University of Alberta

Application of ROC Curve Analysis to Metabolomics Data Sets for the Detection of Cancer in a Mouse Model

by

Jennifer Moroz

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Medical Physics

Department of Physics

Fall 2010
Edmonton, Alberta

Examining Committee

Gino Fallone, Oncology

Alasdair Syme, Oncology

Keith Wachowicz, Oncology

Nicola DeZanche, Oncology

Jack Tuszynski, Physics

Abstract

The goal of this study was to show that quantifiable metabolic changes may be used to screen for cancer. NIH III nude mice were injected with human GBM cells (n=22) or with saline (n=14). Daily urine samples were collected pre and post-injection and analyzed using NMR Spectroscopy. 34 metabolites were identified and quantified through targeted profiling with Chenomx Suite 5.1. Univariate statistical analysis showed that 3 metabolites (2-oxoglutarate, glucose and trimethylamine n-oxide) were significantly altered in the presence of tumour, while PCA and PLS-DA analysis modeled the maximum variance between the healthy and tumour-bearing groups. Receiver operating characteristic (ROC) curve analysis was applied to provide a measure of clinical utility. ROC statistics were as high as 0.850 for the analysis of individual metabolites, 0.939 for the analysis of metabolite pairs and 0.996 for the PLS-DA models. These results show that metabolomics has potential to screen for cancer

Acknowledgements

I would like to thank my supervisors Dr. Gino Fallone and Dr. Alasdair Syme for all their time and support throughout the degree. Huge thanks to Dr. Alasdair Syme for his assistance in writing up conference abstracts, writing my first paper and offering encouragement whenever I needed it. Thank-you to Dr. Joan Turner for supplying the GBM xenograft cells. This project would not have been possible without them. I would also like to thank Dr. Carolyn Slupsky for her help with the initial stages of the project. Thank-you to Gail Hippurson and Dan McGinn for their help in the Vivarium. They went far beyond our expectations to ensure that our animals remained healthy and that our experiment ran smoothly from start to finish. Thank-you also to Dr. Ryan McKay for his help and expertise in acquiring spectra on the 800 MHz NMR Spectrometer. A final thank-you to the students and the staff at the Cross Cancer Institute who made my experience as a Master's student all I could have ever wished for.

I would like to acknowledge the financial support from N'SERC, Alberta Ingenuity and the Alberta Cancer Board. Without such support, this study would not have been possible.

Table of Contents

Chapter 5: ROC Curve Analysis

Chapter 6: Experimental Techniques

Chapter 7: Results and Discussion

List of Tables

Chapter 5: ROC Curve Analysis

Chapter 7: Results and Discussion

List of Figures

Chapter 2: Screening for Cancer

Chapter 3: Nuclear Magnetic Resonance Spectroscopy Theory

Chapter 5: ROC Curve Analysis

Chapter 6: Experimental Techniques

Chapter 7: Results and Discussion

List of Abbreviations

α	Pre-defined Univariate Statistics Significance Threshold
ACS	American Cancer Society
ADC	Analog to Digital Converter
APCI	Atmospheric Pressure Chemical Ionization
AQ	NMR Acquisition
AUC	Area Under the ROC Curve
B_o	Static External Magnetic field
B_1	Applied Radiofrequency Magnetic Field
c	PLS-DA Weight Vector for Y Matrix
CAD	Computer-aided Detection
CSA	chemical shift anisotropy
CV	Cross Validation
C_X	Covariance Matrix
Dist	Minimum Distance to the point (0,1)
D.T.	Decision Threshold
DSS	2,2-dimethhyl-2-silapentane-5-sulfanate
EOC	Epithelial Ovarian Cancer
ESI	Electrospray Ionization
FDR	False Discovery Rate
FID	Free Induction Decay
FN	False Negative
FP	False Positive
FWER	Family Wise Error Rate

GBM	Glio-Blastoma Multiforme
GC-MS	Gas Chromatography – Mass Spectroscopy
HCA	Hierarchal Cluster Analysis
J	Youden Index
KNN	K-Nearest Neighbour Analysis
LB	Line Broadening
LC-MS	Liquid Chromatography – Mass Spectroscopy
LDA	Linear Discriminant Analysis
LV	Latent Variable
M	Magnetization
MS	Mass Spectroscopy
NIPALS	Non-linear Iterative Partial Least Squares Algorithm
NLM	Non-linear Mapping Procedures
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
NPV	Negative Predictive Value
OPLS-DA	Orthogonal Partial Least Squares Discriminant Analysis
OSCC	Oral Squamous Cell Carcinoma
t_i	Multi-variate Statistics Loading Vector
Pap	papanicolaou smear
Par	Pareto Scaling
PBS	Phosphate Buffered Saline
PC	Principal Component
PCA	Principal Component Analysis
pFDR	Positive False Discovery Rate
PLS	Partial Least Squares

PLS-DA	Partial Least Squares Discriminant Analysis
PPV	Positive Predictive Value
PSA	Prostate-specific Antigen
PRESS	Prediction Residual Sum of Squares
PREV	Prevalence
Q^2	Goodness of Prediction Parameter
R^2	Goodness of Fit Parameter
RD	Relaxation Delay
RF	Radiofrequency
ROC	Receiver Operating Characteristic
Sen	Sensitivity
Spec	Specificity
SNR	Signal to Noise Ratio
t_1	Fixed Interval During which the Nuclei Precess Freely
t_i	Multi-variate Statistics Scores Vector
t_m	Mixing Time
t_R	Time to Repetition
T_1	Longitudinal or spin-lattice Relaxation Time Constant
T_2	Transverse or spin-spin Relaxation Time for a homogenous magnetic field
T_2^*	Transverse or spin-spin Relaxation overall
tcho	Total Choline
TN	True Negative
TP	True Positive
TPS	3-trimethylsilylpropionic acid
u	PLS-DA Weight Vector for X matrix

UV	Unit Variance Scaling
X	Data Matrix (n x m); n is the number of samples, m is the number of variables
Y	PLS-DA Response Matrix (c x m); c is the number of classes, m is the number of variables
Y.I.	Youden’s Index

List of Variables

B	Net Magnetic Field
B_o	Static External Magnetic Field
B_1	Applied RF Magnetic Field
δ	Chemical Shift
ΔE	Energy Difference Between Spin States
h	Planck's constant (6.26×10^{-34} Js)
$\hbar$	Planck's constant - h_{bar} (1.055×10^{-34} Js)
I	Spin Quantum Number (1/2 for protons)
I	Intensity of the Magnetization in the Presence of NOE
I_0	Intensity of the Magnetization without the NOE
${}^{n}J_{AB}$	J-coupling Constant where n is the Number of Intervening Bonds and A and B Identify the Two Coupled Spins
k_B	Boltzmann constant (1.381×10^{-23} J/K)
M_o	Net Magnetization
M_z	Longitudinal Magnetization
M_{xy}	Transverse Magnetization
$M_{z,0}^{i}$	Magnetization from a Single Molecule
$\overrightarrow{M}_{molecule}$	Net Magnetization for Each i Spin Type
$\overrightarrow{M}_{total}$	Magnetization From all j Types of Molecules Resonating at a Frequency $f_{,i}$ with a Concentration c_j
N	Number of Spin Nuclei in the Sample
N_α	Number of Nuclei in the Spin Up State
N_β	Number of Nuclei in the Spin Down State
$\eta_I\{S\}$	Observed NOE Enhancement

S_D	Constant Describing the Detector Sensitivity and Noise
τ	Pulse Length
T	Temperature in K
T_1	Longitudinal or spin-lattice Relaxation Time Constant
T_2	Transverse or spin-spin Relaxation Time Constant for a homogenous magnetic field
T_2^*	Transverse or spin-spin Relaxation overall
$T_2(\Delta B_o)$	Relaxation Time Constant due to Inhomogeneities in the Main Magnetic Field
μ	Magnetic Moment
γ	Gyrometric Ratio
$\Delta\upsilon_{1/2}$	Linewidth at Half Height
ν	Larmour Frequency in units of Hz
ω	Larmour Frequency in rad
ω_i	Larmour frequency of the metabolite of interest
ω_{ref}	Larmour frequency of an internal standard molecule
θ	Nutation Angle

Chapter 1
Overview of Thesis

1.1 Overview of the Project

The stage at which a tumour is detected serves as an important prognostic indicator (Feng et al., 2009). It has been well established that early detection of tumours will significantly improve long term survival rates as well as decrease the costs of treatment (Kil-Sun et al., 2010; Taplin et al., 2004). Studies of metabolism have shown promise in the field of early detection of metabolic diseases (Kim et al., 2009).

Metabolomics is the time dependent, quantitative measurement of the multivariate metabolic response to a pathological stimulus or genetic modification (Nicholson et al., 1999). It has been shown that quantifiable metabolic changes appear with the development of a disease (Nicholson et al., 2007). The complement of all metabolic changes that take place are expected to be unique to a particular disease; the so called metabolic fingerprint (Spratlin et al., 2009). By measuring the concentrations of metabolites in a biofluid or tissue extract as individuals transition from healthy to diseased, it may be possible to identify patterns of metabolic behaviour associated with the disease (Goldsmith et al., 2010).

To date, there is no standard method to evaluate the clinical significance of a metabolomics data set. If the results of metabolomics studies are to reach clinical acceptance, it is essential to quantify the effectiveness with which a metabolic signature can differentiate samples from healthy and diseased individuals. Receiver operating characteristic (ROC) curve analysis has been widely used in the clinic to evaluate a diagnostic test (Akobeng, 2006; DeLong et al., 1988). We hypothesize that the application of ROC curve analysis to a metabolomics data set will provide a measure of clinical significance, in which the results could be used to identify the metabolic changes associated with the early development of tumour in the body.

The goal of this thesis is to show that quantifiable metabolic changes in urine samples take place with the onset of tumour development in an animal GBM tumour model. The observed metabolic changes will be quantified with ROC curve analysis to provide a measure of their clinical significance. Univariate statistics are used to test the

significance of the changes in urinary metabolite concentrations between the healthy and tumour-bearing populations on an individual and paired (ratios) basis. Multivariate techniques are employed to investigate complicated metabolic relationships to optimally separate the two populations. An age-matched group of sham-injected control animals will serve to identify metabolic changes that are not related to the presence of the tumour.

1.2 Organization of the Thesis

The organization of the thesis will begin with a theoretical description of the experiment and a presentation of the techniques used in similar experiments. A full description of the methods used in our study will follow. The results will then be presented.

Chapter 2 opens with a description of what constitutes a screening test and provides a brief summary of mammography, cytology and the prostate specific antigen tests. The potential for using metabolomics for screening is presented. Six studies which use metabolomics information for the identification of cancer are then summarized. Next, the advantages of using urine samples for screening are described. The chapter concludes with an overview of the experiment with emphasis on the impacts of our study.

Chapter 3 provides a theoretical description of NMR spectroscopy. This includes a description of the magnetization, the process of relaxation, and the nuclear overhauser effect (NOE). The 1-D NOESY pulse sequence used in our experiments exploits the NOE and will also be described.

Chapter 4 presents an overview of the measurements and analytical techniques often used in metabolomics experiments. NMR spectroscopy and mass spectrometry are described. Most studies rely on chemometrics approached to analyze the data. This chapter covers PCA and PLS-DA, including a mathematical description of each. The chapter closes with an overview of the cross validation techniques used to evaluate PCA and PLS-DA models as well as how to interpret the results obtained from them.

Chapter 5 describes ROC curve analysis is greater detail. The chapter focuses on developing an understanding of how to make an ROC curve and how diagnostic information may be extracted from the curve. The area under the ROC curve will be described in terms of clinical significance. The methods used to determine the optimal decision threshold are also described.

Chapter 6 covers the experimental techniques used in this study in detail. It opens with a description of the animal model, urine collection techniques and proper sample handling. The NMR analysis methodology is presented. This includes the timings used in the pulse sequence and optimizing the magnetic field shims. The Chenomx software is then described in detail. The techniques used for data processing and analysis are described.

Chapter 7 presents the results of the experiment. Emphasis is placed on the addition of ROC analysis to the analysis. These results are presented for both univariate and multivariate statistical analysis. The AUC and the optimal decision thresholds are displayed for the metabolite concentrations and the metabolite ratios.

Chapter 2
Screening for Cancer

2.1 Screening

Screening is defined as the use of simple tests across a population to identify individuals who have a disease before they become symptomatic (World Health Organization a; Cook, 2008; Zivian and Gershater, 2008). The development and implementation of screening tools for cancer is particularly important for two principal reasons: first, it has been estimated that that one in two men and one in three women will develop cancer in their lifetime (Eyre, 2009; Welch and Black, 2010) and second, the stage at which a tumour is detected serves as an important prognostic indicator for patient outcome (de Nooijer et al., 2001; Feng et al., 2009; Smith et al., 2009; Taplin et al., 2004). Figure 2.1 shows the 5-year survival rates for four types of cancers. The examples of pancreas, lung and ovary, are cancer which are often diagnosed at a later stage, and suffer from poor survival rates.

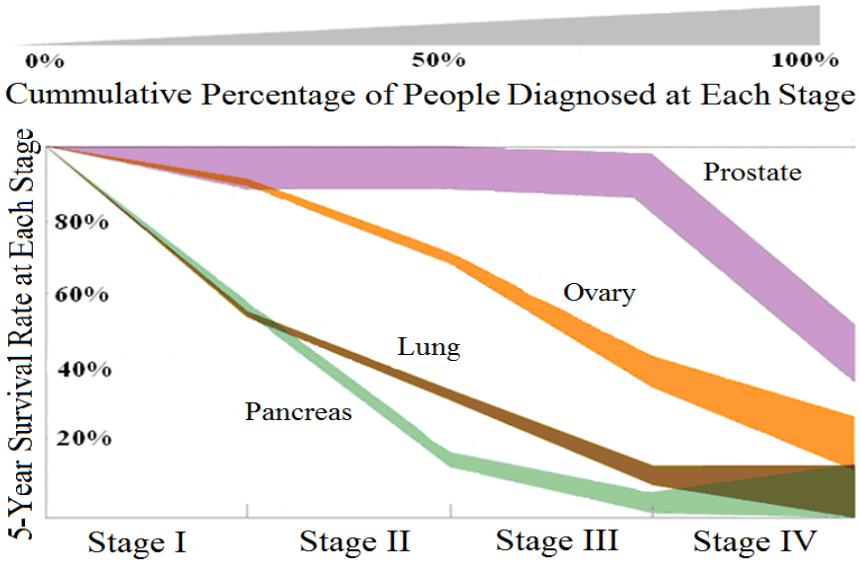

Figure 2.1: 5-year survival rates for four types of cancers. The thickness of the curves relates the percentage of tumour diagnosed at each stage as defined by the gray scale. Pancreas, Lung and Ovarian cancers are examples of cancers which are often diagnosed at a later stage and suffer from poor survival rates. This motivates the need to develop a screening technique for early detection. Permission granted for use of figure by Mr. Thomas Goetz, deputy editor for Wired Magazine and author of article: "Why Early Detection is the Best Way to Beat Cancer", Wired Magazine, 2008, Vol. 17.01

Effective screening programs require a test which is highly sensitive and specific for the disease in question (Cady and Michaelson, 2001), has a gold standard to verify the presence of the disease and has access to treatment (Taplin et al., 2004). The screening test should be safe, reliable and inexpensive (World Health Organization b). Further, the test should be easily accessible to the total population and screen for a disease with high prevalence (Cady and Michaelson, 2001). Taplin et al. (2004) and Cady and Michaelson (2001) have both suggested that non-compliance can be an issue for screening, therefore it is also important to develop a test that is non-invasive and requires minimal time and effort from the patient.

It has been suggested that tumours having 10^8-10^9 cells (approximately weighing 1 g) are often difficult to detect with current methods (Eyre 2009). Detection at this stage would result in improvements in the efficacy of cancer therapies as the tumour has generally not metastasized. Tumours with 10^{12} cells (about 1 kg of tumour) are often associated with death. Patients will often not seek medical expertise for cancers until clinical signs and symptoms present (Welch and Black, 2010). The sizes of tumours at various stages are shown in figure 2.2. As shown, most tumours are detected and treated beyond the point of small, easily treated tumours (10^8-10^9 tumour cells).

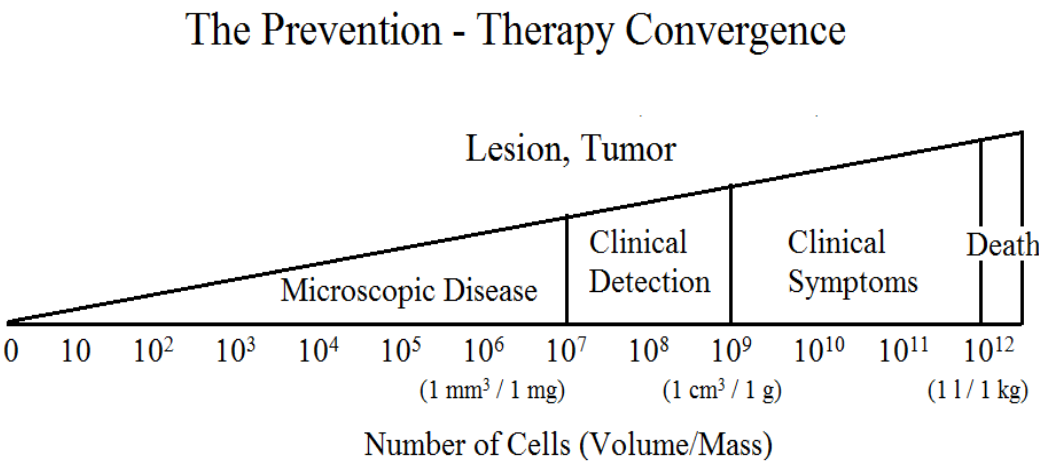

Figure 2.2 Clinical and molecular model of the convergence of invasive cancer and pre-invasive disease, and of cancer therapy and prevention. Permission granted to use figure in thesis from Elsevier. Figure found in: Eyre H 2009 Winning the Cancer Fight: A Look at the Future *Primary Care: Clinics in Office Practice* **36** 859-866.

2.2 Current Methods of Cancer Screening

Current screening techniques include mammography for breast cancer (Cady and Michaelson, 2001), cytology screening for cervical (Smith et al., 2002) and bladder cancer (Issaq et al., 2008), and the prostate-specific antigen (PSA) test for prostate cancer (Isbarn et al., 2009; Spahn et al., 2010). These methods attempt to find evidence of small, but developing tumours before clinical symptoms present (Cook, 2008). The American Cancer Society (ACS) publishes a report summarizing their recommendations for cancer screening each year (Smith et al., 2002; Smith et al., 2009). The annual reports do not contain information for all cancers, but will describe in great detail changes from the previous reports. A brief summary of the three mentioned screening techniques follow.

Breast cancer is the most common cancer in women and is the second leading cause of death from cancers in women (Smith et al., 2009). Mammography is the main imaging modality used to detect breast cancers at an early stage and is thought to decrease mortality rates by 30-40% (Alvarenga et al., 2010; Taplin et al., 2009; van den Biggelaar et al., 2009). Other benefits of mammography include increased survival rates, requirement for less aggressive surgical procedures and adjuvant therapies and an increase in the number of treatment options available (Smith et al., 2009). Cady and Michaelson (2001) suggest that mammography screening will detect smaller sizes of invasive cancers which are often accompanied with patients having low grade histology. This results in significant improvements in outcome. The sensitivity of mammography is predicted to be 70-80% (Nam et al., 2009). However, the accuracy of the test is dependent on the composition of the mammary parenchyma and tumour tissue characteristics (Alvenga et al., 2010). The complicated radiographic structure of breast tissue and the subtle characteristics of early stage tumours make screening more difficult (Zivian and Gershater, 2008). Despite these challenges, the risk of late stage cancer is greater in women who have not had a mammography (Taplin et al., 2009).

Mammography is able to detect cancers that are 0.5-1.0 cm in size in women with non-dense breasts, but tumours are rarely detected smaller than this (Eyre, 2009). Despite breast cancer detection rates increasing by 4.6-15 % (Destounis, 2004; Taplin et al., 2009), a large proportion of detectable lesions are missed – even those at late stage (Taplin et al., 2009; van den Biggelaar et al., 2009) – and not all detected cases are associated with a good prognosis (Smith et al., 2009). It is predicted that approximately

25-30% of visible cancers are missed when mammogram images are read by a single radiologist (Zivian and Gershater, 2008). This could be a consequence of the low incidence of breast cancer lesions observed in mammograms (~ 2.6 lesions in 1000). Newer developments have focused on computer-aided detection (CAD) (Zivian and Gershater, 2008) and breast ultrasound imaging (Alvenga et al., 2010).

Cervical cancers are often detected through cytology testing or papanicolaou (Pap) smears (Smith et al., 2009). Cytology specimens are collected from the entire anal canal using a Dacron swab (Bean and Chhieng, 2010). The cells collected are fixed on a slide for a conventional smear on in a liquid medium for liquid-based cytology (Bean and Chhieng, 2010). The PAP test has been highly effective at screening for pre-invasive cancers (Gustafsson et al., 1997), decreasing the incidence and mortality rates from cervical cancer in Canada between 1981 and 2002 by 39% and 53%, respectively (Lofters et al., 2010), and decreased incidence rates in the USA from 35-40 / 100,000 to 8 / 100,00 (Bean and Chhieng, 2010). This test is capable of reliably detecting precancerous changes with 10^4-10^6 tumour cells (Eyre, 2009), but it is invasive, painful and costly (Issaq et al., 2008). The sensitivity and specificity of a cervical cytology test is 75% and 90% (Bean and Chhieng, 2010).

Screening for prostate cancer is performed via the prostate-specific antigen (PSA) test or a digital rectal examination (Goldsmith et al., 2020; Smith et al., 2002). These tests, however, are not sufficiently specific and have accuracies of 25-50% (Goldsmith et al., 2010). Goldsmith et al. (2010) estimated that PSA testing has sensitivities and specificities in the range of 60-75%. In addition, prostate cancer patients with a PSA value greater than the threshold 20 ng/ml come from a highly heterogeneous group; therefore an elevated PSA is not sufficiently specific to distinguish high risk patients (Sphan et al., 2010). It is essential to discover a better screening technique as the treatment can cause moderate to substantial harms (Smith et al. 2009).

Current screening methods often rely on imaging or pathological specimens from cytology or biopsy (Eyre, 2009). The tests can be invasive, uncomfortable for the patient, and may suffer non-compliance; this deems them unsatisfactory for screening (Wen et al., 2010). In fact, problems with detection account for 40% of late stage cancers in the clinic (Taplin et al., 2004). The 5-year survival rates for oral cancers is around 50% and has not improved over the last 30 years as the tumours are often discovered in late stage. Ninety-eight percent of pancreatic cancers result in death due to the lack of treatment

available and the increased rates of metastases (Sugimoto et al., 2010). These examples emphasize the need for better techniques for screening.

Studies of metabolism hold potential to serve as screening tools (Altmaier et al., 2008; Gowda et al., 2008; Holmes et al., 2000; Griffin and Shockcor, 2004). To date, studies have successfully identified individuals with Alzheimer's disease (Barba et al., 2008; Tukianinen et al., 2008), Duchenne muscular dystrophy (Griffin et al., 2001), myocardial ischemia (Sabotine et al., 2005), type 2 diabetes (Selek et al., 2007), inborn errors of metabolism (Shlomi et al., 2009) and a number of cancers (Claudino et al., 2007) in a laboratory setting.

2.3 Potential of Metabolomics for Screening

Metabolism may be used for early diagnosis, or for real-time monitoring of the effects of a disease (Kim and Maruvada, 2008; Kim et al., 2009). The idea is that the comprehensive analysis of the metabolic content of biofluids - such as urine, serum or cerebral spinal fluid (Beckonert et al., 2007) – or from a tissue extract, will provide information about an individual's current health status (Ala-Korpela, 2008; Bollard et al., 2001; Kaddurah-Daouk et al., 2008; Nordstrom and Lewensohn, 2010). It has been well established that the metabolic content of a sample will be altered in the presence of a disease or toxic effect (Bollard et al., 2005). By analysing these changes, it may be possible to identify the presence of a disease (Saude et al., 2007; Spratlin et al., 2009; Vinayavekhin et al., 2010).

The field of metabolomics emerged in the 1980's to study changes in the concentrations of small endogenous molecules, known as metabolites, and relate these changes to a state of disease or toxicity (Nicholson et al., 2007). A metabolite is defined as any chemical compound involved in, or a product of, metabolism (Lawrence, 2005). These consist of amino acids, oligopeptides, sugars, bile acids and simple fatty acids (Clarke and Haselden, 2008). The complement of all low-molecular weight metabolites found in a biofluid or tissue, which takes part in metabolic reactions, constitutes the metabolome (Kaddurah-Daouk et al., 2008). It is thought to cover 7-9 orders of magnitude of concentration (pmol-mmol) (Dunn and Ellis, 2005). Metabolomics is the comprehensive analysis of all metabolites in an organism, and is used in conjunction with pattern recognition techniques and bioinformatics to detect and follow endogenous

metabolites in biofluids or tissues (Spratlin et al., 2009). In 1999, Nicholson et al. coined the term 'metabonomics' which is the "the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiologic stimuli or genetic modification".

The human body contains a complex and dynamic array of interacting metabolic pathways. In a healthy individual, metabolic processes are kept in balance (homeostasis) (Saude et al., 2007). When a toxic stress is present, whether it is a drug or a disease, cells attempt to maintain homeostasis and metabolic control by altering the composition of the biofluids that perfuse them or are secreted by them (Goldsmith et al., 2010; Holmes and Antti, 2002; Lindon et al., 2003; Nordstrom and Lewensohn, 2010). Metabolites in cells, tissues and biofluids are in dynamic equilibrium, so abnormal processes will be indicated by alterations in the metabolic content (Clarke and Haselden, 2008; Lindon et al., 2001; Maher et al., 2007). By comparing the metabolic content of biofluids between populations of healthy and diseased individuals, it is possible to identify patterns of metabolic behaviour which may be indicative of that disease (Goldsmith et al., 2008; Kim et al., 2009). The toxicological response is defined as the deviation from the control or baseline levels (Bollard et al., 2004). A disease process generates a unique and characteristic signature of metabolic changes which is known as the metabolic phenotype (Holmes and Antti, 2002; Lindon et al., 2001; Serkova et al., 2008). In such studies, a certain degree of homeostasis in the healthy population is assumed (Saude et al., 2007). The metabolome is dynamic, and will respond to stimuli within seconds (Dunn and Ellis, 2006; Kim and Maruvada, 2008; Kim et al., 2008; Weckwerth and Morgenthal, 2005). This makes metabonomics studies good candidates for screening.

Metabonomic studies have shown that external influences can have profound impacts on the metabolic content of biofluids from healthy individuals (Bollard et al., 2005). It is therefore important to follow a standard procedure in terms of the time of day a sample is collected, handling of samples prior to analysis, dietary control, etc. Early studies suggest that the inter-individual differences were larger than the intra-individual differences (Ala-Korpela, 2008; Lens et al., 2003; Kim et al. 2009). However, a more recent study showed that the metabolic content from a single individual can vary greatly over time and even span the range of the entire group (Saude et al., 2007). Despite the magnitude of these changes on an individual basis, studies have found that alterations in response to a disease or toxicological process are detectable (Bollard et al., 2005).

Natural effects on metabolism include diurnal variation, differences between males and females, aging, hormonal effects, diet and changes in the gut microflora (Bollard et al., 2005; Slupsky et al., 2007). Metabolic activity changes throughout the day depending on wake and sleep patterns, which alters the composition and volume of urine excreted. It has been suggested to collect all samples around the same time of day to minimize diurnal effects (Bollard et al., 2005; Lens et al., 2003). Bollard et al., (2005) reviewed metabolic changes associated with gender and age. The results suggest that females experience stronger pharmacological effects than males, likely as a result of lower metabolic activity. Aging animals undergo a number of physiological changes, both physical and biochemical, which results in different amounts of endogenous metabolites being excreted into the urine. Wilson and Nicholson (2007) found that the microbiome in experimental animals was stable and resistant to change post weaning. In general, inter-animal variation is more pronounced than diurnal alterations, which is greater than hormonal effects (Bollard et al., 2005). Also, inter-subject variation exceeds that from the measurement technique used (Dunn and Ellis, 2006; Slupsky et al., 2007). Despite all these changes, Slupsky et al. (2007) found that the analysis of metabolite concentrations in normal urine samples using targeted profiling (matching mathematically modelled pure metabolite peaks to the NMR spectra) produced consistent and reliable results. Variations in experimental mice are caused by the animals' well-being, genetic drift, strain, hormonal differences, the metabolic rate, age and gender. External influences that must be minimized include food and water intake, temperature, light intensity and subjection to stress (Bollard et al., 2005). These influences are reviewed in greater detail in Bollard et al., 2005 and Slupsky et al., 2007.

Metabolomics presents a good screening tool for tumours that readily produce or excrete accessible fluids (Spratlin et al., 2009). Cancer cells possess highly unique metabolic phenotypes characterized by increased glucose uptake (Griffin, 2006; Serkova et al., 2008; Spratlin et al., 2009) and increased lactate production in aerobic conditions (Vizan et al., 2008). Griffin and Kauppinen (2006) claim that metabolic profiles in various tumour preparations show correlations with tumour type, proliferation rates, metabolic activity and cell death. Similarly, Issaq et al., (2008) suggests that metabonomics can provide metabolic patterns characteristic of various benign or malignant conditions. Cell proliferation rates are generally greater in cancers than in normal tissue, thus yielding higher rates of glucose utilization (Griffin and Shockcor, 2006) and increased concentrations of excreted modified nucleosides (Frickenschmidt et

al., 2008). Alterations in the total choline metabolites (tcho) has been used to characterize cases of breast and prostate cancers (Spratlin et al., 2009) and may be linked with malignant cell growth (Griffin, 2004). Alternatively, lactate has been proposed as a diagnostic tool for identifying cases of cancers in patients. Vizan et al., (2008) discovered that increases in lactate levels could identify late stage tumour progression, especially in cases with tumour invasion.

To date, most metabonomics studies have dealt with animals (Lenz et al., 2003), however studies on human populations are now more common (Erb et al., 2008; Guan eta al., 2009; Issaq et al., 2008; Lenz et al., 2003; Odunsi et al., 2005; Woo et al., 2009). Until recently, studies suggested that metabolic information in animals was more stable, thus allowing for increased confidence in the observed changes. Saude et al., (2007) found that the metabolite concentrations in a controlled guinea pig population were similar to that found in a cohort of healthy humans. As long as the animals are given time to acclimatize to the experimental environment (about three days), the metabolome should be stable (Bollard et al., 2005; Lenz and Wilson, 2006). In fact, all animals are known to respond in a similar fashion to humans (Bollard et al., 2005). The techniques used to study metabolism in animal models have been applied to study human cancers (Griffin and Kauppinen, 2006).

2.4 Examples of Metabolomics Studies for Cancer Screening

Metabonomics studies have shown great potential in the field of cancer diagnosis (Claudino et al., 2007). To date, studies have investigated metabolic changes associated with biliary tract cancer (Wen et al., 2010), bladder cancer (Issaq et al., 2008), brain metastases (Simoes et al., 2008), breast cancer (Giskeodegard et al., 2010; Sitter et al., 2010; Whitehead et al., 2005; Woo et al., 2009), cervical cancer (Woo et al., 2009), colorectal cancer (Feng et al., 2009; Ma et al., 2009), liver cancer (Chen et al., 2009), male pancreatic cancer (Beger et al., 2006), oral cancer (Tiziani et al., 2009), ovarian cancer (Guan et al., 2009; Odunsi et al., 2005) and prostate cancer (Jordan and Cheng, 2007; Lokhov et al., 2010). When analysing the metabolic content of urine samples, the test will be non-invasive and easy to perform. A few examples of applications which use metabolomics analysis for cancer detection are reviewed below.

2.4.1 Detection of Malignant Tumour from the Methyl and Methylene Line-widths

The first study using NMR metabolomics for cancer identification was performed by Fossel et al. in 1986. In this study, blood samples were collected from patients with malignant or benign tumours, healthy controls and pregnant women. NMR analysis was carried out at 360 MHz or 400 MHz with a Bruker AM Fourier-transform Spectrometer. Water suppression was achieved using presaturation at the water resonance frequency. The full widths at half height of the methyl and methylene resonances were measured without knowledge of the patients' health status. The results suggested that the average line widths were narrower in patients with cancer, decreasing from 39.5 ± 1.6 Hz for the healthy controls to 36.1 ± 2.6 Hz for patients with malignant tumour, 36.7 ± 2.0 Hz for patients with benign tumours (breast, ovary, uterus and colon cancers), and 29.9 ± 2.5 Hz for patients with untreated cancers. In addition, they found that the line widths obtained from pregnant patients were consistent with the presence of malignant tumour. The study suggested that line widths less than 33 Hz were indicative of the presence of cancer and is independent of the type of cancer.

2.4.2 Epithelial Ovarian Cancer

Odunsi et al., (2005) studied the metabolic changes associated with the presence of epithelial ovarian cancer (EOC). In this study, pre-operative serum specimens were obtained from 38 patients with EOC, 12 patients with benign ovarian cysts and 53 healthy women (21 premenopausal and 32 postmenopausal). ^{1}H NMR Spectroscopic analysis was performed on a 600 MHz spectrometer. Principal component analysis (PCA) of the EOC patients and controls revealed clear separation between the two groups - those with EOC and healthy women – and clustering of the samples in these groups. One model correctly predicted all 38 cancer specimens and 21 premenopausal individuals and a second model correctly separated 37 of the 38 cancer specimens and 31 of the 32 postmenopausal women. They were able to differentiate between women in each of the three categories. They applied Receiver operating characteristic (ROC) curve analysis to the data to evaluate their utility to predict EOC. This analysis showed that by utilizing information from two regions – descriptors at 4.03 ppm and 4.77 ppm –excellent discrimination for EOC is attained (Area under the ROC curve (AUC) = 1.00). Alone, these regions had an AUC of 0.942 and 0.689, respectively.

2.4.3 Breast Cancer

Whitehead et al., (2005) performed ^{1}H NMR analysis on serum samples obtained from 10 mice inoculated with a highly-metastatic mammary carcinoma cell line, 10 mice inoculated with a normally metastatic mammary carcinoma cell line, and 10 healthy controls. Analysis was performed on the linewidths of the composite methyl and methylene peaks. Their results suggest that the average linewidths can be used to distinguish sera from the highly-metastatic and normal-metastatic tumour-bearing animals from the healthy controls with p-values of 0.0002 and 0.0003, respectively. The broadened methylene and narrow methyl resonance were consistent with other studies involving cancerous sera due to abnormal distributions of plasma lipoproteins.

They also applied PCA to the data set to determine if unique metabolic patterns for the different categories of metastatic breast cancer in serum existed. The model correctly separated sera of the tumour-bearing mice from those from the normal controls and was able to distinguish between classes of varying metastatic ability. The regions responsible for the separation were different for the highly-metastic and normally-metastic cases when compared with the control mice. They identified lactate, taurine, choline and sugar moieties as the variables responsible for the separation.

2.4.4 Bladder Cancer

Issaq et al., (2008) analysed the metabolic content of urine samples collected from 41 patients with transitional cell carcinoma and 48 healthy individuals. The patient populations were approximately age-matched, with ages ranging between 51-93 for those with baldder cancer and 20-86 for the healthy controls. Samples were analyzed using high performance liquid chromatography coupled online with a hybrid triple-quad time-of-flight mass spectrometer. The data was subjected to both PCA and orthogonal partial least squares-discriminant analysis (OPLS-DA). PCA correctly predicted 46 of 48 healthy individuals and 40 of 41 bladder cancer samples, while the OPLS-DA model was able to correctly predict 48 of 48 healthy individuals and 41 of 41 bladder cancer urine samples.

2.4.5 Oral Cancer

Tiziani et al., (2009) investigated the potential of ^{1}H NMR metabolomics to diagnose early stage oral squamous cell carcinoma (OSCC). Samples of venous blood were collected from 15 patients with confirmed OSCC (average age 63.8 years) and 10 control samples from normal donors (average age 57 years) who had no personal history of cancer. All samples were collected at approximately the same time during the day.

The group obtained 1-D ^{1}H and 2-D ^{1}H J-resolved NMR spectra using a 500 MHz spectrometer equipped with a cryogenically cooled probe and a 2D COSY experiment was carried out on a 800 MHz Varian spectrometer with a cryogenically cooled probe.

Both PCA and partial least squares discriminant analysis (PLS-DA) were used for statistical data analysis. PCA of the 1-D data shows clear separation between samples from patients with OSCC and healthy individuals and revealed clustering according to the stage of the cancer. The PLS-DA model showed excellent separation between the populations with sensitivity and specificity values greater than 95%. The loading plots showed that levels of valine, ethanol, lactate, alanine, acetate, citrate, phenylalanine, tyrosine, methanol, formaldehyde and formic acid were reduced in the OSCC patients, while glucose, phyruvate, acetone, acetoacetate, 3-hydroxybutyrate and 2-hydorxybutyrate, choline and betaine had enhanced loadings. PLS-DA was also applied to only the data from diseased patients. The loadings plots revealed similar patterns of distinction between the different stages. The increased levels of glucose and reduced levels of lactate are different from most studies involving cancer. They propose that the increase in glucose results from a unique behaviour of oral cancers to interfere with the ability of insulin to modulate the uptake of glucose.

2.4.6 Prostate Cancer

Lokhov et al., (2010) investigated the effects of prostate cancer on the metabolic content of blood plasma. In this study, samples were collected from 30 healthy men and 40 men who had been diagnosed with stage two prostate cancers. Mass spectrometry was carried out to measure the concentrations of positive and negative ions. The two sided Wilcoxon rank sum test was used to define the difference of the peak intensities between the healthy and diseased populations.

Six metabolites showed a statistically significant shift in intensity in the cancerous populations. The effectiveness of these metabolites in a diagnostic situation was assessed using ROC curve analysis. The area under the curve (AUC) was used to indicate the efficiency and clinical applicability of the diagnostic method. They considered curves with an AUC > 0.6 as being clinically applicable and an AUC > 0.8 to indicate a good diagnostic measure. Two metabolites, dimethylhepatonoyl carnitine and arachidonoyl amine, had an AUC > 0.8, with AUC's of 0.97 and 0.86 respectively. These were compared with the conventional screening technique of PSA (prostate specific antigen) which had an AUC of 0.59 for the same sample population.

2.5 Use of Urine in Metabonomics Studies

Urine is an easily accessible, abundant biofluid which contains thousands of metabolites (Goldsmith et al., 2010; Lindon et al., 2006). Samples may be collected non-invasively (Lenz and Wilson, 2006) and multiple samples can be collected without harm to the patient (Saude et al., 2007; Wilson and Nicholson, 2007). This means that samples can be collected from animals before and after inoculation of a disease so that the toxic effect may be monitored on an individual basis (Bollard et al., 2001). Samples may be pooled to average out diurnal influences or day-to-day variation (Lenz and Wilson, 2006). Further, urine provides high resolution NMR spectra due to the content of many small aqueous metabolites and the non-viscous nature of the fluid (Reo, 2002). As a fluid, it is isotropic and magnetically homogenous (Lindon and Wilson, 1989).

Urine provides an integrated view of all metabolic activities in the individual (Lindon et al., 2006; Viant et al., 2007). As a result, it serves as a 'sums of history' of all disease processes in the individual (Griffiths, 2007). The excretatory nature of urine allows it to be used for time-related studies of disease response as it describes what has happened in the individual since the last sample was acquired (Viant et al., 2007; Wilson and Nicholson, 2007). Early markers for disease response are easily identifiable in urine since the kidney's function is to maintain homeostasis (Griffiths and Stubbs, 2007; Wilson and Nicholson, 2007).

Urine samples require minimal preparation; in most cases, the procedure includes pH adjustments and the addition of an internal standard for analysis (Serkova et al., 2008). The pH of urine ranges between 5 and 9, but is generally within the range of 6.5-

7.5 (Nicholson and Wilson, 1989). Most groups will add phosphate buffered saline to bring the pH to a specific value and stabilize the pH (Lindon et al., 2003). Most groups choose a pH around 6.8 (Kim et al., 2008 ; Murdoch et al., 2008; Um et al., 2009) or 7.0 (Jukaininen et al., 2008; Parsons et al., 2007; Troy et al., 2007; Wiljie et al., 2006). Some groups filter samples to remove proteins and macromolecules (Saude and Sykes, 2006; Viant et al., 2007). This step has been shown to halt further metabolic changes.

Samples should be frozen until required for analysis to prevent breakdown due to bacterial contamination (Lauridsen et al., 2007; Viant et al., 2007). Contamination may be observed through decreases in the levels of creatinine and increased creatine (Saude and Sykes, 2006) or increased concentration of acetate and formate (Lenz and Wilson, 2006). Lauidensen et al., (2007) studied the effects of a number of methods for sample storage. He stored samples at 4^{o}C, -25^{o}C and -80^{o}C for variable times ranging from 0 to 26 weeks. An increase in acetate and decreases in citrate and hippurate was observed in the samples without a preservative one week post collection. Samples with an added preservative were stable for a longer period. Samples stored at -25^{o}C and -80^{o}C showed no differences, with and without the added preservative for up to 26 weeks post injection. They recommend that samples be stored at -25^{o}C, but most groups choose to store the samples at -80^{o}C (Holmes et al., 1999; Miyataka et al., 2007; Pohjanen et al., 2006). There are studies (Nordstrom and Lewensohn, 2010; Saude and Sykes, 2006) that have indicated a loss of metabolites in samples stored for long periods of time, however, other studies suggest that the samples are not affected for up to 9 months (Ross et al., 2007). Maher et al., (2007) suggests that urine may be stored at room temperature for 24 hours, which is sufficient for analysis.

2.6 Motivation for our Experiment

The aim of our experiment is to show that metabolomics data has potential to serve as a screening test for cancer at an early stage. To achieve this goal, urine samples are collected from animals bearing a human glioblastoma (GBM) xenograft. Sample collection began one week before cell injection for the purpose of defining a normal baseline, while post cell-injection samples were collected for 5 weeks to observe metabolic changes associated with the presence of a tumour.

Studies have suggested that external sources of variation can mask the metabolic changes associated with tumour development, therefore a group of control animals are examined to identify naturally occurring metabolic changes or alterations that results as a response to the stress of the injection procedure. The control animals are treated the same as the tumour-bearing animals in terms of the environment in which they are housed, the method of urine collection and age of animals throughout the collection procedure. The idea is that by compensating for the metabolic changes in the control population, it may be possible to isolate metabolic changes resulting from only the tumour.

Metabolomics analyses are either performed between a group of healthy individuals and a group of individuals with established disease, or used to study the metabolic changes temporally as individuals transition from healthy to having a disease. Studies of disease in human populations examine changes between a healthy and diseased group. Animal studies have the advantage that an individual may serve as its own control. This reduces inter-individual differences in the data and enables the investigation of subtle metabolic changes associated with early disease progression. Animal studies that compare the metabolic changes on an individual basis often do not perform a parallel analysis of control animals (Keun et al., 2004; Pohjanen et al., 2006; Um et al., 2009) or do not account for the metabolic changes observed in the controls in their further analysis (Kim et al., 2008; Lenz et al., 2004; Tyburski et al., 2009).

Principal component analysis and partial least squares discriminant analysis are common multivariate statistical tools for analysis. Models are constructed with the healthy and diseased populations. Most groups only qualitatively suggest that the model is effective at distinguishing individuals from each population. However, some groups will calculate the sensitivity and specificity for a line drawn manually through the data. If the goal of the analysis is to identify metabolites which may be used to distinguish individuals in each group, univariate statistical analysis is performed. To our knowledge, metabolites with $p<0.05$ are considered significant; no correction is made for testing multiple metabolites.

We propose to track changes on an individual basis by pairing the data before and after cell injection. We envision that this could mirror a human application of this technique in which an individual could submit a urine sample for analysis one or more time per year. Once a “normal” metabolic trajectory is identified, deviation from this path could be used to identify the presence of a disease like cancer.

Univariate analysis, in the form of the non-parametric Wilcoxon rank sum test, will prove that metabolic changes are taking place. Statistical significance will be defined by the Bonferroni correction limit and the false discovery rate (FDR) methodology. The goal of this analysis is to show that useful metabolic information is available using both a conservative compensation (Bonferroni correction) and a newer technique which controls the number of false positive discoveries (FDR method).

The analysis will first consider all post-injection data as one group to confirm that quantifiable metabolic changes are observed in the presence of tumour. After these metabolic changes were established, the tumour-bearing data is divided into early and late stage data to evaluate the ability of our techniques to diagnose cancer at the two stages. It is hypothesized that the pre-injection and early stage data can be distinguished and that the early and late stage data are altered in a similar fashion.

There is no standard method for quantifying the degree of separation between pre-defined classes (i.e. pre-injection and post-injection with cancer) in a metabolomics data set. Recently, some groups have begun to apply ROC analysis to the distributions of the pre and post-injection data (Lokhov et al., 2010; Odunsi et al., 2005) or to statistical models (Sugimoto et al., 2010).

The nature of multivariate statistical analysis is to determine the maximum variation in the data set. It is expected that the maximum variation will be in the direction of class discrimination. However, no method is in place to assess the degree of separation. We propose to apply ROC analysis to the component axes initially, and then rotate the data from a pair of scores vectors in unit angle increments about the origin to determine the angle at which the model is most effective at distinguishing between classes. This analysis will be applied to the pre and post-injection tumour bearing data to simulate a environment for screening, between the post-injection data from the control and tumour-bearing mice to simulate a human study using a age-matched group of health individuals and those who are known to have the disease, and between all ‘tumour-free’ (or ‘healthy’) and tumour bearing samples.

As a final analysis, the tumour-bearing data will be separated into early and late stage data to evaluate the potential of metabolomics for the early identification of tumour. A PLS-DA model will be used between the pre-injection and late stage data. The early stage data will be projected onto the model. ROC analysis will provide a measure of the clinical significance of the metabolic changes.

Chapter 3
Nuclear Magnetic Resonance Spectroscopy Theory

This Chapter overviews the theory related to Nuclear Magnetic Resonance (NMR) Spectroscopy. The Nuclear Overhauser Effect (NOE) and the 1-D NOESY pulse sequence will be discussed briefly. The bulk of information in this chapter was taken from Claridge, 2009 (sections 3.1 - 3.3), Jordan and Cheng, 2008 (sections 3.1 and 3.2), Ross et al., 2007 (sections 3.2 and 3.3) and McKay, 2009 (section 3.3).

3.1 History of Nuclear Magnetic Resonance Spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy in a condensed state of matter was detected independently by Felix Bloch and Edward Purcell in 1945. In the 1950's, scientists discovered that the resonant frequency of nuclei was dependent on its local chemical environment and that nuclei could influence the resonance of other nuclei through intervening chemical bonds. Computers were added to NMR systems in the 1960's. Their addition, in combination with stronger magnets, allowed for significant improvements in sensitivity and analysis of smaller samples. Computers collected a time-domain signal. The Fourier Transform was applied to reveal information about the complement of nuclear spins found in the sample. The mid-1960's brought about the Nuclear Overhausser Effect (NOE); this technique has become the most widely used technique for sensitivity enhancements between nuclei and for studying the spatial proximity of nuclei. The introduction of the superconducting magnet in the 1970's initiated the development of higher magnetic field strengths. Today it is common to find magnets with cryogenically cooled probes which incorporate active shielding to keep stray fields closer to the magnet. NMR Spectroscopy has since found widespread application in chemistry, biology, medicine, materials science and geology.

3.2 Theory of Nuclear Magnetic Resonance (NMR)

3.2.1 Magnetization

Atomic nuclei – protons and neutrons– possess an innate property known as the nuclear spin quantum number, *I* (Pochapsky and Pochapsky, 2006; Ross et al., 2007). Molecules with an odd number of neutrons or protons will have a net spin and contribute to the NMR signal. The spin quantum numbers of these molecules are multiples of ½: i.e. 1/2, 1, 3/2, etc. Spin ½ molecules, such as the protons in water, are most often studied with NMR Spectroscopy. The nuclear spin has $2I+1$ possible spin states. For the example of a spin ½ nucleus, there are two spin states known as α ('spin up', $I = ½$) or β ('spin down', $I = -1/2$). The following discussion will consider only protons.

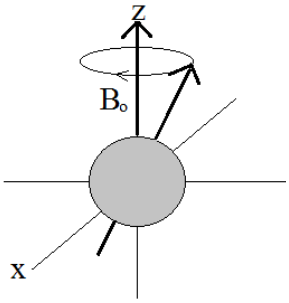

Figure 3.1: Proton in an external magnetic field. When placed in an external magnetic field, a nucleus will precess about the field at the Larmor frequency ($\omega = \gamma B_o$). The proton can be in the 'spin up' (aligned parallel to the field) state or the 'spin down' (aligned anti-parallel to the field) state. The proton is in the spin up state in this figure.

When nuclei are placed in an external, static, magnetic field, B_o, they will either align parallel (spin up) or anti-parallel (spin down) to the field. By convention, the field is aligned along the +z-axis. The actual spin of the nuclei is not aligned with the magnetic field, but presses about it (see figure 3.1). A spinning proton has an angular momentum, $I\,\hbar$, which gives rise to a magnetic moment, μ with magnitude:

$$\mu = \gamma I \hbar \qquad \text{(Equation 3.1)}$$

Where γ is the gyrometric ratio. The magnetic moment is unique to a nucleus and can be interpreted as how 'magnetic' the nucleus is (Claridge 2009).

Precession is a consequence of the applied torque from the magnetic field on the magnetic moment. The rate of precession, which is commonly referred to as the Larmor frequency is given by:

$$\omega = -\gamma B_o \, (rad/s) \qquad \text{(Equation 3.2a)}$$

$$\text{or } \upsilon = -\gamma B_o / 2\pi (Hz) \qquad \text{(Equation 3.2 b)}$$

Since the 'spin up' orientation is at a lower energy, more nuclei will be in this state. The proportion of nuclei in the spin up to the spin down spin states is described by the Boltzmann distribution:

$$\frac{N_\alpha}{N_\beta} = e^{-\frac{\Delta E}{k_B T}} \qquad \text{(Equation 3.3)}$$

Where N_α is the number of nuclei in the spin up state, N_β the number of protons in the spin down state, ΔE is the difference in energy between the two states, k_B is the Boltzmann constant (1.381×10^{-23} J/K) and T is the temperature in Kelvin.

The signal from NMR experiments, known as the magnetization, comes from the difference in the number of nuclei in the spin up and spin down orientations. Due to the statistical averaging of a large number of microscopic magnetic moments, the net magnetization may be represented as a classical macroscopic magnetic moment, **M**, which is aligned with the external magnetic field (figure 3.2). The energy difference is quite small (order of 10^{-25} J for fields of strength 18.8T (Magnetic field of a 800MHz spectrometer)), so the population differences are generally of the order of 1 part in 10^4 at 18.8T. Consequently, NMR Spectroscopy is a relatively insensitive technique when compared to mass spectrometry (Claridge 2009).

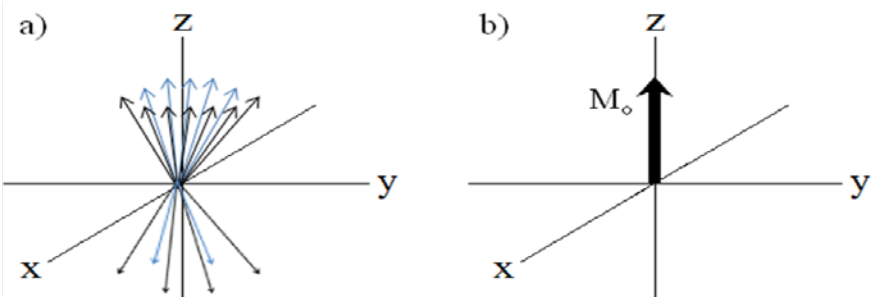

Figure 3.2: Complement of all magnetic moments and Net Magnetization. a) An NMR sample consists of a large number of resonating nuclei. The nuclei resonate at the Larmor frequency, but have different phases. Due to statistical averaging, the transverse component from one nuclei cancels another, while the components along the z-axes add. There are more nuclei in the spin up state, resulting in a net magnetization along the +z-axis (b)).

In an external magnetic field, valence electrons circulate. This creates a local diamagnetic current, which in turn generates a small magnetic field opposite to the main field (Pavia et al., 2009; Ross et al,. 2007). The net magnetic field, i.e. the sum of the main magnetic field and these small local fields, is referred to as B. In effect, the proton is 'shielded' slightly and will resonant at a lower frequency. The Larmor frequency is now dependent on B, not B_o (Jordan and Cheng, 2008; Pavia et al., 2009). These small changes in the Larmor frequency are referred to as the chemical shift, δ_i, (Pochapsky and Pochapsky, 2006) and may be calculated with the following:

$$\delta_i = \frac{\omega_i - \omega_{ref}}{\omega_{ref}} x10^6 \quad \text{(Equation 3.4)}$$

Here, ω_i is the Larmor frequency of the metabolite of interest and ω_{ref} is the Larmor frequency of a standard molecule. The units are parts per million (ppm). Common choices for the internal standard include 3-trimethylsilylpropionic acid (TPS) with the methylene groups deuterated to avoid unwanted peaks (Beckonert et al., 2007; Pavia et al., 2009) or the methyl peak of 2,2-dimethyl-2-silapentane-5-sulfanate (DSS) (Wiljie et al., 2006). The chemical shift is independent of the field strength (and hence the chemical properties) which allows for standard identification among spectrometers.

Nuclei from different molecules experience a different chemical environment. In effect, molecules have unique chemical shifts. This aids in the identification of compounds (Claudino et al., 2007; Jordan and Cheng, 2008). The fact that each compound provides a unique and characteristic spectrum makes NMR spectroscopy attractive for metabolomics experiments (Saielli and Bagno, 2009; Verpoorte et al., 2008; Wishart, 2008).

NMR occurs when a nuclear spin changes spin states after the absorption of a quantum of energy that is equal to the difference between the two spin states. The energy is supplied through a radiofrqeuncy (RF) pulse with a frequency that matches the Larmor frequency of the proton and satisfies the following:

$$\Delta E = h\upsilon = \frac{h\gamma B_o}{2\pi} \quad \text{(Equation 3.5)}$$

Where h is Planck's constant (6.26×10^{-34} Js) and ν is the frequency in Hz.

3.2.2 Excitation and Detection of the Signal

NMR is achieved by tipping the magnetization into the x-y plane by means of a time-dependent RF pulse which has a frequency that matches the Larmor frequency of the protons at the center of the spectrum. If the RF pulse is at the correct frequency, it creates a new time-dependent oscillating effective field, known as the B_1 field, in the transverse plane. By convention, we align the B_1 field with the +x-axis in the rotating reference frame. This field applies a torque to the magnetization, causing it to rotate about the +x-axis onto the x-y plane (figure 3.3).

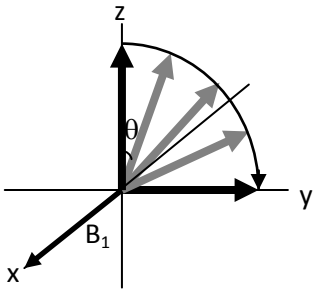

Figure 3.3: Application of an RF excitation pulse along the +x-axis of the rotating reference frame exerts a torque on the net magnetization, causing it to rotate towards the +y-axis. The angle of rotation is given by $\theta = \gamma B_1 \tau$. In most NMR experiments, the B_1 field is constant, so the angle of rotation is dependent on the duration of the pulse.

The angle (in radians) at which the net magnetization rotates (also referred to as the nutation angle), depends on the length of time the pulse is applied, τ, and is given by:

$$\theta = \gamma B_1 \tau \qquad \text{(Equation 3.6)}$$

Microscopically, the absorption of the RF energy will promote some nuclei to the spin down state (Dunn and Ellis, 2006; Goldsmith et al., 2010) and forces all magnetic moments into phase coherence (figure 3.4). Summing the magnetic moments will result in the net magnetization, which has rotated away from the +z-axis by the angle given by equation 2.6. As an example, a 90^o pulse rotates the net magnetization to the x-y plane; here the number of nuclei in the spin up and spin down states are equal and all spins point in the +y direction.

Only magnetization in the x-y plane can be detected by the receiver coils. In the laboratory frame of reference, the net magnetization is precessing about the external magnetic field at its Larmor frequency. The rotating nuclei induce an electrical current in the receiver coils which can be registered as the time-domain signal. The magnitude of the current is proportional to the net magnetization. This will be covered in more detail in section 3.2.4.

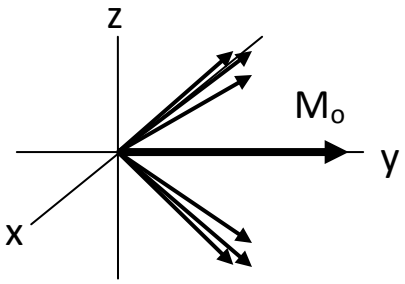

Figure 3.4: The RF pulse forces the microscopic nuclei into phase coherence. If the pulse is a 90^o pulse, the net magnetization lies along the +y-axis and equal numbers of nuclei are in the spin up and spin down states.

NMR samples are often heterogeneous, so protons experiencing different magnetic fields precess at different rates. In the rotating frame of reference, protons resonating at the Larmor frequency remain stationary along the +y-axis, while metabolites resonating at different frequencies will rotate about the z-axis (figure 3.5). Protons resonating at higher rates spin clockwise, while those resonating at lower frequencies spins counter clockwise. This process is referred to as dephasing.

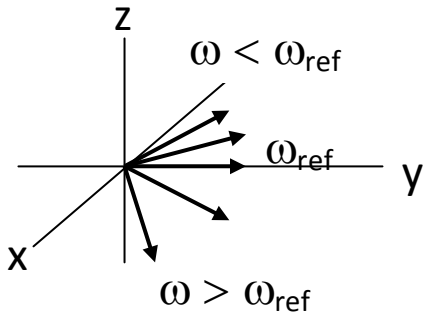

Figure 3.5: Dephasing of the Magnetization due to different Larmor frequencies. Protons experience different magnetic fields due to electron shielding. This results in nuclei precessing at different Larmor frequencies due to different effective magnetic fields. Protons that precess faster than the reference proton spin clockwise with respect to the reference signal, while protons that precess at a slower rate spin counter-clockwise. In NMR, the reference proton generally precesses at the slowest rate.

3.2.3 Relaxation

The rotating nuclei attempt to return to their ground states by emitting a quantum of energy. Macroscopically, the net magnetization will decay in the transverse plane and rebuild along the +z-axis until it re-establishes its equilibrium state. This process is known as relaxation and is presented in figure 3.6.

There are two forms of relaxation: longitudinal or spin-lattice relaxation (characterized by the time constant T_1) and transverse or spin-spin relaxation (characterized by the time constants T_2 and T_2^*). The mechanisms of relaxation are the same for both cases, but T_1 is always larger than T_2 (Pochapsky and Pochapsky, 2006).

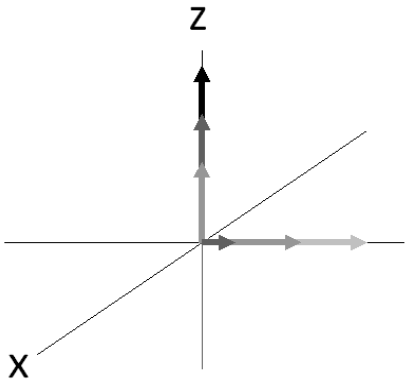

Figure 3.6: T_1 and T_2 Relaxation. This figure shows the magnetization changes over time. At time, t = 0 s, the magnetization is shown in light grey and gradually gets darker as time progresses. The transverse magnetization will decay exponentially with time and rebuild along the +z-axis. The rate at which these processes occur are dictated by the T_1 and T_2 relaxation constants. For small mobile molecules, $T_1 \sim T_2$.

Longitudinal relaxation is the first order relaxation and describes process of rebuilding the magnetization along the +z-axis. Nuclei return to the spin-up state with the release of an energy quantum to the surroundings in the form of heat. The initial magnitude of the acquired signal is proportional to the transverse magnetization. For optimal results, it is recommended to wait approximately $5T_1$ between excitation pulses to allow the magnetization to re-establish equilibrium (achieving 99.3% of the full magnetization).

Because of the small difference in energy between the spin states, nuclear spin transitions are not spontaneous. Relaxation processes require stimulation by the chaotic tumbling of nearby molecules at the proton's Larmor frequency.

According to the Bloch theory of NMR, the rate of recovery of the bulk magnetization is:

$$\frac{dM_z}{dt} = \frac{M_z - M_o}{T_1} \qquad \text{(Equation 3.7)}$$

Where M_z is the magnitude of the magnetization along the z-axis, M_o is the magnetization at thermal equilibrium, and T_1 is the first order time constant (or longitudinal relaxation time) for the process. Solving the differential equation and assuming an initial starting point of no net magnetization (after a 90^o pulse or after inserting the sample into the field), the magnetization is:

$$M_z = M_o\left(1 - e^{-\frac{t}{T_1}}\right) \qquad \text{(Equation 3.8)}$$

The transverse magnetization decays by means of transverse, or T_2, relaxation. This form of relaxation involves the exchange of energy between nearby nuclei; a so-called flip-flop process. Transverse relaxation is a consequence of inhomogeneities in the main magnetic field and differences in the local magnetic fields resulting from inter and intra-molecular interactions. The process is also referred to as spin-spin relaxation since energy is transferred between like nuclei. The nuclei permanently lose their phase coherence as a result of small differences in the local magnetic fields. The loss of phase coherence is characterized by the time constant T_2.

Alternatively, T_2^* relaxation also accounts for additional loss of phase coherence resulting from inhomogeneities in the external magnetic field. In this case, nuclei resonate at different frequencies. If the nuclei do not diffuse in the sample, this loss of phase coherence may be undone. T_2^* is always faster than T_2 relaxation as a result of additional factors. The relaxation rate is defined as:

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_2(\Delta B_o)} \qquad \text{(Equation 3.9)}$$

Where T_2^* is the net relaxation time constant, T_2 is the time constant for relaxation related to inter and intra molecular interactions and $T_2(\Delta B_o)$ is the relaxation time constant due to inhomogeneities in the main magnetic field. T_2^* approaches T_2 in well shimmed magnet fields.

The Bloch equation for the transverse magnetization in the rotating frame is:

$$\frac{dM_{xy}}{dt} = -\frac{M_{xy}}{T_2^*} \qquad \text{(Equation 3.10)}$$

The transverse magnetization is attained by solving the differential equation:

$$M_{xy} = M_o e^{-\frac{t}{T_2^*}} \qquad \text{(Equation 3.11)}$$

The widths of NMR resonances are inversely proportional to T_2^*. For small T_2^*, the frequency differential between like nuclei increases, thus leading to spreading in the frequency regime (Claridge 2009). In general, proteins and macromolecules have broad peaks, while small molecules have sharper peaks (Beckonert et al., 2007) as a result of the T_2^* relaxation time constants.

For single exponential relaxation, the lineshape is approximately Lorentzian with a half-height linewidth equal to:

$$\Delta\upsilon_{1/2} = \frac{1}{\pi T_2^*} \qquad \text{(Equation 3.12)}$$

Small, rapidly tumbling spin ½ nuclei in a low-viscosity solute, $T_1 \sim T_2 \sim T_2^*$, which are on the order of 1-3s. In contrast, larger, slower tumbling molecules have T_1 times that are much larger than T_2. Macromolecules often have very short T_2 times (order of ms), resulting in broader NMR peaks.

3.2.3.1 Mechanisms for Relaxation

There are four main mechanisms by which stimulated emission occurs: the dipole-dipole interactions, chemical shift anisotropy (CSA), spin rotation and quadrupolar mechanisms. For small, spin ½ nuclei, dipole-dipole interactions are the dominant mechanism for relaxation. Such interactions are important for signal enhancement of urinary metabolites when using the 1-D NOESY pulse sequence.

In dipole-dipole interactions, the magnetic fields from two nuclei interact as they near one another. These interactions include attraction and repulsion, depending on the relative orientation of the dipoles. In single, slowly tumbling nuclei, the dipolar orientations are invariant, but their relative positions in space change with time. As the molecule tumbles, the local magnetic fields experienced by one nucleus by another changes. The nuclei will relax if the molecule is tumbling at the appropriate frequency. Protons in a dilute solution tend to relax slower due to the limited numbers of neighbouring magnetic dipoles. In effect, these protons tend to saturate and will not fully recover before the next excitation pulse. This leads to smaller spectral areas. Alternatively, a proton may relax by interacting with an unpaired electron. The magnetic moment for electrons is over 600 times greater than that for a proton, so this mechanism for relaxation is highly effective and is often referred to as paramagnetic relaxation.

The remaining three processes are not important for our application and will only be described briefly here. Chemical shift anisotropy results from an unsymmetrical, or anisotropic, electron distribution in the chemical bonds. The local field experienced by the nucleus is dependent on the orientation of the bond with respect to the applied static

field and will average out in rapidly tumbling molecules. Spin-rotation relaxation is most effective for small, symmetrical molecules or freely rotating methyl groups. The molecular magnetic moments are generated by the rotating electronic and nuclear charges and fluctuate as a result of molecular collisions. The efficiency of the process is dependent on the tumbling rates, with higher frequencies improving the efficiency. Quadrupolar relaxation is only relevant for nuclei with a nuclear spin quantum number greater than ½ and is often the dominant mechanism for relaxation for these molecules. Quadrupolar nuclei possess both a magnetic dipolar and an electric quadrupolar moment. The quadrupolar moment is influenced by electric field gradients about the nucleus and will relax if the changes occur at the correct frequency. The efficiency of quadrupolar relaxation depends on the magnitude of the quadrupolar moment and the magnitude of the electric field gradient.

3.2.4 The NMR Signal

The rotating magnetization induces a current in the receiver coil. The magnitude of the induced current is proportional to the magnetization in the transverse plane and decays exponentially with a time constant T_2^* as the nuclei relax. The signal is known as the free induction decay (FID). The FID is a superposition of the cosine-modulated and T_2-damped contributions from all nuclei. The net magnetization is represented as:

$$\overrightarrow{M}_{molecule}(t) \approx \sum_i M^i_{z,0}\left[\cos(2\pi f_i t)\hat{x} + \sin(2\pi f_i t)\hat{y}\right]e^{-\frac{t}{T_2}} \quad \text{(Equation 3.13 a)}$$

$$\overrightarrow{M}_{total}(t) = \sum_j c_j \overrightarrow{M^j}_{molecule}(t) \quad \text{(Equation 3.13 b)}$$

Where $M^i_{z,0}$ is the magnetization from a single molecule, $\overrightarrow{M}_{molecule}$ is the net magnetization for each i spin type, and $\overrightarrow{M}_{total}$ is the magnetization from all j types of molecules resonating at a frequency $f_{,i}$ with a concentration c_j. The Fourier transform of equation 3.13 b leads to the spectrum. The spectrum of each molecule is a series of defined resonant lines, and the spectrum of the mixture is a linear combination of the concentration-weighted summation of these molecular spectra (Ross et al., 2007). The linearity of the summation allows for accurate quantification of metabolites.

3.2.4.1 Signal to Noise Ratio

The detected NMR signal is dependent on the gyrometric ratio of the nuclei, γ, the number of spins in the sample, N, the external magnetic field, B_o, the sensitivity of the detector and noise. The signal is linearly dependent on the number of spins as each magnetic moment induces a current in the receiver coil through electromagnetic induction. There is a square dependence on the magnetic field. One factor comes directly from the high-temperature approximation of Boltzmann's law, while the other comes from the induced current in the detector coil. The noise-free signal may be represented as:

$$Signal = \gamma_X^3 N B_o^2 S_D \qquad \text{(Equation 3.14)}$$

Where S_D describes the detector sensitivity and noise. The use of a cold head probe, which operates at a few Kelvin, will improve the S_D parameter. The low temperatures serve to dramatically reduce resistance in the wires, and hence reducing electronic noise. Another method is to reduce the size of the detector (Ross et al., 2007), but it is not recommended to use a NMR tube smaller than 5mm in diameter for metabolomics experiments (McKay, 2009). In addition, the scan time may be increased to allow for full relaxation or more transients (Dunn and Ellis, 2006).

The signal to noise ratio may be enhanced by averaging multiple transients (Evilia, 2001). The noise is predicted to be 'white' or uncorrelated, and would increase by the square root of the number of scans (Szantay, 1992). Conversely, the signal is stable, and would increase linearly with each scan (Ross et al., 2007). To a first approximation, the signal to noise ratios of the FID is proportional to $B_o^{3/2}$ (Szantay, 1992).

During acquisition, the receiver coil digitizes the FID using an analog-to-digital converter (ADC) (Szantay, 1992). The digitization rate should be greater than double the highest frequency (or double the spectral width (Evilia, 2001)) to prevent back fold of the data. This is commonly referred to as the Nyquist criterion. Under sampling will cause high frequency signals to be falsely represented at a lower frequency. A low-pass filter can be used before digitization to prevent fold over from noise components by selectively attenuating the signal from the high frequency components, which are often dominated by noise. Szantay (1992) predicted that noise could be up to 41% greater at the spectrum edges than at the center due to fold over.

The receiver coil does not collect the FID from time 0 s, but begins after a short dead time after the RF pulse has been turned off (Szantay, 1992). The idea is to prevent pulse breakthrough, which will contaminate the first few available data points in the FID, leading to baseline distortions.

Szantay (1992) advises scientists to separate consecutive scans by $5T_1$ and to sample for at least $5T_2^*$. This will result in long scan times, but will prevent wiggles in the baseline caused by cutting off the spectrum too soon. Short acquisition times are more detrimental to sharp resonance peaks since these resonances often have longer T_2 relaxation constants.

The signal to noise ratio and peak shape can be improved by multiplying the FID with an exponentially decaying window function prior to the Fourier Transform. The exponential emphasises the contributions from the lower frequency resonances as it attenuates high frequencies and has the form $e^{-\pi LBt}$, where LB is the line broadening (Ross et al., 2007). Application of the window function will accelerate the apparent decay of the FID, thus leading to line broadening. The optimal exponential weight is one that matches the time constant of the exponential decay; this is known as the matched filter (Szantay, 1992).

3.2.4.2 Peak Splitting Due to J-Coupling

Scalar coupling, or J-coupling, is the interaction between active spins sharing a common electronic orbital. The nuclei involved are often separated by less than 5 chemical bonds in the same molecule. The magnetic field experienced by one nucleus, A, from the other, B, is dependent on the relative orientation of the magnetic moments. The coupling causes splitting of the resonances (Rule and Hitchens, 2006). The idea is that the polarization of one nuclear spin will influence the polarization of the surrounding electrons. This subsequent polarization will alter the local magnetic field experienced by the coupled nuclei; whether the field increases or decreases depends on the polarization of the first nucleus. (Rule and Hitchens, 2006).

Separation between the spectral peaks is defined by the degree of electronic overlap between the two spins and is characterized by a coupling constant, $^nJ_{AB}$, where n is the number of intervening bonds and A and B identify the two coupled spins (Pavia et al., 2009; Rule and Hitchens, 2006). $^nJ_{AB}$ is usually given in Hz and is equal to the

observed frequency shift between split resonance (Rule and Hitchens, 2006). It is independent of the magnetic field strength (Pochapsky and Pochapsky, 2006).

J-coupling may be weak or strong, depending on the relative magnitude of the coupling constant and the observed separation between the peaks. The strength of the coupling is related to the magnitude of the coupling constant and the difference in the resonances of the nuclei. At high magnetic field strengths, dispersion of nuclei is on the order of kHz. Consequently, the spin system will be weakly coupled for our experiment. J-coupling through multiple chemical bonds is greatly decreased (Rule and Hitchens, 2006).

Spin ½ nuclei have two possible spin states. As previously suggested, the populations of nuclei in each spin state are approximately equal. The orientations of surrounding coupled nuclei will influence the net field experienced by a nucleus. Depending on these orientations, the nuclei may experience a stronger or weaker magnetic field (Rule and Hitchens, 2006). The different magnetic field will lead to a shift in the resonance of the nuclei, causing the spectral peak to split. The separation between the peaks is equal to the J-coupling (Rule and Hitchens, 2006). The shifts in frequency of both nuclei are identical, so the peak is centered about its chemical shift (Rule and Hitchens, 2006).

For the example of a two nuclei system, half of the nuclei will experience a larger field, while the other half experiences the weaker field. The spectral peak will split in half, with each half having an equal intensity; this is known as a doublet. Triplet resonances occur when two nuclei have identical chemical shifts and couple with another nucleus. The two like nuclei are said to be magnetically equivalent and will not couple with each other. The resultant triplet has peak intensities with ratios 1:2:1. For n equivalent coupled nuclei, the spectral peak will split into (n+1) peaks with intensities following the binomial distribution (Minch, 1994). The distance between consecutive peaks is $^{n}J_{AB}$ (Rule and Hitchens, 2006).

3.2.4.3 Causes of Spectral Distortion

Factors that cause spectral distortion include clipping of the FID vertically, truncating too many data points at the start of the FID (dead time), truncating the end of the FID too early, or using too short an acquisition and relaxation delay (Szantay, 1992).

Vertical truncation will result if the receiver coil gain is set too high. This results in an ADC or memory overload. High concentration samples are at risk of having an ADC overload due to the large magnetization. The FID can induce currents in the receiver coil, which in turn produce weak RF fields. These fields work against the FID, which results in a faster apparent T_2 relaxation time, and thus broadens the resonance. This process is known as radiation damping. The large water signal is prone to this effect as its tails may extend over a large range of frequencies (Szantay, 1992).

Truncating the early parts of the spectrum could have profound effects on the phasing (Szantay, 1992). Generally, this problem is overcome through manual phasing as long as the dead time was short enough to prevent excessive loss of information. Nuclei with short T_2 time constants are especially affected.

In zero-filling, a series of zeros are appended to the end of a FID to improve the resolution without having to increase the acquisition time. Nonetheless, truncating the FID too early, or zero filling when the FID has not reached zero magnitude will produce wiggles in the baseline. The truncation acts as a step function at the end of the FID, thus introducing a plethora of frequency signals (Szantay, 1992). Despite improvements in resolution, zero-filling by more than a factor of 2 is unnecessary (Evilia, 2001).

3.3 Nuclear Overhauser Effect (NOE) and NOESY Pulse Sequence

3.3.1 Requirement for Water Suppression

Urine samples contain metabolites at mM concentrations and water protons at a concentration of 110 M (Lippens et al., 1995). At these concentrations, metabolites are virtually undetectable unless water suppression is used to significantly reduce the dynamic range (Goldsmith et al., 2010; Mo and Raferty, 2008; Potts et al., 2001; Vion Dury, 1993). In addition, the large water resonance may result in baseline distortions, and thus make quantification of metabolites difficult (Aranibar et al., 2006). Methods of water suppression rely on null excitation, relaxation or selective irradiation of the solvent signal (Lauridensen et al., 2008).

Solvent suppression is most effective on symmetric, tall and narrow peaks, as a majority of the solvent resonances are confined to a narrow frequency window and overlap with other metabolites is less likely (McKay, 2009). Therefore, a homogenous magnetic field is a requirement for effective water suppression (Viant et al., 2007).

The standard approach for water suppression involves the pre-saturation of the solvent resonance with a weak irradiation during the relaxation delay (Aranibar et al., 2006; Viant et al., 2007). This technique will reduce the intensity of proton resonances in fast exchange with water protons (Viant et al., 2007; Vion Dury, 1993). Therefore, it is preferred to use an irradiation with lower power, as high power pulses will also attenuate the signals from neighbouring signals (McKay, 2009). This technique is preferred for samples with small molecules (Mo and Raftery, 2008). However, pre-saturation is not sufficient when a strong water signal is present, therefore, it is often paired with a 1-D NOESY experiment (Viant et al., 2007). The pulse sequence exploits the nuclear overhauser effect (NOE) to selectively saturate the water signal, while maintaining the signals from the lower concentrations metabolites (Claridge 2009) and is shown in figure 3.7. Other choices for water suppression in metabolomics experiments include WATERGATE and RECUR (Wishart, 2008).

3.3.2 Nuclear Overhauser Effect (NOE)

The nuclear overhauser effect (NOE) is a consequence of through-space interactions between nuclei. Such interactions can cause an enhancement in the intensity of one resonance after the saturation of another resonance. Unlike J-coupling, NOE interactions need not share an electronic orbital, but must be close in space. The strength of the NOE is dependent on the spatial distance between the spins and falls off as the inverse sixth power of the spacing (α $1/r^6$) (Willimason, 2008).

The system of a NOE interaction consists of two unique spin nuclei, I and S, such as those from two distinct metabolites. The S spins will relax to equilibrium through T_1 relaxation, as well as through dipole-dipole coupling with nearby nuclei (Kumar et al., 1981; Williamson, 2008).

The NOE is defined as the fractional change in intensity of one NMR resonance when another resonance is perturbed as a result of cross relaxation (Kumar et al., 1981). The perturbation often takes the form of saturating one resonance using weak irradiation or inversion by applying a 180^0 degree RF pulse (Claridge 2009). The enhancement, η_I, is calculated using:

$$\eta_i(S) = \frac{I - I_0}{I_0} x 100 \qquad \text{(Equation 3.15)}$$

Here I is the intensity of the magnetization in the presence of NOE, I_0 is the intensity of the magnetization with out the NOE and $\eta_I\{S\}$ represents the observed NOE for spin I when spin S was perturbed. The enhancement is about 50% for small molecules.

3.3.3 NOESY Pulse Sequence

The standard 1-D NOESY pulse sequence is:

$$RD - 90^o - t_1 - 90^o - t_m - 90^o - AQ$$

A schematic of the pulse sequence is shown in figure 3.7. The relaxation delay, RD, is typically a few seconds duration. During this time the water resonance is selectively irradiated with the B_1 RF pulse to saturate the signal (Bollard et al., 2001; Lippens et al., 1995). t_1 is a fixed interval during which the nuclei precess freely. This allows for measurements of frequency related data, as well as for frequency labelling (Macura and Ernst, 2002). The second 90^o pulse serves to align the magnetization along the –z-axis, and thus enabling the build up of NOE through cross-relaxation (Macura and Ernst, 2002). t_m is the mixing time. During this time interval, the water resonance is selectively irradiated for the second time to allow for the build up of NOE from the nuclei of interest (Foreshed et al., 2005; Lippens et al., 1995; Potts et al., 2001). The third 90^o RF pulse returns the magnetization to the x-y plane so that the FID may be acquired (Macura and Ernst, 2002). The water resonance will return to equilibrium during acquisition (AQ), but will be suppressed to zero again during the relaxation delay (Lippens et al., 1995).

The time to repetition, t_R, is dependent on the pre-saturation interval and acquisition time. It is important that the relaxation delay is sufficiently long to suppress the water peak (Lauridensen et al., 2008). The net magnetization from water should be near zero at the start of the acquisition so that the lower concentration metabolites may be detected (Lippens et al., 1995). Also, the acquisition time should be set to allow for near complete relaxation of the transverse magnetization (Lauridensen et al., 2008).

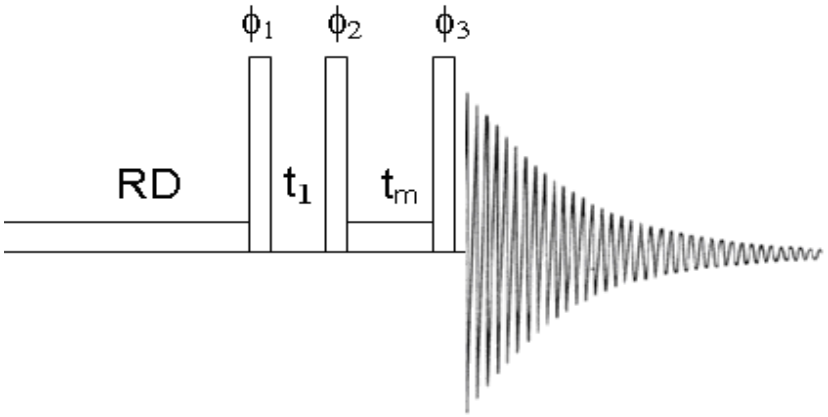

Figure 3.7: The standard 1-D NOESY pulse sequence. RD represents the relaxation delay during which time the water resonance is selectively irradiated with a weak RF pulse. t_1 is a fixed interval used to allow for phase information to be acquired. t_m is the mixing time. The water resonance is again irradiated to allow for NOE build up of the protons of interest.

The 1-D NOESY pulse sequence is robust, requiring little optimization (Potts et al., 2001). This attribute makes it easy to use (Lauridensen et al., 2008), thereby making it attractive for high throughput applications, such as screening. It requires minimal first order phase correction and baseline adjustments (Potts et al., 2001) as it provides efficient water peak suppression (Lauridensen et al., 2008). Water suppression has been enhanced from the simple pre-saturation experiment, while adding minimal time to the experiment (Bollard et al., 2001). It does, however, distort the baseline about the water resonance (Aranibar et al., 2006), so it is preferred to quantify metabolite resonances far from the water resonance whenever possible.

In one study, Potts et al. (2001) showed that the greatest source of variability in NMR experiments results from the water suppression technique and timings used as well as variations in the baseline and differential attenuation of resonances near the water resonance. PCA scores maps revealed that experiments using varying magnetic field strengths were comparable (Potts et al., 2001; Malz and Jancke, 2005). In addition, Lewis et al. (2007) found that analytical precision is independent of the chemical properties of the target molecules. Regarding spectral pre-processing, both Aranibar et al. (2006) and Malz and Jancke (2005) found that the best results were obtained after manual phasing and baseline corrections.

Chapter 4
Analytical Techniques for Metabolomics Experiments

This Chapter will outline analysis and chemometric techniques often used to study metabolism. The main analytical techniques include ^{1}H NMR Spectroscopy and mass spectrometry, paired with liquid or gas chromatography. Once the metabolite content is known, multivariate techniques, such as principal component analysis or partial least squares discriminant analysis are performed to reduce the dimensionality of the data set and to discovery patterns in the data.

4.1 Nuclear Magnetic Resonance Spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy is a popular analytical technique since biofluid sample collection (i.e. urine, saliva) can be non-invasive (Saude et al., 2007), it is non-destructive (Dunn and Ellis, 2005; Goldsmith et al., 2010), provides a rapid, high-throughput methodology (Dunn and Ellis, 2005; Verpoorte et al., 2008), is highly reproducible (Kim et al., 2008; Viant et al., 2007), is not biased towards the detection of certain metabolites (i.e. it is non-selective) (Goodacre et al., 2004; Kaddurah Daouk et al., 2008), and can simultaneously analyze all metabolites present in the biofluid (Griffin and Kauppinen, 2007; Vion-Dury et al., 1993). The output spectrum is simply the sum of all spectra of individual metabolites present in the sample (Verpoorte et al., 2008). NMR experiments generally require minimal sample preparation, often consisting only of an adjustment of the pH and addition of an internal standard (Ross et al., 2007). NMR spectroscopy is a cost effective method after the spectrometer has been purchased, thus making it ideal for screening large populations (Lenz and Wilson, 2007).

NMR spectra provide qualitative and quantitative information of the metabolic content of a biofluid or tissue extract (Forshed et al., 2005; Um et al., 2009). The concentrations of metabolites may be determined by comparing the area of each peak with that of the internal standard, such as 3-trimethylsilylpropionic acid (TPS) (Beckonert

et al., 2007) or 2,2-dimethyl-2-silapentane-5-sulfonate (DSS) (Wiljie et al., 2006). Once a pulse sequence is selected for the experiment, it is pertinent to keep the timings, such as the relaxation delay or mixing time, constant for quantitative purposes. A new set of Nuclear Overhauser Effect enhancements (NOE) are obtained for each mixing time in a NOESY experiment (Kumar et al., 1981). It is well established that the concentration is directly proportional to the area of the peak (Dunn and Ellis, 2005). Depending on the strength of the magnetic field, NMR spectroscopy can measure metabolites in the mM and μM ranges (Nordstrom and Lewensohn, 2010; Slupsky et al., 2007).

Larger magnetic fields are known to improve signal dispersion and sensitivity, thus resolving metabolite peaks to a better extent (Lenz and Wilson, 2007). This is beneficial when analyzing low concentration metabolites. The sensitivity of the instrument can be further improved with the use of a cryoprobe (commonly referred to as a cold probe for Varian systems) (Lenz and Wilson, 2007). This hardware addition decreases electronic noise by operating at liquid helium temperatures, thus allowing for improved signal to noise ratios – by up to a factor of 4 (Lenz and Wilson, 2007).

A few drawbacks with NMR spectroscopy include the heavy overlap of metabolite resonances (Holmes et al., 2000; Ludwig et al., 2009), chemical shift variations due to changes in concentration, pH and ionic strength (Miyataka et al., 2007; Spratlin et al., 2009), poor water suppression (Keun et al., 2002), baseline distortions (Keun et al., 2002) and chemical exchange between metabolites, particularly with water (Verpoorte et al., 2008).

Resonance overlap is a problem with ^{1}H NMR spectra since the scale runs between 0 and 10 ppm (Nicholson et al., 2007; Potts et al., 2001) and thousands of metabolites may contribute to the signal (Claudino et al., 2007; Goldsmith et al., 2010). Several endogenous metabolites contain more than one resonance (Ala Korpela et al., 2008), which can complicate analysis with the conventional method of binning where the spectra are generally divided into regions of width 0.04 ppm. In addition, overlapping signals make profiling less accurate.

Differences in the pH, concentration and ionic strengths will cause variations in the chemical shifts of resonances due to alterations in the acid-base equilibrium and solute-solute interactions (Spratlin et al., 2009). Miyataka et al. (2007) observed that the chemical shift of citrate varied significantly between samples, and that the shift was approximately linear with the sample pH. No other metabolite demonstrates such a

dramatic change in chemical shift with pH (Miyataka et al., 2007). Failure to consider variations in the chemical shift of citrate with pH could introduce artefacts if spectral binning is used.

Dilute samples often suffer from poor spectral quality due to incomplete water suppression (Keun et al., 2002). It is essential to suppress the water resonance as much as possible so that nearby peaks can be resolved. Metabolites of interest are often present in mM concentrations while water protons are 110M (Saude et al., 2006). Potts et al., (2001) found that the water suppression scheme used greatly influenced the baseline and degree of attenuation of signals near the water resonance. The NOESY pulse sequence showed differences based on varying mixing times (Potts et al., 2001). Poor water suppression can elevate the baselines on either side of the peak, which adds to the total peak height and thus affects quantification (Saude et al., 2006). Further, incomplete water suppression can cause dynamic range problems if the water resonance is not suppressed to reflect a mM concentration (Keun et al., 2002).

Proton exchange between OH, NH or SH chemical groups with water can have profound effects on the quantification of these metabolites as the resonance peaks will also be suppressed and broadened due to an increased T_2 relaxation rate (Verpoorte et al., 2008). Peak broadening degrades the spectral resolution (Reo, 2002). Saude et al. (2006) studied the effects of various water suppression techniques. They found that the tnnoesy pulse sequence provided the greatest degree of water suppression across the entire spectrum and had the highest degree of quantitative accuracy.

4.1.1 Spectral Processing

The accuracy of metabolite concentration quantification is largely dependent on the combination of pre-processing techniques, such as phasing or baseline correction, and the profiling technique.

A common methodology is to bin the spectra. This technique involves dividing the spectrum into smaller segments, often of width 0.04 ppm, and defining each bin as a variable (Ross et al., 2007). However, it is not clear if an alteration in the magnitude of a single bin is a result of a metabolic change or an artefact resulting from an altered chemical shift (Slupsky et al., 2007). Some groups attempt to minimize variations in chemical shift by combining adjacent bins so that the full resonance is contained within

the larger bin (Ross et al., 2007). Combing bins will degrade the analytical resolution as multiple resonances may be located in that bin (Forshed et al., 2005; Madsen et al., 2010). This is true in cases where one metabolite shifts to another bin as a result of pH or ion concentration. The nature of spectral binning places greater emphasis on the higher concentration metabolites. Therefore, low concentration metabolites may not be detected (Weljie et al., 2006). Binning is prone to baseline errors (Wiljie et al., 2006). Incomplete water suppression can elevate the baseline on either side of the water resonance and cause further baseline distortions (Saude et al., 2006).

Alternatively, targeted profiling, which consists of matching pre-established metabolites from a library to the spectrum, provides better quantitative accuracy (Kkaddurah-Daouk et al., 2008). Saude and Sykes (2006) showed that targeted profiling provides accuracies in excess of 90% for their samples. Slupsky at al. (2007) measured the coefficient of variability for metabolite near the limit of detection and found that the precision is acceptable for clearly identifiable metabolites. The coefficients ranged from 1% for the larger concentration metabolites to 41% for metabolites near the limit of detection. It is suggested that the larger coefficients are a result of artefacts from the baseline, overlap of resonances and noise (Slupsky et al., 2007).

Studies have evaluated the reproducibility of NMR spectra between facilities and between spectrometers. One such study was completed by Keun et al. (2002). To assess the analytical reproducibility of metabolomics protocols, sample preparation and NMR data acquisition were performed at two sites, one site using a 500 MHz spectrometer and the other using a 600 MHz system. Half of each urine sample was sent for analysis at each facility. They found that despite using spectrometers operating at different frequencies, the datasets were extremely similar. Observed differences were most often related to small changes in the chemical shift. However, the main difference was contributed to the efficiency of the water resonance suppression. The results show that the normal physiological variation within the control group (26%) or pre-dose group (17%)) accounted for five times as much of the total variability as the different spectrometers did (3%). Further, samples considered outliers at one location were not outliers at the other. Keun et al. (2002) suggest that this is due to sporadic differences in sample preparation.

4.1.2 Alternative NMR Techniques

Other NMR techniques include ^{13}C or ^{31}P NMR experiments or magic angle spinning for tissue extracts (Griffin and Shockcor, 2006; Serkova et al., 2008).

^{13}C or ^{31}P imaging allows for improved spectral resolution, but at the expense of lower sensitivity due to the lower natural abundance of these elements (Dunn and Ellis, 2005). ^{13}C imaging spreads the peaks over a spectral width of 200 ppm as compared to ^{1}H which encompasses the range between 0 and 10 ppm. These techniques can be used in 2-D experiments to aid in the identification of metabolites (Serkova et al., 2008).

2-D NMR experiments are useful in improving signal dispersion and exploiting connectivity between signals (Beckonert et al., 2007). Most experiments combine experiments with ^{1}H, ^{13}C, ^{15}N and ^{31}P nuclei to allow for identification of biochemical substances (Beckonert et al., 2007). A few advantages of using a 2-D experiment include the reduced overlap of spectral peaks, thus allowing for direct measurement of these peaks (Lewis et al., 2007) and additional information regarding the multiplicity and coupling patterns of resonances (Beckonert et al., 2007). However, this technique requires longer acquisition times, of the order of hours (Beckonert et al., 2007; Lewis et al., 2007), which makes it unsuitable for screening purposes. In addition, 2-D cross peaks intensities are often influenced by a greater number of variables, including uneven excitation throughout the sample, non-uniform relaxation, evolution time and mixing times (Lewis et al., 2007).

Magic angle spinning NMR spectroscopy is used to profile the metabolites in a tissue extract (Griffin and Shockcor, 2004). This method involves spinning the samples at 54.7^{o} with respect to the main magnetic field to reduce the physical effects that lead to line broadening (Griffin and Kauppenin, 2007). The line broadening is a consequence of dipole coupling and tissue anisotropy in the semi-solid sample which scales as $(3\cos^{2}\theta-1)$ (Seierstad et al., 2008). Studies have shown that aligning the sample at an angle of 54.7^{o} with respect to the external magnetic field and then spinning the sample about its own axis reduces these interactions (Seierstad et al., 2008). Though tissue extracts make the procedure more invasive, there is a gain in robustness as tissues are under greater homeostatic regulation and thus providing more consistent measurements of metabolism (Viant et al., 2007).

^{1}H NMR spectroscopy provides diagnostic and prognostic information on cancers. For this reason, most studies involving NMR applications use ^{1}H. Applications of its use for cancer include breast, brain, head and neck, lymphomas, liver and prostate cancers (Griffin and Kauppinen, 2007), thought these studies have not attempted to screen for these cancers.

4.2 Mass Spectrometry

Mass spectrometry (MS) identifies metabolites based on the mass-to-charge ratio of the charged particles (Dunn and Ellis, 2005; Spratlin et al., 2009). MS is a rapid, sensitive and selective technique which offers a plethora of qualitative and quantitative information (Dunn and Ellis, 2005). It is estimated that the sensitivity is two orders of magnitude greater than NMR and measures metabolites with concentrations well below the mM range (Kaddurah Daouk, 2008). Mass spectrometers operate by ion formation, separation of the ions by means of their mass-to-charge ratio and detection of the separated ions (Dunn and Ellis, 2005).

MS is often paired with a chromatographic technique to separate metabolites. Liquid and gas chromatography (LC-MS and GC-MS) are the most commonly used techniques (Clarke and Haselden, 2008; Verpoorte et al., 2008). Metabolites are identified by comparing the retention time or retention index and mass spectrum of the sample peaks with a commercial database containing pure metabolites under ideal circumstances. (Dunn and Ellis, 2005; Verpoorte et al., 2008). However, the accuracy depends on the efficiency of the ion formation process (Griffin and Kauppenin, 2006). Care must be taken regarding the methods of extraction, quenching, and sample storage conditions as they can potentially alter metabolite structure, thereby introducing greater variability between samples (Spratlin et al., 2009).

GC-MS is considered the gold standard for metabolomics (Dunn and Ellis, 2005). The technique first separates volatile and thermally stable compounds by GC, then detects ions with electron-impact mass spectrometers (Dunn and Ellis, 2005). GC-MS is biased towards detecting volatile, low molecular weight metabolites and requires derivatisation at room or elevated temperatures (Dunn and Ellis, 2005). In fact, the spectrometer will only detect ions which may be derived. The process can be time-consuming, costly and runs the risk of losing metabolites (Spratlin et al., 2009). Sample

stability may also be a concern; the presence of water can cause breakdown of the molecules. However, extensive sample drying can reduce this reaction, but could also result in the loss of volatile metabolites (Dunn and Ellis, 2005).

LC-MS is another combined system which separates metabolites by LC followed by electrospray ionization (ESI) or atmospheric pressure chemical ionization (APCI) (Dunn and Ellis, 2005). This technique may be performed at lower analysis temperatures and does not require sample volatility. Electrospray instruments operate in positive and negative ion modes, so that a larger complement of metabolites may be analyzed (Dunn and Ellis, 2005). Moreover, ionized metabolites are detected in either the positive or negative ion mode, but not both. Metabolite identification is more difficult and time-intensive in LC-MS since spectral libraries are not commonly available. ESI is not effective with high concentrations of salts, acids or bases (Dunn and Ellis, 2005). A majority of metabolomics studies using LC-MS focus on clinical applications (Dunn and Ellis, 2005).

Quantification of the metabolites generally requires external calibration or a response ratio (peak area of metabolite / peak area of internal standard) (Dunn and Ellis, 2005; Verpoorte et al., 2008). However, external calibration is laborious and not all metabolites are available (Dunn and Ellis, 2005). A large numbers of unknown peaks will degrade the quality of the analysis. One solution is to identify only those metabolites that show a significant change between samples (typically of the order of 20-40 metabolites) (Verpoorte et al., 2008). Each MS technique has a bias towards certain compound classes resulting from the ionization technique used, chromatography and detector capabilities (Weckwerth and Morgenthal, 2005).

4.3 Statistical Significance for Univariate Statistics

When simultaneously testing a large number of variables, there is an increases risk of obtaining a false positive result (Broadhurst and Kell, 2006). For this reason, it is important to update the significance limit, α, to reflect the larger number of variables. There are two approaches used in this thesis: the Bonferroni correction and the false discovery rate (FDR).

The Bonferroni correction controls the overall error rate – known as the family-wise error rate (FWER) – at α/N, where α is a predefined significance threshold (i.e. 0.05

or 0.01) and N is the number of variables. The correction can be thought of as the probability of obtaining one or more false positives among all variables tested, not just the probability of obtaining a false positive for one variable (Broadhurst and Kell, 2006). That is, testing each variable individually at level α/N, will maintain a false positive rate less then α. It is important to consider when N is large (Broadhurst and Kell, 2006). For example, analysis on the metabolite ratios (n = 561) at an error rate of $\alpha = 0.05$ would result in 28 false positives (0.05 x 561), whereas, the Bonferroni correction is expected to have zero (on average, there will be a probability of 0.05 false positives). The Bonferroni correction is a more stringent test for significance when testing multiple hypotheses simultaneously (Broadhurst and Kell, 2006).

An alternative method, known as the False Discovery Rate (FDR), was proposed by Benjamini and Hochberg. In contrast to conventional methods, which define a limit of significance prior to testing, this technique controls the number of false positives within the set of defined significant discoveries. The number of false discoveries accepted may be controlled at a higher level, thus increasing the power of the analysis. This property is desired for screening purposes as it is advantageous to identify as many potential discoveries possible, while still controlling the number of false discoveries (Benjamini and Hochburg, 1995). Following is a brief derivation of their method.

Consider a study with m (null) hypotheses, of which, m_o are true and $m\text{-}m_o$ are statistically significant (null hypothesis rejected). Assume that R hypotheses are suggested to be statistically significant. There are two possibilities: 1) the hypothesis was correctly rejected (true positive result or TP) or 2) the hypothesis was incorrectly rejected (false positive result or FP). The false discovery rate may be defined as the number of false positives (FP) divided by the total number of rejected hypothesis (R = FP + TP):

$$FDR = P(R > 0)E\left(\frac{FP}{R}\middle| R > 0\right) \qquad \text{(Equation 4.1)}$$

Where E represents an expectation value of the false discovery rate and P(R>0) ensures that the FDR is controlled when no variables are significant. Two important properties from this definition are: 1) In the limit where no result is statistically significant, the FDR method will produce the same results as the FWER (if FP = 0, then FDR = 0 and if FP > 0, then FDR = 1) and 2) when $m_o < m$, the FDR is smaller than or equal to the FWER, and thus more statistically significant variables will be discovered while maintaining control over the number of false positives.

Consider testing H_1, H_2,…H_m hypotheses based on their corresponding p-values, P_1, P_2,…P_m. Rank the p-values from smallest to largest such that $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ where $P_{(i)}$ corresponds with hypothesis $H_{(i)}$. We define the following Bonferroni-type multiple-testing procedure:

Let k be the largest i for which $P_{(i)} \leq \frac{i}{m}\alpha$; (Equation 4.2)

then reject all $H_{(i)}$ i = 1, 2, …, k

The above definition controls the FDR at level α. Significant discoveries are determined by comparing the p-value, p_i with $i\alpha/m$, i = 1, 2,…,m., (i.e. $p_i \leq i\alpha/m$). If it is smaller, then all preceding (smaller p-values) hypotheses are defined as significant. The maximum i at which the inequality is satisfied caps the number of defined significant discoveries such that the error rate is maintained at α.

It is important here to understand the difference between the false positive rate and the false discovery rate. The false positive rate is the rate at which true null hypotheses are incorrectly defined as significant, whereas the false discovery rate describes the expected proportion of false positives within the set of all tests defined as significant, i.e. the discoveries (Storey, 2002).

4.4 Chemometric Techniques

Chemometric techniques are used to reduce the complexity of large data sets. The goal is to reduce the dimensionality of the data set so that hidden patterns of behaviour can be extracted, with minimal loss of information (Goodacre, 2007). Such techniques are categorized as unsupervised or supervised.

Unsupervised methods apply data reduction to visualize inherent patterns of behaviour and identify similarities in the data structure (Bollard et al, 2005; Lindon et al., 2001). These techniques, however, are sensitive to subtleties in the experimental design, making interpretation more difficult (Kaddurah Daouk et al., 2008). Examples of unsupervised techniques include principal component analysis (PCA), nonlinear mapping procedures (NLM) and hierarchal cluster analysis (HCA) (Holmes and Antti, 2002).

Supervised techniques build a multi-variate statistical model based on known sample classes (Bollard et al., 2005; Erb, 2008). Class information is used to optimize the separation between the classes (Claudino et al., 2007). The goal of these methods is

to find a mathematical transformation that correctly classifies a majority, if not all, of the samples (Goodacre, 2007). Examples of supervised methods include partial least squares (PLS), linear discriminant analysis (LDA), K-nearest neighbour analysis (KNN), partial least squares discriminant analysis (PLS-DA) and orthogonal PLS-DA (Lindon et al., 2001).

Chemometric techniques aim to remove redundant data in the system, so that non-random characteristics, which may be obscured by noise or random variables, may be identified, or interpreted with some methods (Lindon et al., 2001). It is expected that the level of redundancy is high in data sets with a large number of variables (Antoniewics et al., 2006); it does not matter what the variables represent, as long as they were measured independently from one another (Shlens, 2005). As a result, the system could be described with a fewer number of variables, known as latent variables (LV). The LV's represent the greatest sources of variation in the data (Trygg et al., 2007). The data should follow a normal distribution for optimal results (Hagburg, 1998). This is not always true in an experimental situation; therefore, groups may apply a $\log_{10}$ transformation to the variables distributions prior to analysis (Chang, 2009).

Most multivariate methods involve the diagonalization (eigensystems analysis) of a symmetric matrix (Hagburg, 1998; Lindon et al., 2001). In the cases of PCA and PLS-DA, the symmetric matrix is the covariance matrix of the data set (Shlens, 2005; Hagberg, 1998). The eigenvectors of the symmetric matrix represent the direction in K-space, where K is the number of original variables, which contributes to the maximum variation (Shlens, 2005). The eigenvalues are used to calculate the percentage of variation that its corresponding eigenvector describes in K-space and are ordered from largest to smallest. The first eigenvector, associated with the maximum eigenvalue, represents the direction with the maximum variation. This vector is assumed to be least affected by noise (Hagburg et al., 1998). Each successive eigenvector describes the maximum variation remaining in the data set and is orthogonal to all previous eigenvectors. Each sample is projected onto the model axes – i.e. the latent variables – and given a new set of coordinates, known as its scores (Keun et al., 2004).

The data set is often mean-centered and scaled prior to analysis. Mean-centering involves calculating the mean for each variable and then subtracting the mean from all variables (Lindon et al., 2001). The result is a distribution centered at 0, but having the same shape. All multi-variate statistical techniques require mean-centered input data.

Scaling techniques are applied to enhance the information from the lower magnitude variables (Trygg et al., 2007). Common methods of scaling include unit variance scaling (also known as auto scaling) and pareto scaling. Unit variance scaling is achieved by dividing each variable by its standard deviation (Bollard et al., 2001; Lindon et al., 2001). Unit variance scaling gives every variable an equal weighting, therefore the model is not biased towards the higher magnitude variables (as these generally have larger variances). Pareto scaling provides a compromise between enhancing the lower magnitude variables, without over expressing noise (User Guide to SIMCA P+). Pareto scaling is generally recommended for metabolomics analyses, and involves dividing the variables by the square root of its standard deviation (Trygg et al., 2007). There is debate about which scaling method produces the most accurate and reproducible results, however it is well established that projection based techniques are sensitive to scaling methods (Trygg et al., 2007). In a review article, Trygg et al., (2007) recommends using either no scaling or pareto scaling, while Lindon et al., (2001) and Antoniewicz et al., (2006) recommend unit variance scaling. Both methods (pareto and unit variance) have been used in metabolomics studies and will be used in our analysis.

4.4.1 Principal Component Analysis (PCA)

Principle component analysis (PCA) was introduced in 1901 by Karl Pearson (Gorban et al., 2007) and is now used to analyze data from numerous fields of study, ranging from neuroscience to computer graphics (Shlens, 2005). With the increasing complexity of experimental analysis, it has been termed the "workhorse of chemometrics" (Trygg et al., 2007). The goal of PCA analysis is to account for as much total variation as possible with the least number of variables (often known as latent variables, or principal components (PC)) (Bollard et al., 2001; Serkova et al., 2008) and to reveal the 'true' dimensionality of the data set by removing redundant information (Antoniewicz et al., 2006).

PCA suggests that the observation matrix, **X**, of rank m, where m is the number of measured variables can be written as a sum of m matrices of rank 1 (Geladi and Kowalski, 1986) (i.e. each matrix describes one variable),. It is assumed that the matrix is mean-centered (all variables have a mean of zero).

$$X = M_1 + M_2 + ... + M_m \qquad \text{(Equation 4.3)}$$

These rank 1 matrices can be expressed as an outer product of a score vector, t_i, (n x 1) and a loading vector, p_i (m x 1), where n is the number of samples and m is the number of variables. The scores vector stores the projection of each sample onto each principal component. Likewise, the loading vector relates the relative contributions of each variable to that principal component – this is determined by projecting a unit vector along the principal component axis onto the axis defined by each variable (Geladi and Kowalski, 1986). Equation 4.3 may be re-expressed as:

$$X = t_1 p_1^T + t_2 p_2^T + ... + t_a p_a^T + E = XP^T + E \qquad \text{(Equation 4.4)}$$

Where a = m if all factors are used and E is the matrix of residuals. When noise dominates the later vectors, a < m and only the important variables are emphasized. The matrix of residuals contains all information that is not explained in the scores and loading matrix (i.e. noise) (Geladi and Kowalski, 1986).

One method for PCA modelling involves computing an eigenvalue decomposition of the covariance matrix or a singular value decomposition of a data matrix after mean centering (Shlens, 2005). PCA will determine if there is another set of orthonormal vectors, which are linear combinations of the original variables, that better describes the variation observed in the data set. In essence, the technique performs a coordinate rotation to align the axes with the direction of maximum variation (Shlens, 2005). Theoretically, PCA will provide the optimum transformation for minimizing the least squares term for a given data set (Geladi and Kowalski, 1986).

The following section will summarize the key mathematical steps required to construct a PCA model - full details can be found elsewhere (Shlens, 2005). There are four main assumptions related to PCA (Shlens, 2005):

1. The assumption of linearity: The observed data is expected to be constructed from a linear combination of certain basis vectors; the principal components. It is this basis set that we wish to find when performing PCA.
2. Assumptions related to the sample mean and covariance: The nature of PCA means that the eigenvectors of the covariance matrix are determined under the Gaussian assumption. Skewed distributions could bias the final results as the mean and standard deviations will be affected.

3. Large variances correspond with the most important dynamics of interest: Variables with larger variances have a greater weight on the final results as it is assumed that these variables contain the most information about the system. The importance of each principal component may be assessed by ranking the eigenvalues from largest to smallest and determining the percentage of described variation. In the terms of PCA analysis, the signal to noise ratio is defined as the variance along the axis of interested (signal) divided by the variance in a perpendicular direction (noise). Only when the signal to noise ratio is large, can it be assumed that the first few PCs correspond with the dynamics of interest and the latter PCs correspond with noise.

4. Orthogonality of the Principal Components: This property allows for data reduction, via diagonalization, onto a plane spanned by a set of basis vectors – the principal components. Orthogonal vectors mean that redundancy is kept to a minimum; therefore, fewer variables (PC's) are required for an accurate description of the data.

Data is collected from a number of samples and stored in an n x m matrix, X, where m is the number of independently measured variables and n is the number of samples. In the current state, it is not clear which variables contribute differences between samples. PCA will identify a new orthogonal basis of vectors which are linear combinations of the original variables and projects the data onto this new basis; a so-called 'change of basis'.

The covariance matrix is calculated to investigate the degree of correlations between pairs of variables. A small covariance means that the variables have low redundancy (relatively independent), while a large covariance means the variables are highly correlated. Completely independent variables will have a covariance of 0. The covariance between two matrices, A and B, having zero means, is defined as:

$$\mathbf{C}_{\mathbf{X}} = \frac{1}{n-1}\mathbf{X}^{T}\mathbf{X} \qquad \text{(Equation 4.5)}$$

Where n-1 is a constant of normalization. The covariance matrix is an m x m symmetric matrix. The diagonal terms represent the variance of a single variable and the off-diagonal terms describe the covariance between a pair of variables. Noise and redundancy between variables are captured in the covariance terms. Large diagonal terms correlate with dynamics of interest, while small values represent noise. Similarly, large off-diagonal term indicate high level of redundancy between those variables.

The set of principal components are determined from the diagonalization of the covariance matrix. For mathematical purposes, we define a linear transformation matrix, P, such that:

$$X' = XP \qquad \text{(Equation 4.6)}$$

Where X' is an n x m matrix in which the original data has been projected onto a new set of basis vectors. If the optimal basis is used, the covariance matrix of X' should be a diagonal matrix. To determine this basis, we substitute Equation 4.6 into Equation 4.5 and rearrange the matrices:

$$C_{X'} = \frac{1}{n-1} P^T (X^T X) P \qquad \text{(Equation 4.7)}$$

From linear algebra, the matrix X^TX is diagonalized by the orthogonal matrix of its eigenvectors, E, such that:

$$X^T X = EDE^{-1} \qquad \text{(Equation 4.8)}$$

We select the matrix P such that its columns are the eigenvectors of X^TX. Therefore, E = P and $X^TX = PDP^{-1}$. Since X^TX is symmetric, $P^{-1} = P^T$. Substituting this into Equation 4.7 and making the conversion $P^{-1} = P^T$, we get:

$$\begin{aligned} C_{X'} &= \frac{1}{n-1} P^T (X^T X) P \\ C_{X'} &= \frac{1}{n-1} P^{-1} (PDP^{-1}) P \\ C_{X'} &= \frac{1}{n-1} (P^{-1}P) D (P^{-1}P) \\ C_{X'} &= \frac{1}{n-1} D \end{aligned} \qquad \text{(Equation 4.9)}$$

The matrix $C_{X'}$ is diagonalized as desired. As such, PCA identifies the eigenvectors of X^TX and calls them the principal components of the data matrix.

PCA is recommended as a starting point for multivariate data analysis as it will indicate if differences are present between populations (Ala Korpel, 2007; Trygg et al., 2007). The model does not take into account class information, but will determine the gross variability (Barker and Rayens, 2003). The resultant models are not capable of distinguishing between groups, nor can it differentiate within class variation from between class variation. However, it is possible to classify an unknown sample by grouping it with the class with the closest average scores (Barker and Rayens, 2003).

4.4.2 Partial Least Squares Discriminant Analysis (PLS-DA)

PLS-DA was established for chemical applications in the late 1970's by Svante Wold and Harald Martens after initial work by Herman Wold in the 1960's in the field of economics (Barker and Rayens, 2003; Geladi and Kowalski, 1986). The technique combines the PLS algorithm with the linear discriminant analysis procedure developed by Ronald Fisher in 1936 (Barker and Rayens, 2003). The method by Fisher involves a linear projection of the input data where the between class variance is maximized with respect to the within class variance (Rosipal and Kramer, 2006). The PLS algorithm calculates the between groups sums of squares and the cross products matrix and utilizes the results for separation (Barker and Rayens, 2003). Traditionally, discriminant analysis was performed if structural differences between groups or subgroups were desired (Rubingh et al., 2006). However, it may be beneficial to extend the study beyond classification to study the sources of discrimination. PLS-DA was designed to maximize the separation between pre-defined groups relative to the pooled within group variation (Barker and Rayens, 2003).

PLS-DA is applied when a quantitative relationship between the data matrix, X, and the response matrix, Y, is desired (Trygg et al., 2008). It may be described as the regression extension of PCA (Bollard et al., 2001; Murdoch et al., 2008). The PLS regression is capable of quantifying a relationship between the independent and dependent variables that best describe differences in the populations under investigation. The PLS regression assumes a linear relationship between the two matrices (Antoniewicz et al., 2006).

The goal of PLS analysis is to maximize the covariance (or separation) between the data matrix and the response matrix (Holmes and Antti, 2002; Rugingh et al., 2006). The Y matrix may contain qualitative data, such as the age of a patient or a measured value, or it may be comprised of orthogonal unit vectors which define the pre-defined class (Bollard et al., 2001; Trygg et al., 2008). As an example, the class unit vectors may be [1, 0] and [0, 1] for the two class case. Antoniewicz et al., (2006) showed that a PLS model is capable of differentiating between informative variance and irrelevant data.

Following is a brief derivation of the PLS model. Unless specified, all information in the following section is taken from Geladi and Kowalski (1986), Hoskuldsson (1988) and Rosipal and Kramer (2006). For further details regarding the PLS regression, refer to Wold (1975) or Barker and Rayens (2003).

Consider an experimental setting in which N independent, and M response variables are measured for n samples. Two data matrices are defined: the data matrix, X $\subset R^N$ (n x N), and the response matrix, Y $\subset R^M$ (n x M), where N is the number of measured variables, M is the number of response variables and n is the number of samples. These matrices are mean-centered, meaning that each variable has a mean of zero. The data and response matrices may be represented by matrices with lower dimensions and having the form:

$$X = TP^T + E$$
$$Y = UQ^T + F \qquad \text{(Equation 4.10)}$$

Where T and U are the scores matrices (n x a; a is the number of extracted scores vectors), P (N x a) and Q (M x a) are the corresponding loading matrices, and E (n x N) and F (n x M) are the matrices of residuals. The individual score and loading vectors from each extracted scores vector makes up the columns of the scores and loadings matrices and are denoted by the lower case letters (i.e. t_i, u_i, p_i and q_i where i = 1,2,…a). PLS analysis is most effective when it can best describe Y such that ||F|| is as small as possible and it still achieves a useful relationship between the X and Y matrices (Geladi and Kowalski, 1986).

The classical PLS model is derived from the non-linear iterative partial least squares (NIPALS) algorithm. The model operates iteratively between the two matrices to determine the relationship between the variables in each. The NIPALS algorithm starts with a random initialization of one of the scores vectors u_i (could be the first column of Y (Hoskuldsson, 1988)) and repeats the following steps until convergence:

$$1)\ \mathrm{w} = \frac{\mathrm{X}^T u}{u^T u}$$
$$2)\ \mathrm{w} = \frac{\mathrm{w}}{\|w\|}\ (\text{Scale w to length } 1)$$
$$3)\ \mathrm{t} = \mathrm{Xw}$$
$$4)\ \mathrm{c} = \frac{\mathrm{Y}^T t}{t^T t} \qquad \text{(Equation 4.11)}$$
$$5)\ \mathrm{c} = \frac{\mathrm{c}}{\|\mathrm{c}\|}\ (\text{Scale c to length } 1)$$
$$6)\ \mathrm{u} = \mathrm{Y}c$$

Here, w and c are weight vectors. They are used to ensure that the t-values obtained are orthogonal (Geladi and Kowalski, 1986). If the algorithm converges at 6) (that is the u from 1) and 6) differ by less than a predetermined significance threshold, ε), then the x- and y-loadings (p and q, respectively) are calculated as follows. Otherwise, it repeats the iteration with the new u.

$$p = \frac{X^T t}{t^T t}$$
$$q = \frac{Y^T u}{u^T u} \qquad \text{(Equation 4.12)}$$

The next step is to regress u upon t and calculate the residual matrices, E and F.

$$b = \frac{u^T t}{t^T t}$$
$$E = X - tp^T \qquad \text{(Equation 4.13)}$$
$$F = Y - uq^T$$

The algorithm renames the E and F matrices as X and Y and repeats the process for all dimensions of X. In calculating the scores and loadings from the residuals, mutual orthogonality is guaranteed (Rosipal and Kramer, 2006). It is assumed that the score vectors of X, $\{t_i\}_{i=1}^{a}$, are good predictors of Y and that an inner relation exists between the scores vectors t and u. For this reason, the Y matrix is 'deflated' by subtracting off information along the direction of each score vector, t_i. This algorithm converges quickly (Geladi and Kowalski, 1986), so models can be generated promptly.

The PLS factors have four properties (Geladi and Kowalski, 1986):

1. The loading vectors, p_i and q_i, have unit length. i.e. $||p_i|| = ||q_i|| = 1$
2. The scores vectors, t_i and u_i, are centered about zero
3. The scores of the X data matrix, t_i, are orthogonal
4. The weights, w_i, are orthogonal

When PLS is used for discrimination or classification, the samples are divided into g classes such that $\{x_i \in X \subset R^N\}_{i=1}^{n}$ for the set of n samples. We define the (n x g-1) class membership matrix (Rosipal and Kramer, 2006):

$$Y = \begin{pmatrix} 1_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & \cdots & 0_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_{g-1}} & 0_{n_{g-1}} & \cdots & 1_{n_{g-1}} \\ 0_{n_g} & 0_{n_g} & \cdots & 0_{n_g} \end{pmatrix} \qquad \text{(Equation 4.14)}$$

Where $\{n_i\}_{i=1}^{g}$ denotes the number of samples in each class, $\sum_{i=1}^{g} n_i = n$, 0_{n_i} and 1_{n_i} are (n_i x 1) vectors of all zeros and ones, respectively. Let G represent the between class sums-of-squares and H represent the within class sums-of-squares where:

$$G = \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{\bar{x}})(\bar{x}_i - \bar{\bar{x}})^T$$
$$H = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_i^j - \bar{x}_i)(x_i^j - \bar{x}_i)^T \qquad \text{(Equation 4.15)}$$

Here x_i^j represents an N-dimensional vector for the jth sample of the ith class, and

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_i^j$$
$$\bar{\bar{x}} = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{n_i} x_i^j \qquad \text{(Equation 4.16)}$$

Fisher's method of discrimination involves the linear projection of the input data onto the eigenvectors, μ, of the following eigenvalue problem such that the between class variance is maximized when compared with the within class variance (Rosipal and Kramer, 2006):

$$E^{-1} H\mu = \lambda\mu \qquad \text{(Equation 4.17)}$$

The output of a PLS-DA model is a set of component axes, known as latent variables, which maximize the covariance between the measured data and the class membership matrix (response variables) (Bollard et al., 2001; Rubingh et al., 2006). The set of latent variables represent a new prediction matrix with a greater emphasis on the dynamics of interest. The response variables are regressed onto these new predictors (Rosipal and Kramer, 2006) to obtain a new set of coordinates. These coordinates are known as the samples scores. Similar to PCA, unknown samples may be projected onto the model to determine their most probable class membership (Bollard et al., 2005).

Not all components are used in the model as the higher order components represent noise and are generally left out. The number of components used can be determined from the residual matrix of the response variables, F (Geladi and Kowalski, 1986). A common method is to compare the values of ||F|| at the end of each iteration, and stop when the difference becomes small when compared to a previously established measurement error. An alternative approach is to stop when ||F|| drops below a pre-defined threshold value. If the model is to be used for prediction, the prediction residual sum of squares (PRESS) is used. The number of components is determined from the minimum PRESS which is the sum of the squared difference between the observed and predicted values left out of the model (Lindon et al., 2001; User Guide for SIMCA P+).

PLS-DA should be used for dimensional reduction when the objective is to identify patterns of behaviour responsible for defining pre-defined classes. PLS performs better than PCA in cases where the within group variation exceeds the between group variation, thus allowing for the identification of subtle changes associated with a disease (Barker and Rayens, 2003).

4.5 Model Validation

Multivariate statistical models can be sensitive to chance correlations and there is a risk of overfitting the data (Rubingh et al., 2006; Westerhuis et al., 2008). Emphasis on previously known information could result in modeling 'noise' from the measurements (Weckwerth and Morgenthal, 2005). For this reason, cross validation and permutation testing should be performed to evaluate the performance and stability of a model (Rubingh et al., 2006).

The strength of a model may be evaluated using the goodness of fit, R^2, and the goodness of prediction, Q^2 parameters. The R^2 value relates the fraction of variation accounted for by the regression and describes the accuracy of prediction for that model (Antiewicz et al., 2006). In general, having a large R^2 value (close to 1.0) is attractive, but not sufficient for a good model as it could be overfit (Broadhurst and Kell, 2006; Westerhuis et al., 2008). In contrast, the Q^2 value describes the fraction of variation in the training set (measured data,X for PCA and the class membership data, Y, for PLS-DA) predicted by the model as determined from cross validation (Weljie et al,. 2006; Westerhuis et al., 2008). There is no standard requirement for the R^2 and Q^2 values (Westerhuis et al., 2008), but values greater than 0.5 are preferred (Erb et al., 2008).

4.5.1 Cross Validation

Cross validation (CV) identifies the optimal model parameters and tests the predictability of the model with part of the data. It may be used to determine the optimal number of latent variables used. Having too few components will result in a significant loss of information, while having too many components will overfit the data; the larger latent variables explain more noise (Antiewicz et al., 2006).

The method involves splitting the data into two sets: the training set and the prediction set. The training set is used to develop a model, which is then tested with the prediction set (Rubingh et al., 2006). The idea is that a model created with part of the data should be able to correctly classify the data in the prediction set if there is in fact a unique change between populations. This provides a measure of the predictability for a new data set (Rubingh et al., 2006). A common method is referred to as the 'x fold validation' where the data is divided into x groups of the same size. One group is removed from the data set and a model is created using the remaining (x-1) groups. The procedure is repeated until each group has been left out once and only once (Rubingh et al., 2006). The model is used to predict the values of the response variable for the 'left out' samples and residuals (error) are calculated (Antiewicz et al., 2006).

4.5.2 Permutation Testing

Permutation testing is used to assess the significance of a classification (Rubingh et al., 2006). This technique involves permuting the classification of samples and recalculating the model for each permutation. The idea is that the original classifications should produce the best model. By randomly reassigning samples to a class, it is possible to test for an overfitted model by evaluating the R^2 and Q^2 values. For each permutation, the discrimination between the permuted classes is compared to the discrimination between the original classes (Rubingh et al., 2006). If the original model produces the maximum R^2 and Q^2 values of all models tested, then the model is good. A good model will have the best discrimination between classes, and should also have the largest R^2 and Q^2 parameters of all models. The Q^2 value is the more important of the two as the R^2 value can be optimistic in over fit models (Broadhurst and Kell, 2006).

4.6 Model Interpretation

Multivariate statistical models identify the directions in K-space which best explain the variation in the data set and closely approximate the measured data, but in a lower dimensional plane (Trygg et al. 2007). Samples are projected onto the model plane. The position of the sample in the new plane is defined by the scores, which is a vector relating the projection of the sample along each latent variable (Keun et al., 2004; Trygg and Lundstedt, 2007). The scores may be used to describe variation in the sample direction (Serkova et al., 2008). The loadings relate the influence of individual variables in calculating the scores (Keun et al., 2004). The loadings describe the relationship between variables. Variables with the same sign are directly related (both change in the same direction), and those with larger loadings have a greater impact on the variance along its corresponding latent variable (Trygg et al., 2007). It has been suggested that variables with large loadings are well related, so treating these variables as a single group can simplify interpretation (Lindon et al., 2001).

Similarities in the data structure are best observed in a scores plot (Bollard et al., 2005; Reo, 2002). A scores plot is a 2-dimensional plot of two latent variables, which visually displays the locations of samples under the new coordinate system. Objects located close to one another are said to cluster and will have similar data structures or biological composition (Um et al., 2009). Grouping, trends and outliers are easily observed in this plot (Serkova et al., 2008). Values lying outside the Hotelling's T2 ellipse, which represents the 95% confidence interval, are considered strong outliers, and may impact the model. It is suggested to remove these variables from the analysis (Trygg and Lundstedt, 2007).

The loadings plot may be used to identify variables responsible for the observed differences between samples (Bollard et al., 2005; Murdoch et al., 2008). Variables farther from the origin have the largest impact on the explained variation. The direction of data in the scores and loadings plot are correlated (Trygg et al., 2007). In fact, the direction of the change, to the left or right, corresponds to the change experienced in the respective group (Murdoch et al., 2008). This means that populations located to the right of the origin will have an increase in variables located to the right in the loadings plot and a decrease in variables located to the left. The same is true for samples to the left. By determining the direction of optimal separation, it is possible to identify the variables most responsible for the differences.

Chapter 5
ROC Curve Analysis

This Chapter will provide a theoretical description of Receiver Operating Characteristic (ROC) curve analysis and describe how the curve is constructed. The chapter will open with the motivation for using ROC curve analysis over a single measurement of the sensitivity and specificity. Next, a brief description on how to construct a ROC curve from these measurements will follow. Methods for interpreting the results of ROC curve analysis will conclude the chapter.

5.1 Defining Diagnostic Accuracy

A screening test makes a decision about the presence or absence of disease based on the result of a measurement. The result of this measurement might be a numerical value which has a characteristic range for the healthy population and a different characteristic range for the diseased population, though there will likely be overlap between these two ranges. In a diagnostic situation, clinicians will define a decision threshold to classify the samples (Bewick et al., 2004; Metz, 1989; Obuchowski, 2005). Samples with a measurement that is larger than the decision threshold will be classified as having the disease (positive test), while other samples are classified as healthy (negative test).

One of four situations can occur: 1) a sample from an individual who has the disease is correctly diagnosed (true positive or TP), 2) a sample from an individual who has the disease is incorrectly diagnosed as healthy (false negative or FN), 3) a sample from a healthy individual is incorrectly defined as having a disease (false positive or FP) or 4) a sample from a healthy individual is correctly classified as healthy (true negative or TN) (Weinstein et al., 2005, Zivian and Gershater, 2008). The four cases are summarized in a contingency table, shown in Table 5.1. Here, the observer's decision is indicated by the columns and the results of the test are displayed as rows. It is assumed that the true health status of all individuals is known so that the true positive, false positive, false negative and true negative variables may be calculated (Obuchowski, 2005).

Table 5.1: Contingency Table of Possible Test Results		
	Observer's Decision	
Test Result	Disease Present	Disease Absent
Positive Test	True Positive (TP)	False Positive (FP)
Negative Test	False Negative (FN)	True Negative (TN)

Table 5.1: Contingency Table describing four situations experienced in clinic. A true positive occurs when a diseased individual tests positive, a false positive occurs when a healthy individual tests positive, a false negative occurs when a diseased individual tests negative and a true negative occurs when a healthy individual has a negative test result.

These descriptors do not contain sufficient information for the clinic. For this reason, we define two new terms; the sensitivity and specificity. These descriptors are attractive as they do not vary greatly between patient populations (Zivian and Gershater, 2008). The sensitivity is defined as the number of individuals having a disease, who are correctly diagnosed (positive test), divided by the number of individual having the disease. The specificity is defined as the number of healthy individuals, who are correctly diagnosed (negative test), divided by the total number of healthy individuals (Bewick et al., 2004; Cook, 2008; Hojung et al., 2009; Obuchowski, 2005; Weinstein et al., 2008). In mathematical terms:

$$Sensitivity = \frac{TP}{TP + FN} \qquad \text{(Equation 5.1)}$$

$$Specificity = \frac{TN}{TN + FP} \qquad \text{(Equation 5.2)}$$

The sensitivity may be thought of as a measure of how well a diagnostic test, such as a screening test, will perform in a sample of sick patients. Similarly, the specificity would explain how well a diagnostic test performs in a healthy population (Zivian and Gershater, 2008). The sensitivity and specificity are inversely related (Akobeng, 2007; Gardner and Grenier, 2006, Park et al., 2004, Zivian and Gershater, 2008). This means that by improving the sensitivity of a test, the specificity will decrease. Consequently, it is important to weigh the 'relative costs' of false positives and false negatives (Altman and Bland, 1994; Delong et al., 1988; Obuchowski, 2005).

Diseases with high prevalence and mortality rates require high sensitivities (assuming that treatment is available) at the cost of specificity. Alternatively, conditions with a low prevalence or a risky treatment will use a test with high specificity at the cost of a lower sensitivity (Akobeng, 2006). For screening purposes, it is essential to identify as many sick individuals as possible, so a high sensitivity is required (Park et al., 2004).

5.2 Motivation for ROC Curve Analysis

Consider an experiment that involves the collection of data from two distinct groups; one group is composed of healthy individuals, while the second group has a specific disease. The data collected will create a distribution of measurements for each variable tested. For the purpose of this example, it is assumed that the average measurement from the diseased population is greater than that for the healthy group and the distributions overlap one another. The distributions are assumed to be Gaussian in shape, though this may not be the case for small data sets. An example of two such distributions, for a large sample study, is shown in Figure 5.1.

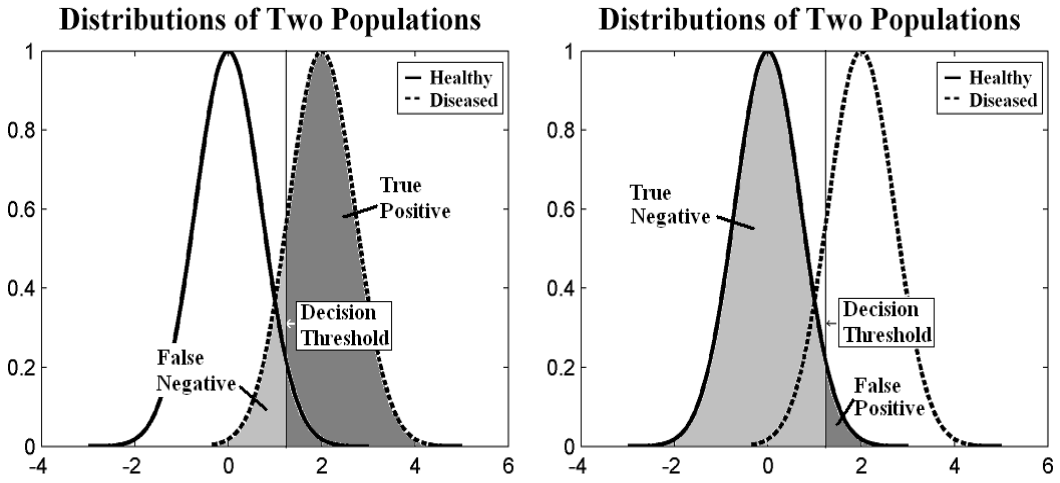

Figure 5.1: Distributions for two populations; one healthy and one having a particular disease. One decision threshold is drawn (vertical line) and the regions corresponding to false positives, false negatives, true positives and true negatives are labelled. The proportions of each are dependent on the location of the decision threshold and the shapes of the two distributions.

The sensitivity and specificity are dependent on the location of the decision threshold. Increasing the threshold to a larger value allows for greater specificity (more true negative results) at the expense of a lower sensitivity (more false negatives). The decision threshold is generally chosen to reflect the relative costs of a false positive versus a false negative, however, it is not obvious which cut point is best (Zivian and Gershater, 2008). For instance, increasing the specificity by a small amount could greatly affect the sensitivity if the distribution of sick individuals is narrow.

The impact of changing the decision threshold is not intuitive when evaluating only the sensitivity and specificity. One solution is to consider the overall accuracy, defined as the total number of correct diagnoses – true positives and true negatives – divided by the number of individuals tested (Zivian and Gershater, 2008). Despite the helpful title, the observer could be biased by the prevalence of a disease.

As an example, consider a disease with low prevalence. The observer analysing test results is likely to use a conservative threshold to minimize the number of false positives. An extreme choice is to label all tests negative. Because of the low prevalence, the test will have a high specificity (100%) and accuracy (large proportion of true diagnoses), but a low sensitivity (0%). If 1 in 10 individuals have the disease, the accuracy would be 90%. On the opposite side of the spectrum, a disease with high prevalence (say 8 in 10) could lead an observer to call all tests positive. This test would have a high sensitivity (100%) and accuracy (80%), but a low specificity (0%). These examples show that even useless tests can achieve strong results for two of specificity, sensitivity and accuracy, but offer little diagnostic information.

5.3 The Receiver Operating Characteristic (ROC) Curve

Receiver Operating Characteristic (ROC) curves were first developed in the 1950's to evaluate radar signal detection (Altman and Bland, 1994). The goal was to quantify how well an electronic receiver detects electronic signals in the presence of noise (Zivian and Gershater, 2008). Lee Lusted realized that ROC curve analysis held potential in medical applications and introduced the technique to medicine in 1971 (Centor, 1991; Hilden, 1991). Since its inception, ROC curve analysis has become a standard method for evaluating a diagnostic test's accuracy (Akobeng, 2006; DeLong et al., 1988; McClish, 1989; Wang et al., 2005) by jointly evaluating the fraction of positive results in each population (Gardner and Greiner, 2006). Analysis of the resulting curve allows scientists to evaluate tradeoffs between the sensitivity and specificity at different thresholds (McClish, 1989; Pepe, 2000). In addition, this technique is independent of the prevalence, so we avoid the issue presented in the previous section (Park et al., 2004; van den Biggelaar et al., 2009, Wang et al., 2005).

An ROC curve is a plot of the sensitivity along the y-axis and (1-specificity) or the false positive fraction along the x-axis (Akobeng, 2006; Obuchowski, 2005; Park et al., 2004; Pepe, 2000). ROC curves are not dependent on a single decision threshold as

all options are considered (Cook, 2008; DeLong et al., 1988; Weinstein et al., 2005). Each point on the graph, known as an operating point (Park et al., 2004), represents the sensitivity and specificity for a different decision threshold (Altman and Bland, 1994; van den Biggelaar et al., 2009). The empirical curve is created by connecting the operating points with straight line segments (Obuchowski, 2005; Park et al., 2004). The smooth ROC curve is bowed towards the top left hand corner – point (0,1) – therefore, the empirical curve will underestimate the true area under the curve. However, the estimation is quite accurate for continuous distributions (Obuchowski, 2005).

ROC curves all pass through the points (0,0) and (1,1), marking the most conservative and strictest decisions thresholds respectively (Obuchowski, 2005). There are (h-1) additional points on the curve to represent each of the h unique decision thresholds (Obuchowski, 2005). The operating points can be thought of as observations from h Radiologists or as the measurements made by one Radiologist as they change their threshold from the most conservative (all test results are positive) to the strictest (all test results are negative) (Zivian and Gershater, 2008).

There are three main classification systems: 1) a 5-point system, 2) a confidence scale system and 3) continuous distributions (Obuchowski, 2005). The first choice is often used to evaluate the diagnostic potential of an imaging modality, such as mammography or MR imaging (Akin et al., 2010; van den Biggelaar et al., 2009). The radiologist will assign images a BI-RADS (Breast Imaging Reporting and Data System) score based on the degree of suspected malignancy (1 = negative examination, 2 = benign finding, 3 = probably benign finding, 4 = suspicious abnormality and 5 = highly suggestive of malignancy) (van den Biggelaar et al., 2009). The confidence scale system operates in a similar fashion, except that the Radiologist scores the exam based on the probability of malignancy (i.e. 0-100%) (Obuchowski, 2005). The continuous distributions test allows users to systematically update the decision threshold across all possible cut-offs (DeLong et al., 1988; Gardner and Greiner, 2006). ROC analysis does not work on binary (yes/no) tests as these only have one cut point (Park et al., 2004).

Drawing a straight line segment between the points (0,0) and (1,1) creates the chance diagonal. This line represents a test that has no ability to discriminate between individuals from two distinct populations and randomly assigns a diagnosis (Park et al., 2004; Weinstein et al., 2005). The random assignment means that there are an equal number of true positives and true negatives for all thresholds considered (Akobeng,

2006). ROC curves that lie in the upper left-hand corner of the plot contain some diagnostic information (Weinstein et al., 2005).

Diagnostic information may be measured using one of the following summary measures:

1. Youden's index: Calculates the maximum vertical distance between the ROC curve and the chance diagonal (Bewick et al., 2004). This parameter is dependent on the decision threshold and may be calculated using:
   $$J = \max\{\text{sensitivity} + \text{specificity} - 1\}$$ (Equation 5.3)
2. Accuracy or the probability of correct diagnosis: This parameter utilizes information about the prevalence of the disease (Prev) and will change as the prevalence changes (Obuchowski, 2005). The accuracy is calculated from:
   $$\text{Accuracy} = \text{Prev} \times \text{sensitivity} + (1 - \text{Prev}) \times \text{specificity}$$ (Equation 5.4)
3. The area under the ROC curve (AUC): The AUC is a combined measure of the sensitivity and specificity and is perhaps the most popular choice (Cook, 2008; McClish, 1989; Obuchowski, 2005; Park et al., 2004; Zivian and Gershater, 2008). Values range between 0.5 and 1.0 for diagnostically useful tests.

The AUC may be interpreted as 1) the average sensitivity for all values of specificity, 2) the average specificity for all values of sensitivity, or 3) the probability of correctly diagnosing two individuals – one healthy and one having the disease – based on a single measurement (i.e. concentration of a metabolite) (Obuchowski, 2005). The third choice is independent of the prevalence as it is fixed at 50% by definition (Gardner and Greiner, 2006). Despite this, the third definition is intuitive in the sense that samples are essentially ranked prior to calculating the ROC curve.

The AUC for a diagnostic test is bounded below by 0.5 for tests with no ability to distinguish two populations, and above by 1.0, which represents a test with perfect discrimination (Zivian and Gershater, 2008). There is no single standard for rating the effectiveness of a diagnostic test. Some applications define a test with an AUC > 0.9 as being an excellent test and an AUC > 0.8 as a good test (Broadhurst and Kell, 2006), while Akobeng (2006) and Gardner and Grenier (2006) propose using:

AUC > 0.9	High Accuracy
AUC > 0.7 and AUC ≤ 0.9	Moderate Accuracy
AUC > 0.5 and AUC ≤ 0.7	Low Accuracy
AUC = 0.5	Chance

The AUC contains information for all sensitivity and specificity pairs available. This may not be sufficiently informative for a diagnosis. The ROC curve offers a graphical representation of the tradeoffs between the sensitivity and specificity as the decision threshold is altered (Gardner and Grenier, 2006). The optimal point on the ROC curve may be determined through the Youden index – equation 3.3 –or from the point closest to the top left-corner of the plot (0,1) by minimizing the following expression (Akobeng, 2006):

$$\text{Optimal Point} = \min\{(1-\text{sensitivity})^2 + (1-\text{specificity})^2\} \qquad \text{(Equation 5.5)}$$

There is a possibility that the Youden index and the distance from the point (0,1) provide different optimal decisions thresholds. One example where the two techniques could produce different results is shown in Figure 5.2. Here, The ROC curve is parallel to the chance diagonal, but has an artefact on one side. The minimum distance from the point (0,1) to the ROC curve differs from the optimal point given by the Youden index (maximum vertical distance from the ROC curve to the chance diagonal) which occurs as the artefact. .

The AUC may be estimated non-parametrically by fitting trapezoids under the empirical ROC curve (Bewick et al., 2004; Metz, 1988; Obuchowski, 2005; Park et al., 2004). The estimate tends to underestimate the AUC (Metz, 1988), however it produces results that are similar to those obtained using a parametric approach (assuming two normal population distributions and fitting a smooth curve to the data) when the variables are continuous (DeLong et al., 1988; Park et al., 2004).

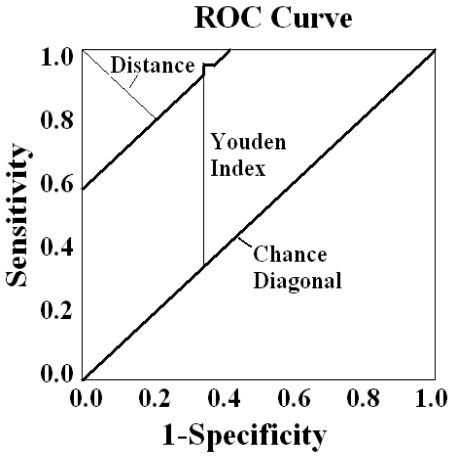

Figure 5.2: Example where the Youden index and minimum point from (0,1) give different optimal decision thresholds. A ROC curve where the region around the point (0,1) is approximately linear and has the same slope as the chance diagonal, however there is a small feature near the line Sensitivity = 1.0. The Youden Index and minimum distance techniques produce different results for the optimal ROC point.

5.4 Predictive Values

Positive and negative predictive values measure the predictive capabilities of a diagnostic test with consideration of the prevalence of the disease (Cook, 2008). The positive predictive value (PPV) is the probability of a disease given a positive results and a negative predictive value (NPV) is the probability of no disease given a negative result (Bewick et al., 2004; Gardner and Greiner, 2006; Weinstein et al., 2005). As suggested by Weinstein et al., (2005), the predictive values help clinicians plan treatments depending on the probability that the individual does or does not have the disease based on the test result.

There are two formal expressions used to calculate the predictive values, depending whether or not the prevalence of the disease is known. The expressions are as follows:

$$PPV = \frac{Se \times P}{Se \times P + (1 - Sp) \times (1 - P)} = \frac{TP}{TP + FP} \qquad \text{(Equation 5.6)}$$

$$NPV = \frac{Sp \times (1 - P)}{(1 - Se) \times P + Sp \times (1 - P)} = \frac{TN}{TN + FN} \qquad \text{(Equation 5.7)}$$

Where Se is the sensitivity, Sp is the specificity and P is the prevalence (fraction of those having the disease in the sample population) (Gardner and Greiner, 2006; Weinstein et al. 2005). Low prevalence results in a decreased PPV (Bewick et al., 2006). Scientific experiments tend to bias the prevalence to greater values to improve the statistical power of the study; therefore, the latter equality is preferred.

5.5 Limitations of ROC curve Analysis on Small Populations

Analysis on small populations could bias the results if the experimental population does not accurately reflect the desired population (not independently sampled). According to Metz (1988), the variance in the ROC curve is inversely proportional to the number of samples in the study. Individual samples will either over or under estimate the true ROC curve data points, but will create a distribution of ROC curves.

Chapter 6
Experimental Techniques

This Chapter will discuss the methodologies associated with the experimental work of this dissertation. The chapter organization follows the sequence of steps from urine collection to the final data analysis. All work involving animals was approved in advance by the Cross Cancer Institute's Animal Care Committee. Animals were monitored daily for signs of distress throughout the experiment.

6.1 Tumour Model

NIH III nude mice (n = 36) were ordered at 6 weeks of age from Charles River Laboratories, USA. Animals were housed in the Cross Cancer Institute's Vivarium with four mice per cage. The Vivarium allowed for a controlled experimental environment with a temperature of 21^{o}C, relative humidity 30-55% and a 12 hour light – 12 hour dark cycle (6am – 6pm light, 6pm – 6am dark). Animals were fed a diet of laboratory autoclavable rodent diet 5010 (Labdiet, Leduc, Canada) and were given free access to autoclaved tap water.

Animals were given three days to acclimatize to the new facility before notching their ears for identification. During the procedure, the mice were anesthetised with 1.5%-2.0% isoflurane mixed with oxygen. A sterile ear clipper punched a hole in one ear.

The mice were whole body irradiated to 350cGy in a ^{137}Cs irradiator one week post arrival to minimize any residual immune response to the tumour cell injections (Laroque et al., 2009). Tumour cells were injected one week later.

The human GBM cell line M006xLo was used to initiate tumours (Franko et al. 1998). Three million cells, suspended in 0.1mL of phosphate buffered saline (PBS), were injected subcutaneously over the animal's right thigh (n = 22) (Laroque et al., 2009). The mice were anesthetised with 1.5%-2.0% isoflurane in oxygen during the injection. The tumours became palpable 3-4 days post injection and were visible to the eye by 5 days. Control animals (n = 14) were injected with 0.1mL PBS in place of the GBM cells.

6.2 Sample Collection

Sample collection began once the animals were whole body irradiated and continued for five weeks post cell injection. To minimize the effects of diurnal variation (Tiziani et al., 2009), urine samples were collected into Eppendorf safe-lock microfuge tubes (Sigma-Aldrich Co, Oakville, Canada) daily between 10:30am and 12:00pm by restraining the mouse and gently massaging its bladder. Samples were immediately stored in a freezer at -80^{o}C until required for NMR analysis. The literature suggests that immediate freezing at -80^{o}C is sufficient to prevent bacterial contamination of the urine samples (Lauridsen et al., 2007). The longest samples were stored before NMR spectrum acquisition was 7 months, but most samples were analysed within 2 months post collection. According to Ross et al. (2007), samples are not affected for up to 9 months when stored at -40^{o}C, so the time frame of storage should have minimal effects on the metabolic composition. Samples selected for NMR analysis were required to exceed 30μL in volume as this appeared to be the lower limit of detection for NMR data acquisition.

For the purpose of our analysis, samples were categorized as either pre-injection or tumour-bearing (collected between 6 and 35 days post-injection). Not every mouse contributed a sample each day, and some samples had insufficient volumes for inclusion in the study. Consequently, the number of analysed samples in the pre-injection group varied from 1-6, and from 1-7 for the tumour-bearing category. One cage (4 mice) was injected with the GBM cells one week late (three weeks post arrival). The data from these animals were scattered amongst the data from the other animals in the scores plots of the preliminary PCA and PLS-DA models, indicating that the extra week had no observable effect on metabolism, so these animals were included in the study. The number of pre-injection samples for these mice ranged between 5 and 8. The tumours continued to grow over the course of the five week urine collection period, but even at the earliest time points, the tumour was macroscopic in size.

6.3 Sample Preparation for NMR Analysis

Selected samples were removed from the freezer and allowed to thaw at room temperature for approximately one hour. Available urine, ranging between 30 and 150μL, was pipetted (Ranin, Mettler Toledo Co, Oakland, USA) into a 3000kDa

Nanosept filter (VMR International, Mississauga, Canada) with the volume brought to 500μL with distilled water (Keun et al., 2004; Ross et al., 2007). Proteins and larger macromolecules were removed from the sample by centrifuging the samples at 4°C for 10 minutes at 13000 rpm (Forshed et al., 2005; Holmes et al., 2000; Maher et al., 2007). Samples were prepared for NMR analysis by adding 90μL of an internal standard (Chenomx, Inc. Edmonton, Canada) which contains 5mM DSS and 0.2% NaN_3 in D_2O at a pH of 7.0. 200μL of distilled water was added to the filtered samples to bring them to an appropriate volume (~650μL ± 50μL).

The pH was measured using an Acumen Microprobe pH meter (Fisher Scientific, Ottawa, Canada). The meter was calibrated using pH 4.01 and pH 10.00 buffer solutions. Initial pH values for the urine samples were most often between 6.20 and 6.75. The pH was adjusted to 6.84 ± 0.04 by adding small volumes of 0.1M HCL or Noah (Kim et al. 2008; Murdoch et al., 2008; Slupsky et al. 2007; Um et al., 2009).

A 600μL aliquot of sample was extracted into a Wilma 535-pp NMR tube (Sigma-Aldrich Co., Oakville, Canada) and stored in a fridge at 4°C overnight (Murdoch et al., 2008; Slupsky et al., 2007). This volume is highly recommended as it permits good water suppression without over diluting the sample (McKay 2009).

6.4 NMR Analysis

Samples were run on an Oxford 2.2K 800 MHZ NMR Magnet (Oxford Instruments, Inc., Oxfordshire, UK) utilising a 5mm HCN cold probe (Varian Inc., Palo Alto, USA). Spectra were acquired using a 1-D NOESY pre-saturation pulse sequence with a saturation delay of 990ms, a mixing time of 100ms and an acquisition time of 4s. Four steady state scans were performed prior to data acquisition. 32 scans were acquired for a total run time of 3:03 per sample. The spectra cover a width of 12ppm. These timings are required for accurate quantification when using Chenomx Software (Chenomx NMR Suite 5.1 User guide).

The 1-D NOESY pre-saturation pulse sequence is displayed in Figure 6.1. This pulse sequence is often used for metabolomics experiments because it provides satisfactory water suppression and allows for high throughput (Beckonert et al, 2007; Betram et al., 2007; Bollard et al., 2001). Spectra are acquired using the Varian VNMR software.

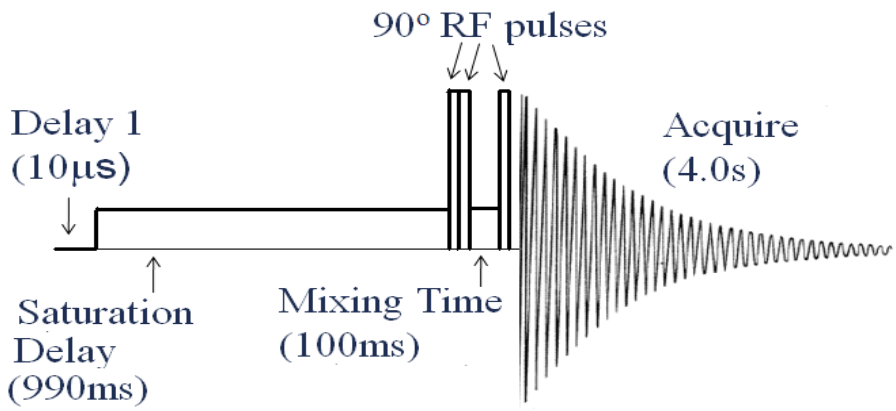

Figure 6.1: Schematic of the 1-D NOESY pulse sequence. The saturation delay and mixing times are constant for all experiments. The length of the 90° RF pulses were optimized by first determining the time required for a 360° pulse (acquire a null signal), and dividing this result by 4. A total of 4 steady state scans and 32 transient scans were used for most samples for a total time of 3:03.

All samples were verified to have volumes of 600±25 μL using a gauge scale. Briefly, the volume was measured by means of the depth of the meniscus in the sample. A reference table, located on the wall at NANUC, provides a gauge value for centering the sample. Using this information and a reference cylinder, the bottom of the tube is aligned with the gauge value. Figure 6.2 shows a schematic for centering a 600μL sample. A sample of this volume has a reference value of 66.

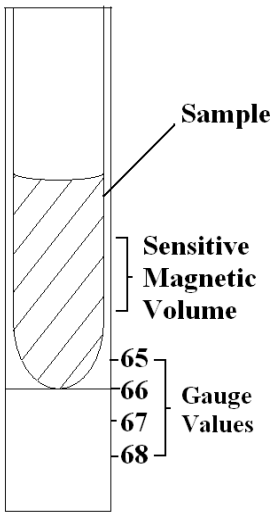

Figure 6.2: Centering the sample. The sample is centered in the magnet by measuring the height of the meniscus and identifying the gauge level. For a sample with a volume of 600 μL, the gauge value is 66. Aligning the bottom of the tube with the gauge at 66 will center the sample in the magnet. If done properly, there should be equal amounts of sample above and below the magnet's sensitive volume.

The sample is inserted into the magnet and allowed to reach an acquisition temperature of 25^{o}C (Bales et al., 1984; Kim et al., 2008; Saude et al., 2006). The first sample of the day was used to calibrate the lock signal. The flip angle for the water resonance is optimized by varying the pulse width of the RF pulses until a null signal is obtained. The null constitutes a 360^{o} flip, so dividing the time by 4 provides the optimal pulse width for excitation of the water resonance.

Next, an optimizing algorithm is applied to optimize the z-gradient shims. These shims are not optimized, but only require minor adjustments to the z1, z2, x and y gradients. The width of the methyl peak of DSS at half maximum is often smaller and the peak shape is more symmetric than obtained with manual shimming; this improves the spectral resolution. The change in the peak shape after manually shimming the field is shown in figure 6.3 a and b.

After shimming, the transmitter offset (tof) and saturation frequency is optimized. This involved setting up an array of values between -294 Hz and -288 Hz and visually inspecting the value at which the water peak is split down the middle. The magnet is tuned appropriately when the tails of the water peak on either side have equal magnitudes. This is shown in figure 6.3 c and d.

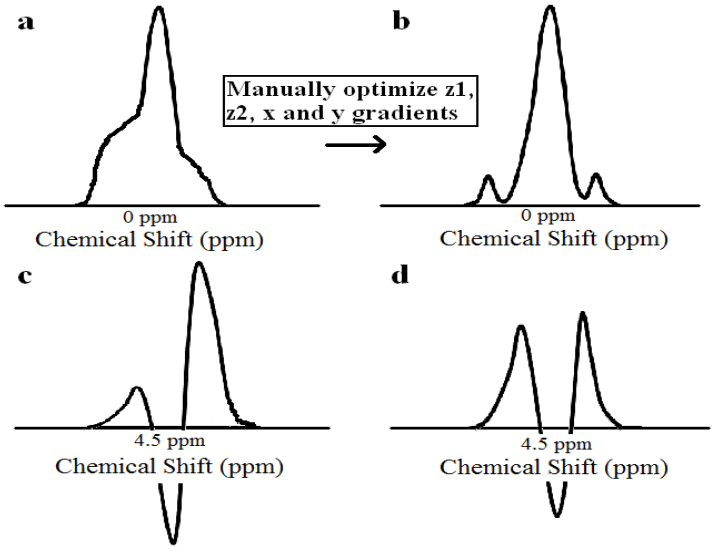

Figure 6.3: Shimming the magnet and optimizing the transmitter offset (tof). a shows the methyl peak of DSS after application of a shim optimization algorithm. It is observed that the peak shape is not optimal, so the field must be manually shimmed using the z1, z2, x and y gradients. The optimized peak shape is shown in b. c and d show the effects of changing the tof. In c, the tof is set too low, such that the right hand tail has a greater amplitude than the left hand tail. The tof has been optimized in d, where both tails are approximately the same height.

The receiver gain was initially set to 18 dB for all scans and decreased to 12 dB when an ADC overload occurred. One study found that dynamic changes in the receiver gain have a negligible effect on the results due to the long word length of the analog digital converter (Malz and Jancke, 2005).

Prior to acquisition, phasing adjustments were made to the entire spectrum to flatten the spectrum baseline on either side of the water resonance. Magnetic field shimming was performed on the methyl resonance peak of the DSS internal standard at 0.0ppm (Saude et al., 2006). The magnetic field was shimmed using the z-1, z-2, x and y gradients to achieve a line width of less than 1.0 Hz. In most cases, the peak width at 0.55% and 0.11% peak height were less than 12 Hz and 20 Hz.

Water suppression on dilute samples was poor, and often resulted in an ADC overload. In this case, the receiver gain was changed from 18dB to 12dB and the number of scans increased to 64. These values were chosen as a compromise between spectrum quality and time of acquisition. The acquisition time for these experiments was 5:47.

6.5 Metabolite Profiling

Identification and quantification of metabolites was carried out with the Chenomx Suite 5.1 (Chenomx Inc., Edmonton, Canada) software package. Quantification is achieved using targeted profiling in which mathematically modelled pure compound NMR resonances from the Chenomx library (pH 6-8) are fit to the acquired spectra (Weljie et al., 2006). A detailed procedure for spectral processing follows. Complete details can be found in the Chenomx Suite 5.1 user guide available at www.chenomx.com. A DSS concentration of 0.125mM is used for all samples.

6.5.1 Spectrum Processing

The 'Chenomx Processor' prepares the NMR spectrum for profiling by correcting phasing and baseline artefacts (Chenomx NMR Suite 5.1 User manual). Upon opening the spectrum, the user must input the sample pH and decide whether or not to apply line broadening. Entering the pH will help with profiling metabolites, such as citrate, which are known to have a pH dependent chemical shift and peak shape (Miyataka et al., 2007; Weljie et al., 2006). For the purpose of our experiment, we chose

to apply line broadening at a later stage. Briefly, the steps applied are phasing, baseline correction, line broadening, reference deconvolution and water deletion. The initial spectrum is shown in Figure 6.4a.

Proper phasing and baseline corrections are essential for accurate profiling (Wishart, 2008). Phasing the spectrum compensates for the missing data points at the start of the free induction decay (FID) (Szantay, 1992). The software automatically applies phase adjustments to the spectra by applying the automatically-determined phase angles. Manual changes must be performed to optimize the phasing adjustments. This includes zeroth order corrections to the phasing of the methyl resonance of DSS and first order corrections to the spectra in the approximate range of 7.0 to 9.0 ppm. The baseline should be symmetric on either side of the water resonance and have a smooth appearance throughout (Vitols, December 2006). After proper phasing, the spectrum will look similar to that shown in Figure 6.4b.

Baseline correction removes baseline distortions resulting from incomplete water suppression (Wishart, 2008). The spline function applies a cubic spline baseline correction based on automatically-determined breakpoints (Cheng et al., 2007). Small adjustments to the spline baseline are made near the urea and water resonances to better fit the actual baseline and to remove very wide signals (line widths of the order of 1 ppm). The effect of baseline correction is shown in Figure 6.4 b and c.

Line broadening was applied to the spectrum to obtain a final approximate peak width of 1.25-1.30 Hz for the methyl resonance at 0.0ppm. The width of 1.25-1.30 Hz provides the closest fit of the modelled compounds to the spectra (Chenomx NMR Suite 5.1 User Guide). Line broadening is performed by applying an exponential FID weighting function to the data which aids in the removal of noise at the tail end of the FID and emphasises the information at the early time points (Szantay, 1992). Line broadening with an exponential weighting factor results in an artificial damping of the FID so that the T_2 values of the metabolites appear faster. Line broadening will improve the accuracy of integration as more data points are used to define the peak shape (Szantay, 1992) and instrumental noise is averaged.

Reference deconvolution was applied to the methyl resonance (including the DSS satellites) to a line width of 1.30 Hz. This step reconstructs an ideal spectrum by removing lineshape distortions, resulting from field inhomogeneities and improper shimming, based on the shape of the methyl resonance of DSS (Vitols and Mercier,

2006). The method involves the deconvolution of the entire spectrum with the experimental line shape of DSS, and then reconvolves the spectrum with an ideal lorentzian line shape (Vitols and Mercier, 2006). Reference Deconvolution is a linear process in that it only uses direct and inverse Fourier transforms to maintain the quantitative relationships between compounds in the spectrum. Reference deconvolution assumes that all lineshapes are systematically distorted in the same way. It is useful for reducing the effects of varying shimming techniques between users and to improve the quantitative accuracy of compounds in regions with heavy overlap of signals. Figure 6.4 d and e shows how the methyl peak is affected by application of reference deconvolution.

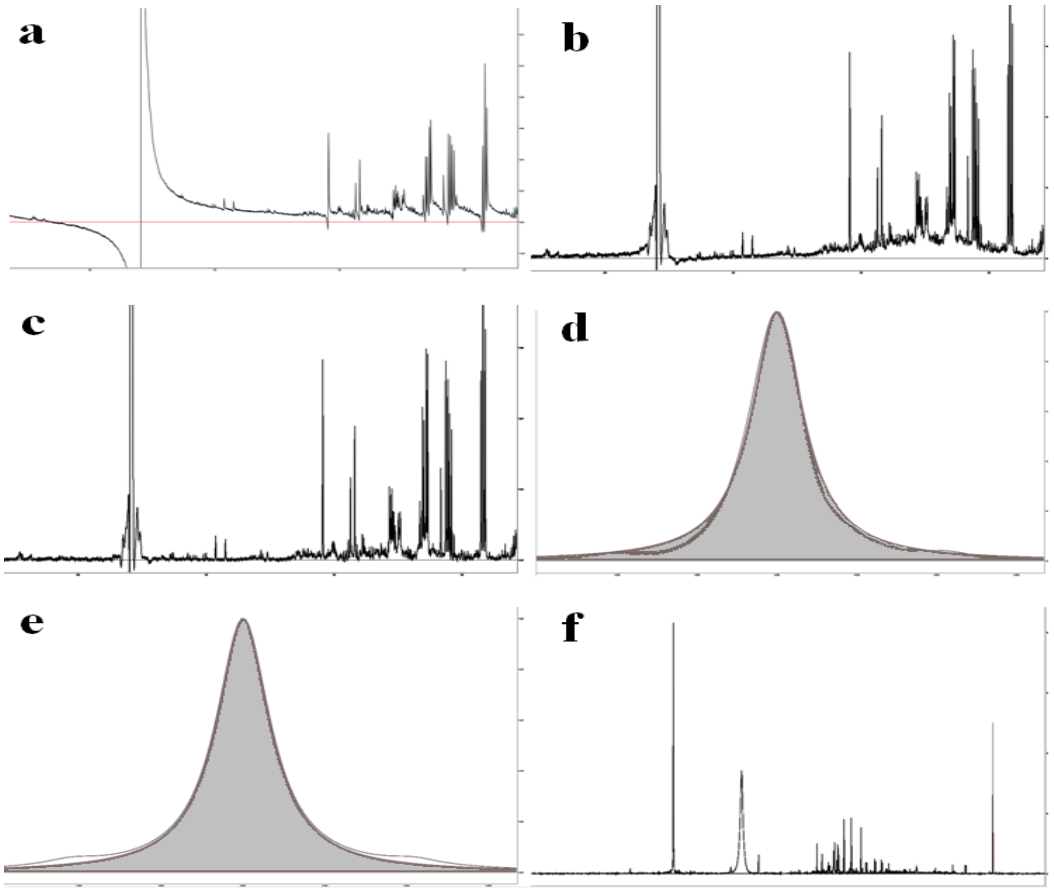

Figure 6.4 Procedure for pre-processing the spectra in Chenomx a, b and c all show the same region of the spectra around the water resonance (~4 – 4.8 ppm). d and e show only the methyl peak of DSS at 0 ppm. a. Initial spectrum when imported into Chenomx. b. After correcting for phasing errors. c. Post baseline correct. d. Methyl peak of DSS before reference deconvolution. e. Methyl peak of DSS after reference deconvolution. f. Final spectrum after processing.

Water deletion removes the sections of the spectrum covered by the water resonance. Removal of the water resonance is not required, but will remove the sections of the spectrum near the water resonance which are distorted from the incomplete water suppression. These metabolite peaks are not profiled because they may also be suppressed (McKay, 2009). The water resonance was removed from all samples for consistency. The final spectrum, after water deletion, is shown in Figure 6.4f.

6.5.2 Metabolite Profiling

Metabolite identification and quantification is performed using 'the Profiler' software. This software includes a library of mathematically modelled NMR resonances for pure metabolites for a specific pH range (Saude et al., 2006). Profiling involves matching the compounds NMR signatures to the patterns found in the spectrum. Figure 6.5 shows the steps in profiling a metabolite. Details of the procedure follow.

There are three important lines displayed in the Profiler: the black line represents the acquired spectrum, the red line is the addition line which adds up the contributions of all metabolite peaks at each chemical shift, and the green line is the subtraction line which calculates the difference between the spectrum and the addition line. Quantitative accuracy is greatest when the subtraction line is used for profiling. A modelled peak is manually adjusted, amplitude and chemical shift, to best fit the spectral peak. This is achieved by obtaining a subtraction line with zero magnitude. The amplitude of the profiled peak is directly proportional to the concentration of that metabolite. In effect, increasing the amplitude means having a higher concentration of the metabolite in that sample. Figure 6.5a shows the creatinine singlet that has only been partially profiled. It is evident that the profiled concentration is incorrect as the addition line (red) does not encompass the area under the peak and the subtraction line (green) is not flat. In Figure 6.5b, the metabolite has been correctly profiled. This time, the subtraction line is flat.

Metabolites are known to have pH dependent chemical shifts, so the modelled peaks are free to move to higher or lower chemical shift locations within a pre-defined region (Vitols and Rosewell, March 2006). The resonance peak may be shifted to any chemical shift within this region, but is bound by an upper and lower limit set by the software. The software allows users to adjust the chemical shift of each resonance cluster individually (Vitols and Rosewell, March 2006). In general, peaks originating from one metabolite will shift in the same direction, but the magnitude may vary between clusters.

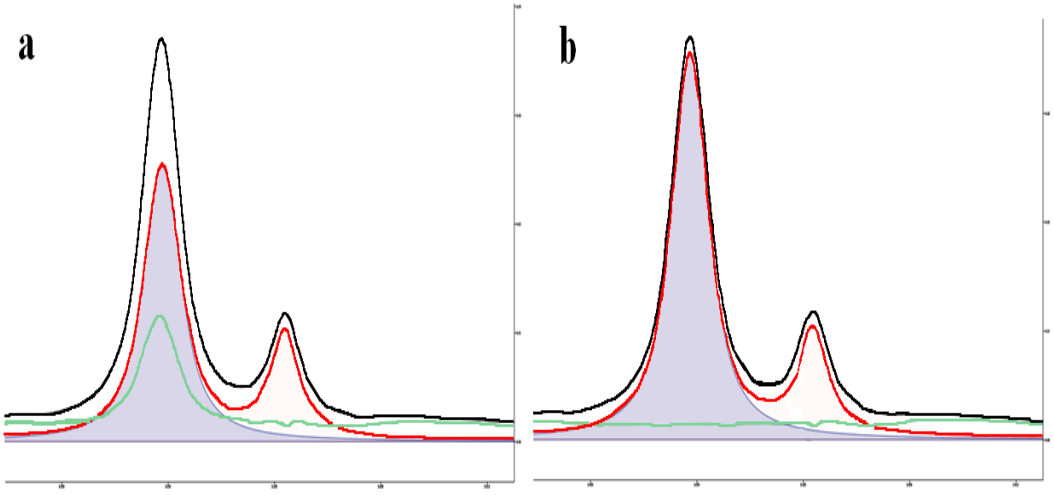

Figure 6.5 Profiling the Creatinine peak at 3.0ppm. This figure shows the three lines used for profiling: the black line is the NMR spectrum signal, the red line is the addition line which shows the sum of all metabolites profiled at the chemical shift, and the green line is the subtraction line which is used for profiling. When it is flat, the fit it good. a. Bad fit as the subtraction line is not flat. b. Good fit.

A library with common urinary metabolites was first used for profiling. This library contains a list of 80 metabolites, though not every metabolite was found in the mouse urine. After profiling these metabolites, a second library, containing all metabolites available, was selected to profile the remaining peaks. These libraries do not contain all metabolites found in urine, but those present were identified and quantified. Each metabolite was selected individually and fit to the spectrum. Metabolites that did not have an obvious fit to the spectrum were assumed to be absent, and their concentrations reset to 0 μM.

Identification and quantification was performed by adjusting the concentrations and chemical shifts of the modelled peaks to fit the spectrum. Metabolites that were fully resolved from nearby peaks and located far from the water and urea resonances (outside 3 and 7 ppm) were profiled first. Next, metabolites with a prominent peak in a crowded area of the spectrum (generally a singlet) were fit. All possible metabolites with a chemical shift in the region were tried to verify that the correct metabolite was profiled. Metabolites that exhibit more than one resonance were fit to each cluster individually. Quantification for these metabolites was performed on a well-resolved, prominent peak far from the water resonance.

Some metabolites are consistently found close together. To profile these metabolites, previous knowledge about the relative chemical shifts for each resonance was utilized. For example, lactate and threonine have a structural similarity, which causes their methyl signal to overlap (Vitols and Fu, April 2006). However, lactate is often found down-field (higher ppm value) of threonine. Another example is between creatine and creatinine; creatinine is often found further down-stream and is generally the larger of the two (Vitols and Fu, April 2006). Such pairs of metabolites were fit concurrently to ensure an accurate measurement for both.

A total of 43 metabolites were profiled, but only 34 were used in the analysis. Glycerol and DSS were removed as they came from external sources. Urea is not quantitative since its protons exchange with water. As a result, the intensity of the peak is strongly dependent on the quality of the water suppression (Ross et al., 2007). Allantoin is found near the urea resonance and is in chemical exchange with urea. This causes broadening and variable attenuation of the two peaks, making quantification unreliable. The glycolate and methylamine resonances are singlets found in a crowded portion of the spectrum. Quantification is complicated by the fact that more than one prominent singlet is present in the area. The chemical shifts of these singlets vary between spectra and overlap at times. Only one singlet in the area was profiled. Since our knowledge of the other peaks is limited, we cannot assume that the profiled resonance is always up (or down) stream from the others. For this reason, we chose to remove these metabolites from the analysis. Phenylacetylglycine, pyroglutamate and tyrosine are all low concentration metabolites found in a crowded area of the spectrum. Since they are near the noise level, quantification is dependent on the user's judgement, and thus eliminated.

6.6 Data Analysis

6.6.1 Preparation of Data

Metabolite concentrations (n=34) and metabolite ratios (n=561) were used in the analysis. The metabolite concentrations must be normalized prior to analysis to bring all samples to the same approximate concentration. Three methods are commonly used in metabolomics studies: integral normalization, normalization with respect to creatinine

and quotient normalization (Ross et al., 2007). In contrast, the metabolite ratios require no normalization since all metabolites in the sample are equally affected by dilution.

Integral normalization is the standard normalization of biofluids (Ross et al., 2007). This technique assumes that the integrals of the spectra are a function of only the overall concentration of the samples. Changes in the concentrations of individual metabolites are assumed to be small compared to changes associated with overall concentrations. In particular, the up-regulation of metabolites should balance the down-regulation of other metabolites to maintain a constant integral area. The integral normalization procedure sums the areas of all variables (bins or metabolites) and divides each variable by the sum. Most groups choose to remove the areas around the water and urea resonances (Ross et al., 2007). A number of variations of integral normalization are found in the literature. While some groups keep the total area at unity (Lenz et al., 2004; Maher et al., 2007; Parsons et al., 2007; Potts et al., 2001), others choose to multiply by 100 for a total area of 100 (Keun et al., 2004). This method works well when the relative change in metabolite concentrations is small. However, it is not robust when these changes are influenced by the alteration in the concentration of another metabolite or when one metabolite change significantly dominates. Under these circumstances, the spectra do not scale correctly which makes the technique inappropriate for metabolomics experiments (Dieterle et al., 2006).

Creatinine normalization is similar to integral normalization, except that that the integral of the creatinine peak is used in place of the overall integral. In simpler terms, the ratio of each metabolite with respect to creatinine is studied. The technique, originating from clinical chemistry, is commonly used for studies involving human or animal urine samples (Kim et al., 2008; Tyburski et al., 2009; Um et al., 2009). It is assumed that creatinine excretion into the urine is constant. Difficulties with this technique involve accurate quantification of the creatinine peak when it is overlapping with another resonance, pH dependent chemical shifts for the peak at 4.05 ppm (when using binning) and metabolic responses that alter the excretion of creatinine. The excretion of creatinine is known to be influenced by the mass of muscle tissues in the body (Stretch and Baracos, in press). Muscle loss, related to cancer, is a well established effect (Stretch and Baracos, in press), so creatinine normalization should not be used for cancer-related studies.

Quotient normalization assumes that changes in the concentrations of single metabolites influence part of the spectrum, whereas changes in the overall concentration influence the entire spectrum (Dieterle et al., 2006). A most probable quotient between the spectrum and a reference spectrum is determined. The entire spectrum is then normalized with respect to this quotient. The reference spectrum can be a 'golden' reference spectrum or the calculated mean or median spectrum from the controls. It has been shown that the choice of reference spectrum is not critical, but the median spectrum from the controls is the most robust for studies with only a few animals. It is recommended to remove all areas of the spectrum that are dominated by noise and perform an integral normalization prior to quotient normalization to bring all spectra to the same absolute magnitude (Ross et al, 2007). One study showed that the normalization method works better than integral normalization for control animals where only small changes in metabolite concentrations occur (Dieterle et al., 2006). This normalization technique was chosen for its robustness.

The total area under each spectrum was normalized to an arbitrary value of 100 following the steps for integral normalization. A median spectrum was created by calculating the median value of the fractional area occupied by each of the 34 metabolites of interest from all of the pre-injection spectra. The distribution of quotients was determined by dividing each spectrum by the reference spectrum. The most probable quotient for a particular sample was chosen as the median quotient in this distribution. An example of a distribution is shown in Figure 6.6. Here, it is evident that most quotients fall in the range between 0.388 and 3.58. For this example, the most probable quotient is 1.03. The concentrations for all metabolites in each sample were scaled with respect to this value.

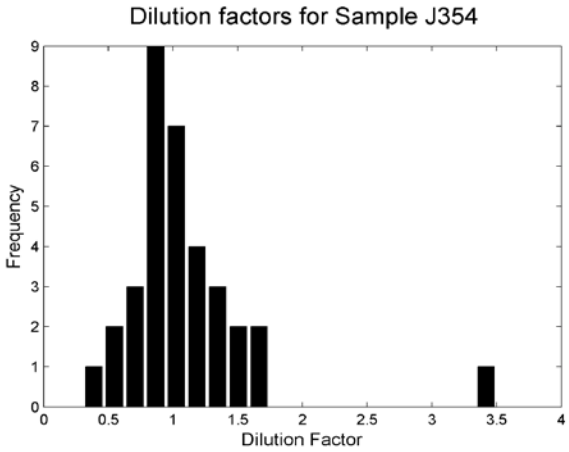

Figure 6.6: Distribution of quotients for spectrum normalization. Quotient normalization requires that the most probable quotient is determined for each sample. The distribution of dilution factors shown here suggests that the most probable quotient for the sample is around 1.0. The median quotient was selected as for normalization of the sample and is equal to 1.03.

Metabolite ratios were calculated for all pairs of metabolites. The smaller of the reciprocal ratios (i.e. A/B vs. B/A) was chosen for analysis to minimize the shift in the ratio as a result of measurement errors in the reported concentration for low concentration metabolites. For instance, metabolites at low concentration will generally have a larger relative error as a result of the numerical resolution of the software. Having this metabolite in the denominator will have a significant impact on the magnitude of the ratio and could bias the calculation of the mean and standard deviation. This effect is smaller when this metabolite is in the numerator. This choice will place greater emphasis on the higher concentration metabolites, which are often measured with greater accuracy.

The mean and standard deviation for each metabolite or ratio was calculated for the pre and post injection samples. Observations that were more than three standard deviations from the mean were defined as outliers and removed from the analysis. The mean and standard deviation for all variables for each mouse were calculated from the contributing samples. Averaging the samples by mouse decreases noise in the spectra related to metabolic variation in the animals and instrument instabilities. Distributions of the pre-injection and the tumour-bearing samples were constructed from the collection of averaged values. The means were $\log_{10}$ transformed prior to analysis. Even after transformation, the distributions are not completely normal.

6.6.2 Univariate Statistical Analysis

The non-parametric Wilcoxon rank sum test (MATLAB statistics library) was used for the univariate statistical analysis as each animal contributed a sample before and after cell injection. Statistical significance was defined using the Bonferroni correction ($p < 0.05 / N$ where N is the number of metabolites or ratios) and the false discovery rate (FDR) for the ratios.

Based on the distributions of the healthy and diseased populations, an ROC curve was constructed from the sensitivity and specificity values obtained at 101 threshold values. These thresholds were determined from the minimum and maximum values form the distributions. The minimum and maximum thresholds were rounded down and up to the nearest thousandth to entirely encompass the distributions. The remaining thresholds were equally spaced between these two end-points. The ROC curve was constructed by plotting the sensitivity along the y-axis, and (1 – specificity) on the x-axis.

Data points were connected with a linear segment for better visualization of the curve. The AUC was calculated using the trapezoidal method for integration (Odunsi, 2005, Weinstein, 2005). This method often underestimates the area since the curve is typically bowed towards the top-left corner; however, the estimate should be fairly accurate since 101 threshold values were used.

A plot of the AUC vs. the p-value was drawn to identify trends in the AUC parameter with respect to the significance of various metabolites / ratios (Broadhurst and Kell, 2006). Vertical lines were drawn on the plots to indicate regions with Bonferroni and FDR significance limits.

6.6.3 Multivariate Statistical Analysis

Multivariate statistical analysis was performed with the SIMCA P+ software, version # 12.0.1 (UMetrics, Sweden). Principal component analysis (PCA) and partial least squares discriminate analysis (PLS-DA) models were generated for the $\log_{10}$ transformed metabolite concentrations using unit variance scaling (UV) or Pareto scaling (par). Seven-fold cross validation was performed for all models. The goodness of fit, R^2, and the goodness of prediction, Q^2, parameters were recorded for every model. There is no standard requirement for these parameters, though values in excess of 0.5 are preferred (Westerhuis et al., 2008). Scores plots for the models were used to ensure that class separation between the pre-injection and tumour bearing animals was present.

The SIMCA software calculates the scores for each sample and stores them in a prediction list. The scores describe the samples location in k-space where k is the number of principal components used to create the model. Samples not used in creating the model are automatically projected on it. The scores for these samples are recorded in the prediction list.

The prediction list for each model was imported to an excel document for ROC curve analysis. Analysis of the scores revealed that the optimal separation between classes was not along a single component axis, but along a combination of two. For this reason, ROC curves were applied to the scores plot at variable angles between 1^o and 180^o in 1^o increments. At each angle, the minimum and maximum values were determined and set as the first and last ROC curve thresholds after rounding down and up to the nearest thousandth. By definition, angles refer to the counter clockwise direction.

The sensitivity and false positive fractions were determined at each threshold and were used to construct the curve. Trapezoidal integration was again used to calculate the AUC. ROC curve analysis was applied to all permutations of component axes to determine the angle at which the AUC was a maximum.

ROC curve analysis was performed with three groups of samples: 1. between the pre and post-injection data for the tumour-bearing animals only, 2. between the post-injection samples for the tumour-bearing and control animals and 3. Grouping all pre-injection samples (tumour-bearing and controls) and the control post-injection data as one population and comparing with the post-injection tumour-bearing samples.

6.6.4 Inclusion of Control Data

Control data was treated similarly to the tumour-bearing data. Univariate statistical analysis on the metabolite concentrations revealed that one metabolite was Bonferroni significant in the control animals, while a few others had p-values less than 0.05 or 0.01. These results suggested that additional influences were competing with the metabolic changes associated with the presence of the tumour. To identify tumour-related changes, these external influences must be carefully considered.

To our knowledge, there is no standard method to include control data in the analysis. Changes experienced in the control animals may be corrected by removing any metabolite that is significant (Bonferroni, FDR or having $p < 0.05$ or $p < 0.01$) in both the control and tumour-bearing animals and travelling in the same direction (increased or decreased in both groups). A more stringent technique is to remove any metabolite that is significant in the control animals, regardless of how it changes in the tumour-bearing population.

Chapter 7
Results and Discussion

This Chapter shows the results of univariate and multivariate analysis. As discussed in the methods chapter, univariate statistical analysis was performed to verify that statistically significant metabolic changes were present. Analysis was performed on both metabolite concentrations and the ratios of pairs of metabolites. ROC analysis was applied to the distributions of the pre-injection and post-injection data. The AUC was calculated to assess the clinical significance of metabolic changes. Multivariate statistical analysis took the form of PCA and PLS-DA. For each technique, unit variance and Pareto scaling was applied prior to modelling. ROC analysis was applied by rotating the scores plots by unit angles between 1 and 180^{o}. The optimal angle was defined as the angle at which the maximum AUC was obtained.

7.1 Pre-Analysis

Metabolite concentrations were exported from Chenomx Suite 5.1 into an excel file and sorted by animal and class. For the purposes of our analysis, classes were defined as pre-injection, post-injection with tumour and post-injection controls. The latter two classes were composed of all samples collected between 6 and 35 days post-injection.

The samples had variable concentrations and required normalization prior to analysis. The concentrations for each sample, for all animals, were plotted to show the spread in the normalized metabolite concentrations. The examples of 3-indoxylsulfate, glucose, taurine and trimethlyamine are displayed in figure 7.1. It is observed that the variability of a metabolite concentration for a single animal is large. However, the range of values is approximately consistent between the pre-injection data from the tumour-bearing mice (black circles) and the control mice (black crosses) as expected. Subtle changes in the concentrations post-injection are observed for glucose and trimethylamine.

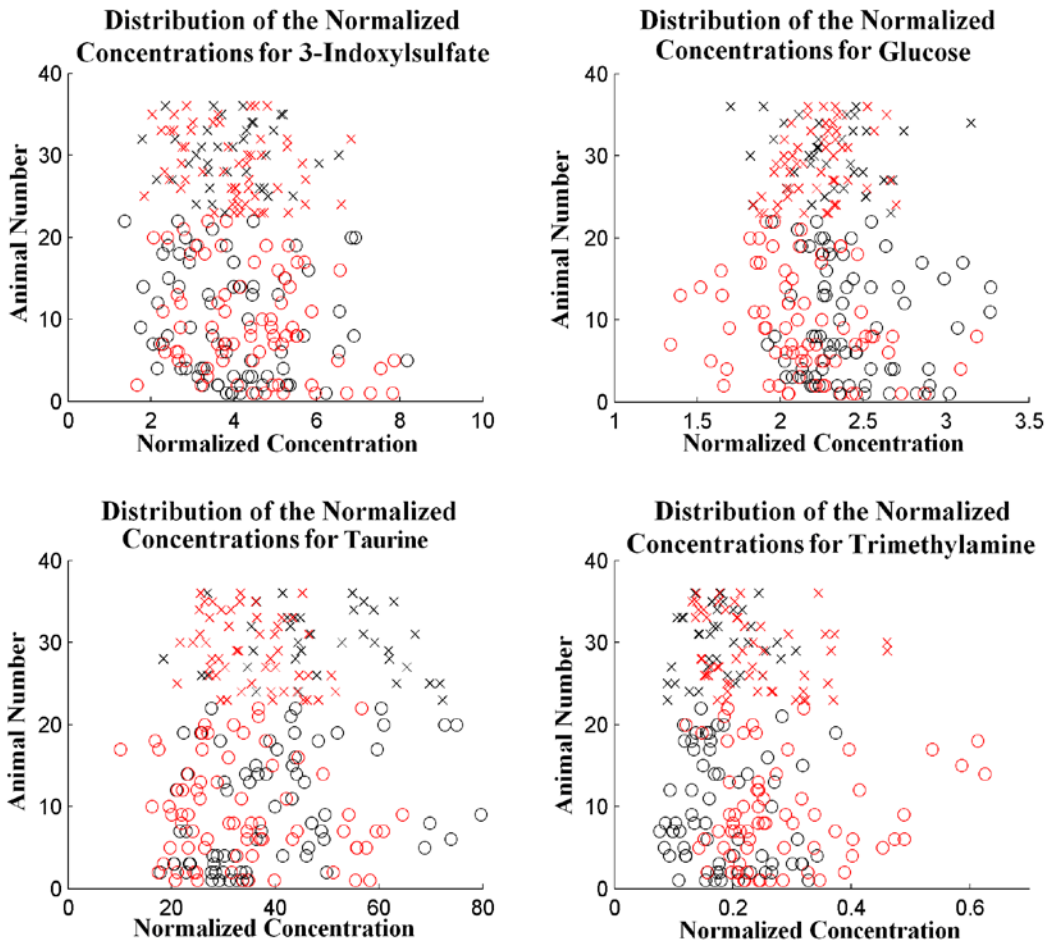

Figure 7.1: Plots of the normalized metabolite concentrations for 3-indoxylsulfate, glucose, taurine and trimethlyamine. The black and red circles represent the pre and post-injection data from the tumour-bearing mice, respectively, while the black and red crosses represent data from the control animals. The intra-animal variation is large for all animals. However, the range of values is consistent among all animals: tumour-bearing and controls.

Paired analysis may be applied since data was collected from the same animals before and after cell injection. This has the advantage of averaging out the day-to-day variations in the urine content and compares metabolic changes experienced on an individual level. For this reason, the pre and post-injection data for each animal was averaged. Distributions of metabolite concentrations were created from the contribution of all averaged concentrations. Prior to averaging, samples defined as outliers – those found more than three standard deviations from the class mean – were removed.

7.2 Univariate Statistical Analysis on Metabolite Concentrations

The distributions of the pre and post-injection data from the tumour-bearing animals were compared using the non-parametric Wilcoxon rank sum test. This test was chosen because the distributions, even after $\log_{10}$ transformation, were often not normal. The role of the $\log_{10}$ transformation is to make a skewed distribution more normal in shape so that statistical tests with an underlying assumption of normality may be applied. The results indicated that 3 (out of 34) metabolites - glucose, trimethylamine n-oxide and 2-oxoglutarate – changed in a manner that satisfied the Bonferroni correction (p-value < 1.47×10^{-3}) and 2 more metabolites – trimethylamine and fumarate – satisfied the FDR criterion for significance. For brevity, variables achieving statistical significance after the Bonferroni correction is applied will be referred to as 'Bonferroni significant' and those considered significant under FDR analysis will be referred to as 'FDR significant'.

The distributions for the pre and post-injection data for glucose, after a $\log_{10}$ transformation, are shown in figure 7.2. There is a notable shift to lower concentrations in the tumour-bearing samples, though there remains heavy overlap between the populations. This observation is consistent with other studies involving cancer where breast cancers had lower levels of glucose than healthy or malignant tissue which could be attributed to the greater rate of aerobic glycolysis in tumour cells (Spratlin et al., 2009; Whitehead et al., 2005).

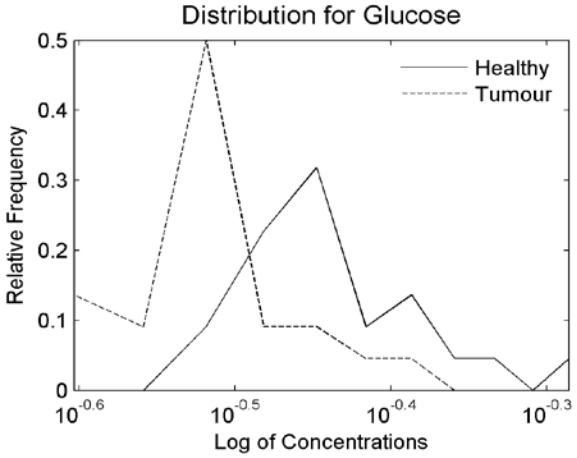

Figure 7.2: Distribution of urinary glucose in experimental animals after applying a $\log_{10}$ transformation. Each point represents the average value of glucose concentration in single animals' samples. The solid line represents the pre-injection (healthy) population and the dashed line represents the post-injection tumour-bearing animals. There is a notable decrease in the concentration post-injection. The distributions were often not normal, so statistical significance was determined with the non-parametric Wilcoxon sum-ranked test.

ROC curve analysis was applied to these distributions to quantify the degree of separation between healthy and tumour-bearing animals. Figure 7.3 shows the ROC curves for the three Bonferroni significant metabolites. The AUC for these metabolites are 0.789, 0.865 and 0.850 for 2-oxoglutarate, glucose and trimethylamine n-oxide, respectively. The other two FDR significant metabolites had AUCs of 0.810 and 0.748 for trimethylamine and fumarate. The three Bonferroni significant metabolites and trimethylamine had the largest AUC's of all metabolites studied. A single non-significant metabolite, betaine (AUC = 0.756), has an AUC larger than fumarate (AUC = 0.748), though this difference is not large.

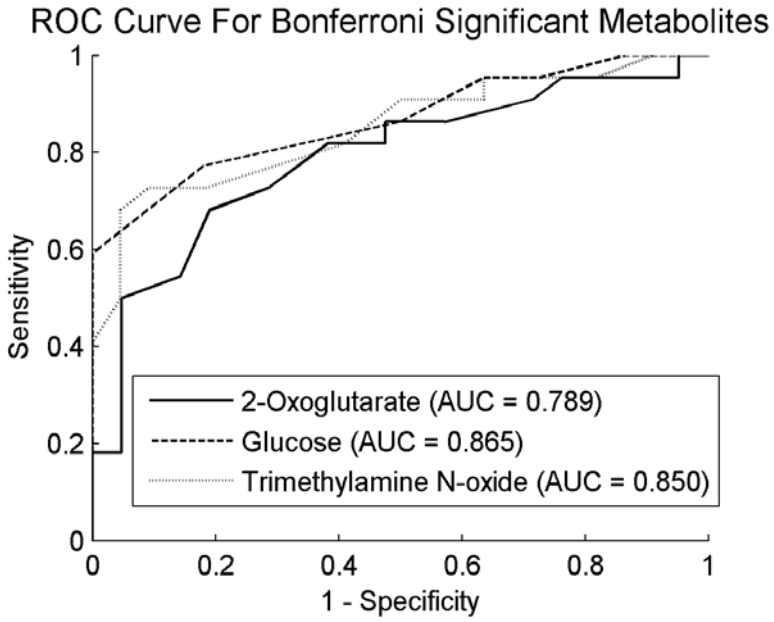

Figure 7.3: ROC curves for the three most significant metabolite concentrations. Each curve was generated by calculating the sensitivity and specificity at 101 different threshold values ranging from the absolute minimum to the absolute maximum ratio in the distributions. The AUC was calculated by fitting trapezoids under the curves.

The optimal decision threshold for each metabolite were determined from the maximum Youden index and from the point that lies closest to the top left hand corner of the graph (point (0,1)). In general, the optimal decision threshold was consistent between the two methods, though there were some discrepancies. Table 7.1 summarizes the optimal decision threshold for the two methods and provides the sensitivity and specificity for each. The decision thresholds here are for the $\log_{10}$ transformed data.

Table 7.1: The Optimal Decision Threshold for FDR Significant Metabolites

Metabolite	Youden Index				Minimum Distance from (0,1)				AUC
	D.T.	Sen	Spec	Y.I.	D.T.	Sen	Spec	Dist	
2-Oxoglutarate	0.725	0.682	0.810	0.491	0.725	0.682	0.810	0.371	0.789
Fumarate	-0.966	0.455	0.952	0.407	-1.084	0.727	0.667	0.431	0.748
Glucose	0.317	0.591	1.000	0.591	0.342	0.773	0.818	0.291	0.865
Trimethylamine	-0.710	0.857	0.591	0.448	-0.643	0.714	0.727	0.395	0.810
Trimethylamine N-oxide	0.603	0.727	0.909	0.636	0.603	0.727	0.909	0.287	0.850

Table 7.1: Optimal decision thresholds, 'D.T.' as determined through the Youden Index (Y.I.) and minimum distance to the point (0,1) (Dist) methods for the five FDR significant metabolites. The sensitivity (sen) and specificity (spec) for these thresholds are indicated in the following columns.

A convenient method to evaluate the clinical and statistical significance of all metabolites involves plotting the AUC against the p-value, as shown in figure 7.4 (Broadhurst and Kell, 2006). The first vertical line, at a p-value of 1.471×10^{-3}, represents the threshold p-value for Bonferroni significance and the second line, at a p-value of 3.47×10^{-3} is the FDR significance limit.

Performing the same analysis on the control animals indicates that only one metabolite, taurine, is Bonferroni significant. FDR analysis showed that this metabolite was also the only FDR significant variable. This metabolite is not statistically significant in the tumour-bearing population, and might result from aging or as a response to the injection procedure. To increase the number of variables considered significant in the control population, significance was defined for variables with $p < 0.05$ and $p < 0.01$. Another two metabolites lie to the left of $p = 0.01$ and four more lie to the left of $p = 0.05$. These limits are commonly used to define statistical significance when testing a single variable, but have also been used in the analysis of metabolomics data sets (Kim et al., 2008; Saude and Sykes, 2007; Tyburski et al., 2009; Um et al., 2009). For this reason, we chose to consider metabolites satisfying both conditions for further analysis. The AUC against the p-value for the control data is displayed in figure 7.5. In this analysis, taurine was the only metabolite to have an AUC greater than 0.9. A total of three metabolites – 2-oxoglutarate, trimethylamine and trimethylamine n-oxide –had an AUC greater than 0.7 in both populations.

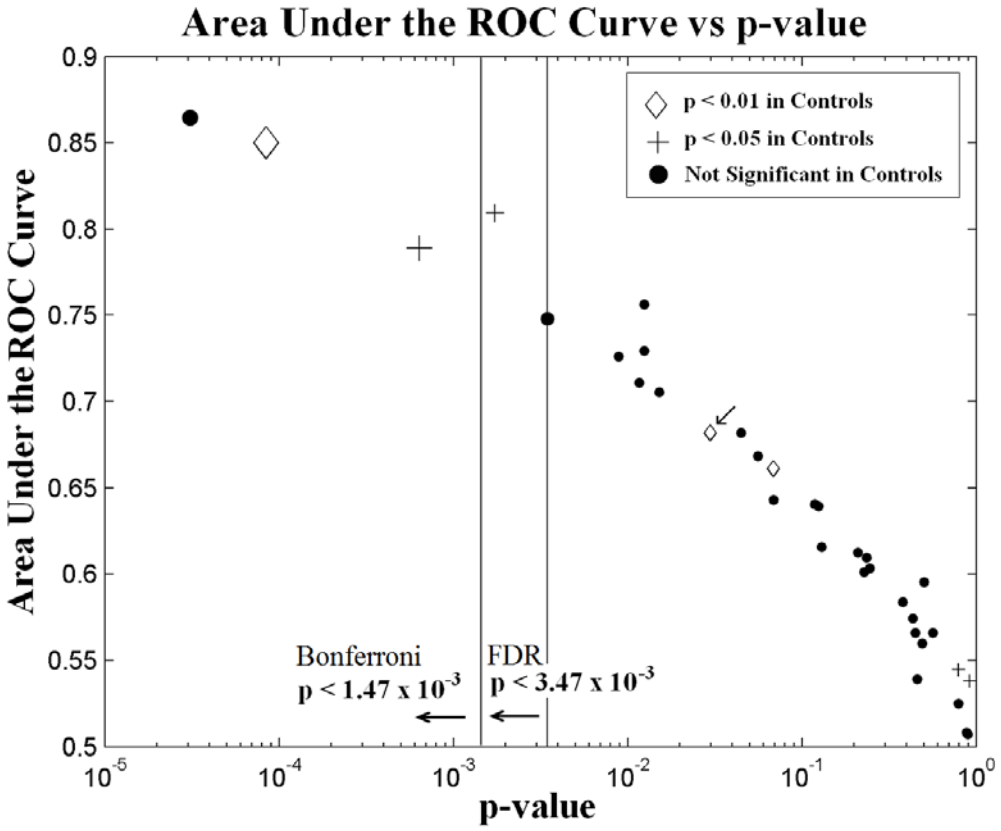

Figure 7.4: AUC versus p-value for metabolite concentrations. The figure shows a total of 3 significant metabolites to the left of the Bonferroni limit (vertical line at a p-value ~ 1.47×10^{-3}) and 5 to the left of the FDR limit (vertical line at a p-value of 3.47×10^{-3}). These metabolites have larger markers to emphasize statistical significance. The diamonds indicate metabolites with $p < 0.01$ in the control animals, the crosses indicate metabolites in the control animals with $p < 0.05$ and the circles represent metabolites that are not significant in the control population. The single Bonferroni significant metabolite in the control population is identified with the arrow. No other metabolites were FDR significant in the controls.

The goal of metabolomics analysis is to identify potential variables that can discriminate between healthy and diseased populations. For this reason, it may be advantageous to exclude variables that change in a statistically significant manner in both populations. Metabolites with a p-value less than 0.01 in the control population are displayed as the diamonds in figure 7.4, while those with a p-value less than 0.05 are identified as crosses. For this analysis, taurine was the only Bonferroni significant metabolite in the control population and is identified with the single arrow in figure 7.4.

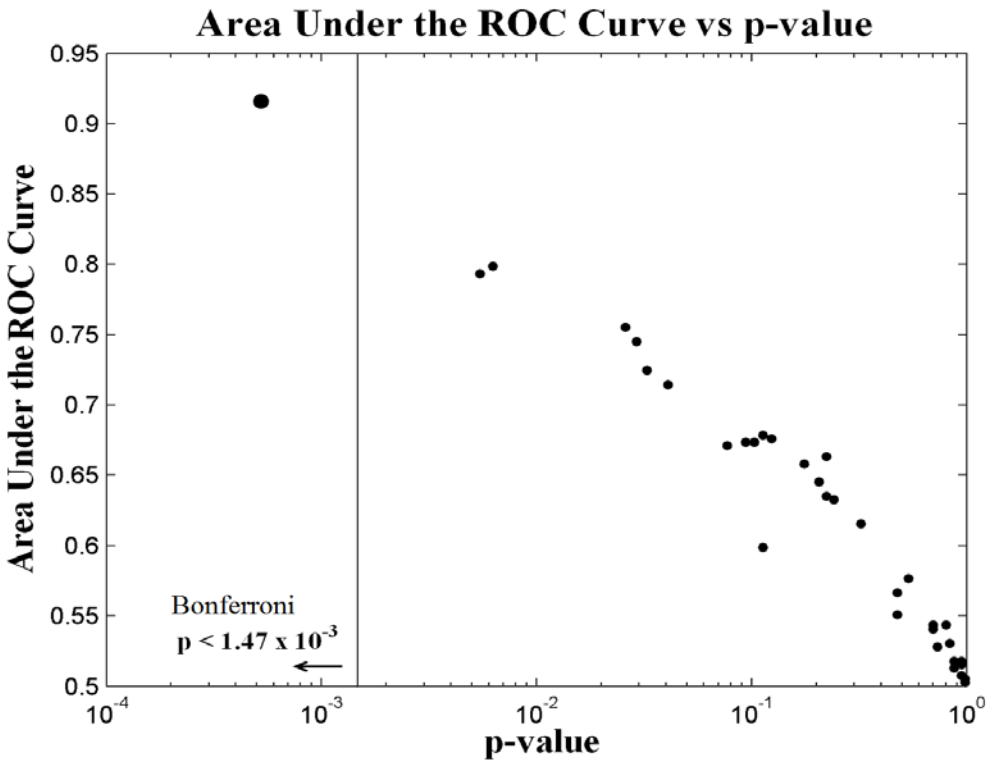

Figure 7.5: AUC versus p-value for metabolite concentrations in the control animals. The figure shows that only one metabolite (taurine) achieves Bonferroni and/or FDR significance (located to the left of the vertical line at a p-value ~ 1.47×10^{-3}). Taurine has an AUC of 0.916. No other metabolites have an AUC greater than 0.8, though trimethylamine n-oxide has AUC = 0.793 and o-phosphocholine has AUC = 0.799.

Of the seven metabolites with $p < 0.05$, four are also significant in the tumour-bearing group and change in the same direction (2-oxoglutarate, taurine, trimethylamine and trimethylamine n-oxide). Removal of these metabolites results in a loss of two Bonferroni-significant metabolites (trimethylamine n-oxide with $p = 8.44 \times 10^{-5}$ and 2-oxoglutarate with $p = 6.37 \times 10^{-4}$) and one FDR significant metabolite (trimethylamine $p = 1.73 \times 10^{-3}$) in the tumour-bearing group. Consequently, only glucose is Bonferroni significant. Fumarate loses its FDR significance with the removal of the 4 metabolites. Only one Bonferroni significant metabolite (trimethylamine n-oxide) is removed when considering metabolites with $p < 0.01$ in the control animals. The reduced number of metabolites is not sufficient to change the Bonferroni limit enough so that trimethylamine becomes Bonferroni significant. Therefore, only two metabolites are considered significant in the tumour-bearing population. Removing the Bonferroni significant metabolite – taurine – has no impact on the analysis as it is not statistically significant in the tumour-bearing population.

Our results show that the presence of a tumour in a host animal alters the metabolic composition of urine. However, the tumour does not affect all metabolites. The magnitude of the change in concentration ranged from negligible to highly significant. For instance, the mean concentrations of 1-methylnicotinamide, 3-indoxylsulfate and fucose remain essentially unchanged with p-values exceeding 0.879 and AUC values less than 0.538. On the other hand, the mean concentration of glucose, after applying the $\log_{10}$ transformation to make the distributions normal, decreased by 16.9% and had a p-value of 3.09×10^{-5}. This metabolite satisfies the condition for a good diagnostic test with an AUC of 0.865. Two other metabolites, trimethylamine and trimethylamine n-oxide, may also be considered good diagnostic tests as they have AUC's of 0.810 and 0.850, respectively. Of these two, only trimethylamine n-oxide was Bonferroni significant, but both were FDR significant.

Analysis on the control animals revealed that some metabolites will change in the same direction (increase or decrease) as in the tumour-bearing population. Removing these metabolites from the analysis will enhance our confidence in the results, but could result in the loss of too many metabolites. Therefore, it is advisable to set a significance threshold prior to removal. In this analysis, only two FDR significant metabolites will remain after removal of metabolites with $p < 0.05$ in the control population. It is unlikely that the behaviour of these two metabolites will be specific to cancer only. However, one investigator (Odunsi et al., 2005) showed that the use of 2 variables was sufficient for the diagnosis of epithelial ovarian cancer in human sera. Others have used six (Lokhov et al., 2010) and ten (Whitehead et al., 2005) metabolites for differentiation.

7.3 Univariate Statistical Analysis on Metabolite Ratios

Univariate analysis was also performed on the ratios of pairs of metabolites. The Wilcoxon rank sum test identified 144 metabolite ratios that changed in a manner that was FDR significant, of these 19 were Bonferroni-significant. A total of 53 ratios (12 Bonferroni significant, 41 FDR significant) produced an AUC between 0.8 and 0.9, and 7 metabolites (all Bonferroni significant) had an AUC greater than 0.9. The AUC versus p-value data for the ratios are plotted in figure 7.6.

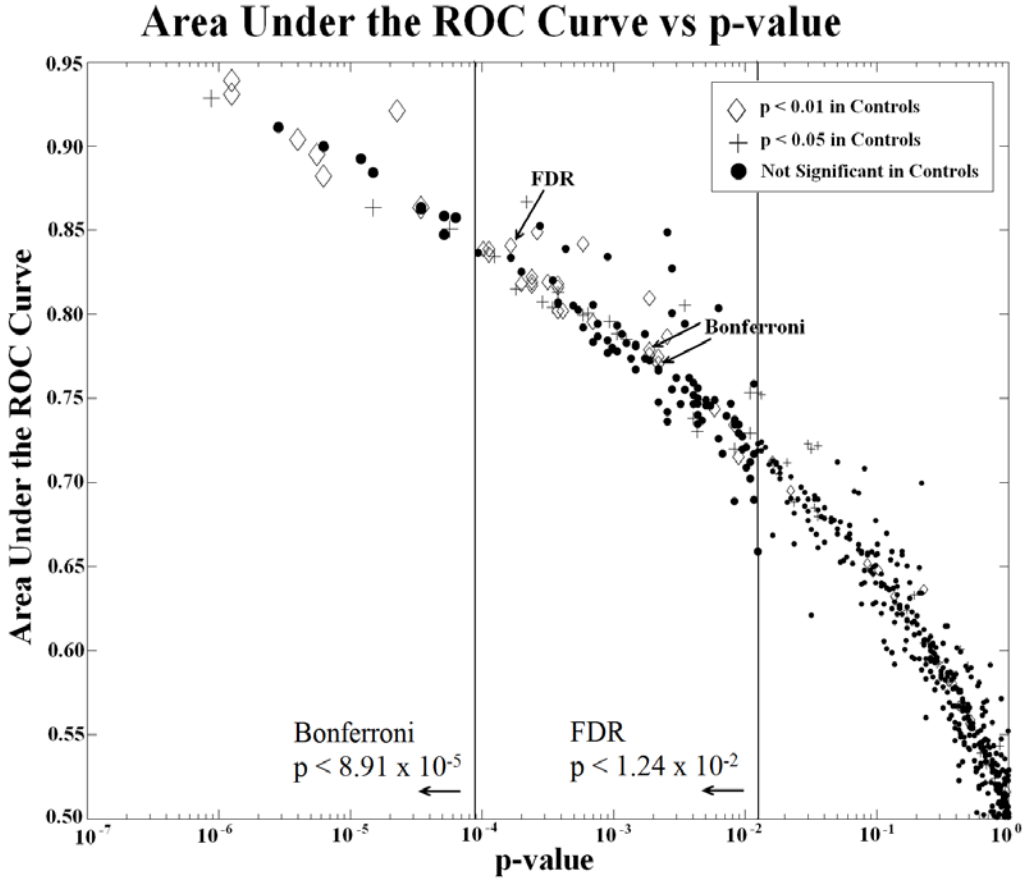

Figure 7.6: Plot of AUC versus the p-value for the ratio of pairs of metabolites. 19 metabolite ratios are Bonferroni-significant (left line) and another 125 are significant to the left of the FDR limit (right line). 53 ratios have an AUC between 0.8 and 0.9, and 7 ratios have an AUC > 0.9 (all Bonferroni significant). Ratios with Bonferroni and FDR significance are drawn with larger markers to emphasize the different levels of significance. There are only two Bonferroni significant metabolites in the control animals (indicated by the arrows). The data points displayed as diamonds and crosses are the metabolite ratios that have $p < 0.01$ and $p < 0.05$ in the control population.

Performing the same analysis on the control population revealed that two metabolite ratios experienced a change that was Bonferroni significant and another two were FDR significant. Both Bonferroni significant ratios were FDR significant in the tumour- bearing population and changed in the same direction and are identified by the arrows in figure 7.6. Of the two FDR significant ratios, only one was FDR significant in the tumour-bearing population and is indicated with an arrow in figure 7.6. The results of the analysis on the control animals are summarized in figure 7.7. The two vertical lines

represent the Bonferroni significance limit and the FDR significance limit when using the Benjamani and Hochberg methodology (Benjamani and Hochberg, 1995). Variables that achieve Bonferroni or FDR significance in the control population are displayed as diamonds and diamonds, and are indicated with an arrow in figure 7.6.

To account for urinary metabolic changes not related to cancer, we again focus on the metabolite ratios in the control population where $p < 0.05$ and $p < 0.01$. Metabolites are removed if they have a p-value less than 0.05 (or 0.01) in the control data and experience shifts in the same direction in both populations.

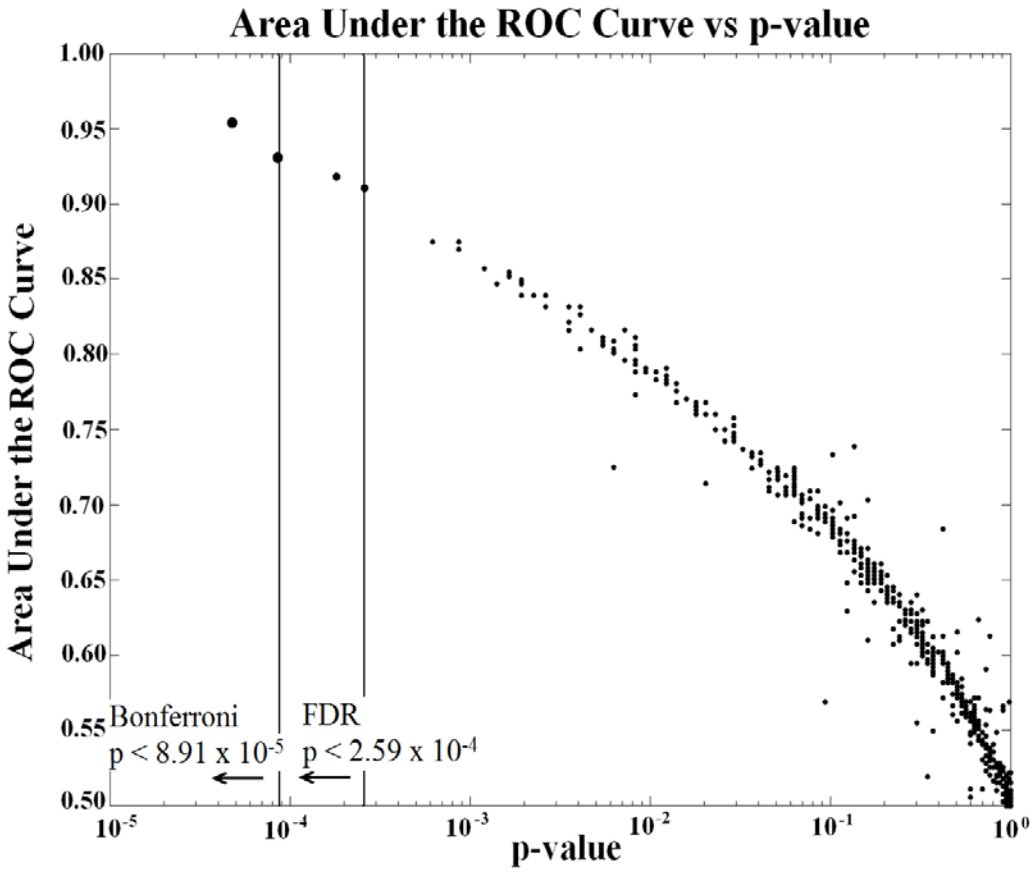

Figure 7.7: Plot of AUC versus the p-value for the ratio of pairs of metabolites for the control animals. Two metabolite ratios are Bonferroni-significant (first line) and another two are significant to the left of the FDR limit when using the Benjamani and Hochberg method (second line). Ratios with Bonferroni and FDR significance are drawn with larger markers to emphasize the different levels of significance.

Removal of metabolites with $p < 0.05$ in both populations, and moving in the same direction, ($n = 64$) resulted in a loss of 53 significant metabolites (43 FDR and 10 Bonferroni). After re-calculating the new Bonferroni significance threshold ($p = 1.006 \times 10^{-4}$), a total of 72 metabolite ratios achieved FDR significance with 10 of these being Bonferroni significant. One metabolite moved from the FDR to the Bonferroni significance region after adjusting the limit. Removal of all metabolites with $p < 0.05$ in the control population ($n = 92$) removed a further 28 non-significant ratios from analysis, and allowed one more metabolite to achieve Bonferroni significance. The affected ratios are displayed as the diamonds and crosses in figure 7.6. Two Bonferroni significant ratios have an AUC greater than 0.9, while the remaining eight (or nine after removal of significant ratios in the control population) have AUC exceeding 0.837.

Removal of any metabolite ratio with $p<0.01$ in the control population, and shifting in the same direction in the tumour-bearing group, resulted in a loss of 32 significant ratios in the tumour-bearing analysis: all are FDR significant, with 7 of them being Bonferroni significant. The more conservative approach, in which all FDR significant metabolite ratios in the control population are removed from the analysis, resulted in the removal of 46 metabolite ratios. The 32 FDR significant ratios mentioned above were affected, as well as the addition of another 14 non-significant ratios. After updating the Bonferroni limit to reflect the smaller number of variables, 12 ratios were Bonferroni significant; one of these ratios was only FDR significant before. All significant ratios had an AUC greater than 0.837, with three having an AUC in excess of 0.9. These ratios are the diamonds in figure 7.6.

The optimal decision thresholds for each metabolite ratio were determined from the ROC curve using both methods. In general, the two techniques agreed within 3 decision thresholds (101 were used to calculate the curve). Table 7.2 summarizes this data for nine Bonferroni significant metabolites which are likely due to the presence of tumour.

The frequency with which individual metabolites appeared in each category of significance (Bonferroni, FDR, $p < 0.01$ or $p < 0.05$) was determined. The results are displayed in Table 7.3 for both the tumour-bearing and the control animals. The first column for each group is the p-value for the metabolite concentration determined previously. The remaining columns describe the various levels of significance used.

Table 7.2: Determination of the Optimal Decision Threshold for Metabolite Ratios

Metabolite in Numerator	Metabolite in Denominator	Youden Index				Minimum Distance from (0,1)			
		D.T.	Sen	Spec	Y.I.	D.T.	Sen	Spec	Dist
Citrate	Acetate	1.53	0.636	0.952	0.589	1.46	0.773	0.810	0.297
Citrate	Betaine	1.29	0.864	0.773	0.636	1.29	0.864	0.773	0.265
Succinate	Betaine	0.14	0.864	0.909	0.773	0.14	0.864	0.909	0.164
2-Oxoisocaproate	Trimethylamine N-oxide	-0.32	0.864	0.818	0.682	-0.32	0.864	0.818	0.227
3-Methyl-2-oxovalerate	Trimethylamine N-oxide	-0.55	0.864	0.818	0.682	-0.55	0.864	0.818	0.227
Acetate	Trimethylamine N-oxide	-0.76	0.857	0.727	0.584	-0.81	0.762	0.818	0.300
Hippurate	Trimethylamine N-oxide	-0.86	0.773	0.818	0.591	-0.86	0.773	0.818	0.291
Niacinamide	Trimethylamine N-oxide	-1.34	0.727	0.909	0.636	-1.34	0.727	0.909	0.287
Valine	Trimethylamine N-oxide	-1.67	0.818	0.818	0.636	-1.67	0.818	0.818	0.257

Table 7.2: Optimal decision thresholds, 'D.T.' for nine Bonferroni significant ratios as determined through the Youden Index (Y.I.) and minimum distance to the point (0,1) (Dist) methods. Based on the analysis of control animals, these nine ratios are likely due to the presence of the tumour in the host animal. The sensitivity (sen) and specificity (spec) for these thresholds are indicated in the following columns.

Inspection of the metabolite ratios (table 7.3) revealed that 38.9% (56 out of 144) FDR significant metabolite ratios contained at least one of the Bonferroni-significant metabolites, while 60.4% (87 out of 144) of these metabolites contained at least one of the five FDR significant metabolites . Further, 15 (of 19) Bonferroni significant ratios contained one of the three Bonferroni significant metabolites and 16 Bonferroni significant metabolites contained one of the five FDR significant metabolites. Fifty-seven FDR significant ratios contained two metabolites that were not considered FDR significant, with three of these ratios achieving Bonferroni significance. This means that 39.6% of the FDR significant metabolite ratios contain two non-significant metabolites.

Analysis of metabolite ratios has the ability to extract information similar to that discovered with the analysis of metabolite concentrations, but also identifies significant changes in the pair of two non-significant metabolites. In a majority of these ratios (50 of the 57 or 87.7%), the two mean metabolite concentration changed in opposite directions. Another two ratios had one metabolite increase in concentration, while the other metabolite maintained the same value post-injection. Further, a total of 53 ratios of pairs of metabolites had an AUC exceeding 0.8 (7 having and AUC greater than 0.9), while only 3metabolites (concentrations) had an AUC above 0.8.

Table 7.3: Frequency of Metabolites in the Significant Ratios

Metabolite	Tumour-Bearing Population					Control Population				
	p-value	Bon	FDR	0.01	0.05	p-value	Bon	FDR	0.01	0.05
1-Methylnicotinamide	9.16E-01		5	5	6	3.26E-02			2	5
2-Oxoglutarate	6.37E-04	2	26	24	29	4.09E-02			5	13
2-Oxoisocaproate	2.09E-01	1	7	7	8	8.36E-01				1
2-Oxovalerate	1.19E-01	1	5	5	9	4.76E-01			2	2
3-Indoxylsulfate	8.79E-01		2	1	5	8.00E-01	1	1	2	3
3-Methyl-2-oxovalerate	1.30E-01	1	6	5	10	8.72E-01				1
3-Phenylpropionate	3.79E-01	1	5	5	6	4.76E-01			3	6
4-Hydroxybenzoate	5.57E-02	2	8	8	11	5.35E-01			1	6
Acetate	1.24E-01	2	8	8	9	3.23E-01				1
Betaine	1.24E-02	4	11	10	12	1.13E-01			4	6
Choline	4.48E-02	1	8	7	11	7.30E-01			1	2
Citrate	8.87E-03	2	20	18	22	1.75E-01			1	3
Creatine	4.32E-01		4	3	10	2.23E-01			1	7
Creatinine	2.45E-01		5	5	9	7.69E-02			4	10
Dimethylamine	1.24E-02	1	11	10	18	1.03E-01	1	1	5	7
Formate	7.87E-01		7	6	11	2.58E-02			5	12
Fucose	8.97E-01		3	2	7	9.82E-01			1	2
Fumarate	3.47E-03		15	15	21	2.41E-01			7	12
Glucose	3.09E-05	2	7	7	11	2.23E-01			3	7
Hippurate	4.89E-01	1	6	3	8	6.96E-01			1	3
Lactate	2.36E-01		6	4	15	9.45E-01				
Methionine	5.04E-01		4	3	6	6.96E-01				2
Niacinamide	7.87E-01	1	5	5	8	9.82E-01		1		4
Nicotinamide N-oxide	4.46E-01		4	3	5	2.06E-01			2	5
O-Phosphocholine	6.89E-02	1	8	8	10	6.26E-03			5	6
Succinate	1.51E-02	1	16	14	21	8.72E-01				2
Taurine	2.99E-02		7	7	11	5.22E-04	1	3	17	22
Threonine	1.16E-02		7	7	9	1.24E-01			2	6
Trigonelline	5.65E-01		6	5	14	9.45E-01			1	3
Trimethylamine	1.73E-03	1	19	17	23	2.91E-02			5	13
Trimethylamine N-oxide	8.44E-05	12	25	25	26	5.44E-03	1	2	11	12
Valine	2.27E-01	1	6	6	8	9.45E-01			3	5
Xylose	6.89E-02		2	2	3	1.13E-01			1	2
trans-Aconitate	4.60E-01		4	2	6	9.35E-02			1	3

Table 7.3: This table shows the frequency with which individual metabolites appeared in the significant metabolite ratios. The first column in each section is the p-value obtained from the metabolite concentration analysis. The following columns relate to the Bonferroni significant metabolite ratios (Bon; $p < 8.913 \times 10^{-5}$), FDR significant metabolite ratios (FDR; 1.24×10^{-2} for tumour-bearing data and 2.59×10^{-4} for control data), metabolite ratios with $p < 0.01$ (0.01) and with $p < 0.05$ (0.05).

Metabolic ratios provide information about the simple interactions between pairs of metabolites and have the potential to identify significant changes that were not obvious in the analysis of metabolite concentrations individually. Analysing metabolite ratios increases the number of tested hypotheses from N to N(N-1)/2 where N is the number of identified metabolites. As a result, more variables achieve statistical significance. The larger number of tested hypotheses will, however, increase the risk of having false discoveries. The Bonferroni correction or use of the FDR method is required to limit the number of false discoveries to an acceptable level.

Control animals were included in the study to investigate the effects of non-tumour-related metabolic changes in an age and gender-matched population. The treatment of all animals was consistent with respect to the light/dark cycle (12 hours light, 12 hours dark), temperature and humidity of environment, diet and method of urine collection. Metabolic changes in these animals may result from a number of factors including aging or the animals' response to whole body irradiation or the injection.

The fact that significant metabolic changes were observed in the control animals complicates the interpretation of the tumour-bearing data. If one were only interested in tracking an individual metabolite concentration or a single metabolite ratio for diagnostic purposes, consideration must be given to how the metabolite or ratio changes in an appropriately-matched healthy subject. The results of this analysis show that, in general, metabolic changes are more pronounced in the tumour-bearing group than in the controls. Despite some similar metabolic changes in both populations, there remain a number of changes which are likely due to the presence of the tumour. The analysis shows that a simple approach, such as removing metabolites or ratios that are common to both the tumour-bearing and control groups, is still able to extract significant (Bonferroni or FDR) metabolic changes which may be characteristic to the tumour-bearing population.

While this treatment of control data will improve our confidence in the significance of the results, it likely underestimates the number of true discoveries. For instance, if the tumour has a strong influence on the concentration of one metabolite and an external source, such as aging or in response to an environmental or procedural stress, has a weak influence on the same metabolite in the same direction, then the metabolite could be unnecessarily removed. As metabolomics moves to the clinic, it is essential to define a healthy base line so that metabolic changes associated with tumour development may be identified. This requires a large cohort of patients from each group.

7.4 Multivariate Statistical Analysis on Metabolite Concentrations

PCA models were created for the pre and post-injection tumour-bearing data to investigate variation in the data set. Five data sets were used for this analysis: i) all metabolites, ii) all metabolites excluding those that had a $p < 0.05$ in the control animals, iii) all metabolites excluding those that had a $p < 0.05$ in both control animals and tumour bearing animals, iv) all metabolites excluding those that had a $p < 0.01$ in the control animals and v) all metabolites excluding those that had a $p < 0.01$ in both control animals and tumour bearing animals. For each data set, two models were constructed; the first using unit variance (UV) scaling, and the other using Pareto scaling. This was done to evaluate the ability of the different scaling techniques to discriminate between the healthy and tumour-bearing groups.

The control data was not used to generate the models as we are interested in metabolic changes associated with the presence of tumour. Instead the data was projected onto the model. This allows for a visual assessment of how similar the post-injection tumour-bearing and the post-injection control data are. If the two groups are well separated, then tumour-related metabolic changes are present.

PLS-DA models were also constructed for the pre and post-injection tumour-bearing metabolite concentration data with the objective of finding separation between the two populations. The same five data sets defined above were used. Similar to the PCA models, two PLS-DA models were created for each data set: one with Pareto scaling and one with UV scaling. A total of 10 models were thus evaluated. The control data was again not used to create the model, but was projected onto it.

A summary of the model parameters – R^2 and Q^2 values as well as the number of components used – is shown in table 7.4. There are no standard R^2 and Q^2 values to determine the strength of the model (Westerhuis et al., 2008), though values greater than 0.5 are attractive (Erb et al., 2008).

Each latent variable is said to describe the maximum variation in the data, which has not already been accounted for in a previous component. This suggests that applying ROC analysis along each component axis will indicate the level of useable tumour-related information contained in that component. Ideally, the largest AUC should be obtained along the first component axis.

Table 7.4 Summary of PCA and PLS-DA Models

	PCA				PLS-DA			
Model	scaling	R^2	Q^2	Components	scaling	R^2	Q^2	Components
All data	UV	0.565	0.172	4	0.421	0.837	0.567	3
	Par	0.619	0.260	4	0.481	0.810	0.556	3
$p < 0.01$ in both removed	UV	0.486	0.186	3	0.418	0.824	0.505	3
	Par	0.621	0.243	4	0.481	0.795	0.517	3
$p < 0.01$ in controls removed	UV	0.509	0.196	3	0.381	0.651	0.427	2
	Par	0.638	0.239	4	0.449	0.602	0.427	2
$p < 0.05$ in both removed	UV	0.261	0.159	1	0.423	0.771	0.422	3
	Par	0.727	0.176	5	0.478	0.729	0.412	3
$p < 0.05$ in controls removed	UV	0.256	0.152	1	0.360	0.615	0.354	2
	Par	0.558	0.222	3	0.425	0.542	0.326	2

Table 7.4: Summary of the PCA and PLS-DA models created after applying a $\log_{10}$ transformation to the normalized data set, mean centering and scaling (indicated in the second column where UV = unit variance and par = pareto scaling). A total of 10 models were constructed to represent the five data sets and the two common types of scaling. The models coefficient of variability, R^2, and coefficient of predictability, Q^2, is shown, as well as the number of components used for the optimal model.

Three class comparisons took place. In the first, only the samples (pre-injection and tumour-bearing) from the animals receiving cancer cell injections were considered. In the second, the post-injection tumour-bearing samples were compared with the post injection samples from the control animals. In the third, all pre-injection data (i.e. samples from both the tumour-bearing and control animals) and the control post sham-injection data were grouped as 'healthy' and compared with the post-injection tumour-bearing data. The results of this analysis on all models are summarized in table 7.5. In all cases, the scores plot using the first two component axes showed separation between the pre-injection and tumour-bearing data, though there remains some overlap. The plots also revealed that the optimal separation did not occur along a single component axis, but a combination of two. This motivated us to perform ROC analysis after rotation of the data about the positive horizontal axis. ROC analysis was performed along the positive horizontal axis after each rotation. Figure 7.8 shows a scores plot for a PCA model with all data and UV scaling.

Table 7.5 ROC Curve Analysis Along the PCA and PLS-DA Model Components

		Pre and Post Cell Injection				Post-injection Data				Healthy Versus Cancer			
		PCA		PLS-DA		PCA		PLS-DA		PCA		PLS-DA	
Data set	LV	UV	par	UV	par	UV	par	UV	par	UV	par	UV	par
Cancer	1	0.613	0.646	0.972	0.924	0.539	0.570	0.867	0.789	0.564	0.615	0.935	0.889
	2	0.916	0.857	0.650	0.704	0.882	0.779	0.580	0.632	0.926	0.858	0.654	0.711
	3	0.525	0.566	0.706	0.734	0.651	0.649	0.575	0.565	0.587	0.621	0.587	0.612
	4	0.519	0.528			0.570	0.648			0.528	0.609		
Remove $p < 0.01$ both	1	0.625	0.655	0.941	0.895	0.547	0.576	0.789	0.745	0.583	0.626	0.890	0.844
	2	0.888	0.817	0.681	0.737	0.831	0.742	0.568	0.594	0.894	0.822	0.664	0.715
	3	0.509	0.556	0.727	0.734	0.645	0.649	0.607	0.610	0.570	0.618	0.575	0.578
	4		0.560				0.645				0.625		
Remove $p < 0.01$ controls	1	0.626	0.654	0.942	0.877	0.552	0.575	0.735	0.687	0.587	0.626	0.866	0.806
	2	0.875	0.810	0.685	0.743	0.813	0.740	0.563	0.607	0.882	0.815	0.666	0.717
	3	0.511	0.543			0.633	0.643			0.553	0.608		
	4		0.596				0.696				0.664		
Remove $p < 0.05$ both	1	0.598	0.507	0.931	0.851	0.520	0.558	0.742	0.659	0.551	0.550	0.853	0.778
	2		0.763	0.683	0.750		0.731	0.571	0.593		0.762	0.661	0.715
	3		0.567	0.707	0.704		0.646	0.658	0.635		0.621	0.515	0.523
	4		0.745				0.656				0.729		
	5		0.572				0.683				0.610		
Remove $p < 0.05$ controls	1	0.637	0.646	0.917	0.832	0.544	0.568	0.705	0.651	0.586	0.612	0.830	0.761
	2		0.718	0.691	0.772		0.635	0.591	0.589		0.696	0.665	0.714
	3		0.665				0.675				0.694		

Table 7.5: Summary of the ROC analysis applied along the components axes (labelled LV for latent variables) for all models. Three groups of data were considered: the first involved only the tumour-bearing animals (pre and post cell injection), the second used the post-injection data for both the tumour-bearing and the control animals (post-injection data) and the third combined all pre-injection data with the post-injection data for the controls, and compared this data with that for the post-injection tumour-bearing animals (healthy versus cancer). This analysis revealed that the majority of tumour-related information was contained in the first component for PLS-DA models, but was often found in the second component axis for the PCA models. The PLS-DA models had an AUC greater than 0.9 when comparing the pre and post-injection tumour-bearing data using UV scaling. PCA models for 'cancer', 'remove $p < 0.01$ both' and 'remove $p < 0.01$ control' data sets had an AUC greater than 0.8 along PC a majority of the time.

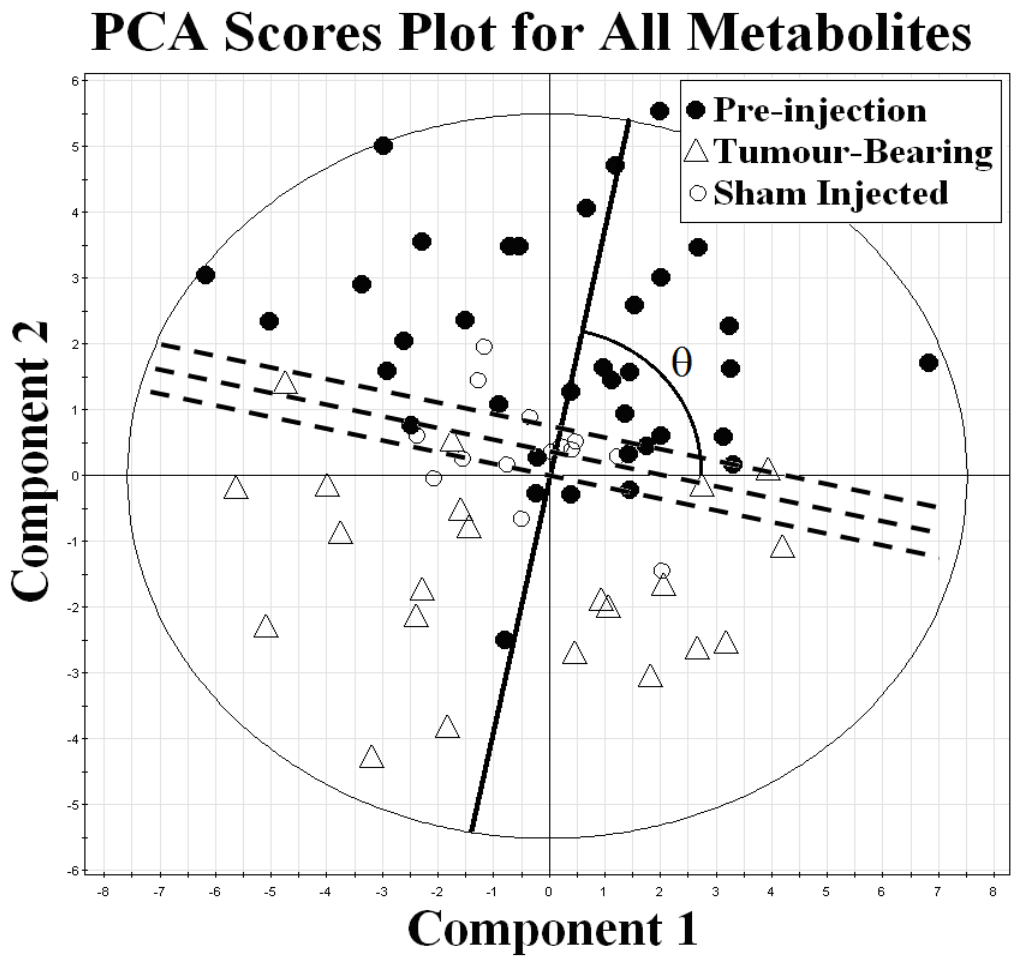

Figure 7.8: Scores plot of PC 1 and PC 2 for the PCA model with all data and UV scaling. There is an evident shifting of the data post-injection along PC 2. ROC curve analysis along this PC provided an AUC of 0.916. However, there is some degree of separation along the first PC, meaning that the optimal direction of ROC analysis is not along a single component axis. This motivated us to perform ROC analysis on the scores plots after rotation of the data. The solid black line provides one example of this analysis where the data is rotated through an angle, θ, and ROC analysis is performed in the direction of the line. Three decision thresholds (dashed lines) are shown.

ROC curve analysis was performed by rotating the axes in one degree increments between 1^o and 180^o and calculating the ROC along the rotated axis (thick black line in Figure 7.8). For each angle, the minimum and maximum decision thresholds were updated to reflect the range of the rotated data. The thresholds used to calculate the ROC curve parameters are perpendicular to the rotated axis – three of these thresholds are displayed as the dashed lines in Figure 7.8. To determine the ability of each model to discriminate between populations, all pairs of latent variables were used.

ROC Analysis was applied to the data set by comparing the post-injection tumour-bearing data with three different groups. The first was the pre-injection data from the tumour-bearing mice. This analysis will indicate the ability of our technique to distinguish between health and tumour-bearing mice on an individual animal basis. Second, the post-injection data from the control mice was used to evaluate the degree of separation between the healthy and tumour-bearing animals post-injection. This is analogous to a human study in which metabolism between pre-defined groups are compared. The third analysis involved the grouping of all pre-injection data (from both the controls and tumour-bearing mice) and the post-injection data from the control animals. This decreases the prevalence of cancer in the test population.

Figure 7.9 shows the ROC curves along PC 1, PC 2 and the optimal rotated curve for the pre and post-injection tumour-bearing data in the PCA model with all metabolites and UV scaling. It is clear that the rotated data provides better separation between the pre-defined groups. The results for the ten PCA models are shown in table 7.6.

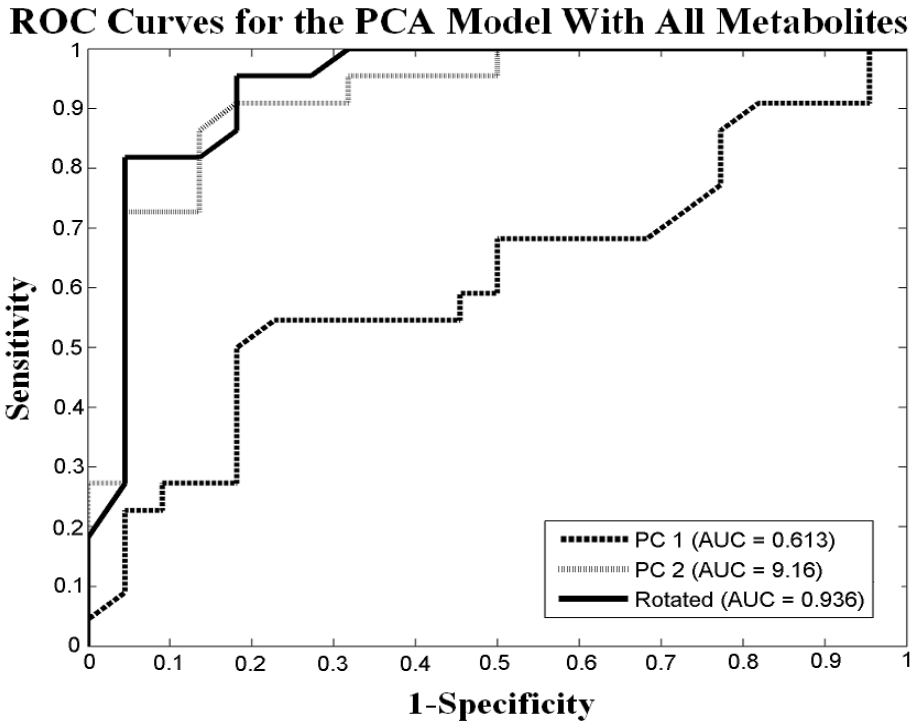

Figure 7.9: ROC curves for the pre and post-injection data from the tumour-bearing animals for the PCA model using all metabolites. The curves displayed are those along PC 1, PC 2 and the optimal ROC curve after rotation of the data. The AUC increases from 0.613 and 0.916 for PC 1 and PC 2, respectively, to 0.936 after rotating the data.

Table 7.6 ROC Curve Analysis on the PCA Models

	Model Summary			Best pair of LV	Pre and Post Cell Injection		Post-injection Data		Healthy versus Cancer	
Data Set	scaling	R^2	Q^2		AUC	Angle (deg)	AUC	Angle (deg)	AUC	Angle (deg)
All data	UV	0.565	0.172	1 & 2	0.936	79	0.883	92	0.930	78
	Par	0.619	0.260	1 & 2	0.892	111	0.812	112	0.882	103
Removed p < 0.01 both	UV	0.486	0.186	1 & 2	0.910	79	0.859	74	0.907	73
	Par	0.621	0.243	1 & 2	0.871	108	0.787	114	0.856	106
Removed p < 0.01 control	UV	0.509	0.196	1 & 2	0.917	77	0.852	80	0.911	77
	Par	0.638	0.239	1 & 2	0.873	71	0.768	65	0.852	74
Removed p < 0.05 both	UV	0.261	0.159	-	-	-	-	-	-	-
	Par	0.727	0.176	2 & 4	0.847	59	0.820	45	0.842	64
Removed p < 0.05 control	UV	0.256	0.152	-	-	-	-	-	-	-
	Par	0.558	0.222	2 & 3	0.784	136	0.727	121	0.787	126

Table 7.6: Results of ROC analysis after rotation of the data for all models created. Three population comparisons took place and were defined earlier (figure 5.3). The data was rotated about the origin for angles between 1^o and 180^o in 1^o increments. At each angle, a ROC curve was constructed based on the distributions of data from each defined group (pre-injection, post-injection control, health data or post-injection with tumour). The optimal angle of rotation was defined as the angle at which the maximum AUC was obtained. These angles, along with the optimal AUC are displayed.

The loadings plots of a PCA model indicate which metabolites contribute most significantly to each principal component. The location of variables in the loading plot corresponds directly with the samples in the scores plot and describes how the metabolites changes between populations (Murdoch et al., 2008). The distance to the origin corresponds to the relative contribution of a particular metabolite.

A loading plot for the PCA model using all metabolites and UV scaling is shown in figure 7.10. It is interesting to note that the six FDR significant metabolites from the univariate analysis of individual metabolites (2-oxoglutarate, fumarate, glucose, trimethylamine and trimethylamine n-oxide from the tumour-bearing population and taurine from the control population) are all located far from the origin. The scores plot for this model shows that the pre-injection samples are located primarily in the first quadrant, meaning that they have greater proportions of 1-methylnicotinamide, 2-oxovalerate, 3-methyl-2-oxovalerate, 4-hydroxybenzoate, glucose, Nicotinamide, valine, xylose and lower levels of 2-oxoglutarate, citrate, fumarate, lactate and succinate than the tumour-bearing samples. Of interest, glucose is located in the top right corner of the plot,

meaning that there is a net decrease in the concentration post-injection. This is consistent with the previous anlaysis. Fumarate, the other FDR significant metabolite remaining after removal of metabolites with $p<0.05$ in the control population, is found in the lower left hand corner, and thus representing an increase in concentration post-injection. This also was consistent with the univariate analysis.

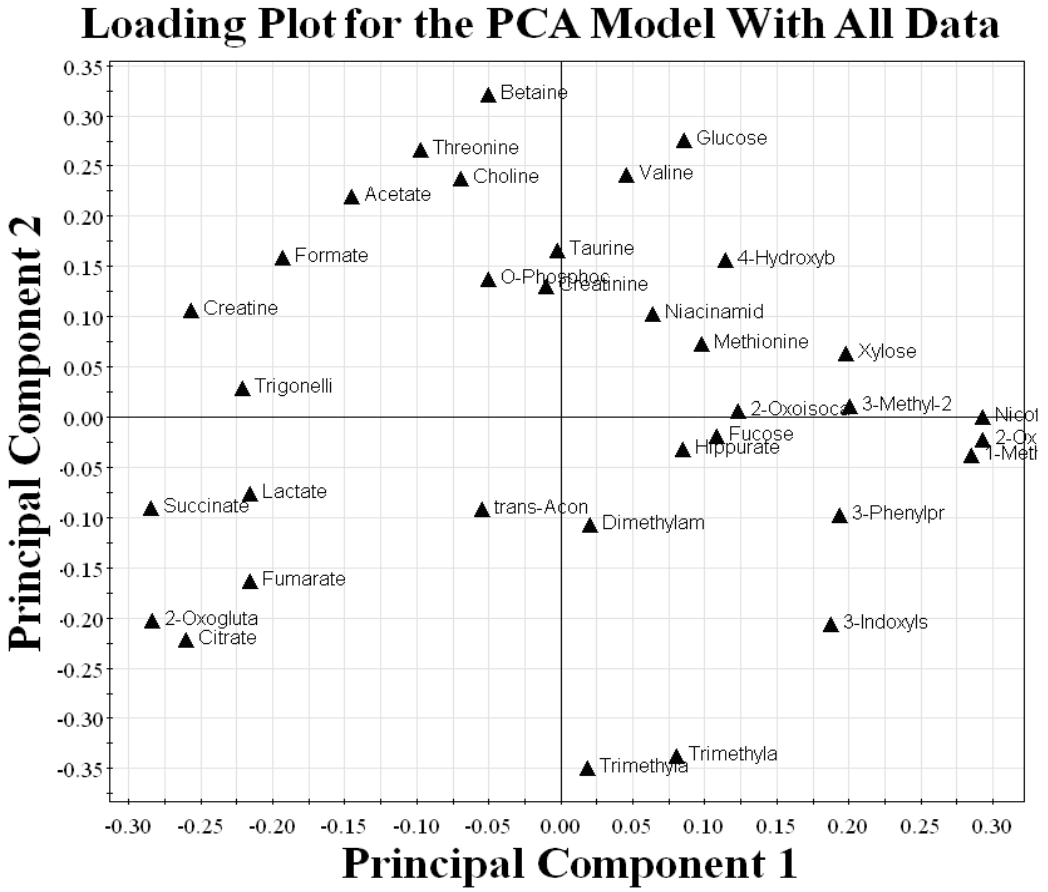

Figure 7.10: Loading Plot for the PCA model using all metabolites and UV scaling. The distance between the origin and each metabolite represents its contributions to each component axis. For instance, 1-Methylnicotinamide (1-Meth) has a strong weighting along component axis 1, but little impact along component axis 2. The score and loading plots are directly related, so sample in the first quadrant has greater levels of metabolites with positive weights along each component axis. As an example, the pre-injeciton samples (generally found in the first quadrant) will have higher levels of glucose, valine and 4-Hydroxybenzoate (4-Hydroxyb) and lower levels of citrate, 2-oxoglutarate (2-oxogluta) and fumarate than the tumour-bearing samples. The FDR significant metabolites from the anlaysis of individual metabolites (2-oxoglutarate, fumarate, glucose, trimethylamine and trimethylamine n-oxide in the tumour-bearing group and taurine from the controls) are all located far from the origin.

This same analysis was applied to the PLS-DA models. A scores plot for the first two latent variables for the PLS-DA model with all metabolites and UV scaling is shown in figure 7.11. It is expected that optimal separation should be near 0^o for these models as PLS-DA is designed to identify the maximum separation between pre-defined classes and align the first latent variable in this direction.

The ROC curves were calculated for all ten models for the three class comparisons: 1) pre and post-injection tumour-bearing groups, 2) post-injection controls and post-injection with tumour and 3) healthy versus cancer. The results are summarized in table 7.7. The ROC curves created using the scores plot in figure 7.11 are displayed in figure 7.12. The AUC for all these curves exceeds 0.9. The AUC for the healthy versus cancer group is smaller than between the pre and post-injection tumour-bearing data due to the slight overlap of the post-injection control data with the post-injection tumour-bearing data. This was observed for all models.

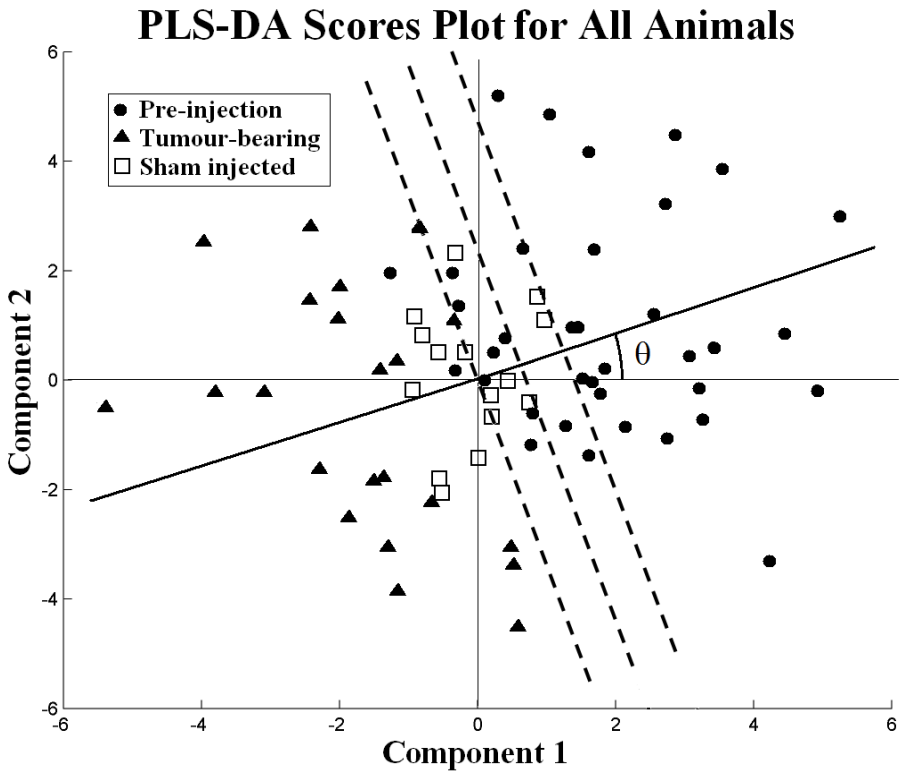

Figure 7.11: PLS-DA model for metabolite concentrations between the pre-injection and post-injection tumour-bearing data when all metabolites were considered. There is clear separation between the healthy (pre-injection and the tumour-bearing population). Applying ROC curve analysis to the plot at an angle, θ, of 21^o gave an AUC of 0.996. The control data clusters together between the healthy and tumour-bearing populations.

Table 7.7 ROC Curve Analysis on the PLS-DA Models

	Model Summary				Pre and Post Cell Injection		Post-injection Data		Healthy versus Cancer	
Data Set	scaling	R_x^2	R_y^2	Q^2	AUC	Angle (deg)	AUC	Angle (deg)	AUC	Angle (deg)
All data	UV	0.421	0.837	0.567	0.996	21	0.912	17	0.969	20
	Par +	0.481	0.810	0.556	0.981	23	0.869	33	0.946	33
Removed p < 0.01 both	UV	0.418	0.824	0.505	0.995	30	0.838	18	0.938	28
	Par	0.481	0.795	0.517	0.977	38	0.792	30	0.915	38
Removed p < 0.01 control	UV	0.381	0.651	0.427	0.980	26	0.808	20	0.920	27
	Par	0.449	0.602	0.427	0.964	36	0.768	33	0.899	35
Removed p < 0.05 both	UV *	0.423	0.771	0.422	0.980	25	0.763	15	0.898	28
	Par #	0.478	0.729	0.412	0.958	37	0.709	19	0.866	36
Removed p < 0.05 control	UV	0.360	0.615	0.354	0.972	33	0.765	27	0.893	29
	Par	0.425	0.542	0.326	0.937	42	0.716	48	0.850	47

Table 7.7: Summary of the results of ROC analysis on the ten PLS-DA models for the three class comparisons. The AUC for the PLS-DA components 1 and 2 is presented in the table, although in some cases components 1 and 3 produced a marginally better AUC; the model labelled with a '+' had a better result (0.4% larger) for the cancer comparison, the model with the '*' marker had better results for both the cancer (0.2% larger) and post-injection (5.1% larger) comparisons and the model labelled with '#' had a better result for the post-injection comparison (11.2% larger). The optimal AUC produced after rotation of the scores plot and the angle of rotation are displayed.

Three ROC curves from the models created after removal of the FDR significant metabolites in the control population from the analysis are shown in figure 7.13. The results show that all models have excellent ability to distinguish the pre-injection and tumour-bearing data with AUC's in excess of 0.937.

PLS-DA is a standard method used in the analysis of metabolomics data sets. The technique is designed to separate pre-defined populations and provides information about more complicated patterns of metabolic behaviour. In our study, scores plots of the models showed good separation between the pre-injection, post-injection with tumour and post-injection controls. The fact that the control data, which was projected onto the model, did not cluster with the pre-injection data suggested that metabolic changes are taking place. However, the post-injection data from both groups were well separated, meaning that we can distinguish the populations based on the complement of all metabolic changes taking place. ROC curve analysis revealed excellent separation between the controls and the tumour bearing data, post injection (AUC = 0.912).

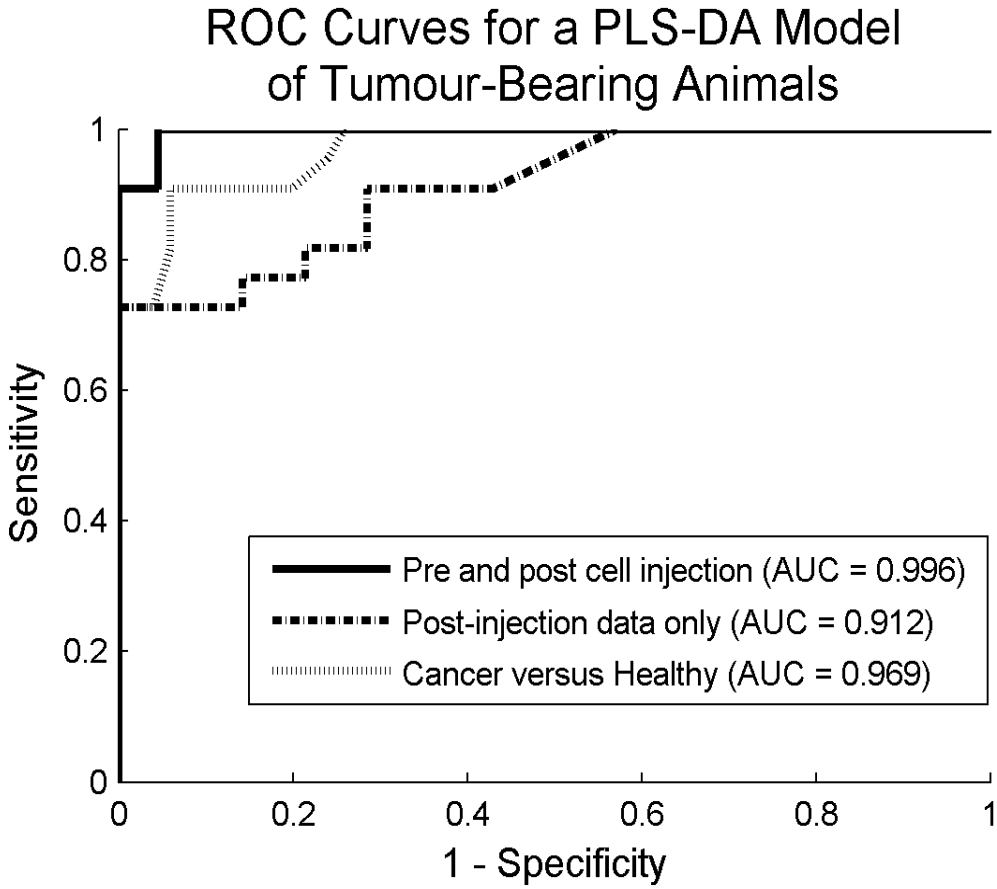

Figure 7.12: ROC curves for the scatter plot shown in figure 7.11. This figure shows that the model successfully separates healthy and tumour-bearing samples. The AUC is greatest when considering only the pre-injection and post-injection samples from the tumour-bearing mice. Despite similar metabolic changes in the control animals, the model is still able to identify the tumour-bearing samples.

The metabolite concentrations must be scaled prior to constructing the PLS-DA model as high concentration metabolites will dominate the results if not scaled properly. Two common methods for scaling are Pareto scaling, which scales high concentration metabolites down and limits the up scaling of noise, or UV (also referred to as auto scaling), which places equal weightings on all metabolites. Our results show that there is little change between the models using either scaling method. However, UV scaling did produce models with larger R^2 and Q^2 values.

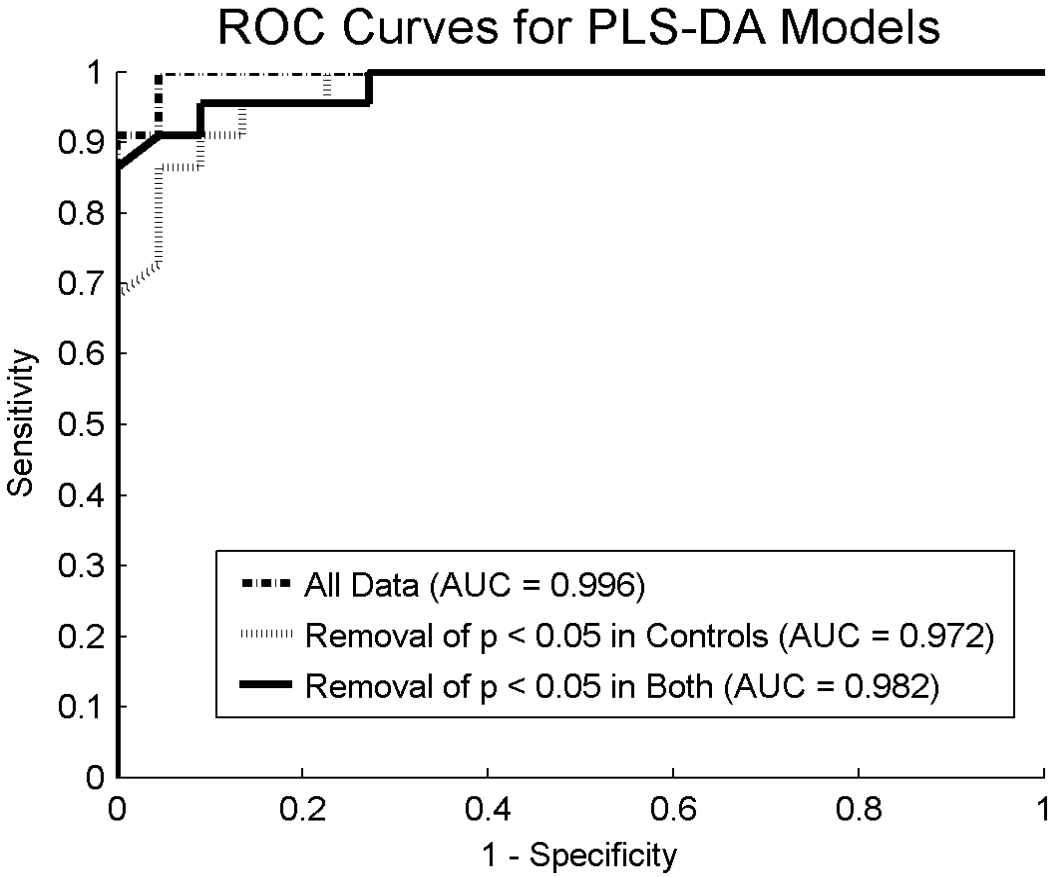

Figure 7.13: ROC curves of the tumour-bearing data for the PLS-DA models (UV scaling) of the three data sets; all metabolites present, removal of metabolites with $p < 0.05$ in both populations and removal of metabolties with $p < 0.05$ in the control population. For all cases, the comparison was between the pre and post injection tumour-bearing data. The figure shows that the AUC decreases as metabolites are removed from the analysis, but only by a few percent. The magnitude of the decrease is larger when considering the models with pareto scaling or when analyzing the 'post-injection data' or 'healthy versus cancer' class comparisons.

PLS-DA models have been criticised for over-fitting the data (Westerhuis et al., 2008). For this reason it is important to validate the model via permutation testing. Permutation testing was performed on all models in SIMCA with 999 permutations. Valid models were required to have a Q^2 values greater than any permuted model and a R^2 value greater than a majority of the permuted models as displayed in the validation plot. All models passed on the basis of the R^2 results, but one model – removal of metabolites with $p < 0.05$ in the control populations with Pareto scaling – failed the Q^2 condition; this case had one permutation which provided a Q^2 value exceeding the actual model's Q^2.

The permutation testing results are displayed in figures 7.14 and 7.15 for a good model (all metabolites present and UV scaling) and the model that failed. As observed in figure 7.14, the R^2 and Q^2 values are both larger than 0.5 and are largest for the original model (fraction of correct classifiers of 1.0). In contrast, the failed model only has an R^2 value greater than 0.5 and has a large number of permuted models having a R^2, and one Q^2 value, greater than that for the original model.

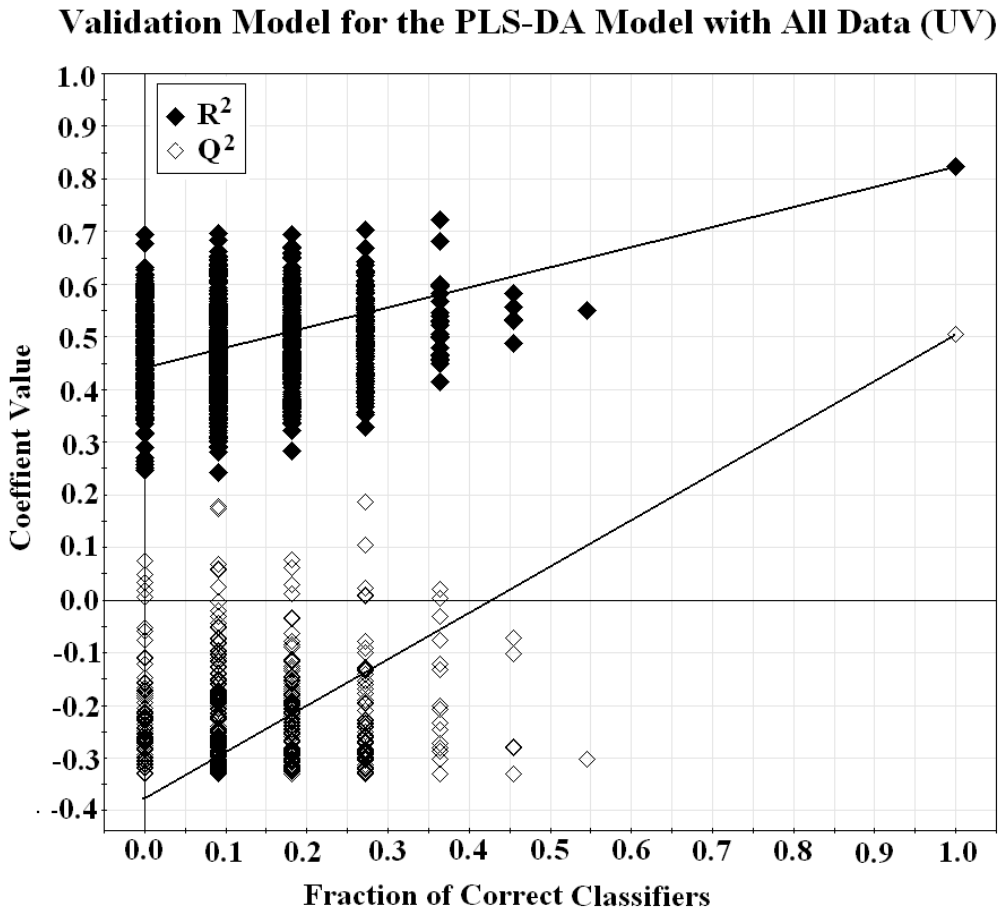

Figure 7.14: Cross validation of the PLS-DA model for all data using UV scaling. This model satisfies the conditions for a good model: 1) The Q^2 value of the original model is the maximum, and 2) The R^2 value for the original model is greater than a majority of the permuted models. In additional, the Q^2 value of 0.567 exceeds the desired value of 0.5

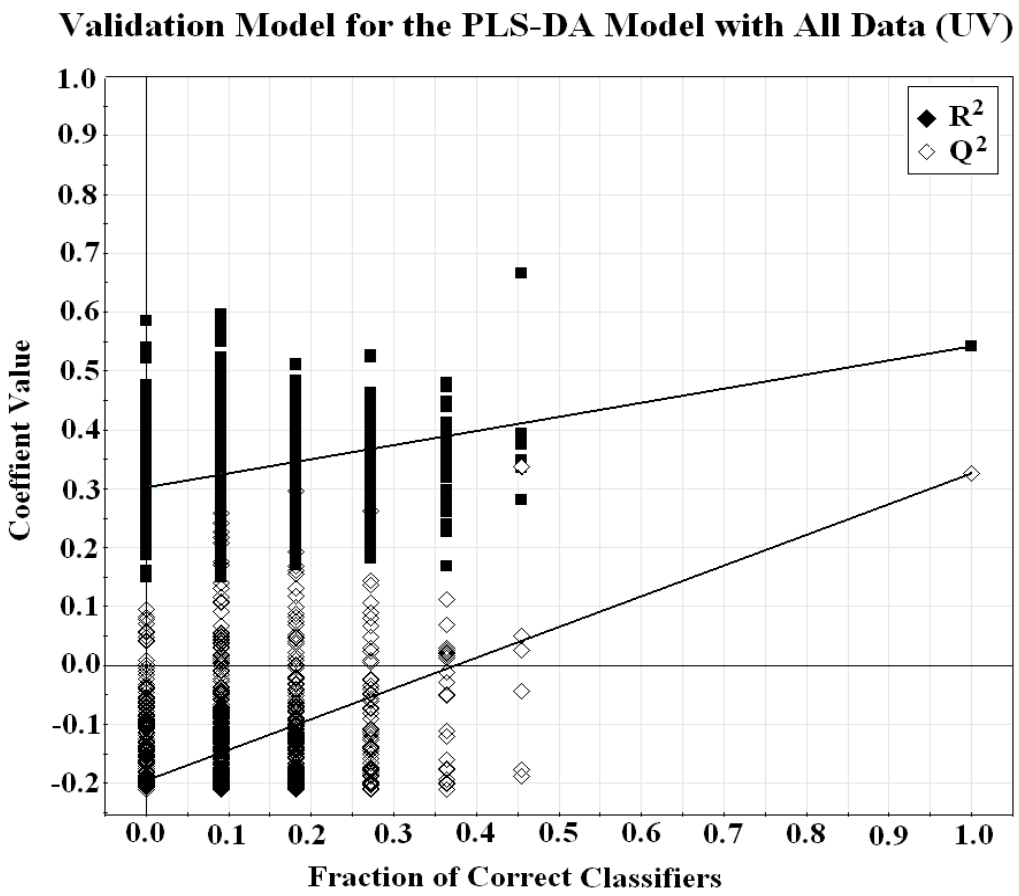

Figure 7.15: Cross validation for the PLS-DA model with all metabolite with $p < 0.05$ in the control population removed. This model does not satisfy the conditions for a good model: 1) The Q^2 value of the original model is not a maximum, however 2) The R^2 value for the original model is greater than a majority of the permuted models.

The loading plot for the model with all metabolites and UV scaling is displayed as figure 7.16. It is observed that the five FDR significant metabolites from the univaraiate analysis of individual metabolites all have large weights along latent variables 1 and 2. This supports the previous results in suggesting that these metabolites have potential to discriminate healthy from tumour-bearing urine samples.

To evaluate the ability of this technique to detect the onset of disease at a relatively early stage, we constructed PLS-DA models, using both UV and Pareto scaling, with the pre-injection and large tumour samples. Samples in the large tumour category were collected an average of 22.1 ± 9.6 days post cell injection. Data for the small tumours (6-8 days post-injection) are projected onto the models to investigate the level of similarities in the data structure of the early and large stage tumour data. PLS-DA was selected for its ability to optimize the separation between classes.

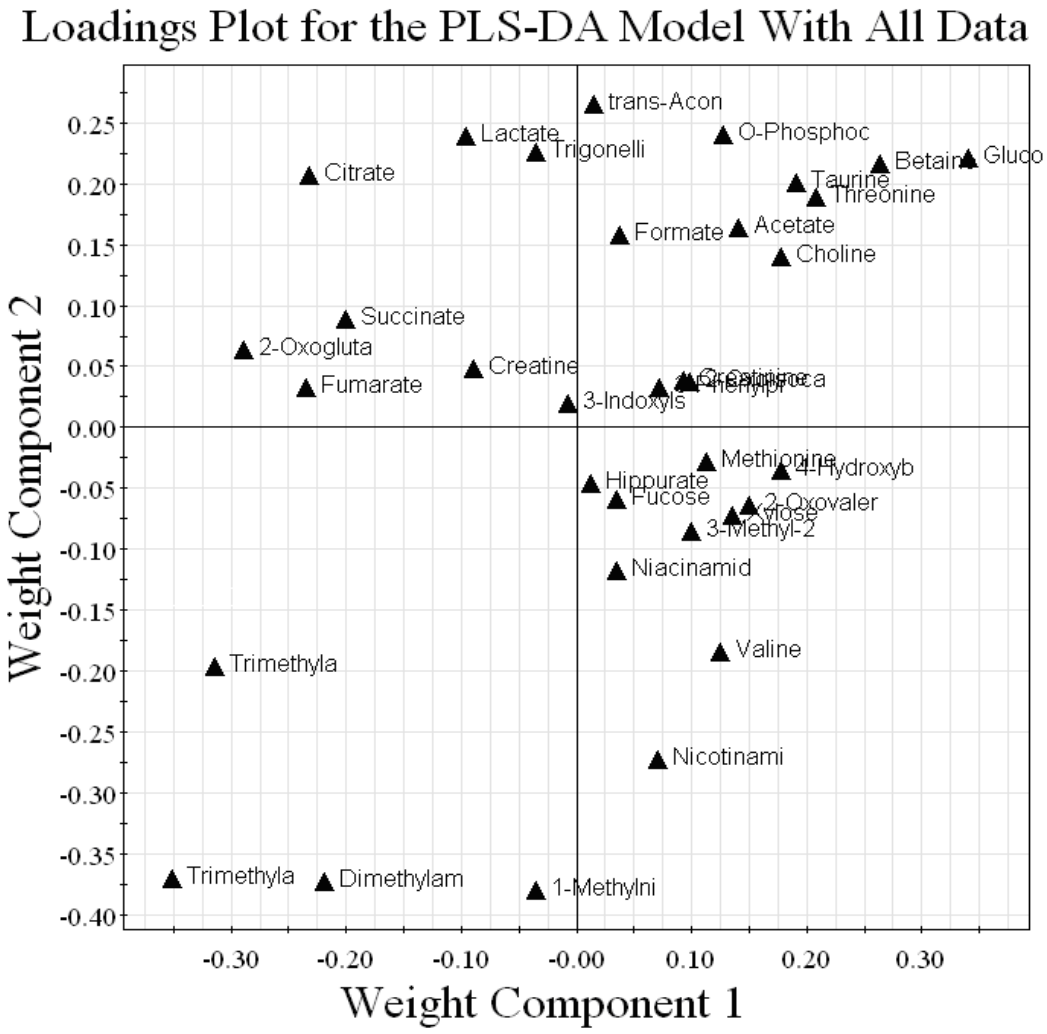

Figure 7.16: Loading Plot for the PLS-DA model using all metabolites and UV scaling. The distance between the origin and each metabolite represents its contributions to each component axis. For instance, 1-Methylnicotinamide (1-Methylni) and trans-Aconitate have strong weights along component axis 2 and small weight along component axis 1. The score and loading plots are directly related, therefore, it is simple to determine what metabolites are more abundant in each population. As an example, the pre-injeciton samples (clustering on the left hand side of the plot) will have higher levels of choline, glucose, betaine, taurine and threonine and lower levels of dimethylamine, trimethylamine and trimethylamine n-oxide than the tumour-bearing samples. The five FDR significant metabolites from the anlaysis of individual metabolites (2-oxoglutarate, fumarate, glucose, trimethylamine and trimethylamine n-oxide) in the tumour bearing populations are have large weights along weight components 1 and 2.

The scores plot for the models with UV scaling is shown in figure 7.17. In both models, the small tumour data clustered with the large tumour data, suggesting that even at early stage, similar metabolic changes are present.

Four class comparisons were considered for this analysis: the first evaluated the separation between the pre-injection and late stage tumour data, the second was between the pre-injection and late stage data, the third between the pre-injection and all tumour data (early and late stage date remained separate) and the fourth compared the early and late stage data. The fourth class comparison was performed to indicate if further metabolic changes occurred as the tumour developed. Table 7.8 summarises the models R^2 and Q^2 values, as well as the best AUC and angle of rotation for the four cases.

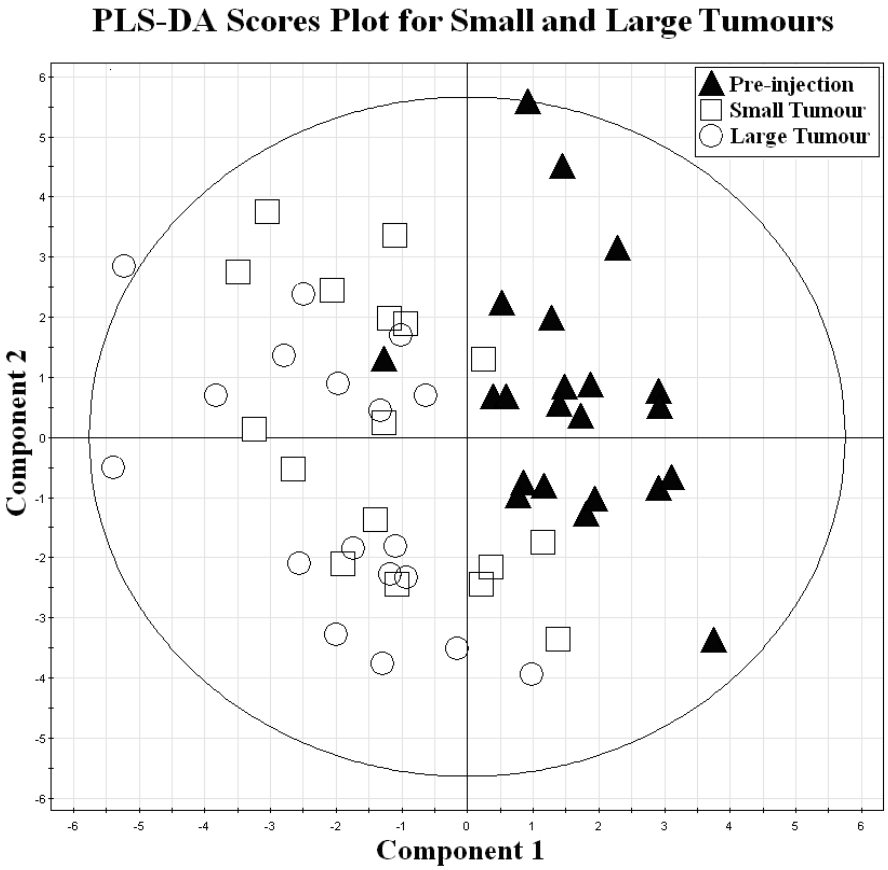

Figure 7.17: PLS-DA model for the pre-injection and large tumour (22.1 ± 9.6 days post cell injection) samples. The AUC for the pre-injection and large tumour samples is 0.995 at 31^o with respect to the horizontal axis, thus showing that there is good separation between the two populations. The small tumour data (6-8 days post cell injection) clusters with the large tumour data.

Table 7.8: PLS-DA Analysis for the Pre-injection and Large Tumour Data

scaling	R_x^2	R_y^2	Q^2	Pre vs. All Tumour		Pre vs. Small Tumour		Pre vs. Large Tumour		Small vs. Large Tumour	
				AUC	Angle (deg)	AUC	Angle (deg)	AUC	Angle (deg)	AUC	Angle (deg)
UV	0.355	0.738	0.552	0.981	22	0.971	23	0.995	31	0.752	29
Par	0.416	0.673	0.493	0.975	30	0.961	30	0.987	28	0.699	45

Table 7.8: Summary of the PLS-DA models created for the pre-injection and large tumour data (22.1 ± 9.6 days post-injection) after applying a $\log_{10}$ transformation to the normalized data set, mean centering and scaling (indicated in the second column where UV = unit variance and par = pareto scaling). The small tumour data was projected onto the model. The small tumour data clusters with the large tumour data, revealing similarities in the metabolic content of these samples. Four groups of data were considered: the first involved the pre-injection and all tumour data (small and large), the second used the pre-injection and small tumour data, the third used the pre-injection and large tumour data and the fourth compared the small and large tumour data.. The optimal AUC produced after rotation of the scores plot through angles between 0^o and 180^o, and the angle of rotation for all groups considered are displayed.

The model for the pre-injection and large tumour samples shows that even at an early stage in development, tumours produce similar changes in the metabolic content of urine. The model with UV scaling has an AUC of 0.971, at an angle of 23^o, when separating the pre-injection and small tumour samples. The large AUC, in combination with agreement of the optimal angle of rotation between the pre-injection and early stage data and the pre-injection and late stage data, suggests that if the metabolic changes between a healthy group and a populations with late stage tumour growth are identified, screening for cancer at early stage is possible.

Inclusion of the large tumour samples increases this value to 0.981. However, the AUC was smaller than it was for the single tumour-bearing class. The discrepancy is likely caused by the larger number of tumour-bearing data points in the new model since the tumour-bearing population was divided into early and late-stage.

Applying ROC curve analysis to the small and large tumour classes shows that a maximum AUC of 0.752 (UV scaling) or 0.699 (Pareto scaling) is achieved between the groups. The fact that the AUC in this case is well above 0.5 indicates that the changes in urinary metabolism continue to evolve as the tumour grows; however, the extent of the observed metabolic changes are smaller than those observed between the pre-injection samples and the small tumours. Furthermore, the optimum angle of rotation was similar between this class comparison and between pre-injection and tumour-bearing data.

The loadings plot for the model with UV scaling is shown in figure 7.18. The figure shows that four of the five FDR significant metabolites from the univariate statistical analysis of individual metabolites (2-oxoglutarate, glucose, trimethylamine and trimethylamine n-oxide) all have large weights along the first latent variable. This means that they are responsible for the majority of the variation along this axis and therefore supports the results obtained earlier. It is interesting to note that most of the metabolites are located in the same quadrants as in figure 7.16 as expected.

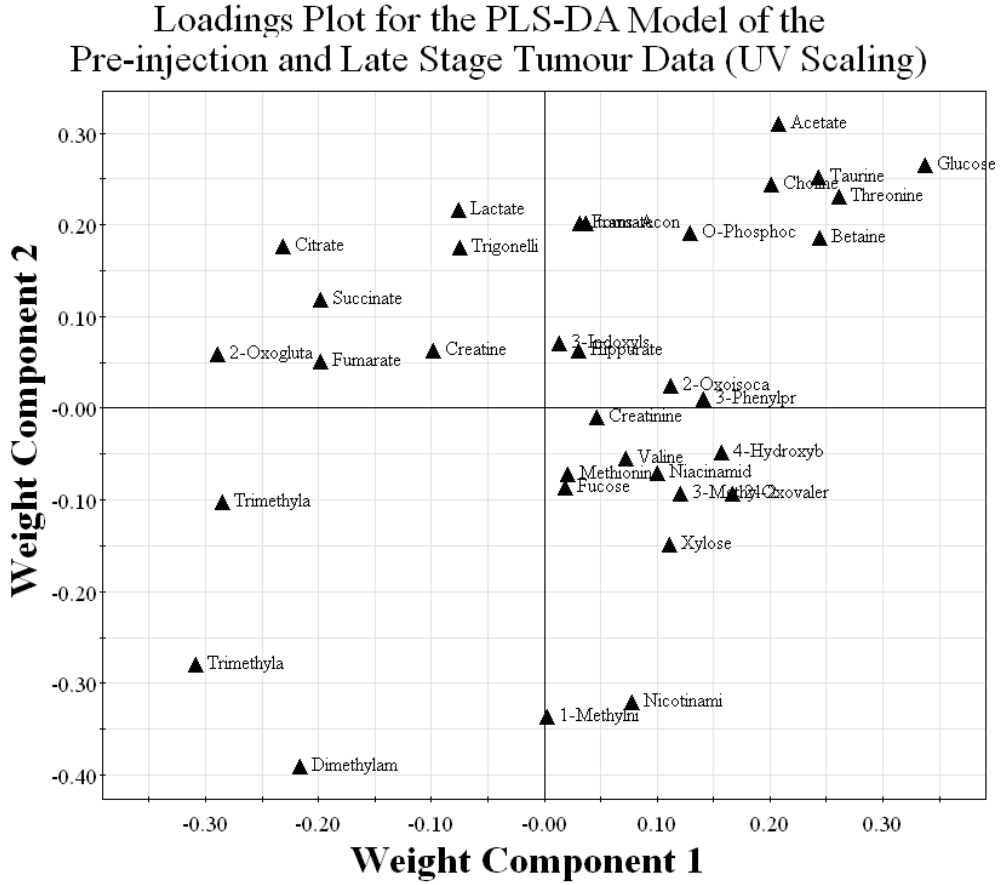

Figure 7.18: Loading Plot for the PLS-DA model of the pre-injection and late stage tumour data using UV scaling. This plot is similar to that shown in figure 7.16 with few differences related to the locations of the metabolites in each quadrant. In general, metabolites are in the same quadrant as in figure 7.16. Four of the five FDR significant metabolites from the anlaysis of individual metabolites (2-oxoglutarate, glucose, trimethylamine and trimethylamine n-oxide) are responsible for the majority of the variation observed along component axes 1 as indicated by the maximum distances from the origin. Based on the results of the ROC curve analysis (maximum AUC at an angle of 22^{o} for the pre-injection and all tumour data), acetate, dimethylamine, taurine and threonine are additional metabolites used to discriminate between the two classes.

ROC curve analysis has found widespread use in medical applications, but has rarely been used in the field of metabolomics. ROC curves can simultaneously present information about the sensitivity (fraction of diseased patients correctly diagnosed as having the disease) and the specificity (the fraction of healthy individuals correctly diagnosed as being healthy) of a diagnostic test.

The use of ROC curve analysis in metabolomics research will provide a measure of the clinical significance of a metabolic change. ROC curves with an AUC exceeding 0.8 are considered good tests for diagnostics, while curves with an AUC greater than 0.9 are excellent. Glucose and trimethylamine n-oxide were two individual metabolites in which the concentration changed in a Bonferroni significant manner post-injection and produced AUC values in excess of 0.8. Seven metabolite ratios produced an AUC value in excess of 0.9 and PLS-DA analysis of the data resulted in two models with AUC's in excess of 0.975 (pre and post cell injection samples for the tumour-bearing animals). More complicated analysis techniques are better able to identify metabolic behaviours that are indicative of the presence of cancer. The use of rotated axes for ROC curve construction seems intuitive for scatter plot analysis, however, to our knowledge, this technique has not been published.

7.5 Prospects of the Study

Our study involved the collection of multiple urine samples from mice before and after injection of the GBM-xenograft cells. Averaging the samples on a per mouse basis, before and after, should have the effect of reducing 'noise' in the metabolic signature as a result of day to day variation. All mice used in this study provided urine samples before and after cell injection, so a paired test was used. The distributions of the pre and post-injection data, which were created from the pooling of averaged data from all mice, were often not normal, so the non-parametric Wilcoxon rank sum test was used. In pairing the data from each mouse, we have effectively reduced the variation between individuals, which has been suggested to exceed that found within an individual (Bollard et al., 2005; Slupsky et al., 2007). The results obtained from the analysis of the paired data should emphasize metabolic changes associated with tumour on an individual basis and simulate a clinical-like environment in which urine samples, from one individual, are compared between screening tests over the years.

One limitation of the study is the highly controlled environment. All mice were age matched and were housed in a highly-controlled environment in terms of the room temperature, humidity, light/dark cycle (12 hours each), method of urine collection, and the diet. Animals were fed the same diet and exposed to the same environmental stressors (WBI one week post arrival, cell / PBS injection two weeks post arrival and method of urine collection). This level of control is not realistic when testing a human population. External metabolic influences, unrelated to the tumour, will complicate interpretation of the results. In studies of human populations, a larger number of patients are required to acquire an understanding of the non-tumour related metabolic behaviours and to determine what constitutes a healthy sample.

Having a small number of animals in the study increases the chance that a metabolite concentration or ratio will achieve statistical significance when there is no significant change. Random errors in the reported metabolite concentration, resulting from noise, could overestimate the significance of a variable and lead to a false discovery. Application of the FDR method or Bonferroni correction will reduce the occurrence of false discoveries, but these methods in themselves rely on the relative change of the means of the metabolite before and after the cell injection.

The high presence of cancer in the test population tends to bias the findings towards significance. Some metabolites or metabolite ratios were shown to change in the same direction in both the control and tumour-bearing populations. Such changes would significantly impact the specificity of a diagnostic test in a clinical environment. For this reason, we chose to remove any variable with $p<0.01$ or $p<0.05$ in the control population. This choice is conservative, but does emphasize that significant metabolic changes, likely due to the presence of tumour, are still present. The remaining variables (identified as the black circles in figures 7.4 and 7.6) could be potential indicators for cancer screening.

The idea of monitoring a metabolic signature for changes is not new. These data support the idea that a simple, biofluid-based program holds potential as a screening tool for cancer. Significant human data would be required to identify a “normal” age-dependent metabolic trajectory. Identification of deviations from this path that correlate with the presence of cancer in general (and possibly specific forms of cancer) would potentially allow earlier diagnosis and more successful treatment.

Chapter 8
Conclusion

In this work, we have shown that metabolic changes occur in the presence of tumour. Urine samples were collected daily from NIH III nude mice (n=22) before and after injection with GBM xenografts. The metabolic content of the samples were measured with an 800 MHz NMR spectrometer using the 1-D NOESY pulse sequence. A total of 34 metabolites were identified and quantified via targeted profiling with Chenomx Suite 5.1. A set of control animals (n=14) were studies to identify metabolic changes that occur naturally and are not likely altered by the tumour.

Univariate statistics identified a number of statistically significant metabolic changes in both the metabolite concentrations and the ratio of pairs of metabolites. Analysis of the control animals revealed that some metabolic changes were occurring naturally or as a response to the injection procedure. We chose to acknowledge these changes by removing any metabolite with $p<0.05$ or $p<0.01$, in only the control population or in both populations, from the analysis. Despite the conservative choice of significance, there remained useful metabolic information which may be used for screening for cancer.

The analysis of a single metabolite may not be sufficiently specific for cancer screening. After removal of significant metabolites in the control population, only two FDR significant metabolites remained. The analysis of metabolite ratios provided more information, as shown through a larger number of statistically significant variables and improved ROC statistics. The results showed that 43/85 FDR significant ratio did not include any FDR significant metabolites. This added information may be useful for identifying tumour from other diseases.

Multivariate statistical analysis provided the most diagnostic information. The scores plots showed clustering of like samples (i.e. healthy with healthy and tumour-bearing with tumour-bearing) and separation between the different classes. PLS-DA models were generally better than PCA models for discriminating the two groups. ROC analysis of the PLS-DA models showed that most models had AUC in excess of 0.9 between the pre and post injection data of the tumour-bearing animals and greater than

0.7 for the post-injection data of the controls and tumour-bearing mice. Further, the FDR significant metabolites were far from the origin on the loadings plots, thus supporting the claims of univariate statistical analysis. PLS-DA models of the pre-injection and last stage data suggest that early detection of cancer is possible.

This work supports the idea that a simple, NMR-based analysis of a biofluid such as urine could be beneficial in screening for cancer. The approach used in this study would be similar to asking patients to submit a urine sample one or more times per year for analysis. Over time, a "normal" metabolic trajectory could be obtained which would serve as a baseline against which each subsequent NMR sample could be compared. If a sample showed significant deviation from the baseline in a manner that was consistent with a "cancer trajectory", the patient could be referred for additional testing. This baseline metabolic signature would require significant characterization to account for the many factors (i.e. age, sex, other disease presence, diet etc.) that cause inter and intra-individual variation. Although this would represent a significant challenge in data analysis, the raw materials (urine, multivariate analysis software) are cheap, the technology is readily available and the concept is simple. Furthermore, if this technique proved effective at diagnosing various cancers at an earlier stage than they are with current techniques, there could be significant improvements in patient outcomes.

Future work related to this study will compare the metabolic information obtained outside the tumour (i.e. from urine) with that which can be obtained directly in vivo using 9.4T magnetic resonance spectroscopy. The quality of in vivo spectroscopy is significantly inferior to that of NMR, however, it will be interesting to observe if any similar metabolic patterns can be observed in the two types of signals. In addition, the spectroscopic data from the urine will be compared to ex vivo analysis of the cultured cells to identify correlations with the in vivo data.

In conclusion, the application of ROC analysis to a metabolomics data set provides useful information for screening purposes. Plots of the AUC against the p-value relate information about the clinical and statistical significance of all metabolic changes simultaneously. Such figures may be used to assess the capacity of using certain metabolites for screening. The application of ROC curves to a PLS-DA scores plot is new. Optimal separation does not occur along a single latent variable, but a combination of two. ROC curve analysis was applied to the scores plot after rotation of the data set in unit angle increments. The optimal angle of rotation was defined as the angle at which

the maximum AUC was obtained. It is expected that this direction will contain the best diagnostic information related to the tumour. Application of the Youden index or the minimum distance from the point (0,1) will identify the optimal decision threshold.

List of References:

Ala-Korpela M. Potential Role of Body Fluid ^{1}H NMR Metabonomics as a Prognostic and Diagnostic Tool. *Expert Rev Mol Diagn.*2007; 7: 761-773.	p. 47
Ala-Korpela M. Critical Evaluation of ^{1}H NMR Metabonomics of Serum as a Methodology for Disease Risk Assessment and Diagnostics. *Clin Chem Lab Med.* 2008; 46: 27-42.	p. 8, 9, 37
Akin O, Riedl C C, Ishill N M, Moskowitz C S, Zhang J and Hricak H. Interactive Dedicated Training Curriculum Improves Accuracy in the Interpretation of MR Imaging of Prostate Cancer. *Eur Radiol.* 2010; 20: 995-1002.	p. 61
Akobeng A K. Understanding Diagnostic Tests 3: Receiver Operating Characteristic Curves. *Acta Paediatrica. 2007;* 96: 644-647.	p. 1, 58, 60-63
Altmaier E, Ramsay S L, Graber A, Mewes H W, Weinberger K M and Suhre K. Bioinformatics Analysis of Targeted Metabolomics – Uncovering Old and New Tales of Diabetic Mice Under Medication. *Endocrinology*. 2008; 149: 3478-3489.	p. 8
Altman D G and Bland J M. Diagnostic Tests 3: Receiver Operating Characteristic Plots. *Brit Med J.* 1994; 309: 188.	p. 58, 60, 61
Alvarenga A V, Infantosi A F C, Pereir W C A and Azevedo C M. Assessing the Performance of Morphological Parameters in Distinguishing Breast Tumours on Ultrasound Images. *Med Eng Phys.* 2010; 32: 49-56.	p. 6, 7
Antoniewicz M R, Stephanopoulos G and Kelleher J K. Evaluation of Regression Models in Metabolic Physiology: Predicting Fluxes in Isotopic Data without Knowledge of the Pathway. *Metabolomics.* 2006; 2: 41-52.	p. 45, 46, 50, 54, 55
Aranibar N, Ott K, Roongta V and Mueller L. Metabolomic Analysis Using Optimized NMR and Statistical Methods. *Anal Biochem.* 2006; 355; 62-70.	p. 32, 33, 35
Bales J R, Higham D P, Howe I, Nicholson J K and Sadler P J. Use of High-Resolution Proton Nuclear Magnetic Resonance Spectroscopy for Rapid Muli-Component Analysis of Urine. *Clin Chem.* 1984; 30: 426-432.	p. 69

Barba I, Fernandez-Montesinos R, Garcia-Dorado D, and Pozo D. Alzheimer's Disease Beyond the Genomic Era: Nuclear Magnetic Resonance (NMR) Spectroscopy-Based Metabolomics. *J Cell Mol Med.* 2008; 12: 1477-1485.	p. 8
Barker M and Rayens W. Partial Least Squares for Discrimination. *J Chemometrics.* 2003; 17: 166-173.	p. 49, 50, 54
Bean S M and Chhieng D C. Anal-Rectal Cytology: A Review. *Diagn CytopatholI.* 2009; 38: 538-546.	p. 7
Beckonert O, Keun H C, Ebbels T M D, Bundy J, Holmes E, Lindon J C and Nicholson J K. Metabolic profiling, Metabolomic and Metbonomic Procedures for NMR Spectroscopy of Urine, Plasma, Serum and Tissue Extracts. *Nat Protoc.* 2007; 2: 2692-2703.	p. 8, 22, 26, 37, 40, 67
Beger R, Schnackengberg L, Holland R, Li D and Dragan Y. Metabonomic Models of Human Pancreatic Cancer Using 1D Proton NMR Spectra of Lipids in Plasma. *Metabolomics.* 2006; 2: 125-134.	p. 11
Benjamini Y and Hochberg Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met.* 1995; 57: 289-300.	p. 42-45, 90
Bertram H C, Melmendal a, Petersen B O, Madsen J C, Pedersen H, Nielsen N C, Hoppe C, Molgaard C, Michealsen K f and Duus J O. Effect of Magnetic Field Strength on NMR-Based Metabonomic Human Urine Data. Comparative Study of 250, 400, 500, and 800 MHz. *Anal Chem.* 2007; 79: 7110-7115.	p. 67
Bewick V, Cheek L and Ball J. Statistics Review 13: Receiver Operating Characteristic Curves. *Critical Care.* 2004; 8: 508-512.	p. 57, 58, 62-64
Bollard M E, Holmes E, Lindon J C, Mitchell S C, Branstetter D, Zhang W, Nicholson J K. Investigation into Biochemical Changes Due to Diurnal Variation and Estrus Cycle in Female Rats Using High Resolution ^{1}H NMR Spectroscopy of Urine and Pattern Recognition. *Anal Biochem. 2001;* 295: 194-202.	p. 8, 9, 15, 34, 35, 46, 50, 53, 67
Bollard M E, Stanley E G, Lindon J C, Nicholson J K and Holmes E. NMR-Based Metabonomic Approaches for Evaluating Physiological Influences on Biofluid Composition. *NMR in Biomedicine.*2005; 18: 143-162.	p. 8-11, 53, 56, 112
Broadhurst D I and Kell D B. Statistical Strategies for Avoiding False Discoveries in Metabolomics and Related Experiments. *Metabolomics.*2006; 2: 171-196.	p. 42, 43, 54, 55, 62, 79, 85
Cady B and Michaelson J S. The Life-Sparing Potential of Mammographic Screening. *Cancer.* 2001; 91: 1699-1703.	p. 5, 6

Centor R. Signal Detectability: The Use of ROC Curves and Their Analysis. *Med Decis Making.* 1991; 11: 102-106.	p. 60
Chen J, Wang W, Lv S, Yin p, Zhao X, Lu X, Zhang F and Xu G. Metabonomics Sgudy of Liver Cancer Based on Ultra Performance Liquid Chromatography Coupled to Mass Spectrometry with HILIC and RPLC Separations. *Anal Chem ACTA.* 2009; 650: 3-9.	p. 11
Cheng D, Banack C D and Shah S L. Robust Baseline Correction Algorithm for Signal Dense NMR Spectra. *Journal of Magnetic Resonance*. 2007; 187: 288-292.	p. 71
Chenomx NMR Suite 5.1 User Guide. Chenomx Inc. Edmonton, Alberta. 2008; 11-67.	p. 67, 70, 71
Claridge T D W. High-Resolution NMR Techniques in Organic Chemistry, 2nd ed. Amsterdam, London, Elsevier Inc., 2009.	p. 19-35
Clarke C and Haselden J. Metabolic Profiling as a Tool for Understanding Mechanisms of Toxicity. *Toxicol Pathol.* 2008; 36: 140-147.	p. 8, 9, 41
Claudino W M, Quattrone A, Biganzoli L, Pestrin M, Bertini I and Leo A D. Metabolomics: Available Results, Current Research Projects in Breast Cancer, and Future Applications. *J Clin Oncol.* 2005; 25: 2840-2846.	p. 8, 11, 22, 37, 44
Cook N R. Statistical Evaluation of Prognosis versus Diagnostic Models: Beyond the ROC Curve. *Clin Chem.* 2008; 54: 17-23.	p. 4, 6, 58, 61, 62, 64
De Nooijer J, Lechner L and de Vries H. Early Detection of Cancer: Knowledge and Behaviour among Dutch Adults. *Cancer Det and Prev.* 2002; 26: 362-369.	p. 4
DeLong E R, DeLong D M and Clarke-Pearson D L. Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics.* 1988; 44: 837-845.	p. 1, 58, 60, 61, 63
Destounis S V, DiNitto P, Logan-Young W., Bonaccio E, Zuley M L and Willison K M. Can Computer-aided Detection with Double Reading of Mammograms Help Decrease the False Negative Rate? Initial Experience. *Radiology*. 2004; 232: 578-584.	p. 6
Dieterle F, Ross A, Schlotterbeck G and Senn H 2006. Account for Dilution of Complex Biological Mixtures. Application in ^{1}H NMR Metabonomics. *Anal Chem.* 2006; 78: 4281-4290.	p. 76, 77
Dunn and Ellis. Metabolomics: Current Analytical Platforms and Methodologies. *Trends Anal Chem.* 2005; 24: 285-294.	p. 8-10, 22, 29, 36, 37, 40-42

Erb G, Elbayed K, Piotto M, Raya J, Neuville A, Mohr M, Maitrot D, Kehrli P and Namer I J. Toward Improved Grading of Malignancy in Oligodendrogliomas Using Metabolomics. *Mag Res Med.* 2008; 59: 959-965.	p. 11, 44, 54, 95
Evilia R F. Quantitative NMR Spectroscopy. *Anal Letters.* 2001; 34: 2227-2236.	p. 29, 32
Eyre H J. Winning the Cancer Fight: A Look at the Future. *Prim Care Clin Office Pract.* 2009; 36: 859-865.	p. 4-7
Feng B, Yue F, Zheng M H.Urinary Markers in Colorectal Cancer. *Advances in Clin Chem.* 2009; 47: 45-57.	p. 1, 4, 11
Forshed J, Torgrip R J O, Aberg K M, Karlberg B, Lindberg J and Jacobson S P. A Comparison of Methods for Alignment of NMR Peaks in the Context of Cluster Analysis. *J Pharm and Biomed Anal.* 2005; 38: 824-832.	p. 34, 36, 39, 67
Fossel E, Carr J and McDonagh J. Detection of Malignant Tumours: Water-Suppressed Proton Nuclear Magnetic Resonance Spectroscopy of Plasma. *The New England Journal of Medicine.* 1986; 315: 1369-1376.	p. 12
Franko A, Parliament M, Allalunis-Turner J and Wolokoff B. Variable presence of hypoxia in M006 human glioma spheroids and in spheroids and xenografts of clonally derived sublines. *Brit J Cancer*. 1998; 78: 1261–1268.	p. 65
Frickenschmidt A, Frohlich H, Bullinger D, Zell A, Laufer S, Gleiter C, Liebich H and Kammerer B. Metabonomics in Cancer Diagnosis: Mass Spectrometry-Based Profiling of Urinary Nucleosides from Breast Cancer Patients. *Biomarkers.* 2008; 13: 435-449.	p. 10
Gardner I A and Greiner M. Receiver-operating Characteristic Curves and Likelihood Ratios: Improvements Over Traditional Methods for the Evaluation and Application of Veterinary Clinical Pathology Tests. *Veterinary Clinical Pathology.* 2006; 35: 8-17.	p. 58, 60-64
Geladi P and Kowalski B R. Partial Least Squares Regression: A Tutorial. *Analytica Chemica Acta.* 1986; 185: 1-17.	p. 46, 47, 50-54
Giskeodegard G, Grinde M, Sitter B, Axelson D, Lundgren S, Fjosne H, Dahl S, Gribbestad I and Bathen T. Multivariate Modeling and Prediction of Breast Cancer Prognostic Factors Using MR Metabolomics. *J Proteome Res.* 2010; 9: 972-979.	p. 11
Goldsmith P, Fenton H, Morris-Stiff G, Ahmad N, Fisher J and Prasad K R. Metabonomics: A Useful Tool for the Future Surgeon. *J Surgical Res.* 2010; 160: 122-132.	p. 1, 7, 9, 15, 22, 32, 36, 37

Goodacre R. Metabolomics of a Superorganism. *J Nutr.* 2007; 137: 259S-266S. p. 44, 45

Goodacre R, Vaidyanathan S, Dunn W B, Harrigan G G and Kell D B. Metaboloimics by Numbers: Acquiring and Understanding Global Metabolite Data. *Trends Biotech.* 2004; 22: 245-252. p. 36

Gorban A N , Kegl B, Wunsch D C, Zinovyev A. *Principal Manifolds for Data Visualization and Dimension Reduction: Preface*. 1st Ed. Springer Publishing Company, Inc. 2007; v. p. 46

Gowda G a N, Zhang S C, Gu H W, Asiago V, Shanaiah N and Taftery D. Metabolomics-Based Methods for Early Detection Diagnostics. *Expert Rev Mol Diagnostics.* 2008; 8: 617-633. p. 8

Griffin J L. Metabolic Profiles to Define the Genome: Can We Hear the Phenotypes. *Phil Trans R Soc Lond B.* 2004; 359: 857-871. p. 11

Griffin J L. The Cinderella Story of Metabolic Profiling: Does Metabolomics Get to Go to the Functional Genomics Ball. *Phil Trans R Soc B.* 2006; 361: 147-161. p. 10

Griffin J L and Kauppinen R A. Tumour Metabolomics in Animal Models of Human Cancer. *J Proteome Res.*2007; 6: 498-505. p. 10, 11, 36, 40, 41

Griffin J L and Shockcor J P. Metabolic Profiles of Cancer Cells. *Nature Reviews Cancer.* 2004; 4: 551-561. p. 8, 10, 40

Griffin J L, Williams H J, Sang E, Clarke K, Rae C, and Nicholson J K. Metabolic Profiling of Genetic Disorders: A Multitissue ^{1}H Nuclear Magnetic Resonance Spectroscopic and Pattern Recognition Study into Dystrophic Tissue. *Anal Biochem.* 2001; 293: 16-21. p. 8

Griffiths J and Stubbs M. Opportunities for Studying Cancer by Metabolomics: Preliminary Observations on Tumors Deficient in Hypoxia-Inducible Factor 1. *Advances in Enzyme Regulation.* 2003; 43: 67-76. p. 15

Guan W, Zhou M, Hampton C, Benigno B, Walker D, Gray A, McDonald J and Fernandez F. Ovarian Cancer Detection from Metabolomics Liquid Chromatography / Mass Spectroscopy Data by Support Vector Machines. *BMC Bioinformatics.* 2009; 10: article 259. p. 11

Gustafsson L., Ponten J., Zack M., and Adami H. International Incidence Rates of Invasive Cervical Cancer after Introduction of Cytological Screening. *Cancer Cause Control*. 1997; 8: 755-763. p. 7

Hagberg G. From Magnetic Resonance Spectroscopy to Classification of Tumors. A Review of Pattern Recognition Methods. *NMR in Biomed.* 1998; 11: 148-156. p. 45

Henderson's dictionary of Biology. Thirteenth edition, edited by Eleanor Lawrence Pearson Education Limited England. 2005. p. 8

Hilden J. The Area Under the ROC Curve and Its Competitors. *Med Decis making.* 1991; 11: 95-101. p. 60

Hojung N, Chung B C, Kim Y, Lee K and Lee D. Combining Tissue Transcriptomic and Urine Metabolomics for Breast Cancer Biomarker Identification. *Bioinformatics.* 2009; 25: 3151-3156. p. 58

Holmes E and Antti H. Chemometric Contributions to the Evolution of Metabonomics: Mathematical Solutions to Characterising and Interpreting Complex Biological NMR Spectra. *Analyst.* 2002; 127: 1549-1557. p. 9, 44, 50

Holmes E, Nicholls A W, Lindon J C, Connor S C, Connelly J C, Haselden J N, Damment S J P, Spraul M, Neidig P and Nicholson J K. Chemometric Models for Toxicity Classification Based on NMR Spectra of Biofluids. *Chem Res Toxicol.* 2000; 13: 471-478. p. 8, 16, 37, 67

Hoskuldsson A. PLS Regression Methods. *J Chemometrics* .1988; 2: 211-228. p. 50-54

Isbarn H, Wanner M, Salomon G, Steuber T, Schlomm T, Kollermann J, Sauter G, Haese A, Heinzer H, Huland H and Graefen M.Long-term Data on the Survival of Patients with Prostate Cancer Treated with Radical Prostatectomy in the Prostate-Specific Antigen Era. *BJUI.* 2009; 106: 37-43. p. 6

Issaq H J, Nativ O, Waybright T, Luke B, Veenstra T D, Issaq E J, Kravstov A and Mullerad M. Detection of Bladder Cancer in Human Urine by Metabolomic Profiling Using High Performance Liquid Chromatography / Mass Spectroscopy. *J Urology.* 2008; 179: 2422-2426. p. 6, 7, 10, 11, 13

Jordan K and Cheng L. NMR-Based Metabolomics Approach to Target Biomarkers for Human Prostate Cancer. *Expert Review of Proteomics.* 2007; 4 389-400. p. 11, 19-32

Jukarainen N, Korhonen S, Laakso M, Korolainen M, Niemitz M, Soininen P, Tuppurainen K, Vepsa la inen J, Pirttila T and Laatikainen R. Quantification of ^{1}H NMR Spectra of Human Cerebrospinal Fluid: A Protocol Based on Constrained Total-Line-Shape Analysis. *Metabolomics.* 2008; 4: 150-160. p. 16

Kaddurah-Daouk R, Kristal B S and Weinshilboum R M. Metabolomics: A Global Biochemical Approach to Drug Response and Disease. *Annu Rev Pharmacol Toxicol.* 2008; 48: 653-683. p. 8, 36, 39, 41, 44

Keun H C, Ebbels T M D, Antti H, Bollard M E, Beckonert O, Schlotterbeck G, Senn H, Niederhauser U, Holmes E, Lindon J C and Nicholson J K. Analytical Reproducibility in H NMR-Based Metabonomic Urinalysis. *Chem Res Toxicol.* 2002; 15: 1380-1386. p. 37-39

Keun H C, Ebbels T M d, Bollard M E, Beckonert O, Antti H, Holmes E, Lindon J C and Nicholson J K. Geometric Trajectory Analysis of Metabolic Responses to Toxicity Can Define Treatment Specific Profiles. *Chem Res Toxicol.* 2004; 17: 579-587. p. 17, 45, 56, 67, 76

Kil-Sun H, Woong J, Hyuck Jae C, Jeong Kon K, Mi-hyun K and Kyuong-Sik C. Differential Diagnostic Performance of Magnetic Resonance Imaging in the Detection of Lymph Node Metastases According to the Tumour Size in Early-Stage Cervical Cancer Patients. *Int J Gynaecol Cancer.* 2010; 20: 841-846. p. 1

Kim K, Aronov P, Zakharkin S, Anderson D, Perroud B, Thompson I and Weiss R. Urine Metabolomics Analysis for Kidney Cancer Detection and Biomarker Discovery. *Mol Cell Proteomics.* 2009; 8: 558-570. p. 1, 8, 9

Kim K B, Chung M W, Um S Y, Oh J S, Kim S H, Na M A, Oh H Y, Cho W S, Choi K H. Metabolomics and Biomarker Discovery: NMR Spectral Data of Urine and Hepatotoxicity by Carbon Tetrachloride, Acetaminophen and D-galactosamine in Rats. *Metabolomics.* 2008; 4: 377-392. p. 9, 16, 17, 36, 67, 69, 76, 85

Kim Y S and Maruvada P. Frontiers in Metabolomics for Cancer Research: Proceedings of a National Cancer Institute Workshop. *Metabolomics.* 2008; 4: 105-113. p. 8, 9

Kumar A, Wagner G, Ernst R and Wuthrich K. Buildup Rates of the Nuclear Overhauser Effect Measured by Two-Dimensional Proton Magnetic Resonance Spectroscopy: Implications for Studies of Proton Conformation. *J Am Chem Soc.* 1981; 103: 3654-3658. p. 33, 37

Larocque M, Syme A, Yahya A, Wachowicz K, Allalunis-Turner J and Fallone G.Temporal and Dose Dependence of T2 and ADC at 9.4T in a Mouse Models Following Single Fraction Radiation Therapy. *Med Phys.* 2009; 36: 2948-2954. p. 65

Lauridsen M, Hansen S H., Jaroszewski J W and Cornett C. Human Urine as Test Material in ^{1}H NMR-Based Metabonomics: Recommendations for Sample Preparation and Storage. *Anal Chem.* 2007; 79: 1181-1186. p. 16, 32, 34, 35, 66

Lenz E M, Bright J, Knight R, Wilson I D and Major H. Cyclosporin A-induced Changes in Endogenous Metabolites in Rat Urine: A Metabonomic Investigation Using High Field ^{1}H NMR Spectroscopy, HPLC-TOF MS and Chemometrics. *J Pharm Biomed Anal.* 2004; 35: 599-608. p. 17, 76

Lenz E M, Bright J, Wilson I D, Morgan S R and Nash A F P. A ^{1}H NMR-Based Metabonomic Study of Urine and Plasma Samples Obtained from Healthy Human Subjects. *J Pharm Biomed Anal.* 2003; 33: 1103 – 1115. p. 9-11

Lenz E and Wilson I. Analytical Strategies in Metabonomics. *J Proteome Res.* 2007; 6: 443-458. p. 11, 15, 16, 36, 37

Lewis I A, Schommer S C, Hodis B, Robb K A, Tonelli M, Westler W M, Sussman M R and Markley J L. Method for Determining Molar Concentrations of Metabolites in Complex Solutions from Two-Dimensional ^{1}H-^{13}C NMR Spectra. *Anal Chem.* 2007; 79: 9385-9390. p. 35, 40

Lippens G, Dhalluin C and Wieruszeski J M. Use of a Water Flip-Back Pulse in the Homonuclear NOESY Experiment. *J Biomol NMR.* 1995; 5: 327-331. p. 32, 34

Lindon J, Holmes E, Nicholson J.Pattern Recognition Methods and Application in Biomedical Magnetic Resonance. *Prog Nucl Mag Res Spect.* 2001; 39: 1-40. p. 7, 44-46, 54, 56

Lindon J, Holmes E and Nicholson J. Metabonomics Techniques and Applications to Pharmaceutical Research and Development. *Pharmaceutical Research.* 2006; 23: 1075-1088. p. 15

Lindon J C, Nicholson J K, Holmes E, Antti H, Bollard M E, Keun H, Beckonert O, Ebbels T M, Reily M D, Robertson D, Stevens G J, Luke P, Breau A P, Cantor G H, Bible R H, Niederhauser U, Senn H, Schlotterbeck G, Sidelmann U G, Laursen S M, Tymiak A, Car B D, Lehman-McKeeman L, Colet J, Loukaci A and Thomas C. Contemporary Issues in Toxicology: The Role of Metabonomics in Toxicology and its Evaluation by the COMET Project. *Toxicol and appl Pharmaco* 2003; 187: 137-146. p. 9, 16

Lofters A K, Moineddin F, Hwang S W and Glazier R H.Low Rates of Cervical Cancer Screening Among Urban Immigrants: A Population-Based Study in Ontario, Canada. *Med Care.* 2010; 48: 611-618. p. 7

Lokhov P, Dashtiev M, Moshkovskii S and Archakov A. Metabolite Profiling of Blood Plasma of Patients with Prostate Cancer. *Metabolomics.* 2010; 6: 156-163. p. 11, 14, 18, 88

Ludwig C, Ward D G, Martin A, Viant M R, Ismail T, Johnson P J, Wakelam M J O and Gunther U L. Fast Targeted Multidimensional NMR Metabolomics of Colorectal Cancer. *Mag Res Chem.*2009; 47: S68-S73. p. 37

Ma Y, Quin h, Liu W, Peng J, Juang L, Zhao X and Cheng Y. Ultra-High Performance Liquid Chromatography-Mass Spectrometry for the Metabolomic Analysis of Urine in Colorectal Cancer. *Digestive Diseases and Sciences.* 2009; 54: 2655-2662. p. 11

Macura S and Ernst R R. Elucidation of Cross Relaxation in Liquids by Tow-Dimensional NMR Spectroscopy. *Molecular Phys.* 2002; 100: 135-147. p. 34

Madsen R, Lundstedt T and Trygg J. Chemometrics in metabolomics – A Review in Human Disease Diagnosis. *Analytica Chemica Acta.* 2010; 659: 23-33. p. 39

Maher A, Zirah S, Holmes E and Nicholson J. Experimental and Analytical Variation in Human Urine in ^{1}H NMR Spectroscopy-Based Metabolic Phenotyping Studies. *Anal Chem.* 2007; 79: 5204-5211. p. 9, 16, 67, 76

Malz F and Jancke H. Validation of Quantitative NMR. *J Pharmaceut and Biomed Anal.* 2005; 38: 813-823. p. 35

McClish D K. Analyzing a Portion of the ROC Curve. *Med Decis Making.* 1989; 9: 190-195. p. 60, 62

McKay R. Recent Advances in Solvent Suppression for Solution NMR: A Practical Reference. *Annual Reports on NMR Spectroscopy.* 2009; 66: 33-76. p. 32-35, 67, 74

Metz C E. Some Practical Issues of Experimental Design and Data Analysis in Radiological ROC Studies. *Invest Radiol.* 1989; 24: 234-245. p. 57, 63, 64

Minch M. J. Orientational Dependence of Vicinal Proton-Proton NMR Coupling Constants: The Karplus Relationship. *Concepts in Magnetic Resonance.* 1994; 6: 41-56. p. 31

Miyataka H, Ozaki T and Himeno S. Effect of pH on ^{1}H-NMR Spectroscopy of Mouse Urine. *Biol Pharm Bull.* 2007; 30: 667-670. p. 16, 37, 38, 70

Mo H and Raftery D. Pre-SAT180, a Simple and Effective Method for Residual Water Suppression. *J Mag Reson.* 2008; 190: 1-6. p. 32, 33

Murdoch T B, Fu H, MacFarlane S, Sydora B C, Fedorak R N and Slupsky C M. Urinary Metabolic Profiles of Inflammatory Bowel Disease in Interleukin-10 Gene-Deficient Mice. *Anal Chem.* 2008; 80: 5524-5531. p. 16, 50, 56, 67

Nam h, Chung B C, Kim Y, Lee K and Lee D.Combining Tissue Transcriptomics and Urine Metabolomics for Breast Cancer Biomarker Identification. *Bioinformatics.* 2009; 25: 3151-3157. p. 6

Nicholson J, Holmes E and Lindon J.The Handbook of Metabonomics and Metabolomics. Ed. Lindon J C, Nicholson J K and Holmes E. Amsterdam, Oxford: Elseiver . 2007; 1-33.	p. 1, 8, 37
Nicholson J K, Lindon J C and Holmes E. `Metabonomics`: Understanding the Metabolic Response of Living Systems to Pathophysiological Stimuli Via Multivariate Statistical Analysis of Biological NMR Spectroscopic Data. *Xenobiotica.*1999; 29: 1181-1189.	p. 1, 9
Nicholson J and Wilson I.High Resolution Proton MRS of Biological Fluids. *Prog NMR Spect.* 1989; 21: 449-501.	p. 16
Nordstrom A and Lewensohn R.Metabolomics: Moving to the Clinic. *J Neuroimmune Pharmacol.* 2010; 5: 4-17.	p. 8, 9, 16, 37
Obuchowski N A. Fundamentals of Clinical Research for Radiologists: ROC Analysis. *Amer J Roent.* 2005; 184: 364-372.	p. 57, 58, 60-63
Odunsi K, Wollman R, Ambrosone C, Hutson A, McCann S, Tammela J, Geisler j, Miller G, Sellers T, Cilby W, Qian F, Keitz B, Intengan M, Lele S and Alderfer J. Detection of Epithelial Ovarian Cancer Using ^{1}H-NMR-Based Metabonomics. *Int J Cancer.* 2005; 113: 782-788.	p. 11, 12, 18, 79, 88
Park S H, Goo J M and Jo C. Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. *Korean J Radiol.* 2004; 5: 11-18.	p. 58, 60-63
Parsons H, Ludwig C, Gunther U and Viant M. Improved Classification Accuracy in 1- and 2-Dimensional NMR Metabolomics Data Using the Variance Stabilising Generalised Logarithm Transformation. *BMC Bioinformatics.* 2007; 8: 234-249.	p. 16, 76
Pavia D L, Lampman G M, Kriz G S and Vyvyan J A. *Introduction to Spectroscopy* 4th ed. Brooks / Cole, Cengage Learning, Belmont, CA, USA. 2009.	p. 22, 30
Pepe M S. An Interpretation for the ROC Curve and Inference Using GLM Procedures. *Biometrics.* 2000; 56: 352-359.	p. 60
Pochapsky T and Pochapsky S. *NMR for Physical and Biological Scientists*. Garland Science. New York, NY. 2007.	p. 20, 22, 24, 31
Pohjanen E, Thysell E, Lindberg J, Schuppe-Koistinen I, Moritz T, Jonsson P and Antti H.Statistical Multivariate Metabolite Profiling for Aiding Biomarker Pattern Detection and Mechanistic Interpretations in GC/MS Based Metabolomics. *Metabolomics.* 2006; 2: 257-268.	p. 16, 17

Potts B C M, Deese A J, Stevens G J, Reily M D, Robertson D G and Theiss J. NMR of Biofluids and Pattern Recognition: Assessing the Impact of NMR Parameters on the Principal Component Analysis of Urine from Rat and Mouse. *J Pharmaceu and Biomed Anal.* 2001; 26: 463-476. p. 32, 34, 35, 37, 38, 76

Reo N. NMR-Based Metabolomics. *Drug Chem Toxicol.* 2002; 25: 375-382. p. 15, 38, 56

Rosipal R and Kramer N. Overview and Recent Advances in Partial Least Squares. *SLSFS 2005.* 2006; 3940: 34-51. p. 50-54

Ross A, Schlotterbeck G, Dieterle F and Senn H. NMR Spectroscopy Techniques for Application to Metabonomics. *Handbook of Metabolomics and Metabonomics.* ed. Lindon J Nicholson J and Holmes E. 2007; 55-112. p. 16, 19-35, 36, 38, 66, 67, 75, 76

Rubingh C M, Bijlsma S, Derks E P P A, Bobeldijk I, Verheij E R, Kockhar S and Smilde A K. Assessing the Performance of Statistical Validation Tools for Megavariate Metabolomics Data. *Metabolomics.* 2006; 2: 53-61. p. 50, 53-55

Rule G S and Hitchens T K. *Fundamentals of Protein NMR Spectroscopy (Focus on Structural Biology)*. Vol. 5. Springer Netherlands. 2006; 135-151. p. 30, 31

Sabatine M, Liu E, Morrow D, Heller E, McCarroll R, Wiegand R, Berriz G, Roth F, Gerszten R. Metabolomic Identification of Novel Biomarkers of Myocardial Ischemia. *Circulation.* 2005; 112: 3868-3875. p. 8

Saielli G and Bagno A. Can Two Molecules Have the Same NMR Spectrum? Hexacyclinol Revisited. *Organic Letters.* 2009; 11: 1409-1412. p. 22

Saude E J, Adamko D, Rowe B H, Marrie T and Sykes B D. Variation of Metabolites in Normal Human Urine. *Metabolomics.* 2007; 3: 439-451. p. 8, 9, 11, 15

Saude E J, Slupsky C M and Sykes B D. Optimization of NMR Analysis of Biological Fluids for Quantitative Accuracy. *Metabolomics.* 2006; 2: 113-123. p. 36, 38, 39, 69, 70, 73

Saude E J and Sykes B D. Urine Stability for Metabolomic Studies: Effects of Preparation and Storage. *Metaboloimcs.* 2007; 3: 19-27. p. 16, 39, 85

Seierstad T, Roe K, Sitter B, Halgunset J, Flatmark K, Ree A, Olsen D R, Gribbestad I S and Bathen T F. Principal Component Analysis for the Comparison of Metabolic Profiles from Human Rectal Cancer Biopsies and Colorectal Xenografts Using High-Resolution Magic Angle Spinning ^{1}H Magnetic Resonance Spectroscopy. *Mol Cancer*. 2008; 7: 33. p. 40

Selek R M et al. A Metabolic Comparison of Urinary Changes in Type 2 Diabetes in Mouse, Rat, and Man. *Physiol Genomics*. 2007: 29: 99-108. p. 8

Serkova N J, Reisdorph N A and Tissot van Patot M C. Metabolic Markers of Hypoxia: Systems Biology Application in Biomedicine. *Toxicol Mech Methods.* 2008; 18: 81-95. p. 9, 10, 15, 40, 46, 56

Shlens J. A Tutorial on Principal Component Analysis. 2005; http://www.snl.salk.edu/~shlens/pca.pdf p. 45-49

Shlomi T, Cabili M N and Ruppin E. Predicting Metabolic Biomarkers of Human Inborn Errors of Metabolism. *Mol Sys Biol.*2009; 5: Article Number 263. p. 8

Simoes R, Martinez-Aranda A, Martin B, Cerdan S, Sierra A and Arus C. Preliminary Characterization of an Experimental Breast Cancer Cells Brain Metastasis Mouse Model by MRI/MRS. *Mag Res Mat Phys Biol Med.* 2008; 21: 237-249. p. 11

Sitter B, Bathen T, Singstad T, Fiosne H, Lundgren S, Halgunset J and Gribbestad I.Quantification of Metabolites in Breast Cancer Patients with Different Clinical Prognosis Using HR MAS MR Spectroscopy. *NMR Biomed.* 2010; 23: 424-431. p. 11

Slupsky C M, Rankin K N, Wagner J, Fu H, Chang D, Weljie A M, Saude E J, Lix B, Adamko D J, Shah S, Greiner R, Sykes B D and Marrie T J. Investigations of the Effects of Gender, Diurnal Variation, and Age in Human Urinary Metabolomic Profiles. *Anal Chem.* 2007; 79: 6995-7004. p. 10, 37-39, 67, 112

Smith R A, Cokkinides V and Brawley O W. Cancer Screening in the United States, 2009: A Review of Current American Cancer Society Guidelines and Issues in Cancer Screening. *Ca Cancer J Clin.* 2009; 59: 27-41. p. 4, 6, 7

Smith R A, Cokkinides V, von Eschenback A C, Levin B, Cohen C, Runowicz C D, Sener S, Saslow D and Eyre H J. American Cancer Society Guidelines for the Early Detection of Cancer. *CA Cancer J Clin.* 2002; 52: 8-22. p. 6, 7

Spahn M, Joniau S, Gontero P, Fieuws S, Marchioro G, Tombal B, Kneitz , Hsu C, Van Der Eeckt K, Bader P, Frohneberg D, Tizzani A and Van Poppel H. Outcome Predictors of Radical Prostatectomy in Patients with Prostate-Specific Antigen Greater than 20 ng/ml: A Euorpean Mulit-Institutional Study of 712 Patients. *European Urology.* 2010; 58: 1-7. p. 6, 7

Spratlin J L, Serkova N J and Eckhardt S G. Clinical Applications of Metabolomics in Oncology: A Review. *Clin Cancer Res.* 2009; 15: 431-440. p.1, 8-11, 37, 41, 83

Storey J D. A Direct Approach to False Discovery Rates. *J Roy Stat Soc B.*2002; 64: 479-498. p. 44

Stretch C, Baracos V. ^{1}H NMR-based Metabolomics of Human Urine: Variations in Metabolic Patterns are Dominated by Muscle Metabolism with Minimal Effects due to Factors such as Age, Diet or Metabolic Features. (In Press). p. 76

Sugimoto M, Wong D T, Hirayama A, Soga T and Tomita M. Capillary Electrophoresis Mass Spectrometry-Based Saliva Metabolomics Identified Oral, Breast and Pancreatic Cancer-Specific Profiles. *Metabolomics.* 2010; 6: 78-95. p. 8, 18

Szantay C. High-Field NMR Spectroscopy as an Analytical Tool for Quantitative Determinations: Pitfalls, Limitations and Possible Solutions. *Trends Anal Chem.* 1992; 11: 332-344. p. 29-32, 71

Taplin S H, Ichikawa L, Ulcickas Yood M, Manos M M, Geiger A M, Weinmann S, Gilbert J, Mouchawar J, Leyden W A, Altaras R, Beverly R K, Casso D, Oakes Westbrook E, Bischoff K, Zapka J G and Barlow W E. Reason for Late-Stage Breast Cancer: Absence of Screening or Detection, or Breakdown in Follow-up? *J Nat Cancer Inst.* 2004; 96: 1518-1527. p. 1, 4-7

Tiziani S, Lopes V and Gunther U. Early Stage Diagnosis of Oral Cancer Using ^{1}H NMR-Based Metabolomics. *Neoplasia.* 2009; 11: 269-276. p. 11, 14, 66

Troy H, Chung Y, Mayr M, Ly L, Kaye W, Stratford I, Harris A, Griffiths J and Stubbs M. Metabolic Profiling of Hypoxia-Inducible Factor-1β-Deficient and Wild Type Hepa-1 Cells: Effects of Hypoxia Measured by ^{1}H Magnetic Resonance Imaging. *Metabolomics.*2009; 1: 293-302. p. 16

Trygg J, Holmes E and Lundstedt T. Chemometrics in Metabolomics. *J Proteome Res.*2007; 6: 469-479. p. 45, 46, 49, 50, 56

Trygg J and Lundstedt T. *The Handbook of Metabonomics and Metabolomics.* Ed. Lindon J C, Nicholson J K and Holmes E. Amsterdam, Oxford: Elseiver. 2007; 171-199. p. 56

Tukiainen T, Tynkkkynen T, Makinen V P, Jylanki P, Kangas A, Hokkanen J, Vehtari A, Grohn O, Hallikainen M, Soininen H, Kivipelto M, Groop P H, Kaski K, Laatikainen R, Soininen P, Pirttila T and Ala-Korpela M. A Multi-metabolite Analysis of Serum by ^{1}H NMR Spectroscopy: Early Systemic Signs of Alzheimer's Disease. *Biochem and Biophys Res Comm.* 2008; 375: 356-361. p. 8

Tyburski J B, Patterson A D, Kraus K W, Slavik J, Fornace A J, Gonzalez F J and Idle J R.Radiation Metabolomics. 2. Dose- and Time Dependent Urinary Excretion of Deaminated Purines and Pyrimidines After Sublethal Gamma-Radiation Exposure in Mice. *Radiat Res.* 2009; 172: 42-57. p. 17, 76, 95

Um S Y, Chung M W, Kim K B, Kim S H, Oh J S, Oh H Y, Lee H J and Choi K H. Pattern Recognition Analysis for the Prediction of Adverse Effects by Nonosteroidal Anti-Inflammatory Drugs Using ^{1}H NMR-Based Metabolomics in Rats. *Anal Chem.* 2009; 81: 4734-4741. p. 16,17, 36, 56, 67, 76, 85

User Guide for SIMCA P+.Umetrics, AN MKS. Company Umetrics Inc. Kinnelon, NJ, USA. 2006. p. 46, 54

van den Biggelaar F J H M, Kessels A G H, van Engelshoven J M A and Flobbe K. Strategies for Digital Mammography Interpretation in a Clinical Patient Population. *Int J Cancer.* 2009; 125: 2923-2929. p. 6, 60, 61

Verpoorte R, Choi Y H, Mustafa N R and Kim H K. Metabolomics: Back to Basics. *Phytochem Rev.* 2008; 7: 525-537. p. 22, 36-38, 41, 42

Viant M, Ludwig C and Gunther U. 1D and 2D NMR Spectroscopy: From Metabolic Fingerprinting to Profiling Metabolomics, Metabonomics and Metabolite Profiling. Ed. Griffiths W. *J The Royal Society of Chemistry.* 2007; 44-70. p. 15, 16, 32, 35, 36, 40

Vinayavekhin N, Homan E A and Saghatelian A. Exploring Disease through Metabolomics. *ACS Chem Biol.* 2010; 5: 91-103. p. 8

Vion-Dury J, Favre R, Sciaky M, Kriat M, Confort-Gouny S, Harle J R, Grazzini N, Viout P, Grisoli F and Cozzone P J. Graphic-Aided Study of Metabolic Modifications of Plasma in Cancer Using Proton Magnetic Resonance Spectroscopy. *NMR in Biomed.* 1993; 6: 58-65. p. 32, 33, 36

Vitols C. Optimizing Spectra for Metabolomics. *Chenomx Inc.* December, 2006. p. 71

Vitols C and Fu H. Targeted Profiling of Common Metabolites in Urine. *Chenomx Inc.* April, 2006. p. 75

Vitols C and Mercier P. Correcting Lineshapes in NMR Spectra. *Chenomx Inc.* June, 2006. p. 71, 72

Vitols C and Rosewell R. Identifying Metabolites in Biofluids. *Chenomx Inc.* March, 2006. p. 73

Vizan P, Mazurek S and Cascante M. Robust Metabolic Adaptation Underlying Tumour Progression. *Metabolomics.* 2008; 4: 1-12. p. 10, 11

Wang S, Chang C, Yang S, Hsu G, Hsu H, Chung P, Guo S and Lee S. 3D ROC Analysis for Medical Imaging Diagnosis. *IEEE Engineering in Medicine and Biology. 27th Annual Conference.* 2005; 7545-7548. p. 60

Weckwerth W and Morgenthal K.Metabolomics: from Pattern Recognition to Biological Interpretation. *DDT.* 2005; 10: 1551-1557. p. 9, 42, 54

Weinstein S, Obuchowski N A and Lieber M L. Fundamentals of Clinical Research for Radiologists: Clinical Evaluation of Diagnostic Tests. *AJR.* 2005; 184: 14-19. p. 57, 58, 61, 62, 64, 79

Welch H G and Black W C. Overdiagnosis in Cancer. *J natl Cancer Inst.* 2010; 102: 605-613. p. 4, 5

Wen H, Yoo S S, Kang J, Kim H G, Park J S, Jeong S, Lee J I, Kwon H N, Kang S, Lee D H and Park S. A New NMR-Based Metabolomics Approach for the Diagnosis of Biliary Tract Cancer. *J Hepatology.* 2010; 52: 228-233. p. 7, 11

Westerhuis J A, Hoefsloot H C J, Smit S, Vis D J, Smilde A K, van Velzen E J J, van Duijnhoven J P M and van Dorsten F A. Assessment of PLSDA Cross Validation. *Metabolomics.*2008; 4: 81-89. p. 54, 79, 95, 105

Whitehead T L, Monzavi-Karbassi B and Kieber-Emmons T. ^{1}H-NMR Metabonomics Analysis of Sera Differentiates Between Mammary Tumour-Bearing Mice and Healthy Controls. *Metabolomics.* 2005; 1: 269-278. p. 11, 13, 83, 88

Wiljie A, Newton J, Mercier P, Carlson E and Slupsky C. Targeted Profiling: Quantitative Analysis of ^{1}H NMR Metabolomics Data. *Anal Chem.* 2006; 78: 4430-4442. p. 16, 22, 37, 39, 54, 70, 71

Williamson M P. *The Nuclear Overhauser Effect*. Modern Magnetic Resonance. Ed. Graham A Webb. Springer. Dordretcht, The Netherlands. 2008; 409-412. p. 33

Wilson and Nicholson. Metabonomics and Global Systems Biology. *Metabolomics, Metabonomics and Metabolite Profiling*. Ed. Griffiths W. J The Royal Society of Chemistry. 2007; 295-316. p. 10, 15

Wishart D. Quantitative Metabolomics Using NMR. *Trends Anal Chem.* 2008; 27: 228-237. p. 22, 33, 71

Wold S. Pattern Recognition by Means of Disjoing Principal Components Models. *Pattern Recognition.*1976; 8: 127-139.	p. 50
Woo H, Kim K, Choi M, Jung B, Lee J, Kong G, Nam S, Kim S, Bai S and Chung B. Mass Spectrometry Based Metabolomics Approaches in Urinary Biomarker Study of Women's Cancers. *Clinica Chimica Acta.* 2009; 400: 63-69.	p. 11
World Health Organization a: http://www.who.int/cancer/detection/en/	p. 4
World Health Organization b: http://www.who.int/cancer/detection/variouscancer/en/index.html	p. 5
Zivian M T and Gershater R. Cancer Imaging: Instrumentation and Applications. Vol. 2. Ed. Hayat M.A. Elsevier Inc. 2008; 109-118.	p. 4, 6, 7, 57, 58-62