any bins that were considered influential in a trend of interest. The integral values of bins were tested for normality using the Shapiro-Wilk test. If the data distribution was normal and the variances of the two groups were equal the Student's *t*-test was performed. If the variances were not equal the Welch-Aspin test was used, a modification of the Student's *t*-test. For non-normal data the non-parametric Mann-Whitney U test was used instead. For both, the mean integral values were compared between groups and a *p*-value obtained for the statistical significance of the difference between groups. For tissue extracts data, samples belonged to the same patients when the two groups being tested were Tumour and Normal. Therefore, the data were paired and required the use of the paired-samples *t*-test or the Wilcoxon test for normal or non-normal data, respectively. All tests of normality and significance were performed in IBM SPSS Statistics 20.0 (IBM Corporation, Armonk, New York, USA). Resulting *p*-values were adjusted for multiple comparisons using the FDR correction in R software, version 2.7.0 (R Foundation for Statistical Computing, Vienna, Austria).

8.6 SHY Analysis of Breast Cancer Plasma and Urine (Chapter 4)

PQ normalised case and control sample matrices were analysed separately. The Pearson's and Spearman's correlation coefficients and their statistical significances (through *p*-values) were calculated for every column in the plasma matrix against every column in the urine matrix using the "corr(x,y)" function in MatLab, version 7.12.0.635 (R2011a) (The MathWorks, Inc., Natick, Massachusetts, USA). In the resulting correlation matrix any correlations with *p*-value >0.001 were set to zero to reduce spurious correlations. The correlation matrix was plotted using the "imagesc" function.

A personal computer (Evesham Technology, Evesham, Worcestershire, UK; processor speed = 3.0 GHz, RAM = 2 GB, processor type = Intel(R) Pentium(R) 4 CPU) was used in an attempt to produce a correlation matrix generated from every data point of full resolution data (37,815 data points per spectrum) but the processing requirements surpassed the capabilities of the personal computer. Approximately a square 6,000 matrix was the maximum correlation matrix size that could be generated. A higher specification personal computer (Dell Inc., Round Rock, Texas, USA; processor speed = 2.93 GHz, RAM = 16 GB, processor type = Intel(R) Core(TM) i7 CPU) was used but a square 37,815 correlation matrix was not able to be generated. The data points pertaining to the region 4.500-6.200 ppm, which contained water and additionally urea for urine, were excluded resulting in 30,852 data points per spectrum but generation of a correlation matrix was not possible. Removing data points that had a value of zero for all samples resulted in 9,060 and 27,450 data points per spectrum for plasma and urine, respectively (Table 8.7, third and fourth columns), and allowed generation of a correlation matrix.

Integrals of the data points were plotted in OriginPro 8 SR4, version 8.9051 (OriginLab Corporation, Northampton, Massachusetts, USA).

8.7 GC Analysis of Breast Cancer Plasma (Chapter 5)

The EZ:Faast (Phenomenex, Macclesfield, UK) system was used to extract and derivatise AAs from breast cancer plasma samples. The sample preparation procedure listed in the information booklet that accompanied the system was followed. The solid phase extraction step was performed *via* a sorbent packed tip that bound AAs but allowed interfering compounds to flow through. AAs on the sorbent were extruded into the sample vial and derivatised with a reagent in aqueous solution. Derivatised AAs concomitantly migrated to the organic layer and were additionally separated from interfering compounds. An aliquot (2 µl) from the organic layer was analysed by GC with FID using an Autosystem XL gas chromatograph (Perkin-Elmer, Waltham, Massachusetts, USA). A ZB-AAA 10 m x 0.25 mm column (Phenomenex, Macclesfield, UK) was used in constant flow mode with a split ratio of 1:15 at 250^{o}C, a helium flow rate of 1.5 ml min^{-1} and a temperature increase of 32^{o}C min^{-1} between 110^{o}C and 320^{o}C.

Using the EZ:Faast system, calibration standard sets used for AA quantification were made and consisted of three samples containing all 26 AAs that could be quantified at increasing concentration (50, 100 and 200 nmol ml^{-1}). Up to three injections of the same standard set were performed within a 24 hour period (the maximum storage time at room temperature recommended by the manufacturers). After every 10 test samples a standard set was run. Three injections of every test sample were performed within a 24 hour period.

Multivariate and univariate statistical analysis was performed as described in Section 8.4 and 8.5 with the exceptions that data were not normalised and UV scaling was performed rather than Pareto.

Appendices

Appendix 1: Arterial Disease

Initial work focussed on arterial disease, as discussed below, but due to samples (plasma, urine and carotid or femoral plaques) being unavailable for a substantial period of time that would be collected specifically for metabolomics studies, investigation into metabolic markers related to breast cancer ensued.

A1.1 Introduction

Atherosclerosis is a disease of large and medium-sized arteries. It is the most important contributor to CVD[227] and heavily affects the burden of myocardial infarction (MI; heart attack) and stroke. Published work provides various, but high, figures for the seriousness of the problem. In Westernised societies atherosclerosis is the main cause of morbidity and mortality, [228] contributing to about 50% of all deaths.[229,230] CVD is the most common, and second most common, cause of death in European men and women respectively, who are under 65 years of age[231] and the disease is expected to become the leading cause of death worldwide in the twenty-first century.[227,231,232] It is estimated that CVDs affect 57 million Americans, and each year causes 954,000 deaths and costs $259 billion.[230] Imaging techniques commonly employed clinically are biased towards the detection of severe, flow-limiting stenoses (narrowing of blood vessels) and are relatively poor at detecting early disease.[233] Consequently, the first presentation of arterial disease is often an acute event such as myocardial infarction or stroke.[233]

The effect of atherosclerosis depends on where it occurs: in the arteries that perfuse the brain it can cause stroke and transient ischaemic attacks (TIA) whereas in the heart it can lead to myocardial infarction and heart failure.[234] Stroke is often preceded by TIA[234] whereby reduced supply of blood and oxygen occurs to the brain temporarily so the symptoms soon go unlike for stroke. It is sometimes called

a mini stroke. Renal impairment, hypertension, abdominal aortic aneurysms and critical ischaemia (CI) can occur if atherosclerosis affects other arteries.[235] Peripheral arterial disease (PAD) normally occurs in legs with symptoms most common in calf muscles with some patients being affected in the thigh and buttock regions.[236] There are various categories associated with progression of the disease. Intermittent claudication is a less severe form with the symptoms being aching or pain in the legs reproducibly brought on by walking and relieved by rest.[237] CI is more severe with pain at rest. Ulceration and gangrene can occur and if revascularisation surgery is not performed in time amputation is required.[236] Over £200 million is the estimated annual cost to the UK of CI.[236] Patients suffering from PAD have more than twice the risk of exhibiting coronary heart disease (CHD) events, such as MI, but only 25% of patients are undergoing treatment.[238]

A1.2 Previous Metabolomics Studies of Arterial Related Diseases

There have been a number of metabolomics studies into arterial disease and associated factors.[145,239-243] Many potential markers have been identified in relation to blood pressure in urine by ^{1}H NMR: formate, sodium and alanine are positively correlated with blood pressure, with hippurate showing the inverse association.[145] 4-hydroxyproline was identified as the metabolite of interest in plasma of patients with non-ST-elevation acute coronary syndrome (NSTEACS) using GC-MS.[240] ST is the segment between S and T time points in an electrocardiogram (ECG). An elevated ST segment is associated with acute MI.[244] The authors have suggested the decrease in plasma hydroxyproline levels in NSTEACS versus control may reflect a status of low collagen synthesis and turnover. Analysis of gas chromatography coupled to time-of-flight mass spectrometry (GC-TOF-MS) serum data identified pseudouridine, 2-oxoglutaric acid, 2-hydroxy, 2-methylpropanoic acid, erythritol, and 2,4,6-trihydroxypyrimidine as showing significant differences in levels between patients with systolic heart failure and controls.[239] 2-oxoglutarate is a major intermediate of the TCA cycle and a change in several other constituents of the TCA cycle has been shown in plasma using LC-MS: 6 of the 23 most-changed metabolites

after exercise in patients with inducible ischaemia are involved in the Krebs cycle.[241] The same study also showed the six most discordant metabolites between cases and controls are citric acid, γ-aminobutyric acid, uric acid, MET193, MET 200 and MET288. One of the first metabolomics human studies had shown that metabolomics can diagnose the presence and severity of CHD[242] but a later study proved the results were due to a confounding factor, statin treatment, which caused decreases in cholesterol, low-density lipoprotein (LDL) and triglycerides, thus affecting the lipid peak contributions in the spectra.[243] Most arterial disease studies have focussed on coronary related problems with many fewer concerned with PAD.[245]

A1.3 NMR Analysis of Plasma

The following chapter will describe ^{1}H NMR analysis of plasma obtained from 75 patients who conformed to a control group or one of two arterial disease groups (Table A.1); one group exhibited coronary problems whilst the other consisted of patients suffering from claudication.

Prior to detailed analysis of the NMR spectra it was immediately apparent that citrate was present as a contaminant: two large doublets centred at 2.52 and 2.65 ppm.[246,247] The most likely source of the 'contamination' was sample collection vessels that used citrate as an anti-coagulant. Vacuettes coated with lithium heparin have become the usual vessels for blood collection but previously citrate and ethylenediaminetetraacetic acid (EDTA) had been used extensively especially for "bio-bank" samples.[246] The region 2.461-2.741 ppm containing the citrate resonances was excluded from subsequent analysis.

Table A1.1 Summary of patients' demographics.

Descriptor / Group	Total	Male / Female	Non-Diabetic / Diabetic	Non-Smoker / Smoker
Claudication	26	18	13	7
				6
			5	4
				1
		8	8	4
				4
			0	0
				0
Coronary	23	18	14	11
				3
			4	3
				1
		5	5	3
				2
			0	0
				0
Control	26	19	17	15
				2
			2	1
				1
		7	4	2
				2
			3	1
				2

A1.3.1 Reference Compound Investigation

A number of different metabolites have been used to reference ^{1}H-NMR spectra, including lactate, alanine, acetate and glucose[71,246,248,249] rather than TSP,[48,250] which has been used traditionally, but spectra and results showing comparison are not readily available. Signals referenced to TSP can exhibit sizeable chemical shift

variation between samples.[251] Additionally, in plasma, due to binding with albumin, TSP cannot be used to quantify the concentration of other compounds,[252] which is its secondary use. Using a CPMG pulse sequence heavy molecules are filtered out so TSP bound to albumin is not detected and because the ratio of unbound to bound TSP changes, variable levels of TSP result in CPMG spectra. The chemical shift of metabolite signals was investigated when spectra were referenced to different compounds (Table A1.2).

Table A1.2 Maximum chemical shift of signals using various reference compounds.

Reference Compound	**Chemical Shift (ppm)**	
	Reference Signal[103]	Maximum Variation of Other Signals*
TSP	0.000	0.011
Lactate (CH_3)	1.317	0.004
Alanine (CH_3)	1.463	0.006
Acetate (CH_3)	1.910	0.004
Glucose (H_1' α)	5.223	0.004

*Not including those in the table.

The maximum chemical shift variation of peaks was greater when referenced to TSP than to any of the other compounds. Most signals exhibited no more than 0.001 ppm variation when referenced using lactate, acetate or glucose.

Excluding TSP, when alanine was used as the reference compound the chemical shift range of other metabolites varied the most and when referenced to other compounds the alanine signal exhibited greater chemical shift range than lactate, acetate or glucose.

Of the 75 samples included in the analysis the one that exhibited the greatest upfield chemical shift of signals when referenced to TSP contained a greater quantity of unbound TSP. Data acquisition was not possible for two prepared plasma samples (not included in Table A1.1) and due to volume loss during the refreezing cycle insufficient quantity was available. Further deuterium oxide (D_2O)/TSP solution was added to provide the standard volume for data acquisition

(Section A1.5.2). The data were not included in MVA. Upfield chemical shift of signals was also displayed by the two samples (Figure A1.1a). Conversely, when referenced to acetate, alanine, glucose or lactate, the three samples showed greater downfield shift of TSP (Figure A1.1b) but not of other signals. These observations can be explained by protein interactions causing chemical shift variation:[251] the greater the amount of protein binding relative to the amount of TSP, the greater the downfield chemical shift of signals when referenced to TSP and *vice-versa*.

Of the reference compounds investigated acetate, glucose and lactate were superior to alanine and TSP. Signal overlap with underlying lipid resonances can occur for the lactate doublet at 1.317 ppm although this problem does not occur for the quartet at 4.103 ppm. The H_1' α-glucose doublet at 5.223 ppm will be used for referencing in future blood vessel disease work because it is readily resolved and of the alternative reference compounds to TSP has found common use.[246,249]

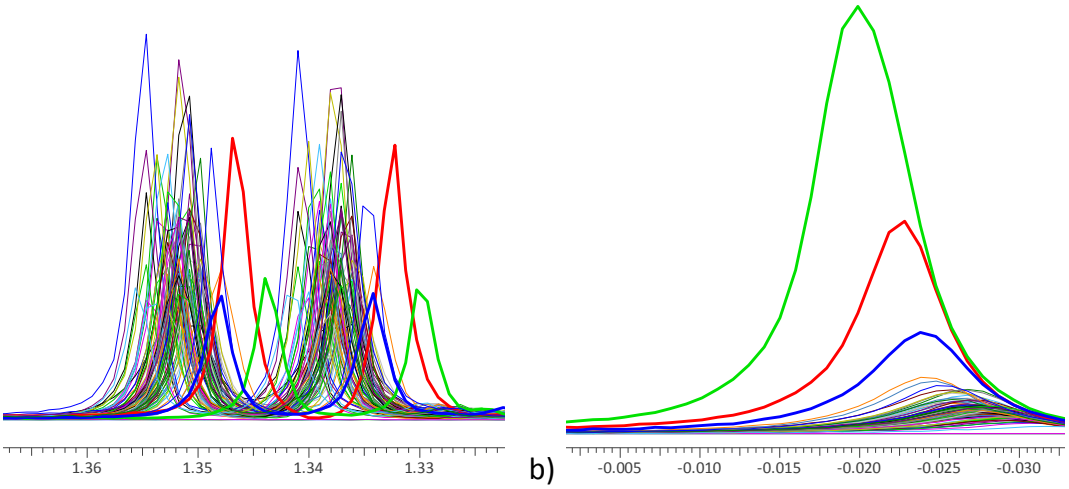


Figure A1.1 Chemical shift of spectra showing a) lactate signal referenced to TSP and b) TSP signal referenced to lactate. Sample included in MVA with the greatest amount of TSP shown in bold red and the two excluded refrozen samples shown in bold blue and green.

A1.3.2 Metabolomics Investigation

A1.3.2.1 Analysis of Disease Type Groups

PCA was performed with samples assessed according to disease type category (claudication, coronary or control). Scores and loadings plots of the five PC model ($R^2\boldsymbol{X}$(cum) = 0.693 and $Q^2\boldsymbol{X}$(cum) = 0.456) are shown in Figure A1.2 and Figure A1.3 respectively.

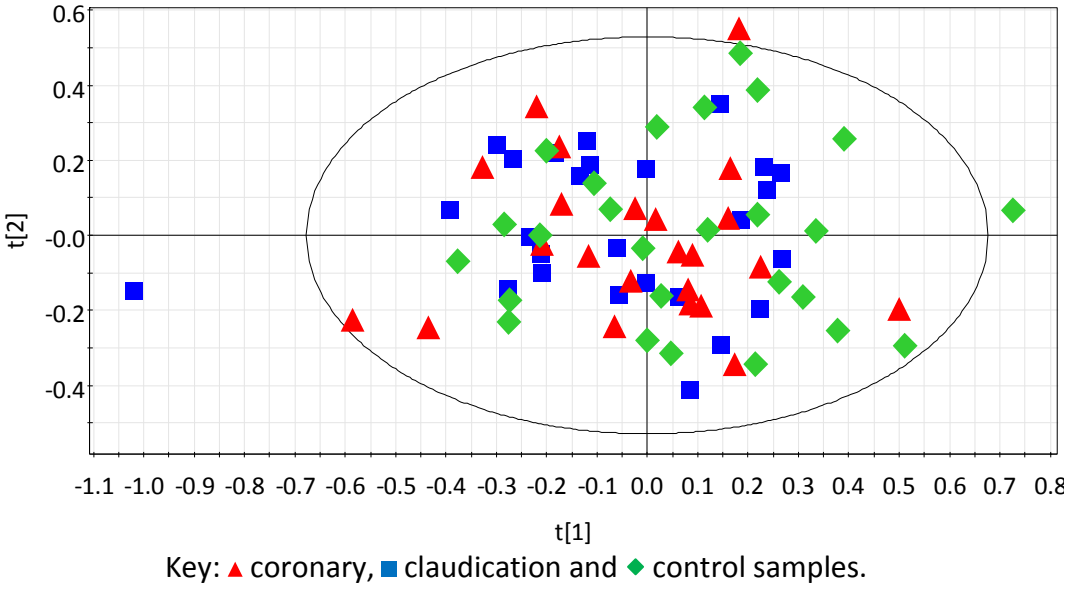


Figure A1.2 PCA scores plot for CPMG plasma data showing the first two model components. R^2X = 0.291 and 0.177, and Q^2X = 0.196 and 0.138 for PC 1 and PC 2, respectively.

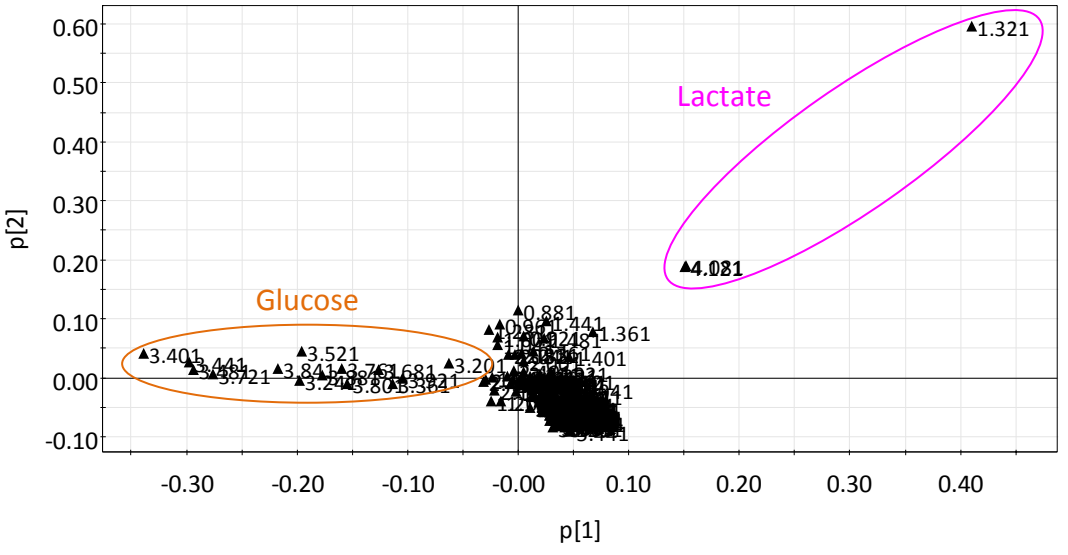


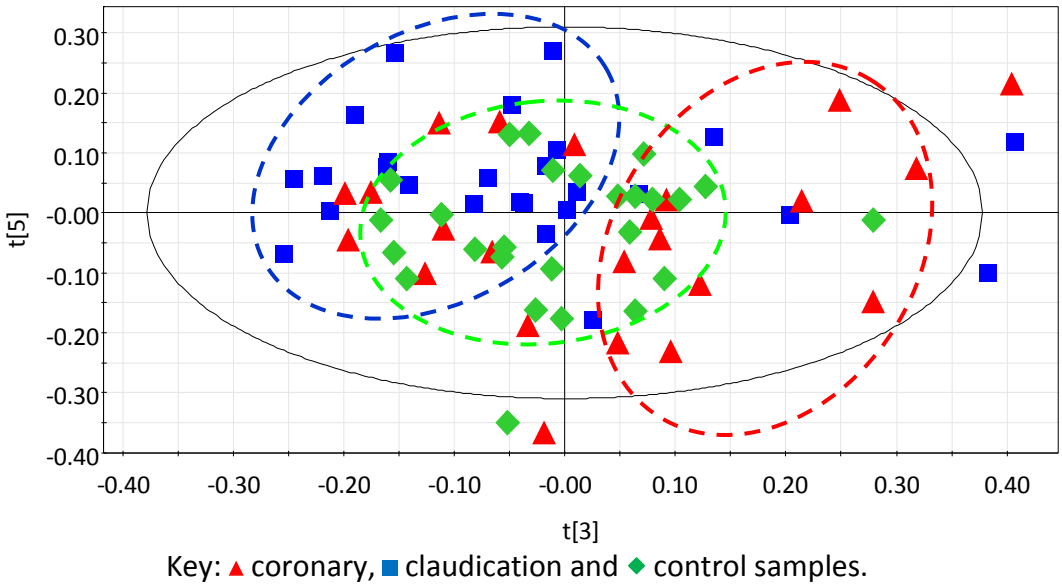
Figure A1.3 PCA loadings plot corresponding to the model displayed in Figure A1.2.

No separation was present in scores space between the three groups of samples. Lactate (all bins that had a ***p***[1] value greater than 0.10; centred at 1.321, 4.081 and

4.121 ppm) and glucose (all bins that had a ***p***[1] value less than -0.005; centred between 3.201-3.921 ppm, inclusive) containing bins dominated the loadings plot therefore the different sample groups were not represented by levels of the two metabolites. PC 5 versus PC 3 scores plot showed very tentative separation between the three sample groups based primarily in PC 3 (Figure A1.4). The loadings plot (Figure A1.5) indicated lipids (0.881 and 1.281 ppm) were highest and lactate lowest in coronary samples that were broadly identified by the red ellipse whilst for claudication samples broadly identified by the blue ellipse the reverse was implied; levels of the metabolites were of intermediate values for the control group. However, there were quite a few coronary samples positioned in scores space primarily populated by claudication samples and two of the samples with greatest ***t***[3] values belonged to the claudication group. Additionally, the $R^2\boldsymbol{X}$ and $Q^2\boldsymbol{X}$ values were low for both PCs (Figure A1.4 legend).

A one component model (data not shown; $R^2\boldsymbol{X}$ = 0.267, $R^2\boldsymbol{Y}$ = 0.075 and $Q^2\boldsymbol{Y}$ = 0.006) was generated using PLS-DA but with such a low $Q^2\boldsymbol{Y}$ value little credence can be attributed to the findings. Separation in scores space similar to that in PC 3 of the PCA model was observed but based on glucose and lactate with the former raised in claudication samples and the latter raised in coronary samples.

Levels of lactate[(253)] and glucose[(66)] in biofluids can be highly dependent on diet. Due to the impracticality of restricting patients' food and fluid intake prior to sample donation, the time of their last meal varied and drink types and quantities consumed were not recorded. Glucose (3.181-3.941 ppm) and lactate (1.301-1.341 ppm and 4.061-4.141 ppm) containing bins were excluded to allow other metabolites that are less influenced by diet to determine scores space positions of samples.



Key: ▲ coronary, ■ claudication and ◆ control samples.

Figure A1.4 PCA scores plot for CPMG plasma data showing PC 5 v. PC 3. R^2X = 0.091 and 0.061, and Q^2X = 0.065 and 0.091 for PC 3 and PC 5, respectively. Ellipses coloured according to the key indicate areas occupied by many samples of the respective groups and are for display only.

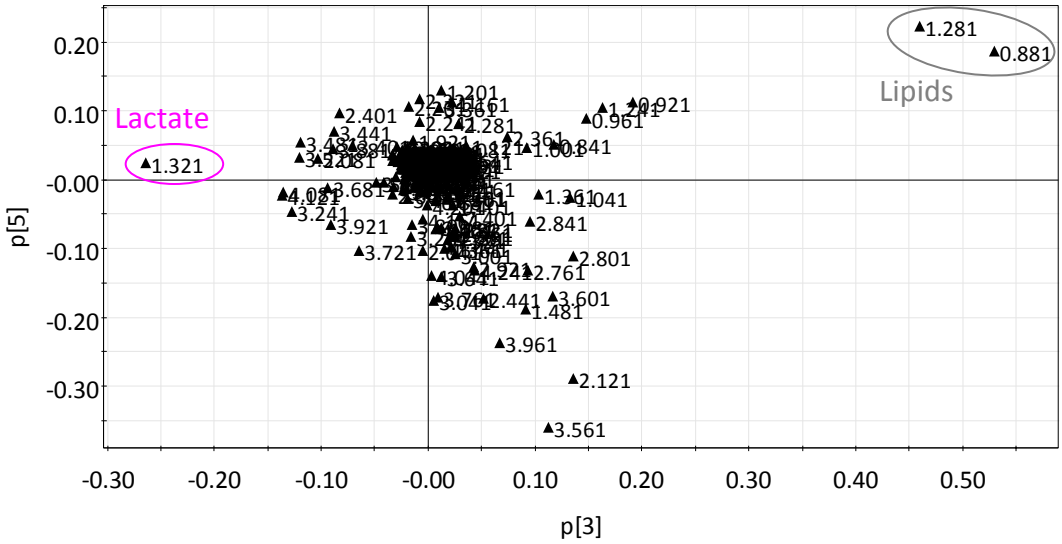


Figure A1.5 PCA loadings plot corresponding to the model displayed in Figure A1.4.

Figure A1.6 and Figure A1.7 show the scores and loadings plots, respectively, of PC 1 and PC 2 upon removal of glucose and lactate regions. Separation of the three groups was not present in these components or when other components of the four PC model (R^2X(cum) = 0.601 and Q^2X(cum) = 0.423) were visualised (data not shown).

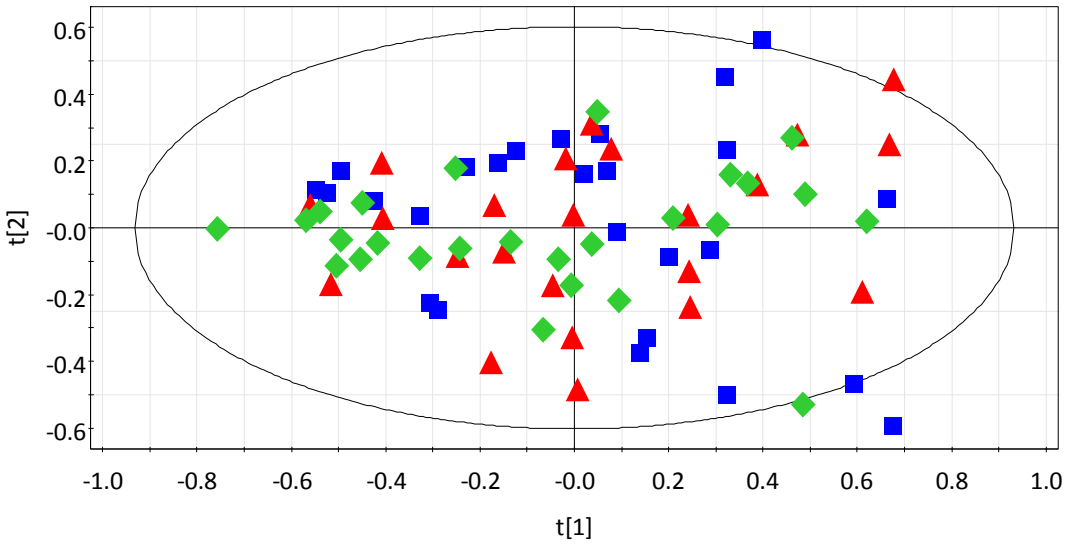


Figure A1.6 PCA scores plot for CPMG plasma data excluding glucose and lactate regions showing the first two model components. R^2X = 0.309 and 0.129, Q^2X = 0.262 and 0.093 for PC 1 and PC 2, respectively.

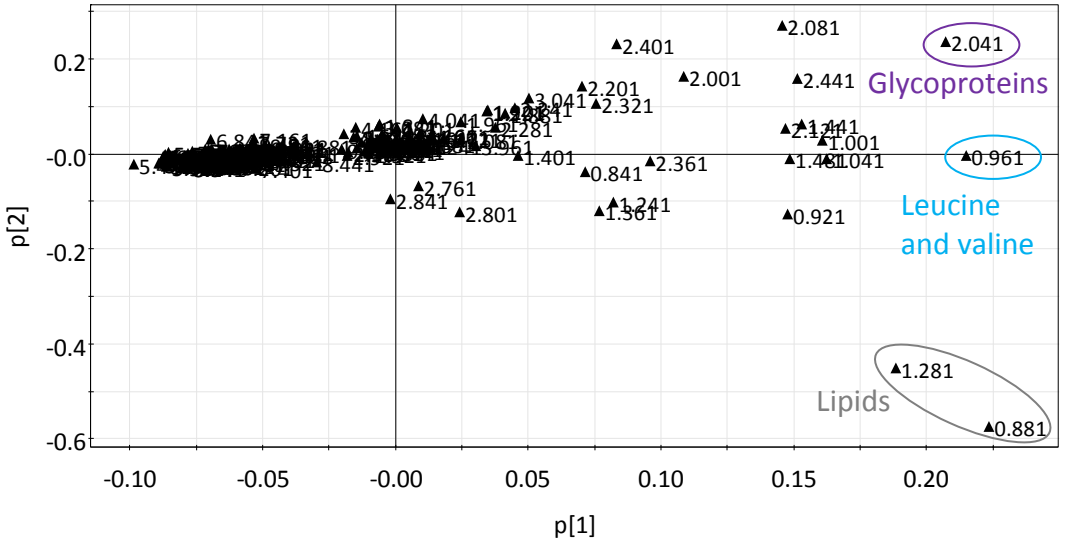


Figure A1.7 PCA loadings plot corresponding to the model displayed in Figure A1.6.

Lipids are important in determining the position of samples in scores space as are the bins, amongst others, centred at 2.041 ppm, the resonance arising from acetyl groups of glycoprotein, and 0.961 ppm, which contains the majority of the leucine triplet and upfield peak of valine doublet. PLS-DA produced a non-discriminatory, in terms of the three disease type groups, one PC model (data not shown; $R^2\boldsymbol{X}$ = 0.213, $R^2\boldsymbol{Y}$ = 0.084 and $Q^2\boldsymbol{Y}$ = 0.002) that had a very low $Q^2\boldsymbol{Y}$ value.

Although both are related to blood vessels, coronary and claudication afflictions are distinct so analysis was performed using all combinations of two disease type groups to investigate whether there were metabolic differences that were previously obscured when analysing all three groups. A summary of the models is shown in Table A1.3.

Table A1.3 Summary of models for all combinations of two disease type groups.

Groups Included	PCA			PLS-DA			
	Number of PCs	R^2X (cum)	Q^2X (cum)	Number of PCs	R^2X (cum)	R^2Y (cum)	Q^2Y (cum)
Coronary and control	3	0.544	0.384	0	/	/	/
Claudication and control	3	0.553	0.378	2	0.408	0.399	0.087
Coronary and claudication	4	0.630	0.445	0	/	/	/

No separation between groups was observed in PCA for any combination of two disease type groups (data not shown). Only the combination of claudication and control groups generated a PLS-DA model (scores plot shown in Figure A1.8 and loadings plot in Figure A1.9).

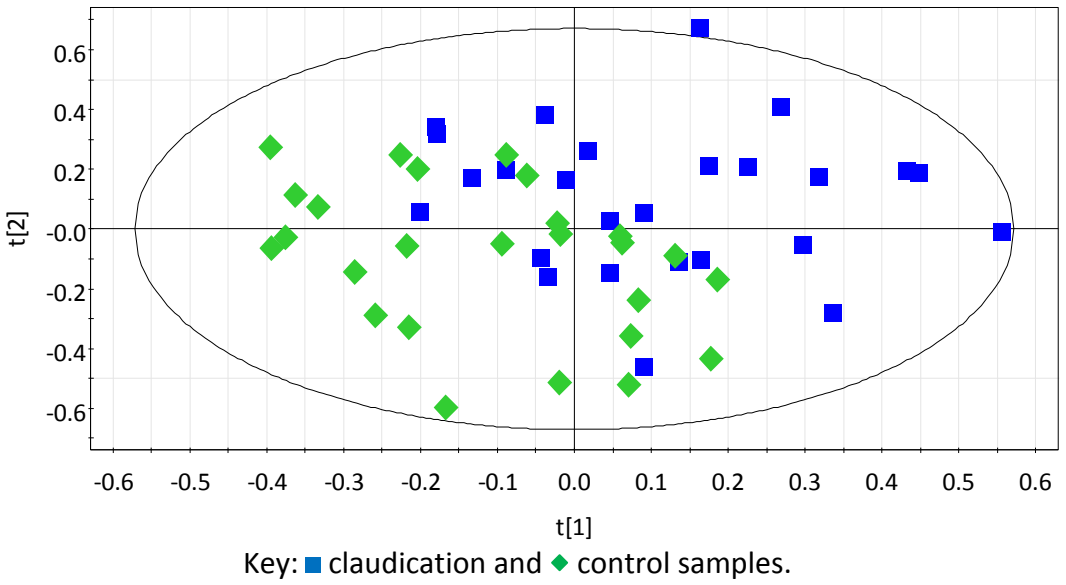


Figure A1.8 PLS-DA scores plot for CPMG plasma data excluding glucose and lactate regions of claudication and control samples.

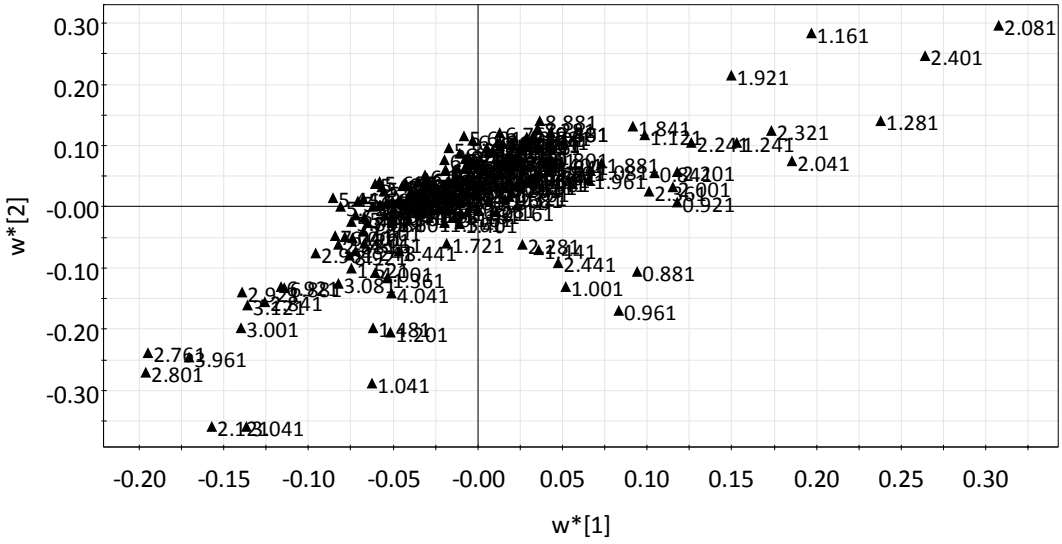
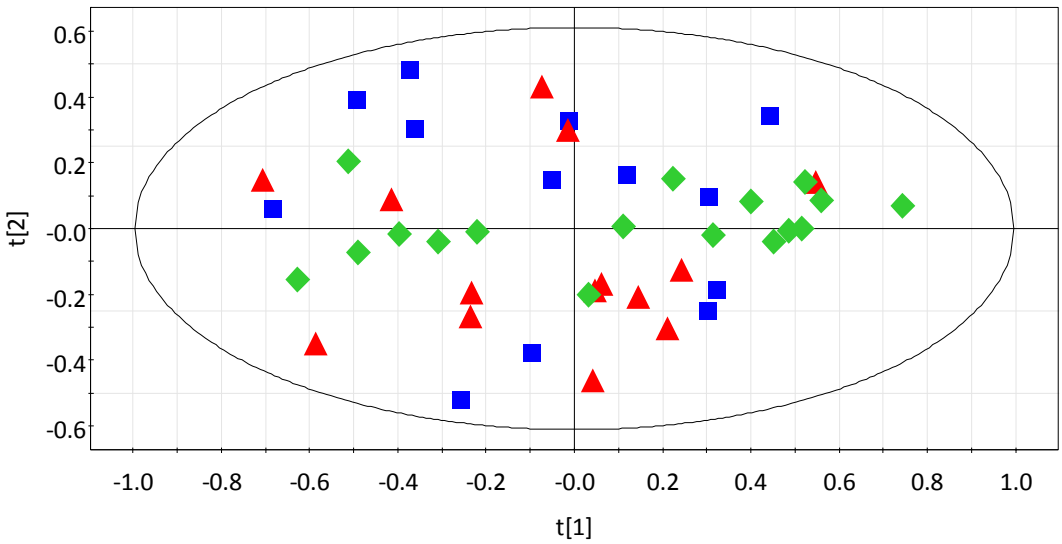


Figure A1.9 PLS-DA loadings plot corresponding to the model displayed in Figure A1.8.

The model was tested by cross-validation through exclusion of one-third of samples that were randomly assigned and their class predicted from the resultant model with the process performed three times to ensure the operation was applied to every sample. Of the 52 samples, 28 (54%) were predicted correctly (***Y***-predicted value >0.50) with 9, 6 and 13 having ***Y***-predicted values of <0.60, 0.60-0.70 and >0.70, respectively, and 24 incorrectly predicted. Not many more samples were correctly predicted than would be expected by chance hence the PLS-DA model in Figure A1.8 was surmised to have poorly fitted the data leading to the conclusion that there was no difference in metabolite levels in claudication and control samples when all samples were assessed. Permutation testing (plots not shown) also indicated the model was poor due to the high number of permuted models that had $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ values in excess of the values for the original model.

A number of confounding factors are known such as race,[254] sex[40,255] and diabetic state.[256,257] All samples were obtained from white European patients but the other two factors were not consistent for all samples (Table A1.1). Of the 75 samples, 55 were from males of whom 44 did not have diabetes. Analysis based on disease type group was performed for the reduced numbers in the sample cohort ("male, non-diabetic").

A three PC model ($R^2\boldsymbol{X}$(cum) = 0.562 and $Q^2\boldsymbol{X}$(cum) = 0.369) was generated using PCA. The scores plot is shown in Figure A1.10 but the loadings plot is not shown due to the similarity (vertical mirror image excepted) with Figure A1.7, for which all samples were included. Control samples tended to be positioned in low PC 2 scores space with coronary and claudication samples possessing greater, either positive or negative, $\boldsymbol{t}$[2] values. The contribution to the predicted variation ($Q^2\boldsymbol{X}$) was low in the second component, being only 0.017, and no separation was present in the first component. Using PLS-DA, a model was unable to be made. The three combinations of two disease type groups were analysed but separation was not present between any of the groups (data not shown; summary of models displayed in Table A1.4).



Key: ▲ coronary, ■ claudication and ◆ control samples.

Figure A1.10 PCA scores plot for CPMG plasma data excluding glucose and lactate regions of "male, non-diabetic" samples showing the first two model components. R^2X = 0.337 and 0.126, Q^2X = 0.293 and 0.017 for PC 1 and PC 2, respectively.

Table A1.4 Summary of models including only "male, non-diabetic" samples for all combinations of two disease type groups.

Groups Included	PCA			PLS-DA
	PCs	R^2X(cum)	Q^2X(cum)	PCs
Coronary and control	1	0.379	0.325	0
Claudication and control	2	0.482	0.300	0
Coronary and claudication	3	0.598	0.345	0

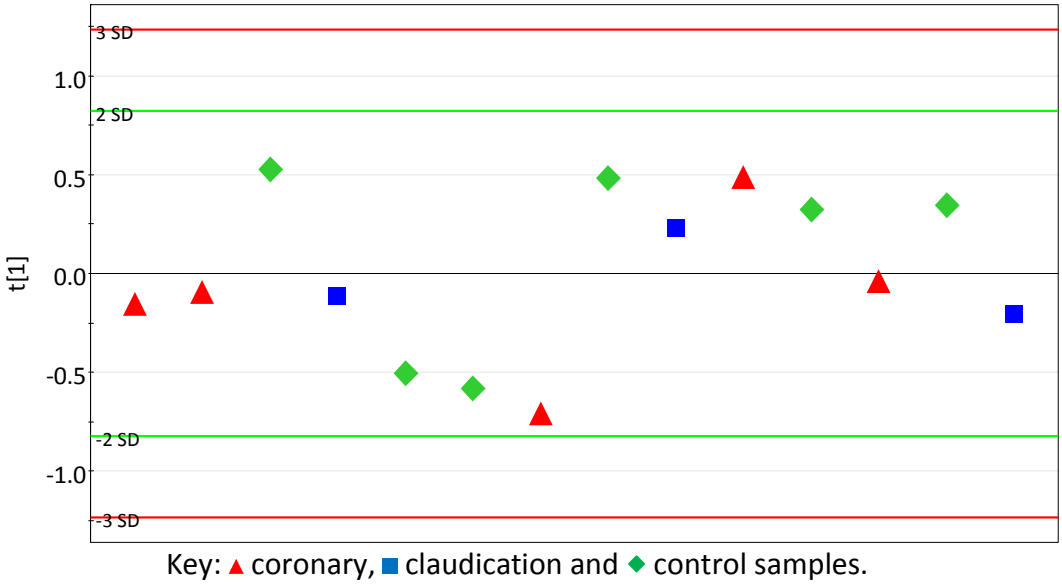
Smoking is another factor that can influence metabolomic profiles[255] and previous studies have excluded patients who smoke.[258] Excluding current smokers from the "male, non-diabetic" cohort reduced the sample number to 33. Analysis as previously reported was performed. No separation between disease type groups of the "male, non-diabetic, non-smoker" cohort was observed; scores and loadings plots are not shown due to similarity with those produced using "male, non-diabetic" samples. Table A1.5 provides a summary of the models.

Table A1.5 Summary of models including only "male, non-diabetic, non-smoker" samples for all combinations of two disease type groups.

Groups Included	PCA			PLS-DA
	PCs	R^2X(cum)	Q^2X(cum)	PCs
Coronary, claudication and control	2	0.493	0.344	0
Coronary and control	1	0.414	0.352	0
Claudication and control	1	0.403	0.318	0
Coronary and claudication	2	0.502	0.304	0

A final approach to analysis of disease type was to remove all samples that were not part of the "male, non-diabetic, never smoked" cohort. Smoking is a risk factor for blood vessel disease and although the risk for PAD decreases with time of abstinence, former smokers still have a greater risk than those who have never smoked.[259] Additionally, the risk increases with total cigarettes smoked[259] so the metabolic profile could be affected by previous smoking. Despite providing a well-defined cohort, including samples only from "male, non-diabetic, never smoked" patients substantially reduced the data to be analysed: of the 14 samples, three, five and six were coronary, claudication and control, respectively.

Figure A1.11 shows the PCA scores plot of the one component model ($R^2\boldsymbol{X}$ = 0.365 and $Q^2\boldsymbol{X}$ = 0.214) for "male, non-diabetic, never smoked" samples indicating no separation was present between the three disease type groups. The loadings plot is not displayed due to similarity with PC 1 of Figure A1.7, for which all samples were included. Due to the small number of coronary, claudication and control samples analysis was not performed between any two of the disease type groups.



Key: ▲ coronary, ■ claudication and ◆ control samples.

Figure A1.11 PCA scores plot for CPMG plasma data excluding glucose and lactate regions of "male, non-diabetic, never smoked" samples.

Irrespective of how defined the cohort was through elimination of potential confounding factors in a step-wise manner, coronary, claudication and control samples could not be distinguished according to their metabolic profile.

A1.3.2.2 Analysis of Other Sample Descriptors

In addition to disease type, sex and smoking and diabetic statuses many other sample descriptors were recorded (Table A1.6). Whether these affected the positioning of samples in scores space and their potential relevance to disease type groups was investigated.

For scores plots, samples were coloured according to data classified either by percentile or value classification (Table A1.6). Where values are commonly used in the medical profession to describe defined parts of the category, value classification for this study has been used, as for ankle brachial index (ABI) (Table A1.7) and body mass index (BMI) [weight (kg)/height (m^2)] (Table A1.8). For percentile classification, as near as possible 20% of samples were in classes 1 to 5. Cholesterol, glucose, glycated haemoglobin, HDL, LDL and triglyceride values were recorded to one

decimal place so multiple values could occur resulting in non-exact percentile distributions.

Table A1.6 Sample categories investigated with classification used in analysis.

Category	Data Classification
ABI	Percentile and Value
Age	Percentile
BMI	Percentile and Value
Cholesterol	Percentile
Glucose	Percentile
Glycated Haemoglobin	Percentile
HDL	Percentile
LDL	Percentile
Triglycerides	Percentile

Percentile classification: as near as possible 20% of samples in classes 1 to 5 with 1 having the lowest level and 5 the highest level.

Table A1.7 ABI classification for this study with corresponding indication of medical condition.[260]

ABI Value	Value Classification	Indication of Medical Condition
0.50-0.79	1	Moderate arterial disease
0.80-0.89	2	Some arterial disease
0.90-0.99	3	Acceptable
1.00-1.19	4	Normal
1.20+	5	Abnormal blood vessel hardening

Table A1.8 BMI classification for this study with corresponding weight category.[254]

BMI Value	Value Classification	Weight Category
<18.50	1	Underweight
18.50-24.99	2	Normal
25.00-29.99	3	Overweight
>30.00	4	Obese

Units: BMI = kg m^{-2}.

Some separation was observed between triglyceride percentiles in PC 2 versus PC 1 and PC 4 versus PC 2 scores plots (Figure A1.12 a and b, respectively) when all samples (data available for 72 samples) were included (four component model, $R^2\boldsymbol{X}$(cum) = 0.604 and $Q^2\boldsymbol{X}$(cum) = 0.417). Most of the class 5 samples (highest 20% of values) are clearly separated from samples with lower triglyceride values. Loadings plots (data not shown for PC 2 versus PC 1 due to similarity with Figure A1.7; PC 4 versus PC 2 shown in Figure A1.13) indicate class 5 samples have the highest lipid levels and generally a lower lipid level is associated with lower class value. Triglycerides are a type of lipid so connection between the two would be expected: NMR spectroscopy based metabolomics can provide an indication of the relative amount of triglyceride in blood plasma for larger data sets. In smaller data sets, *i.e.* "male, non-diabetic, non smoker" samples, a similar trend could not be observed. This provides an example of the need for balance between a well-defined cohort and suitable sample numbers.

Previously, disease type groups had been tentatively separated (Figure A1.4) with lipids being important to the discrimination (Figure A1.5); triglyceride percentile classes were incorporated into the scores plot (Figure A1.14). It can be observed that all control samples (diamonds) except one are within the ellipse (brown) drawn to show positioning of many control samples whilst the $\boldsymbol{t}$[3] range for coronary (triangles) and claudication (squares) samples is much greater despite all three disease type groups containing samples from the most extreme percentile classes. Speculatively, this could infer that patients who are not afflicted by coronary or claudication complications can regulate their metabolic profile better than some patients with similar lipid/triglyceride levels who are afflicted with blood vessel disease.

Classes of other categories (Table A1.6) did not determine sample positioning in any PC combination of scores space and were not associated with disease type groups. PC 2 versus PC 1 scores plots for all samples and "male, non-diabetic, non-smoker" samples of the various categories are referred to in Table A1.9 along with

summaries of models. PLS-DA models, where generated, were not shown because Q^2Y values were too small for the models to be considered reliable.

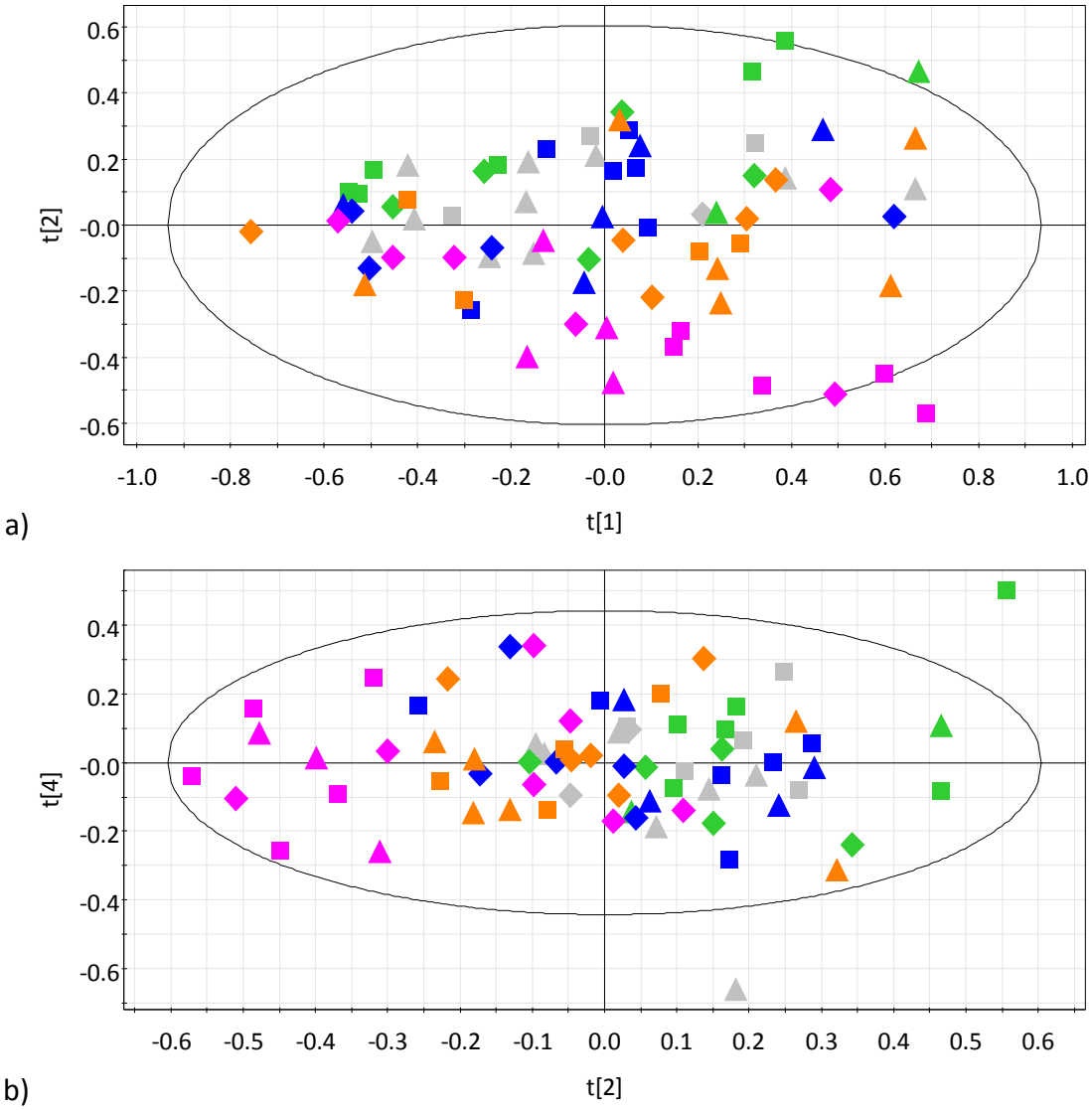


Figure A1.12 PCA scores plot for CPMG plasma data excluding glucose and lactate regions showing a) PC 2 versus PC 1 and b) PC 4 versus PC 2. Coloured according to percentile classification of triglyceride level (Table A1.6). R^2X = 0.309, 0.129 and 0.069, and Q^2X = 0.268, 0.076 and 0.035 for PC 1, PC 2 and PC 4, respectively.

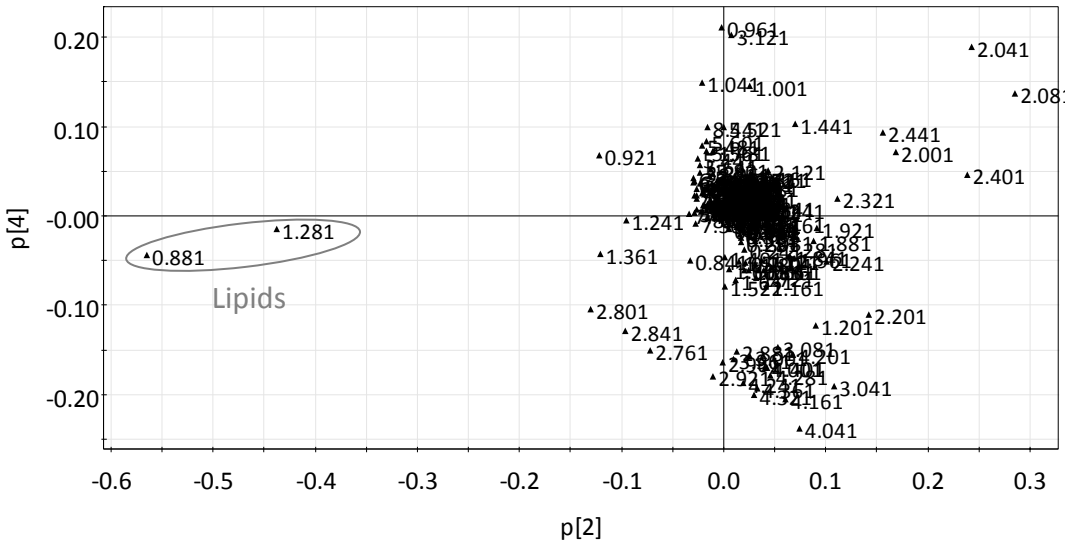
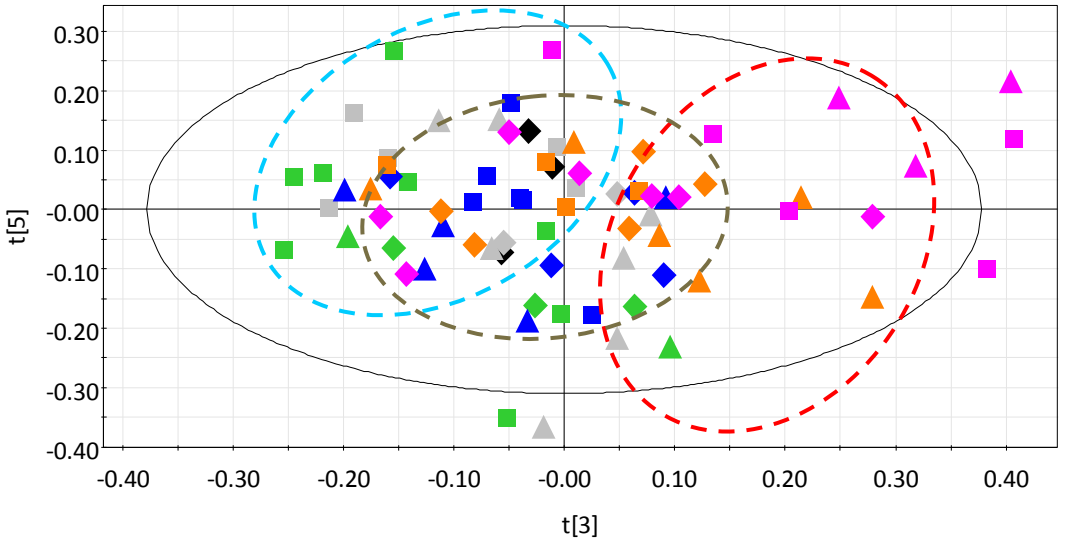


Figure A1.13 PCA loadings plot corresponding to the model displayed in Figure A1.12b).

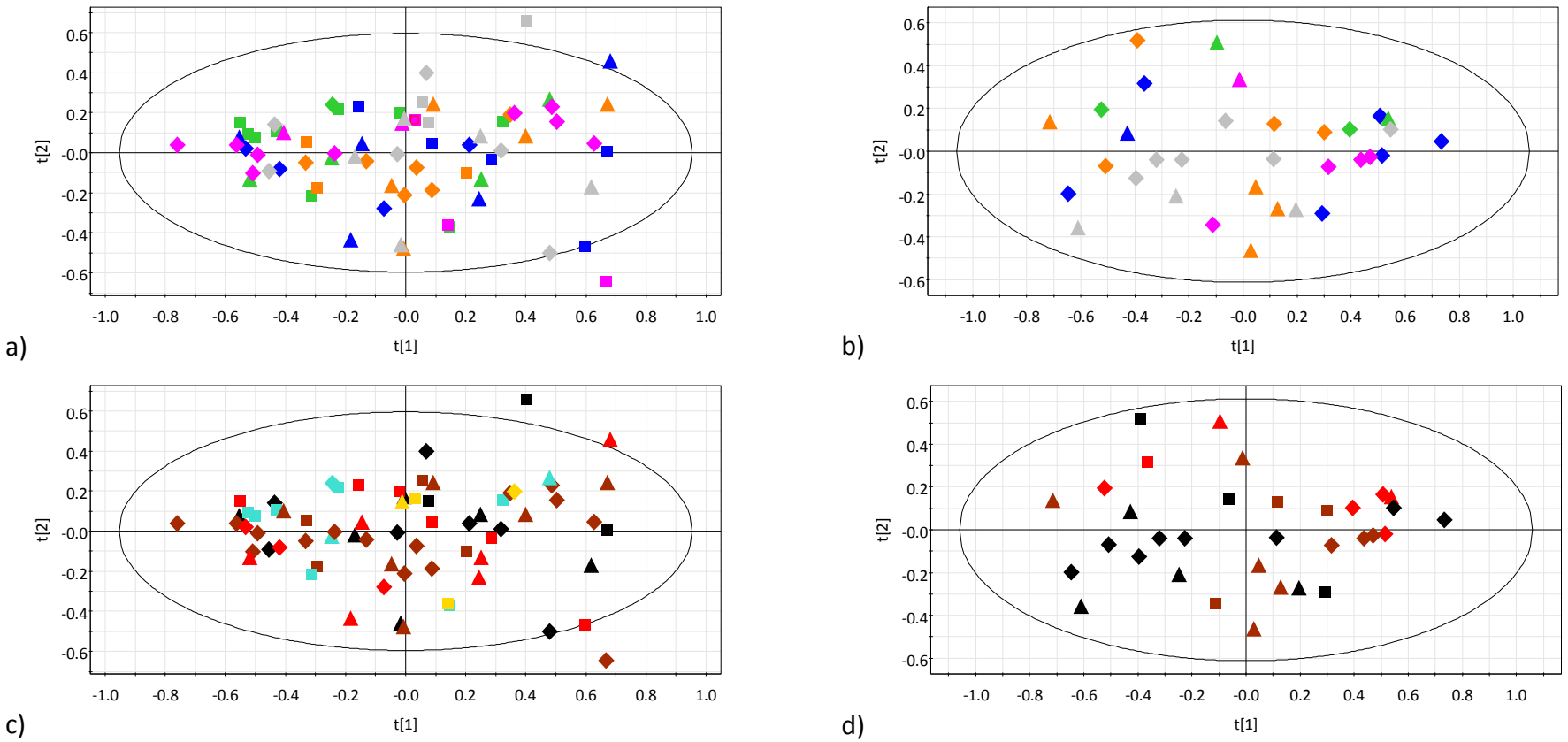


Key: shape - △ coronary, □ claudication and ◇ control samples,
colour - class ● 1, ● 2, ● 3, ● 4, ● 5 and ● no data.

Figure A1.14 PCA scores plot for plasma data including glucose and lactate regions showing PC 5 versus PC 3. Coloured according to percentile classification of triglyceride level (Table A1.6). R^2X = 0.091 and 0.061, Q^2X = 0.065 and 0.091 for PC 3 and PC 5, respectively. Red, turquoise and brown coloured ellipses indicate areas occupied by many samples of coronary, claudication and control groups, respectively, and are for display only.

Table A1.9 Summary of models created using various sample descriptors (categories) as the class identifier. M, ND, NS = "male, non-diabetic, non-smoker" samples.

Category	Samples Included	Number of Samples	Figure	PCA					PLS-DA			
				PCs	R^2X (cum)	Q^2X (cum)	R^2X of PC 2/PC 1	Q^2X of PC 2/PC 1	PCs	R^2X (cum)	R^2Y (cum)	Q^2Y (cum)
ABI	All	70	Figure A1.15 a) and c)	3	0.549	0.425	0.330/0.128	0.290/0.110	0	/	/	/
	M, ND, NS	31	Figure A1.15 b) and d)	1	0.396	0.339	0.000/0.396	0.000/0.339	0	/	/	/
BMI	All	75	Figure A1.16 a) and c)	4	0.601	0.423	0.309/0.129	0.262/0.093	0	/	/	/
	M, ND, NS	33	Figure A1.16 b) and d)	2	0.493	0.344	0.370/0.124	0.309/0.050	0	/	/	/
Age	All	75	Plots not shown	4	0.601	0.423	0.309/0.129	0.262/0.093	2	0.424	0.116	0.008
	M, ND, NS	33	Plots not shown	2	0.493	0.344	0.370/0.124	0.309/0.050	0	/	/	/
Cholesterol	All	73	Plots not shown	4	0.604	0.397	0.309/0.130	0.266/0.097	0	/	/	/
	M, ND, NS	31	Plots not shown	2	0.500	0.348	0.374/0.126	0.314/0.049	0	/	/	/
Glucose	All	74	Plots not shown	4	0.602	0.396	0.308/0.130	0.267/0.088	1	0.278	0.050	0.003
	M, ND, NS	32	Plots not shown	2	0.496	0.348	0.370/0.126	0.312/0.052	0	/	/	/
Glycated Haemoglobin	All	72	Plots not shown	3	0.542	0.402	0.313/0.129	0.276/0.070	0	/	/	/
	M, ND, NS	32	Plots not shown	2	0.496	0.348	0.370/0.126	0.312/0.052	0	/	/	/
HDL	All	74	Plots not shown	4	0.602	0.396	0.308/0.130	0.267/0.088	0	/	/	/
	M, ND, NS	32	Plots not shown	2	0.496	0.348	0.370/0.126	0.312/0.052	0	/	/	/
LDL	All	70	Plots not shown	4	0.601	0.406	0.312/0.122	0.276/0.050	0	/	/	/
	M, ND, NS	30	Plots not shown	2	0.504	0.355	0.374/0.130	0.309/0.066	1	0.371	0.087	0.007



Key: shape - △ coronary, □ claudication and ◇ control,
colour - class: a) and b) ● 1, ● 2, ● 3, ● 4 and ● 5, c) and d) ● 1, ● 2, ● 3, ● 4 and ○ 5

Figure A1.15 ABI: PCA scores plots for plasma data excluding glucose and lactate regions showing PC 2 versus PC 1 for a) and c) all samples and b) and d) "male, non-diabetic, non-smoker" samples. Coloured according to a) and b) percentile classification (Table A1.6) and c) and d) value classification (Table A1.7).

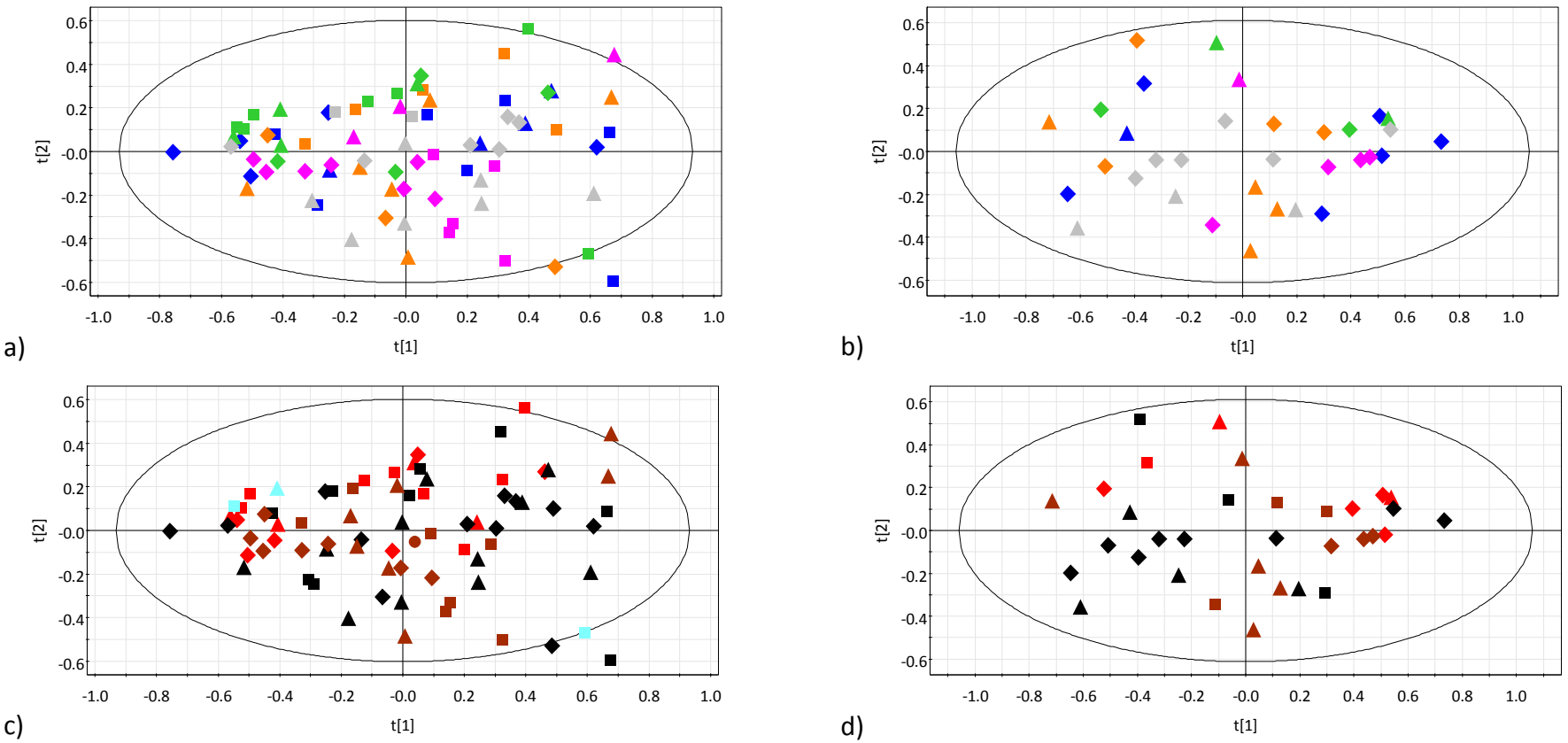


Figure A1.16 BMI: PCA scores plots for plasma data excluding glucose and lactate regions showing PC 2 v. PC 1 for a) and c) all samples and b) and d) "male, non-diabetic, non-smoker" samples. Coloured according to a) and b) percentile classification (Table A1.6) and c) and d) value classification (Table A1.8).

A1.4 Conclusions

The degree of chemical shift variation between samples is dependent on the reference compound. TSP, which has traditionally been used, has been shown to be inferior to acetate, glucose and lactate. Due to the readily resolvable nature of H_1' of α-glucose this was selected as the reference compound for all further work.

Plasma spectra were dominated by diet dependent metabolites, namely glucose and lactate, and distinction between coronary, claudication and control samples was not present in the higher PCs. Tentative separation was visible in lower PCs based on lipid levels with the indication that levels were highest in coronary samples and lowest in claudication samples. Upon removal of glucose and lactate, separation between the three disease type groups was not visible. Through removal of samples, certain known confounding factors were accounted for such as sex, smoking and presence of diabetes. Samples from the optimum cohort in terms of retaining sufficient sample numbers, which was the "male, non-diabetic, non-smoker" cohort, did not provide further information regarding differences between the groups. Consideration has to be made between reducing confounding factors whilst providing a suitable number of samples for analysis.

Clinical laboratory recorded triglyceride levels corresponded to relative lipid levels as shown by scores and loadings plots for larger data sets therefore NMR spectroscopy based metabolomics could provide an indication of the relative amount of triglyceride levels in blood plasma. It was tentatively implied that for patients whose triglyceride levels were similar those not afflicted by coronary or claudication complications could regulate their metabolic profile better than some patients who were afflicted with blood vessel disease.

A1.5 Experimental Methods

A1.5.1 NMR Sample Preparation

Chemicals were purchased from Sigma-Aldrich Company Ltd. (Poole, Dorset, UK). NMR tubes (S-5-500-7, Norell) were purchased from GPE Scientific Ltd. (Leighton Buzzard, Bedfordshire, UK).

A1.5.2 Samples for ^{1}H-NMR Analysis

Samples were stored at -80^{o}C before defrosting at room temperature. Samples were centrifuged (Hettich Mikro 120 (C1204) Centrifuge, angle rotor E2384) at 12,009 *g* for 5 minutes. 350 µl of plasma supernatant was added to 408 µl of a 0.17% weight by volume solution of TSP in D_2O. The mixture was vortexed for 8 s before transferring 600 µl to a 5 mm NMR tube. Samples not analysed immediately were stored at 4^{o}C for a maximum of 1.5 hours.

A1.5.3 NMR Analysis

All 1H-NMR spectra were acquired on a Varian Unity Inova 500 spectrometer (Varian Inc., Palo Alto, California, USA) operating at 499.97 MHz proton frequency, at 20^{o}C. Samples were loaded into the probe and left for 5 minutes to allow temperature equilibration.

A1.5.3.1 CPMG Experiment

The CPMG pulse sequence [RD – 90^{o} – (τ – 180^{o} – τ)$_n$ – acq] was used to obtain metabolic profiles for all plasma samples. The relaxation delay (RD) was 2 s, during which the water resonance was selectively saturated, τ was 1.5 ms and *n* was 150. For each spectrum 512 transients were collected into 16,384 pairs of data points with a spectral width of 8,000.00 Hz.

A1.5.4 NMR Spectral Processing

All spectral data were processed using ACD Labs software 12.01 (Advanced Chemistry Development, Inc., (ACD/Labs), Toronto, Canada). An exponential line broadening of 1.applied to each FID prior to zero filling to 65,536 points and Fourier transformation. The resulting spectra were phased, baseline corrected and referenced (Table A1.10).

Table A1.10 Reference Compounds Used

Reference Compound	Chemical Shift (ppm)
TSP	0.000[142]
Lactate	1.317[103]
Alanine	1.463[103]
Acetate	1.910[103]
α-glucose	5.223[103]

A1.5.5 Additional NMR Data Processing for Multivariate Statistical Analysis

A1.5.5.1 Binning and Dark Regions

Spectral binning was performed using ACD Labs software 12.01. The data set was integrated into fixed bins of 0.04 ppm.

Prior to binning over the spectral range 0.141-0.998, several dark regions were created for spectra to exclude signals from subsequent multivariate analysis (Table A1.11).

Table A1.11 Dark regions used for spectra.

Dark Region (ppm)	Excluded	Comment
2.461-2.741	Citrate	Possible contaminant from collection tube
4.421-5.421	Water region	Variable water suppression efficiency
1.301-1.341 4.061-4.141	Lactate	Excluded later to remove influence from statistical models
3.181-3.941	Glucose	

A1.5.6 Multivariate Statistical Analysis

All multivariate statistical analyses were performed using SIMCA-P+ software, version 12.0.1.0 (Umetrics, Umeå, Sweden). This software was used to scale and mean centre the data.

PCA and PLS-DA was performed. The predictive ability of the PLS-DA models was assessed by exclusion of a third of the samples and their class memberships predicted using models built from the remaining samples, where the samples were randomly allocated into three exclusion groups using a random number generator (Microsoft Excel 2007). Further validation was performed using permutation testing whereby the classes of samples were randomised and a PLS-DA model built. The $R^2\boldsymbol{Y}$ and $Q^2\boldsymbol{Y}$ values should be less than those generated for the original model, which were based on the real classifications. The maximum number of permutations permissible by the software was performed, 999.

Appendix 2: Gas Chromatography

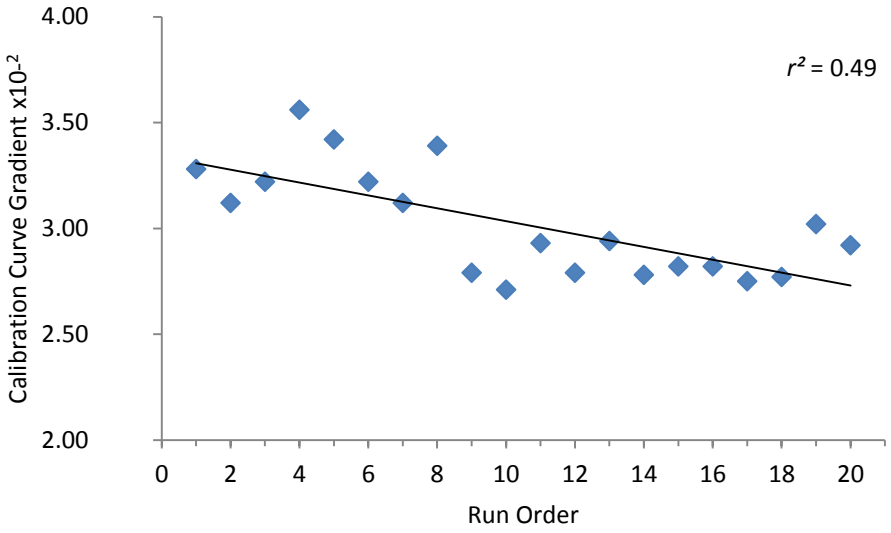
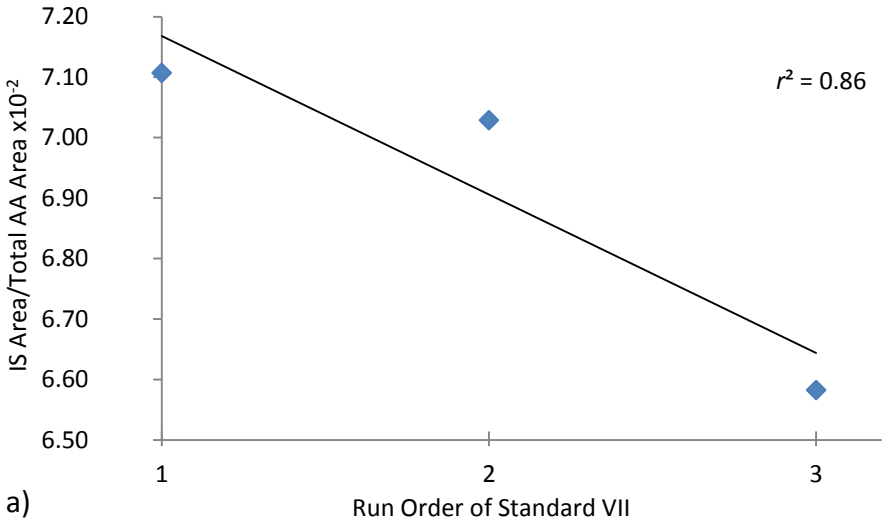
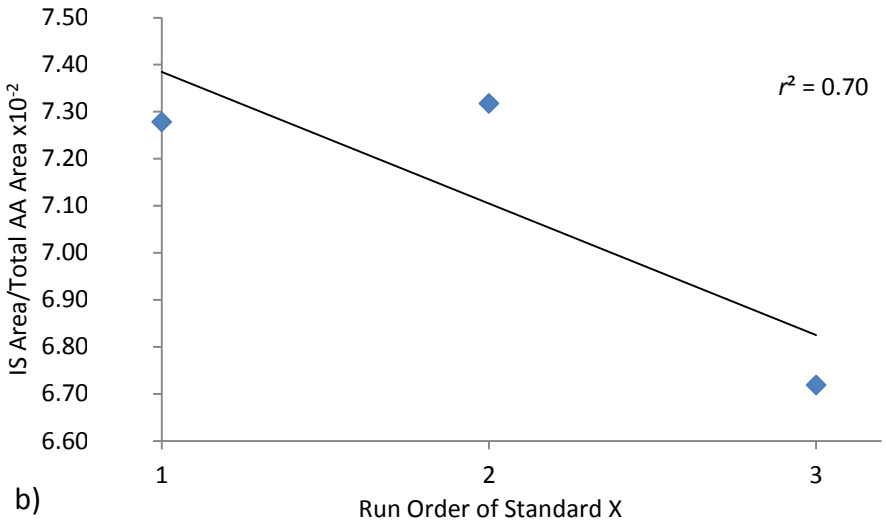


Figure A2.1 Calibration curve gradient of glycine (GLY) for all runs of standards against run order.

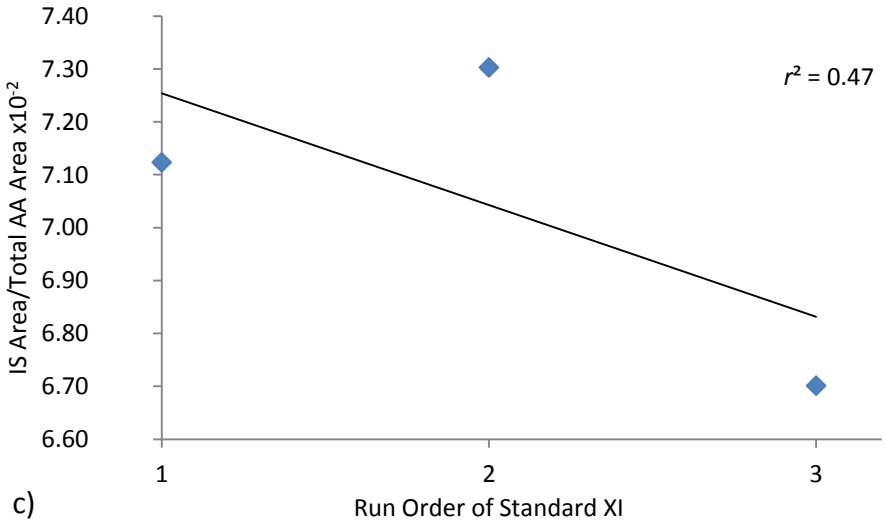


a)

Figure A2.2 Plot of IS area/total AA area against run order of 200 nmol ml^{-1} standards that had different elapsed times between first and last runs: a) Standard VII, 6 h 30 min, b) Standard X, 23 h 55 min and c) Standard XI, 23 h 55 min.



b)



c)

Figure A2.2 Continued.

Table A2.1 Structure and retention times of AAs identifiable by GC analysis of plasma.

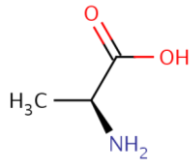
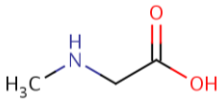
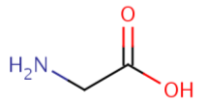
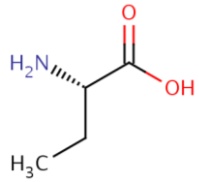
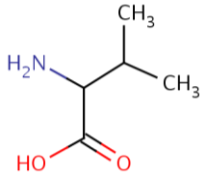
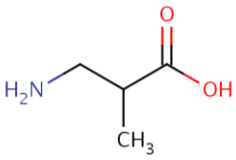
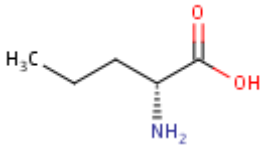
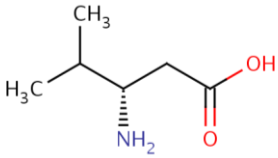
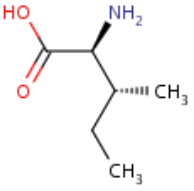
AA		Retention Time (minutes)	Structure
Full Name	**Abbreviation**		
Alanine	ALA	1.59	
Sarcosine	SAR	1.65	
Glycine	GLY	1.69	
α-Aminobutyric acid	ABA	1.79	
Valine	VAL	1.88	
β-Aminoisobutyric acid	β-AIB	1.96	
Norvaline (internal standard)	IS	2.00	
Leucine	LEU	2.08	
allo-Isoleucine	α-ILE	2.11	

Table A2.1 Continued.

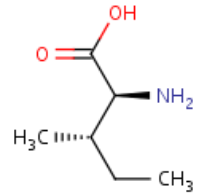
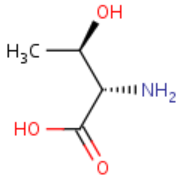
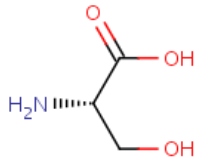
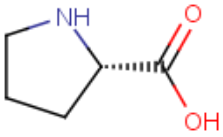
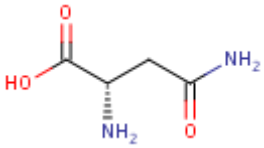
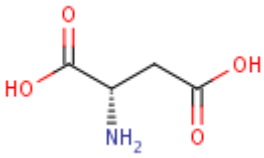
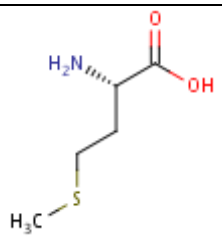
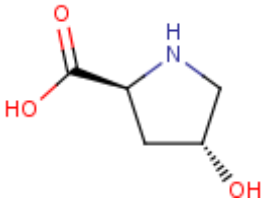
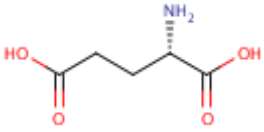
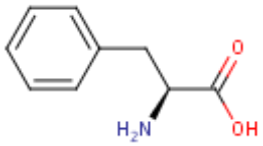
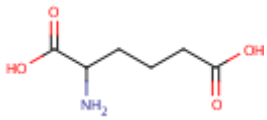
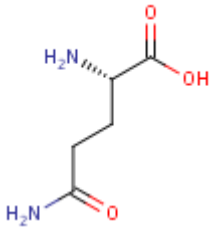
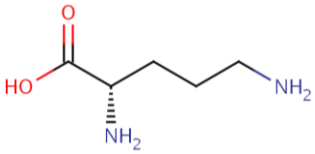
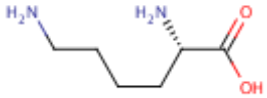
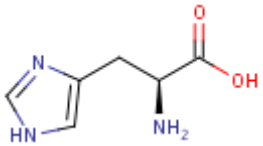
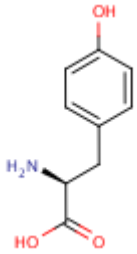
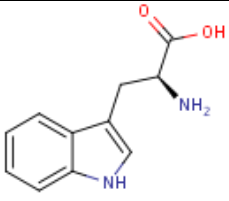
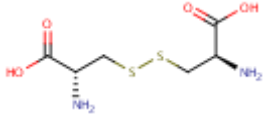
Isoleucine	ILE	2.14	
Threonine	THR	2.34	
Serine	SER	2.38	
Proline	PRO	2.45	
Asparagine	ASN	2.54	
Aspartic acid	ASP	3.05	
Methionine	MET	3.09	
Hydroxyproline	HYP	3.22	
Glutamic acid	GLU	3.39	

Table A2.1 Continued.

Phenylalanine	PHE	3.43	
α-Aminoadipic acid	AAA	3.68	
Glutamine	GLN	4.01	
Ornithine	ORN	4.37	
Lysine	LYS	4.62	
Histidine	HIS	4.80	
Tyrosine	TYR	5.07	
Tryptophan	TRP	5.36	
Cystine	C-C	6.01	

References

1. Jemal A, Bray F*, et al.* Global Cancer Statistics. *CA - A Cancer Journal for Clinicians*, 2011, 61(2): 69-90.
2. GLOBOCAN_website. 'GLOBOCAN 2008'. Last updated 01-10-2012. Accessed 01-02-2013. Available from: http://globocan.iarc.fr/.
3. Davies H, Bignell G R*, et al.* Mutations Of The BRAF Gene In Human Cancer. *Nature*, 2002, 417(6892): 949-954.
4. Green D R and Kroemer G. The Pathophysiology Of Mitochondrial Cell Death. *Science*, 2004, 305(5684): 626-629.
5. Onitilo A A, Engel J M*, et al.* Breast Cancer Subtypes Based On ER/PR And HER2 Expression: Comparison Of Clinicopathologic Features And Survival. *Clinical Medicine & Research*, 2009, 7(1-2): 4-13.
6. Brisken C and O'Malley B. Hormone Action In The Mammary Gland. *Cold Spring Harbor Perspectives in Biology*, 2010, 2(12).
7. Habel L A and Stanford J L. Hormone Receptors And Breast Cancer. *Epidemiologic Reviews*, 1993, 15(1): 209-219.
8. Huang W Y, Newman B*, et al.* Hormone-Related Factors And Risk Of Breast Cancer In Relation To Estrogen Receptor And Progesterone Receptor Status. *American Journal of Epidemiology*, 2000, 151(7): 703-714.
9. Petit T, Dufour P*, et al.* A Critical Evaluation Of The Role Of Aromatase Inhibitors As Adjuvant Therapy For Postmenopausal Women With Breast Cancer. *Endocrine-Related Cancer*, 2011, 18(3): R79-R89.
10. Mokbel K. The Evolving Role Of Aromatase Inhibitors In Breast Cancer. *International Journal Of Clinical Oncology*, 2002, 7(5): 279-283.
11. Goldhirsch A and Gelber R D. Endocrine Therapies Of Breast Cancer. *Seminars in Oncology*, 1996, 23(4): 494-505.
12. Slamon D J, Leyland-Jones B*, et al.* Use Of Chemotherapy Plus A Monoclonal Antibody Against HER2 For Metastatic Breast Cancer That Overexpresses HER2. *New England Journal of Medicine*, 2001, 344(11): 783-792.
13. Le X F, Pruefer F*, et al.* HER2-Targeting Antibodies Modulate The Cyclin-Dependent Kinase Inhibitor p27(Kip1) Via Multiple Signaling Pathways. *Cell Cycle*, 2005, 4(1): 87-95.
14. Hudis C A and Gianni L. Triple-Negative Breast Cancer: An Unmet Medical Need. *Oncologist*, 2011, 16: 1-11.
15. Aitken S J, Thomas J S*, et al.* Quantitative Analysis Of Changes In ER, PR And HER2 Expression In Primary Breast Cancer And Paired Nodal Metastases. *Annals of Oncology*, 2010, 21(6): 1254-1261.
16. Choudhury K R, Yagle K J*, et al.* A Robust Automated Measure Of Average Antibody Staining In Immunohistochemistry Images. *Journal of Histochemistry & Cytochemistry*, 2010, 58(2): 95-107.
17. Allred D C, Harvey J M*, et al.* Prognostic And Predictive Factors In Breast Cancer By Immunohistochemical Analysis. *Modern Pathology*, 1998, 11(2): 155-168.
18. Mudduwa L K B. Quick Score Of Hormone Receptor Status Of Breast Carcinoma: Correlation With The Other Clinicopathological Prognostic

Parameters. *Indian Journal of Pathology and Microbiology*, 2009, 52(2): 159-163.

19. Qureshi A and Pervez S. Allred Scoring For ER Reporting And Its Impact In Clearly Distinguishing ER Negative From ER Positive Breast Cancers. *Journal of the Pakistan Medical Association*, 2010, 60(5): 350-353.
20. Boughey J C, Wagner J*, et al.* Neoadjuvant Chemotherapy In Invasive Lobular Carcinoma May Not Improve Rates Of Breast Conservation. *Annals of Surgical Oncology*, 2009, 16(6): 1606-1611.
21. Li C I, Malone K E*, et al.* Risk Of Invasive Breast Carcinoma Among Women Diagnosed With Ductal Carcinoma In Situ And Lobular Carcinoma In Situ, 1988-2001. *Cancer*, 2006, 106(10): 2104-2112.
22. Rakha E A, El-Sayed M E*, et al.* Prognostic Significance Of Nottingham Histologic Grade In Invasive Breast Carcinoma. *Journal of Clinical Oncology*, 2008, 26(19): 3153-3158.
23. Elston C W and Ellis I O. Pathological Prognostic Factors In Breast-Cancer. The Value Of Histological Grade In Breast-Cancer - Experience From A Large Study With Long-Term Follow-Up. *Histopathology*, 1991, 19(5): 403-410.
24. Frkovic-Grazio S and Bracko M. Long Term Prognostic Value Of Nottingham Histological Grade And Its Components In Early (pT1 N0M0) Breast Carcinoma. *Journal of Clinical Pathology*, 2002, 55(2): 88-92.
25. Vander Heiden M G, Cantley L C*, et al.* Understanding The Warburg Effect: The Metabolic Requirements Of Cell Proliferation. *Science (New York, N.Y.)*, 2009, 324(5930): 1029-1033.
26. Kim J-w and Dang C V. Cancer's Molecular Sweet Tooth And The Warburg Effect. *Cancer Research*, 2006, 66(18): 8927-8930.
27. DeBerardinis R J. Is Cancer A Disease Of Abnormal Cellular Metabolism? New Angles On An Old Idea. *Genetics in Medicine*, 2008, 10(11): 267-277.
28. DeBerardinis R J, Mancuso A*, et al.* Beyond Aerobic Glycolysis: Transformed Cells Can Engage In Glutamine Metabolism That Exceeds The Requirement For Protein And Nucleotide Synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 2007, 104(49): 19345-19350.
29. Jones R G and Thompson C B. Tumor Suppressors And Cell Metabolism: A Recipe For Cancer Growth. *Genes & Development*, 2009, 23(5): 537-548.
30. Derome A E. Modern NMR Techniques for Chemistry Research, 1st Edition. Pergamon Press: Oxford, UK, 1987.
31. Claridge T D W. High-Resolution NMR Techniques in Organic Chemistry, 1st Edition. Elsevier: Oxford, UK, 1999.
32. Lindon J C, Holmes E*, et al.* Metabonomics Techniques And Applications To Pharmaceutical Research & Development. *Pharmaceutical Research*, 2006, 23(6): 1075-1088.
33. Van der Greef J and Smilde A K. Symbiosis Of Chemometrics And Metabolomics: Past, Present, And Future. *Journal of Chemometrics*, 2005, 19(5-7): 376-386.
34. Nicholson J K, Lindon J C*, et al.* 'Metabonomics': Understanding The Metabolic Responses Of Living Systems To Pathophysiological Stimuli Via Multivariate Statistical Analysis Of Biological NMR Spectroscopic Data. *Xenobiotica*, 1999, 29(11): 1181-1189.

35. Nicholson J K and Lindon J C. Systems biology - Metabonomics. *Nature*, 2008, 455(7216): 1054-1056.
36. Nicholson J K, O'Flynn M P*, et al.* Proton-Nuclear-Magnetic-Resonance Studies Of Serum, Plasma And Urine From Fasting Normal And Diabetic Subjects. *Biochemical Journal*, 1984, 217(2): 365-375.
37. Dunn W B, Broadhurst D I*, et al.* Serum Metabolomics Reveals Many Novel Metabolic Markers Of Heart Failure, Including Pseudouridine And 2-Oxoglutarate. *Metabolomics*, 2007, 3: 413-426.
38. Lewis G D, Asnani A*, et al.* Application Of Metabolomics To Cardiovascular Biomarker And Pathway Discovery. *Journal of the American College of Cardiology*, 2008, 52(2): 117-123.
39. Dunn W B and Ellis D I. Metabolomics: Current Analytical Platforms And Methodologies. *TrAC-Trends in Analytical Chemistry*, 2005, 24(4): 285-294.
40. Gibney M J, Walsh M*, et al.* Metabolomics In Human Nutrition: Opportunities And Challenges. *American Journal of Clinical Nutrition*, 2005, 82(3): 497-503.
41. Robertson D G. Metabonomics In Toxicology: A Review. *Toxicological Sciences*, 2005, 85(2): 809-822.
42. Hollywood K, Brison D R*, et al.* Metabolomics: Current Technologies And Future Trends. *Proteomics*, 2006, 6(17): 4716-4723.
43. Fiehn O. Metabolomics - The Link Between Genotypes And Phenotypes. *Plant Molecular Biology*, 2002, 48(1-2): 155-171.
44. Dunn W B, Bailey N J C*, et al.* Measuring The Metabolome: Current Analytical Technologies. *Analyst*, 2005, 130(5): 606-625.
45. Lindon J C and Nicholson J K. Spectroscopic And Statistical Techniques For Information Recovery In Metabonomics And Metabolomics. *Annual Review of Analytical Chemistry*, 2008, 1: 45-69.
46. Lenz E M and Wilson I D. Analytical Strategies In Metabonomics. *Journal of Proteome Research*, 2007, 6: 443-458.
47. Lundstedt T, Hedenstrom M*, et al.* Dynamic Modelling Of Time Series Data In Nutritional Metabonomics - A Powerful Complement To Randomized Clinical Trials In Functional Food Studies. *Chemometrics and Intelligent Laboratory Systems*, 2010, 104(1): 112-120.
48. Fancy S A, Beckonert O*, et al.* Gas Chromatography/Flame Ionisation Detection Mass Spectrometry For The Detection Of Endogenous Urine Metabolites For Metabonomic Studies And Its Use As A Complementary Tool To Nuclear Magnetic Resonance Spectroscopy. *Rapid Communications in Mass Spectrometry*, 2006, 20(15): 2271-2280.
49. Lindon J C, Nicholson J K*, et al.* Metabonomics: Metabolic Processes Studied By NMR Spectroscopy Of Biofluids. *Concepts in Magnetic Resonance*, 2000, 12(5): 289-320.
50. Viant M R, Bearden D W*, et al.* International NMR-Based Environmental Metabolomics Intercomparison Exercise. *Environmental Science & Technology*, 2009, 43(1): 219-225.
51. Beckonert O, Keun H C*, et al.* Metabolic Profiling, Metabolomic And Metabonomic Procedures For NMR Spectroscopy Of Urine, Plasma, Serum And Tissue Extracts. *Nature Protocols*, 2007, 2(11): 2692-2703.

52. Lindon J C and Nicholson J K. Analytical Technologies For Metabonomics And Metabolomics, And Multi-Omic Information Recovery. *TrAC-Trends in Analytical Chemistry*, 2008, 27(3): 194-204.
53. Lindon J C, Holmes E*, et al.* Metabonomics In Pharmaceutical R & D. *FEBS Journal*, 2007, 274(5): 1140-1151.
54. Griffin J L and Shockcor J P. Metabolic Profiles Of Cancer Cells. *Nature Reviews Cancer*, 2004, 4(7): 551-561.
55. Parsons H M, Ludwig C*, et al.* Line-Shape Analysis Of J-Resolved NMR Spectra: Application To Metabolomics And Quantification Of Intensity Errors From Signal Processing And High Signal Congestion. *Magnetic Resonance in Chemistry*, 2009, 47: S86-S95.
56. Saude E J, Slupsky C M*, et al.* Optimization Of NMR Analysis Of Biological Fluids For Quantitative Accuracy. *Metabolomics*, 2006, 2(3): 113-123.
57. Wishart D S. Quantitative Metabolomics Using NMR. *Trac-Trends in Analytical Chemistry*, 2008, 27(3): 228-237.
58. Trygg J, Holmes E*, et al.* Chemometrics In Metabonomics. *Journal of Proteome Research*, 2007, 6: 469-479.
59. Craig A, Cloareo O*, et al.* Scaling And Normalization Effects In NMR Spectroscopic Metabonomic Data Sets. *Analytical Chemistry*, 2006, 78(7): 2262-2267.
60. Anderson P E, Reo N V*, et al.* Gaussian Binning: A New Kernel-Based Method For Processing NMR Spectroscopic Data For Metabolomics. *Metabolomics*, 2008, 4(3): 261-272.
61. Weljie A M, Newton J*, et al.* Targeted profiling: Quantitative Analysis Of ^{1}H NMR Metabolomics Data. *Analytical Chemistry*, 2006, 78(13): 4430-4442.
62. Holmes E, Foxall P J D*, et al.* Automatic Data Reduction And Pattern-Recognition Methods For Analysis Of ^{1}H Nuclear-Magnetic-Resonance Spectra Of Human Urine From Normal And Pathological States. *Analytical Biochemistry*, 1994, 220(2): 284-296.
63. Spraul M, Neidig P*, et al.* Automatic Reduction Of NMR Spectroscopic Data For Statistical And Pattern-Recognition Classification Of Samples. *Journal of Pharmaceutical and Biomedical Analysis*, 1994, 12(10): 1215-1225.
64. Loo R L, Coen M*, et al.* Metabolic Profiling And Population Screening Of Analgesic Usage In Nuclear Magnetic Resonance Spectroscopy-Based Large-Scale Epidemiologic Studies. *Analytical Chemistry*, 2009, 81(13): 5119-5129.
65. Teahan O, Gamble S*, et al.* Impact Of Analytical Bias In Metabonomic Studies Of Human Blood Serum And Plasma. *Analytical Chemistry*, 2006, 78(13): 4307-4318.
66. Salek R M, Maguire M L*, et al.* A Metabolomic Comparison Of Urinary Changes In Type 2 Diabetes In Mouse, Rat, And Human. *Physiological Genomics*, 2007, 29(2): 99-108.
67. Beger R D, Schnackenberg L K*, et al.* Metabonomic Models Of Human Pancreatic Cancer Using 1D Proton NMR Spectra Of Lipids In Plasma. *Metabolomics*, 2006, 2(3): 125-134.
68. Jahns G L, Kent M N*, et al.* Development Of Analytical Methods For NMR Spectra And Application To A ^{13}C Toxicology Study. *Metabolomics*, 2009, 5(2): 253-262.

69. Dieterle F, Ross A*, et al.* Probabilistic Quotient Normalization As Robust Method To Account For Dilution Of Complex Biological Mixtures. Application In ^{1}H NMR Metabonomics. *Analytical Chemistry*, 2006, 78(13): 4281-4290.
70. Webb-Robertson B J M, Lowry D F*, et al.* A Study Of Spectral Integration And Normalization In NMR-Based Metabonomic Analyses. *Journal of Pharmaceutical and Biomedical Analysis*, 2005, 39(3-4): 830-836.
71. Sheedy J R, Ebeling P R*, et al.* A Sample Preparation Protocol For ^{1}H Nuclear Magnetic Resonance Studies Of Water-Soluble Metabolites In Blood And Urine. *Analytical Biochemistry*, 2010, 398(2): 263-265.
72. Keun H C, Ebbels T M D*, et al.* Improved Analysis Of Multivariate Data By Variable Stability Scaling: Application To NMR-Based Metabolic Profiling. *Analytica Chimica Acta*, 2003, 490(1-2): 265-276.
73. van den Berg R A, Hoefsloot H C J*, et al.* Centering, Scaling, And Transformations: Improving The Biological Information Content Of Metabolomics Data. *BMC Genomics*, 2006, 7.
74. Holmes E, Cloarec O*, et al.* Probing Latent Biomarker Signatures And In Vivo Pathway Activity In Experimental Disease States Via Statistical Total Correlation Spectroscopy (STOCSY) Of Biofluids: Application To HgCl2 Toxicity. *Journal of Proteome Research*, 2006, 5(6): 1313-1320.
75. Constantinou M A, Papakonstantinou E*, et al.* ^{1}H NMR-Based Metabonomics For The Diagnosis Of Inborn Errors Of Metabolism In Urine. *Analytica Chimica Acta*, 2005, 542(2): 169-177.
76. Cloarec O, Dumas M E*, et al.* Evaluation Of The Orthogonal Projection On Latent Structure Model Limitations Caused By Chemical Shift Variability And Improved Visualization Of Biomarker Changes In ^{1}H NMR Spectroscopic Metabonomic Studies. *Analytical Chemistry*, 2005, 77(2): 517-526.
77. Parsons H M, Ludwig C*, et al.* Improved Classification Accuracy In 1-And 2-Dimensional NMR Metabolomics Data Using The Variance Stabilising Generalised Logarithm Transformation. *BMC Bioinformatics*, 2007, 8.
78. Manly BFJ. Multivariate Statistical Methods: A Primer, 3rd Edition. Chapman & Hall/CRC: Boca Raton, Florida, USA, 2005.
79. Westerhuis J A, Hoefsloot H C J*, et al.* Assessment Of PLSDA Cross Validation. *Metabolomics*, 2008, 4(1): 81-89.
80. Brindle J T, Nicholson J K*, et al.* Application Of Chemometrics To ^{1}H NMR Spectroscopic Data To Investigate A Relationship Between Human Serum Metabolic Profiles And Hypertension. *Analyst*, 2003, 128(1): 32-36.
81. Lee K R, Lin X W*, et al.* Megavariate Data Analysis Of Mass Spectrometric Proteomics Data Using Latent Variable Projection Method. *Proteomics*, 2003, 3(9): 1680-1686.
82. Wold S, Esbensen K*, et al.* Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, 1987, 2(1-3): 37-52.
83. Wold S, Sjostrom M*, et al.* Multivariate Design. *Analytica Chimica Acta*, 1986, 191: 17-32.
84. Ringner M. What Is Principal Component Analysis? *Nature Biotechnology*, 2008, 26(3): 303-304.

85. Defernez M and Colquhoun I J. Factors Affecting The Robustness Of Metabolite Fingerprinting Using ^{1}H NMR Spectra. *Phytochemistry*, 2003, 62(6): 1009-1017.
86. Eriksson L, Johansson E*, et al.*, Multi- And Megavariate Data Analysis, 2nd Edition. Umetrics: Umea, Sweden, 2006.
87. Eriksson L, Trygg J*, et al.* CV-ANOVA For Significance Testing Of PLS And OPLS Models. *Journal of Chemometrics*, 2008, 22(11-12): 594-600.
88. Stretch C, Eastman T*, et al.* Prediction Of Skeletal Muscle And Fat Mass In Patients With Advanced Cancer Using A Metabolomic Approach. *Journal of Nutrition*, 2012, 142(1): 14-21.
89. Bland M. An Introduction To Medical Statistics, 3rd Edition. Oxford University Press Inc.: New York, USA, 2005.
90. Shapiro S S and Wilk M B. An Analysis Of Variance Test For Normality (Complete Samples). *Biometrika*, 1965, 52: 591-&.
91. Aspin A A. Tables For Use In Comparisons Whose Accuracy Involves Two Variances, Separately Estimated. *Biometrika*, 1949, 36(3-4): 290-296.
92. Broadhurst D I and Kell D B. Statistical Strategies For Avoiding False Discoveries In Metabolomics And Related Experiments. *Metabolomics*, 2006, 2(4): 171-196.
93. Storey J D and Tibshirani R. Statistical Significance For Genomewide Studies. *Proceedings of the National Academy of Sciences of the United States of America*, 2003, 100(16): 9440-9445.
94. Cloarec O, Dumas M E*, et al.* Statistical Total Correlation Spectroscopy: An Exploratory Approach For Latent Biomarker Identification From Metabolic ^{1}H NMR Data Sets. *Analytical Chemistry*, 2005, 77(5): 1282-1289.
95. Crockford D J, Holmes E*, et al.* Statistical Heterospectroscopy, An Approach To The Integrated Analysis Of NMR And UPLC-MS Data Sets: Application In Metabonomic Toxicology Studies. *Analytical Chemistry*, 2006, 78(2): 363-371.
96. Maher A D, Cysique L A*, et al.* Statistical Integration Of ^{1}H NMR And MRS Data From Different Biofluids And Tissues Enhances Recovery Of Biological Information From Individuals With HIV-1 Infection. *Journal of Proteome Research*, 2011, 10(4): 1737-1745.
97. Dunn W B, Broadhurst D I*, et al.* Systems Level Studies Of Mammalian Metabolomes: The Roles Of Mass Spectrometry And Nuclear Magnetic Resonance Spectroscopy. *Chemical Society Reviews*, 2011, 40(1): 387-426.
98. Sumner L W, Amberg A*, et al.* Proposed Minimum Reporting Standards For Chemical Analysis. *Metabolomics*, 2007, 3: 211-221.
99. Rubtsov D V, Jenkins H*, et al.* Proposed Reporting Requirements For The Description Of NMR-Based Metabolomics Experiments. *Metabolomics*, 2007, 3(3): 223-229.
100. Goodacre R, Broadhurst D*, et al.* Proposed Minimum Reporting Standards For Data Analysis In Metabolomics. *Metabolomics*, 2007, 3: 231-241.
101. Beckonert O, Coen M*, et al.* High-Resolution Magic-Angle-Spinning NMR Spectroscopy For Metabolic Profiling Of Intact Tissues. *Nature Protocols*, 2010, 5(6): 1019-1032.

102. Lin C Y, Wu H F*, et al.* Evaluation Of Metabolite Extraction Strategies From Tissue Samples Using NMR Metabolomics. *Metabolomics*, 2007, 3(1): 55-67.
103. Wishart D S, Knox C*, et al.* HMDB: A Knowledgebase For The Human Metabolome. *Nucleic Acids Research*, 2009, 37: D603-D610.
104. Skoog D A, West D M*, et al.* Fundamentals Of Analytical Chemistry, 7th Edition. Harcourt College Publishers: San Diego, California, USA, 1995.
105. Kim J, Bae B*, et al.* Development Of A Micro-Flame Ionization Detector Using A Diffusion Flame. *Sensors and Actuators B-Chemical*, 2012, 168: 111-117.
106. Damadian R, Zaner K*, et al.* Nuclear Magnetic Resonance As A New Tool In Cancer Research - Human Tumors By NMR. *Annals of the New York Academy of Sciences*, 1973, 222(DEC31): 1048-1076.
107. Damadian R. Tumor Detection By Nuclear Magnetic Resonance. *Science*, 1971, 171(3976): 1151-&.
108. Medina D, Hazlewood C F*, et al.* Nuclear Magnetic-Resonance Studies On Human Breast Dysplasias And Neoplasms. *Journal of the National Cancer Institute*, 1975, 54(4): 813-818.
109. Ekstrand K E, Dixon R L*, et al.* Proton NMR Relaxation Times In Peripheral Blood Of Cancer Patients. *Physics in Medicine and Biology*, 1977, 22(5): 925-931.
110. Williams P G, Helmer M A*, et al.* Lipid Domain In Cancer Cell Plasma-Membrane Shown By ^{1}H NMR To Be Similar To A Lipoprotein. *FEBS Letters*, 1985, 192(1): 159-164.
111. Nicholson J K, Buckingham M J*, et al.* High Resolution ^{1}H NMR Studies Of Vertebrate Blood And Plasma. *Biochemical Journal*, 1983, 211(3): 605-615.
112. Nicholson J K, Oflynn M P*, et al.* Proton-Nuclear-Magnetic-Resonance Studies Of Serum, Plasma And Urine From Fasting Normal And Diabetic Subjects. *Biochemical Journal*, 1984, 217(2): 365-375.
113. Nagrath D, Caneba C*, et al.* Metabolomics For Mitochondrial And Cancer Studies. *Biochimica Et Biophysica Acta-Bioenergetics*, 2011, 1807(6): 650-663.
114. Duarte I F and Gil A M. Metabolic Signatures Of Cancer Unveiled By NMR Spectroscopy Of Human Biofluids. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 2012, 62: 51-74.
115. Spratlin J L, Serkova N J*, et al.* Clinical Applications Of Metabolomics In Oncology: A Review. *Clinical Cancer Research*, 2009, 15(2): 431-440.
116. O'Connell T M. Recent advances in metabolomics in oncology. *Bioanalysis*, 2012, 4(4): 431-451.
117. Davis V W, Bathe O F*, et al.* Metabolomics And Surgical Oncology: Potential Role For Small Molecule Biomarkers. *Journal of Surgical Oncology*, 2011, 103(5): 451-459.
118. Moestue S, Sitter B*, et al.* HR MAS MR Spectroscopy In Metabolic Characterization Of Cancer. *Current Topics in Medicinal Chemistry*, 2011, 11(1): 2-26.
119. Oakman C, Tenori L*, et al.* Uncovering The Metabolomic Fingerprint Of Breast Cancer. *International Journal of Biochemistry & Cell Biology*, 2011, 43(7): 1010-1020.

120. Claudino W M, Quattrone A*, et al.* Metabolomics: Available Results, Current Research Projects In Breast Cancer, And Future Applications. *Journal of Clinical Oncology*, 2007, 25(19): 2840-2846.
121. Gu H, Pan Z*, et al.* Principal Component Directed Partial Least Squares Analysis For Combining Nuclear Magnetic Resonance And Mass Spectrometry Data In Metabolomics: Application To The Detection Of Breast Cancer. *Analytica Chimica Acta*, 2011, 686(1-2): 57-63.
122. Asiago V M, Alvarado L Z*, et al.* Early Detection Of Recurrent Breast Cancer Using Metabolite Profiling. *Cancer Research*, 2010, 70(21): 8309-8318.
123. Oakman C, Tenori L*, et al.* Identification Of A Serum-Detectable Metabolomic Fingerprint Potentially Correlated With The Presence Of Micrometastatic Disease In Early Breast Cancer Patients At Varying Risks Of Disease Relapse By Traditional Prognostic Methods. *Annals of Oncology*, 2011, 22(6): 1295-1301.
124. Keun H C, Sidhu J*, et al.* Serum Molecular Signatures Of Weight Change During Early Breast Cancer Chemotherapy. *Clinical Cancer Research*, 2009, 15(21): 6716-6723.
125. Stebbing J, Sharma A*, et al.* A Metabolic Phenotyping Approach To Understanding Relationships Between Metabolic Syndrome And Breast Tumour Responses To Chemotherapy. *Annals of Oncology*, 2012, 23(4): 860-U862.
126. Fossel E T, Carr J M*, et al.* Detection Of Malignant-Tumors - Water-Suppressed Proton Nuclear-Magnetic-Resonance Spectroscopy Of Plasma. *New England Journal of Medicine*, 1986, 315(22): 1369-1376.
127. Holmberg L, Jakobsson U*, et al.* Failure To Detect Early Breast-Cancer Using In-Vitro Nuclear-Magnetic-Resonance Spectroscopy Of Plasma. *British Journal of Cancer*, 1993, 68(2): 389-392.
128. Yabsley W, Homer-Vanniasinkam S*, et al.* Nuclear Magnetic Resonance Spectroscopy In The Detection And Characterisation Of Cardiovascular Disease: Key Studies. *ISRN Vascular Medicine*, 2012, 2012: 11.
129. Ala-Korpela M, Lankinen N*, et al.* The Inherent Accuracy Of ^{1}H NMR Spectroscopy To Quantify Plasma Lipoproteins Is Subclass Dependent. *Atherosclerosis*, 2007, 190(2): 352-358.
130. Slupsky C M, Steed H*, et al.* Urine Metabolite Analysis Offers Potential Early Diagnosis Of Ovarian And Breast Cancers. *Clinical Cancer Research*, 2010, 16(23): 5835-5841.
131. Woo H M, Kim K M*, et al.* Mass Spectrometry Based Metabolomic Approaches In Urinary Biomarker Study Of Women's Cancers. *Clinica Chimica Acta*, 2009, 400(1-2): 63-69.
132. Gribbestad I S, Sitter B*, et al.* Metabolite Composition In Breast Tumors Examined By Proton Nuclear Magnetic Resonance Spectroscopy. *Anticancer Research*, 1999, 19(3A): 1737-1746.
133. Gribbestad I S, Fjosne H E*, et al.* In-Vitro Proton NMR-Spectroscopy Of Extracts From Human Breast-Tumors And Noninvolved Breast-Tissue. *Anticancer Research*, 1993, 13(6A): 1973-1980.

134. Gribbestad I S, Petersen S B*, et al.* ^{1}H-NMR Spectroscopic Characterization Of Perchloric-Acid Extracts From Breast Carcinomas And Noninvolved Breast-Tissue. *NMR in Biomedicine*, 1994, 7(4): 181-194.
135. Beckonert O, Monnerjahn K*, et al.* Visualizing Metabolic Changes In Breast-Cancer Tissue Using ^{1}H-NMR Spectroscopy And Self-Organizing Maps. *NMR in Biomedicine*, 2003, 16(1): 1-11.
136. Sitter B, Bathen T F*, et al.* Quantification Of Metabolites In Breast Cancer Patients With Different Clinical Prognosis Using HR MAS MR Spectroscopy. *NMR in Biomedicine*, 2010, 23(4): 424-431.
137. Cheng L L, Chang I W*, et al.* Evaluating Human Breast Ductal Carcinomas With High-Resolution Magic-Angle Spinning Proton Magnetic Resonance Spectroscopy. *Journal of Magnetic Resonance*, 1998, 135(1): 194-202.
138. Bathen T F, Jensen L R*, et al.* MR-Determined Metabolic Phenotype Of Breast Cancer In Prediction Of Lymphatic Spread, Grade, And Hormone Status. *Breast Cancer Research and Treatment*, 2007, 104(2): 181-189.
139. Giskeodegard G F, Grinde M T*, et al.* Multivariate Modeling And Prediction Of Breast Cancer Prognostic Factors Using MR Metabolomics. *Journal of Proteome Research*, 2010, 9(2): 972-979.
140. Li M, Song Y*, et al.* An HR-MAS MR Metabolomics Study On Breast Tissues Obtained With Core Needle Biopsy. *PLoS One*, 2011, 6(10).
141. Sitter B, Lundgren S*, et al.* Comparison Of HR MAS MR Spectroscopic Profiles Of Breast Cancer Tissue With Clinical Parameters. *NMR in Biomedicine*, 2006, 19(1): 30-40.
142. Nicholson J K, Foxall P J D*, et al.* 750-MHZ ^{1}H And ^{1}H-^{13}C NMR-Spectroscopy Of Human Blood-Plasma. *Analytical Chemistry*, 1995, 67(5): 793-811.
143. Abu-Bedair F A, El-Gamal B A*, et al.* Serum Lipids And Tissue DNA Content In Egyptian Female Breast Cancer Patients. *Japanese Journal of Clinical Oncology*, 2003, 33(6): 278-282.
144. Ray G and Husain S A. Role Of Lipids, Lipoproteins And Vitamins In Women With Breast Cancer. *Clinical Biochemistry*, 2001, 34(1): 71-76.
145. Holmes E, Loo R L*, et al.* Human Metabolic Phenotype Diversity And Its Association With Diet And Blood Pressure. *Nature*, 2008, 453(7193): 396-U350.
146. Curtis C, Shah S P*, et al.* The Genomic And Transcriptomic Architecture Of 2,000 Breast Tumours Reveals Novel Subgroups. *Nature*, 2012.
147. Hirschhaeuser F, Sattler U G A*, et al.* Lactate: A Metabolic Key Player In Cancer. *Cancer Research*, 2011, 71(22): 6921-6925.
148. Brown C D, Higgins M*, et al.* Body Mass Index And The Prevalence Of Hypertension And Dyslipidemia. *Obesity Research*, 2000, 8(9): 605-619.
149. Flegal K M, Graubard B I*, et al.* Excess Deaths Associated With Underweight, Overweight, And Obesity. *JAMA-Journal of the American Medical Association*, 2005, 293(15): 1861-1867.
150. Wang-Sattler R, Yu Y*, et al.* Metabolic Profiling Reveals Distinct Variations Linked To Nicotine Consumption In Humans - First Results From The KORA Study. *PLoS One*, 2008, 3(12).

151. Jonasson J M, Ljung R*, et al.* Insulin Glargine Use And Short-Term Incidence Of Malignancies-A Population-Based Follow-Up Study In Sweden. *Diabetologia*, 2009, 52(9): 1745-1754.
152. Boden G, Sargrad K*, et al.* Effect Of A Low-Carbohydrate Diet On Appetite, Blood Glucose Levels, And Insulin Resistance In Obese Patients With Type 2 Diabetes. *Annals of Internal Medicine*, 2005, 142(6): 403-411.
153. Turner E, Brewster J A*, et al.* Plasma From Women With Preeclampsia Has A Low Lipid And Ketone Body Content - A Nuclear Magnetic Resonance Study. *Hypertension in Pregnancy*, 2007, 26(3): 329-342.
154. Kolwijck E, Engelke U F*, et al.* N-Acetyl Resonances In In Vivo And In Vitro NMR Spectroscopy Of Cystic Ovarian Tumors. *NMR in Biomedicine*, 2009, 22(10): 1093-1099.
155. Hounsell E F, Young M*, et al.* Glycoprotein Changes In Tumours: A Renaissance In Clinical Applications. *Clinical Science*, 1997, 93(4): 287-293.
156. Czuba M and Smith I C P. Biological And NMR Markers For Cancer. *Pharmacology & Therapeutics*, 1991, 50(2): 147-190.
157. Mercier P, Lewis M J*, et al.* Towards Automatic Metabolomic Profiling Of High-Resolution One-Dimensional Proton NMR Spectra. *Journal of Biomolecular NMR*, 2011, 49(3-4): 307-323.
158. Rasmussen L G, Savorani F*, et al.* Standardization Of Factors That Influence Human Urine Metabolomics. *Metabolomics*, 2011, 7(1): 71-83.
159. Dragsted L O. Biomarkers Of Meat Intake And The Application Of Nutrigenomics. *Meat Science*, 2010, 84(2): 301-307.
160. Nicholls A W, Mortishire-Smith R J*, et al.* NMR Spectroscopic-Based Metabonomic Studies Of Urinary Metabolite Variation In Acclimatizing Germ-Free Rats. *Chemical Research in Toxicology*, 2003, 16(11): 1395-1404.
161. Li J V, Saric J*, et al.* Chemometric Analysis Of Biofluids From Mice Experimentally Infected With Schistosoma Mansoni. *Parasites & Vectors*, 2011, 4.
162. Yap I K S, Clayton T A*, et al.* An Integrated Metabonomic Approach To Describe Temporal Metabolic Disregulation Induced In The Rat By The Model Hepatotoxin Allyl Formate. *Journal of Proteome Research*, 2006, 5(10): 2675-2684.
163. Yang W, Wang Y*, et al.* Analysis Of Human Urine Metabolites Using SPE And NMR Spectroscopy. *Science in China Series B-Chemistry*, 2008, 51(3): 218-225.
164. Psihogios N G, Gazi I F*, et al.* Gender-Related And Age-Related Urinalysis Of Healthy Subjects By NMR-Based Metabonomics. *NMR in Biomedicine*, 2008, 21(3): 195-207.
165. Holmes E, Loo R L*, et al.* Detection Of Urinary Drug Metabolite (Xenometabolome) Signatures In Molecular Epidemiology Studies Via Statistical Total Correlation (NMR) Spectroscopy. *Analytical Chemistry*, 2007, 79(7): 2629-2640.
166. Gerchman F, Tong J*, et al.* Body Mass Index Is Associated With Increased Creatinine Clearance By A Mechanism Independent Of Body Fat Distribution. *Journal of Clinical Endocrinology & Metabolism*, 2009, 94(10): 3781-3788.

167. Rule A D, Bailey K R*, et al.* For Estimating Creatinine Clearance Measuring Muscle Mass Gives Better Results Than Those Based On Demographics. *Kidney International*, 2009, 75(10): 1071-1078.
168. Oterdoom L H, Gansevoort R T*, et al.* Urinary Creatinine Excretion, An Indirect Measure Of Muscle Mass, Is An Independent Predictor Of Cardiovascular Disease And Mortality In The General Population. *Atherosclerosis*, 2009, 207(2): 534-540.
169. Lindon J K N and Elaine Holmes. The Handbook of Metabonomics and Metabolomics. 1st Edition. Elsevier: Amsterdam, Netherlands, 2007.
170. Xu J, Cai S*, et al.* Statistical Two-Dimensional Correlation Spectroscopy Of Urine And Serum From Metabolomics Data. *Chemometrics and Intelligent Laboratory Systems*, 2012, 112: 33-40.
171. Crockford D J, Maher A D*, et al.* ^{1}H NMR And UPLC-MSE Statistical Heterospectroscopy: Characterization Of Drug Metabolites (Xenometabolome) In Epidemiological Studies. *Analytical Chemistry*, 2008, 80(18): 6835-6844.
172. Kutyshenko V P, Molchanov M*, et al.* Analyzing And Mapping Sweat Metabolomics By High-Resolution NMR Spectroscopy. *PLoS One*, 2011, 6(12).
173. Kim L S, Axelrod L J*, et al.* Efficacy Of Methylsulfonylmethane (MSM) In Osteoarthritis Pain Of The Knee: A Pilot Clinical Trial. *Osteoarthritis and Cartilage*, 2006, 14(3): 286-294.
174. Engelke U F H, Tangerman A*, et al.* Dimethyl Sulfone In Human Cerebrospinal Fluid And Blood Plasma Confirmed By One-Dimensional ^{1}H And Two-Dimensional ^{1}H-^{13}C NMR. *NMR in Biomedicine*, 2005, 18(5): 331-336.
175. Winning H, Roldan-Marin E*, et al.* An Exploratory NMR Nutri-Metabonomic Investigation Reveals Dimethyl Sulfone As A Dietary Biomarker For Onion Intake. *Analyst*, 2009, 134(11): 2344-2351.
176. Engelke U F H, Kremer B*, et al.* NMR Spectroscopic Studies On The Late Onset Form Of 3-Methylglutaconic Aciduria Type I And Other Defects In Leucine Metabolism. *NMR in Biomedicine*, 2006, 19(2): 271-278.
177. Horvath T D, Matthews N I*, et al.* Measurement Of 3-Hydroxyisovaleric Acid In Urine From Marginally Biotin-Deficient Humans By UPLC-MS/MS. *Analytical and Bioanalytical Chemistry*, 2011, 401(9): 2805-2810.
178. Marshall M W, Kliman P G*, et al.* Effects Of Biotin On Lipids And Other Constituents Of Plasma Of Healthy-Men And Women. *Artery*, 1980, 7(4): 330-351.
179. Fernandez-Mejia C. Pharmacological Effects Of Biotin. *Journal of Nutritional Biochemistry*, 2005, 16(7): 424-427.
180. Sealey W M, Teague A M*, et al.* Smoking Accelerates Biotin Catabolism In Women. *American Journal of Clinical Nutrition*, 2004, 80(4): 932-935.
181. Freedman D S, Otvos J D*, et al.* Relation Of Lipoprotein Subclasses As Measured By Proton Nuclear Magnetic Resonance Spectroscopy To Coronary Artery Disease. *Arteriosclerosis Thrombosis and Vascular Biology*, 1998, 18(7): 1046-1053.
182. Michalaki V, Koutroulis G*, et al.* Evaluation Of Serum Lipids And High-Density Lipoprotein Subfractions (HDL2, HDL3) In Postmenopausal Patients With Breast Cancer. *Molecular and Cellular Biochemistry*, 2005, 268(1-2): 19-24.

183. Bittner V, Johnson B D*, et al.* The Triglyceride/High-Density Lipoprotein Cholesterol Ratio Predicts All-Cause Mortality In Women With Suspected Myocardial Ischemia: A Report From The Women's Ischemia Syndrome Evaluation (WISE). *American Heart Journal*, 2009, 157(3): 548-555.
184. Nishiumi S, Kobayashi T*, et al.* A Novel Serum Metabolomics-Based Diagnostic Approach For Colorectal Cancer. *PLoS One*, 2012, 7(7).
185. Dunn W B, Broadhurst D*, et al.* Procedures For Large-Scale Metabolic Profiling Of Serum And Plasma Using Gas Chromatography And Liquid Chromatography Coupled To Mass Spectrometry. *Nature Protocols*, 2011, 6(7): 1060-1083.
186. Ward J L, Baker J M*, et al.* An Inter-Laboratory Comparison Demonstrates That [H-1]-NMR Metabolite Fingerprinting Is A Robust Technique For Collaborative Plant Metabolomic Data Collection. *Metabolomics*, 2010, 6(2): 263-273.
187. Tugnoli V, Tosi M R*, et al.* Characterization Of Lipids From Human Brain Tissues By Multinuclear Magnetic Resonance Spectroscopy. *Biopolymers*, 2001, 62(6): 297-306.
188. Lehtimaki K K, Valonen P K*, et al.* Metabolite Changes In BT4C Rat Gliomas Undergoing Ganciclovir-Thymidine Kinase Gene Therapy-Induced Programmed Cell Death As Studied By ^{1}H NMR Spectroscopy In Vivo, Ex Vivo, And In Vitro. *Journal of Biological Chemistry*, 2003, 278(46): 45915-45923.
189. Moseley H N B, Lane A N*, et al.* A Novel Deconvolution Method For Modeling UDP-N-Acetyl-D-Glucosamine Biosynthetic Pathways Based On ^{13}C Mass Isotopologue Profiles Under Non-Steady-State Conditions. *BMC Biology*, 2011, 9.
190. Pan X, Wilson M*, et al.* In Vitro Metabonomic Study Detects Increases In UDP-Glcnac And UDP-Galnac, As Early Phase Markers Of Cisplatin Treatment Response In Brain Tumor Cells. *Journal of Proteome Research*, 2011, 10(8): 3493-3500.
191. MacIntyre D A, Jimenez B*, et al.* Serum Metabolome Analysis By ^{1}H NMR Reveals Differences Between Chronic Lymphocytic Leukaemia Molecular Subgroups. *Leukemia*, 2010, 24(4): 788-797.
192. Tessem M-B, Swanson M G*, et al.* Evaluation Of Lactate And Alanine As Metabolic Biomarkers Of Prostate Cancer Using ^{1}H HR-MAS Spectroscopy Of Biopsy Tissues. *Magnetic Resonance in Medicine*, 2008, 60(3): 510-516.
193. Kim J W, Tchernyshyov I*, et al.* HIF-1-Mediated Expression Of Pyruvate Dehydrogenase Kinase: A Metabolic Switch Required For Cellular Adaptation To Hypoxia. *Cell Metabolism*, 2006, 3(3): 177-185.
194. Icard P, Poulain L*, et al.* Understanding The Central Role Of Citrate In The Metabolism Of Cancer Cells. *Biochimica Et Biophysica Acta-Reviews on Cancer*, 2012, 1825(1): 111-116.
195. Medina M A and Decastro I N. Glutaminolysis And Glycolysis Interactions In Proliferant Cells. *International Journal of Biochemistry*, 1990, 22(7): 681-683.
196. Moreadith R W and Lehninger A L. The Pathways Of Glutamate And Glutamine Oxidation By Tumor-Cell Mitochondria - Role Of Mitochondrial NAD(P)$^{+}$-Dependent Malic Enzyme. *Journal of Biological Chemistry*, 1984, 259(10): 6215-6221.

197. Monleon D, Morales J M*, et al.* Benign And Atypical Meningioma Metabolic Signatures By High-Resolution Magic-Angle Spinning Molecular Profiling. *Journal of Proteome Research*, 2008, 7(7): 2882-2888.
198. Lehnhardt F G, Bock C*, et al.* Metabolic Differences Between Human Brain Tumors: A ^{1}H NMR Primary And Recurrent Spectroscopic Investigation. *NMR in Biomedicine*, 2005, 18(6): 371-382.
199. Richardson A D, Yang C*, et al.* Central Carbon Metabolism In The Progression Of Mammary Carcinoma. *Breast Cancer Research and Treatment*, 2008, 110(2): 297-307.
200. Kung H-N, Marks J R*, et al.* Glutamine Synthetase Is A Genetic Determinant Of Cell Type-Specific Glutamine Independence In Breast Epithelia. *PLoS Genetics*, 2011, 7(8).
201. Amores-Sanchez M I and Medina M A. Glutamine, As A Precursor Of Glutathione, And Oxidative Stress. *Molecular Genetics and Metabolism*, 1999, 67(2): 100-105.
202. Yeh C-C, Hou M-F*, et al.* A Study Of Glutathione Status In The Blood And Tissues Of Patients With Breast Cancer. *Cell Biochemistry and Function*, 2006, 24(6): 555-559.
203. Balendiran G K, Dabur R*, et al.* The Role Of Glutathione In Cancer. *Cell Biochemistry and Function*, 2004, 22(6): 343-352.
204. Kuhajda F P, Jenner K*, et al.* Fatty-Acid Synthesis - A Potential Selective Target For Antineoplastic Therapy. *Proceedings of the National Academy of Sciences of the United States of America*, 1994, 91(14): 6379-6383.
205. Somashekar B S, Kamarajan P*, et al.* Magic Angle Spinning NMR-Based Metabolic Profiling Of Head And Neck Squamous Cell Carcinoma Tissues. *Journal of Proteome Research*, 2011, 10(11): 5232-5241.
206. Onda T, Uzawa K*, et al.* Ubiquitous Mitochondrial Creatine Kinase Downregulated In Oral Squamous Cell Carcinoma. *British Journal of Cancer*, 2006, 94(5): 698-709.
207. Baird M F, Graham S M*, et al.* Creatine-Kinase- And Exercise-Related Muscle Damage Implications For Muscle Performance And Recovery. *Journal of Nutrition and Metabolism*, 2012, 2012.
208. Stapleton P P, O'Flaherty L*, et al.* Host Defense - A Role For The Amino Acid Taurine? *Journal of Parenteral and Enteral Nutrition*, 1998, 22(1): 42-48.
209. El Agouza I M, Eissa S S*, et al.* Taurine: A Novel Tumor Marker For Enhanced Detection Of Breast Cancer Among Female Patients. *Angiogenesis*, 2011, 14(3): 321-330.
210. Stewart J D, Marchan R*, et al.* Choline-Releasing Glycerophosphodiesterase EDI3 Drives Tumor Cell Migration And Metastasis. *Proceedings of the National Academy of Sciences of the United States of America*, 2012, 109(21): 8155-8160.
211. Misra D and Bajpai U. Metabolite Characterization In Serum Samples From Normal Healthy Human Subjects By ^{1}H And ^{13}C NMR Spectroscopy. *Bulletin of the Chemical Society of Ethiopia*, 2009, 23(2): 211-221.
212. Moestue S A, Borgan E*, et al.* Distinct Choline Metabolic Profiles Are Associated With Differences In Gene Expression For Basal-Like And Luminal-Like Breast Cancer Xenograft Models. *BMC Cancer*, 2010, 10.

213. Glunde K, Jie C*, et al.* Molecular Causes Of The Aberrant Choline Phospholipid Metabolism In Breast Cancer. *Cancer Research*, 2004, 64(12): 4270-4276.
214. Moestue S A, Giskeodegard G F*, et al.* Glycerophosphocholine (GPC) Is A Poorly Understood Biomarker In Breast Cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 2012, 109(38): E2506-E2506.
215. Marchan R, Stewart J D*, et al.* Reply To Moestue Et Al.: Untangling The Contribution Of Choline Metabolism To The Metastatic Process. *Proceedings of the National Academy of Sciences of the United States of America*, 2012, 109(38): E2507-E2507.
216. Taylor L A, Arends J*, et al.* Plasma Lyso-Phosphatidylcholine Concentration Is Decreased In Cancer Patients With Weight Loss And Activated Inflammatory Status. *Lipids in Health and Disease*, 2007, 6.
217. Broschat K O, Gorka C*, et al.* Kinetic Characterization Of Human Glutamine-Fructose-6-Phosphate Amidotransferase I - Potent Feedback Inhibition By Glucosamine 6-Phosphate. *Journal of Biological Chemistry*, 2002, 277(17): 14764-14770.
218. Gu Y, Mi W*, et al.* Glcnacylation Plays An Essential Role In Breast Cancer Metastasis. *Cancer Research*, 2010, 70(15): 6344-6351.
219. Hilvo M, Denkert C*, et al.* Novel Theranostic Opportunities Offered By Characterization Of Altered Membrane Lipid Metabolism In Breast Cancer Progression. *Cancer Research*, 2011, 71(9): 3236-3245.
220. Noula C, Bonzom P*, et al.* ^{1}H-NMR Lipid Profiles Of Human Blood Platelets; Links With Coronary Artery Disease. *Biochimica Et Biophysica Acta-Molecular and Cell Biology of Lipids*, 2000, 1487(1): 15-23.
221. Adosraku R K, Choi G T Y*, et al.* NMR Lipid Profiles Of Cells, Tissues, And Body-Fluids - Proton NMR Analysis Of Human Erythrocyte Lipids. *Journal of Lipid Research*, 1994, 35(11): 1925-1931.
222. Bathen T F, Krane J*, et al.* Quantification Of Plasma Lipids And Apolipoproteins By Use Of Proton NMR Spectroscopy, Multivariate And Neural Network Analysis. *NMR in Biomedicine*, 2000, 13(5): 271-288.
223. Casu M, Anderson G J*, et al.* NMR Lipid Profiles Of Cells, Tissues And Body-Fluids. I-1D And 2D Proton NMR Of Lipids From Rat-Liver. *Magnetic Resonance in Chemistry*, 1991, 29(6): 594-602.
224. Otvos J D, Jeyarajah E J*, et al.* Quantification Of Plasma-Lipoproteins By Proton Nuclear Magnetic Resonance Spectroscopy. *Clinical Chemistry*, 1991, 37(3): 377-386.
225. Viant M R. Improved Methods For The Acquisition And Interpretation Of NMR Metabolomic Data. *Biochemical and Biophysical Research Communications*, 2003, 310(3): 943-948.
226. Wishart D S, Bigam C G*, et al.* ^{1}H, ^{13}C And ^{15}N Chemical-Shift Referencing In Biomolecular NMR. *Journal of Biomolecular NMR*, 1995, 6(2): 135-140.
227. Libby P. Inflammation In Atherosclerosis. *Nature*, 2002, 420(6917): 868-874.
228. Dhale M A, Divakar S*, et al.* Isolation And Characterization Of Dihydromonacolin-MV From Monascus Purpureus For Antioxidant Properties. *Applied Microbiology and Biotechnology*, 2007, 73: 1197-1202.

229. Lusis A J. Atherosclerosis. *Nature*, 2000, 407(6801): 233-241.
230. Moreno J J and Mitjavila M T. The Degree Of Unsaturation Of Dietary Fatty Acids And The Development Of Atherosclerosis (Review). *Journal of Nutritional Biochemistry*, 2003, 14(4): 182-195.
231. Hansson G K. Mechanisms Of Disease - Inflammation, Atherosclerosis, And Coronary Artery Disease. *New England Journal of Medicine*, 2005, 352(16): 1685-1695.
232. Tang D, Yang C*, et al.* Local Maximal Stress Hypothesis And Computational Plaque Vulnerability Index For Atherosclerotic Plaque Assessment. *Annals of Biomedical Engineering*, 2005, 33(12): 1789-1801.
233. Lee J M S, Shirodaria C*, et al.* Multi-Modal Magnetic Resonance Imaging Quantifies Atherosclerosis And Vascular Dysfunction In Patients With Type 2 Diabetes Mellitus. *Diabetes and Vascular Disease Research*, 2007, 4(1): 44-48.
234. Giles M F and Rothwell P M. Risk Of Stroke Early After Transient Ischaemic Attack: A Systematic Review And Meta-Analysis. *Lancet Neurology*, 2007, 6(12): 1063-1072.
235. Hansson G K and Libby P. The Immune Response In Atherosclerosis: A Double-Edged Sword. *Nature Reviews Immunology*, 2006, 6(7): 508-519.
236. Beard J D. ABC Of Arterial And Venous Disease - Chronic Lower Limb Ischaemia. *British Medical Journal*, 2000, 320(7238): 854-+.
237. Regensteiner J G and Stewart K J. Established And Evolving Medical Therapies For Claudication In Patients With Peripheral Arterial Disease. *Nature Clinical Practice Cardiovascular Medicine*, 2006, 3(11): 604-610.
238. Rosamond W, Flegal K*, et al.* Heart Disease And Stroke Statistics - 2008 Update - A Report From The American Heart Association Statistics Committee And Stroke Statistics Subcommittee. *Circulation*, 2008, 117(4): E25-E146.
239. Lewis G D, Asnani A*, et al.* Application Of Metabolomics To Cardiovascular Biomarker And Pathway Discovery. *Journal of the American College of Cardiology*, 2008, 52(2): 117-123.
240. Vallejo M, Garcia A*, et al.* Plasma Fingerprinting With GC-MS In Acute Coronary Syndrome. *Analytical and Bioanalytical Chemistry*, 2009, 394(6): 1517-1524.
241. Sabatine M S, Liu E*, et al.* Metabolomic Identification Of Novel Biomarkers Of Myocardial Ischemia. *Circulation*, 2005, 112(25): 3868-3875.
242. Brindle J T, Antti H*, et al.* Rapid And Noninvasive Diagnosis Of The Presence And Severity Of Coronary Heart Disease Using ^{1}H-NMR-Based Metabonomics. *Nature Medicine*, 2002, 8(12): 1439-1444.
243. Kirschenlohr H L, Griffin J L*, et al.* Proton NMR Analysis Of Plasma Is A Weak Predictor Of Coronary Artery Disease. *Nature Medicine*, 2006, 12(6): 705-710.
244. Correll D J, Hepner D L*, et al.* Preoperative Electrocardiograms Patient Factors Predictive Of Abnormalities. *Anesthesiology*, 2009, 110(6): 1217-1222.

245. Coolen S A, Daykin C A*, et al.* Measurement Of Ischaemia-Reperfusion In Patients With Intermittent Claudication Using NMR-Based Metabonomics. *NMR in Biomedicine*, 2008, 21(7): 686-695.
246. Barton R H, Waterman D*, et al.* The Influence Of EDTA And Citrate Anticoagulant Addition To Human Plasma On Information Recovery From NMR-Based Metabolic Profiling Studies. *Molecular Biosystems*, 2009, 6(1): 215-224.
247. Wishart D S, Tzur D*, et al.* HMDB: The Human Metabolome Database. *Nucleic Acids Research*, 2007, 35: D521-D526.
248. Wang Y L, Holmes E*, et al.* Experimental Metabonomic Model Of Dietary Variation And Stress Interactions. *Journal of Proteome Research*, 2006, 5(7): 1535-1542.
249. Hong Y S, Coen M*, et al.* Chemical Shift Calibration Of ^{1}H MAS NMR Liver Tissue Spectra Exemplified Using A Study Of Glycine Protection Of Galactosamine Toxicity. *Magnetic Resonance in Chemistry*, 2009, 47: S47-S53.
250. Gronwald W, Klein M S*, et al.* Urinary Metabolite Quantification Employing 2D NMR Spectroscopy. *Analytical Chemistry*, 2008, 80(23): 9288-9297.
251. Alum M F, Shaw P A*, et al.* 4,4-Dimethyl-4-Silapentane-1-Ammonium Trifluoroacetate (DSA), A Promising Universal Internal Standard For NMR-Based Metabolic Profiling Studies Of Biofluids, Including Blood Plasma And Serum. *Metabolomics*, 2008, 4(2): 122-127.
252. de Graaf R A and Behar K L. Quantitative ^{1}H NMR Spectroscopy Of Blood Plasma Metabolites. *Analytical Chemistry*, 2003, 75(9): 2100-2104.
253. Gollnick P D, Bayly W M*, et al.* Exercise Intensity, Training, Diet, And Lactate Concentration In Muscle And Blood. *Medicine and Science in Sports and Exercise*, 1986, 18(3): 334-340.
254. Lawton K A, Berger A*, et al.* Analysis Of The Adult Human Plasma Metabolome. *Pharmacogenomics*, 2008, 9(4): 383-397.
255. McClay J L, Adkins D E*, et al.* ^{1}H Nuclear Magnetic Resonance Metabolomics Analysis Identifies Novel Urinary Biomarkers For Lung Function. *Journal of Proteome Research*, 9(6): 3083-3090.
256. Mas S, Martinez-Pinna R*, et al.* Local Non-Esterified Fatty Acids Correlate With Inflammation In Atheroma Plaques Of Patients With Type 2 Diabetes. *Diabetes*, 59(6): 1292-1301.
257. Quinones M P and Kaddurah-Daouk R. Metabolomics Tools For Identifying Biomarkers For Neuropsychiatric Diseases. *Neurobiology of Disease*, 2009, 35(2): 165-176.
258. Takeda I, Stretch C*, et al.* Understanding The Human Salivary Metabolome. *NMR in Biomedicine*, 2009, 22(6): 577-584.
259. Conen D, Everett B M*, et al.* Smoking, Smoking Cessation, [Corrected] And Risk For Symptomatic Peripheral Artery Disease In Women: A Cohort Study. *Annals of Internal Medicine*, 2011, 154(11): 719-726.
260. Vowden K and Vowden P. Doppler And The ABPI: How Good Is Our Understanding? *Journal of Wound Care*, 2001, 10(6): 197-202.