DEVELOPMENT OF DATA PROCESSING METHODS FOR HIGH RESOLUTION MASS SPECTROMETRY-BASED METABOLOMICS WITH AN APPLICATION TO HUMAN LIVER TRANSPLANTATION

by

OLGA HRYDZIUSZKO

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Centre for Systems Biology
School of Biosciences
University of Birmingham
March 2012

Abstract

Direct Infusion (DI) Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometry (MS) is becoming a popular measurement platform in metabolomics. This thesis aims to advance the data processing and analysis pipeline of the DI FT-ICR based metabolomics, and broaden its applicability to a clinical research. To meet the first objective, the issue of missing data that occur in a final data matrix containing metabolite relative abundances measured for each sample analysed, is addressed. The nature of these data and their effect on the subsequent data analyses are investigated. Eight common and/or easily accessible missing data estimation algorithms are examined and a three stage approach is proposed to aid the identification of the optimal one. Finally, a novel survival analysis approach is introduced and assessed as an alternative way of missing data treatment prior univariate analysis. To address the second objective, DI FT-ICR MS based metabolomics is assessed in terms of its applicability to research investigating metabolomic changes occurring in liver grafts throughout the human orthotopic liver transplantation (OLT). The feasibility of this approach to a clinical setting is validated and its potential to provide a wealth of novel metabolic information associated with OLT is demonstrated.

Acknowledgments

I would like to thank my supervisor, Prof. Mark Viant for his help and support throughout my PhD and the University of Birmingham for funding this research. I extend my thanks to the members of Centre for Systems Biology, in particular to Dr Jan-Ulrich Kreft and Dr Rafik Salama, for their guidance and advice, not necessarily limited to my research and therefore much appreciated. I offer my thanks to Prof. Krystian Kubica and Prof. Małgorzata Kotulska from Wroclaw University of Technology – thank you for believing in me. I thank my family and my friends for reminding me that the world does not end with my research and for making the last four years of my life memorable. Last, thank you Birmingham for your warm welcome, lively atmosphere, amazing people and interesting places.

Table of content

List of figures

List of tables

CHAPTER 1

Introduction

1.1 Introduction to metabolomics

Metabolomics is a relatively new field with the term 'metabolome' first used in a published peer-reviewed journal in 1998 (Oliver, Winson et al. 1998). Shortly afterwards, 'metabonomics', interchangeably used with 'metabolomics', was defined as 'the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modifications'(Nicholson, K. et al. 1999). In simple words, metabolomics is a study of low-molecular-weight (typically less than 1,500 Da) compounds, intermediates and products of metabolism, such as arising from carbohydrates, lipids, nucleotides, amino acids, bile acids or other organic acids and bases etc. These metabolites, measured simultaneously offer an insight into the functioning of metabolic pathways of the whole biological system or of its selected cellular, tissue or organ levels (Fiehn 2002).

It is believed that metabolites, being further down the line from genes or proteins to functions (the end products of cellular regulatory processes), relate the closest (when compared to genes and proteins) with the activities of the biological systems at a functional level and represent an ultimate biological system's response to external and internal stimuli (Fiehn 2002, Goodacre, Vaidyanathan et al. 2004). For example, studying the genes and proteins cannot provide a complete picture on what is happening in the cell, since a vast of the its actual activity, including cell signalling, energy transfer and cell-to-cell communication is regulated at the metabolite level (Schmidt 2004). In addition, metabolites are the first line of response to the changes in the actual cellular environment, for example reflecting the modifications in nutrition or the exposure to drugs and/or pollutants. This central to metabolomics belief was captured by Prof. Bill Lasely from the University of California, when he said "Genomics and proteomics tell you what might happen, but metabolomics tell you what actually did happen".

The above makes metabolomics a useful tool in various fields, with some of its current (and potential) applications including to drug assessment, functional genomics, toxicology, clinical studies, disease diagnosis, nutrition studies, metabolic engineering or environmental studies. To name just few examples, i) analysis of urine in rats exposed to three model hepatotoxins (α-naphthyl isothiocyanate, d-(+)-galactosamine, and butylated hydroxytoluene) has uncovered novel metabolic markers of liver damage, being able to differentiate between biliary and parenchymal injury (Beckwith-Hall, Nicholson et al. 1998) ii) intracellular concentration of metabolites have revealed silent phenotypes in yeast – after deletion of genes having no overt phenotype in terms of growth rate or other fluxes when deleted from the genome (Raamsdonk 2001), iii) metabolomics study has indentified novel biomarkers, pseudouridine and 2-oxoglutarate, predictive of heart failure in patients with systolic heart failure (Dunn, Broadhurst et al. 2007) and iv) metabolic studies have been used to show that the whole-grain fed rats had higher urinary levels of Kreb's cycle intermediates, aromatic acids and hippuric acid as well as reduced levels of plasma and liver glutathione, all indicative of a better health status, i.e. shift in the basal metabolic rate and a lowered oxidative stress (Fardet, Canlet et al. 2007) - with similar methods being transferable to address challenges such as finding new ways of treating and/or preventing diseases brought on by malnutrition, over eating or an unbalanced diet (Wishart 2008).

The exact number of metabolites present in human metabolome remains a subject of debate, largely depending on the metabolome definition; however these number is estimated to be somewhere between as low as 2,000 and as high as 20,000 with metabolites occurring at a wide range of abundances (intensities) (Schmidt 2004, Kaddurah-Daouk, Kristal et al. 2008). For these reasons, it was only when novel analytical technologies appeared, being able to detect and

measure simultaneously low-molecular-weight compounds diverse in terms of their physiochemical properties, metabolomics experienced its rapid growth. Currently, the main metabolites detection platforms include mass spectrometry (MS) and nuclear magnetic resonance mass spectroscopy (NMR), with the former being often coupled with methods of gas or liquid chromatography for metabolites separation prior their quantification (Kaddurah-Daouk, Kristal et al. 2008). The collected data need to be processed and analysed in order to address a given biological aim of the study such as identifying a specific disease signature. The data processing and analysis steps have become a significant and a compound step of the metabolomics experiment drawing from various fields including, and often combining, engineering, chemometrics, computer sciences, statistics and other mathematical approaches.

Metabolomics offers a unique opportunity to investigate genotype-phenotype and genotype-envirotype relationships, and as it has been mentioned earlier it is already applicable to various fields, helping to address significant biological questions. However, it is not without challenges ahead. Being able i) to simultaneously, accurately and reproducibility detect and quantify thousands of metabolites within a biological samples, ii) to process and analyse the collected data and iii) to put the obtained results in the biochemical, and ideally in the 'omics' context to gain a full insight of processes at the cellular level is not an easy task. Therefore, there is room for improvement of every step in the metabolomics experiment (sample preparation, data acquisition or data mining) and the methods for greater metabolome coverage, better metabolite identification, enhanced reproducibility for long-term studies or feature selections methods are still currently being sought and developed (Dunn 2008).

1.2 Direct infusion Fourier transform ion cyclotron resonance mass spectrometry based metabolomics

The huge physiochemical diversity of metabolites as well as their occurrence at a vast range of abundance spanning up to nine order of magnitude combined with a short half-life (Han, Danell et al. 2008) offers a challenge for analytical chemistry, and currently none of the existing analytical methods can alone cover (detect and quantify) all the metabolites forming the human metabolome (Dunn, Bailey et al. 2005, Mayr 2008). Historically, nuclear magnetic resonance (NMR) method marked the beginning of metabolomics, being able to measure in a non-invasive and not-selective way of a wide range of metabolites, offering a high analytical reproducibility whilst maintaining the simplicity of the sample preparation step (Dunn, Bailey et al. 2005). However, only the compounds possessing a property known as nuclear spin (the spinning motion of the nucleus about its own axis) and present in sufficiently high (medium to high) abundance can be detect with this method (less than 40 in a typical experiment) (Goodacre, Vaidyanathan et al. 2004, Mayr 2008). For these reasons, mass spectrometry has become an alternative, and currently largely employed method of choice by the metabolomics investigators. Invented in 1912, it is now a well-developed analytical platform successfully used in various scientific fields, and offering metabolomics a measurement of a wide range of compounds classes at their physiological concentrations as well as the identification of these compounds through their molecular mass (indicative of the molecular formula) or via collection of fragmentation mass spectra (indicative of molecular structure) (Dunn 2008).

Mass spectrometry has been operating in an unchanged since its invention workflow which comprises of sample introduction (gas, liquid or solid form), ion source converting sample into charged ions, mass analyser separating ions based on their mass-to-charge (m/z) ratio and a

detection system; all coupled to a computer system for data acquisition and system's control (Hollywood, Brison et al. 2006, Dunn 2008). Naturally, various enhancements and adaptations of mass spectrometry has been developed and these mainly include coupling mass spectrometry with gas or liquid chromatography for an improved metabolites separation (prior entering mass spectrometer), developments of a range of ionization methods (e.g. atmospheric pressure chemical ionization, chemical, electron impact or electrospray ionization) or advances in mass analyser techniques (e.g. Fourier transform ion cyclotron resonance, linear quadrupole, Orbitrap, Time of flight or Triple quadrupole) (Dunn 2008, Mayr 2008). Out of these, direct infusion (DI) Fourier transform ion cyclotron resonance (FT ICR) mass spectrometry is noticeable for its ultrahigh resolution, sensitivity and mass accuracy (Ohta, Shibata et al. 2007). FT MS analyser has the highest resolving power (resolution) of all mass analysers (>100,000) which combined with the sub-ppm (<1 parts per million error) mass accuracy allows alone, without the utilisation of time-consuming chromatographic separation steps, the measurement of complex mixtures and the identification of metabolites based on the assignments of molecular formulas to each single mass-to-charge ratio (Brown, Kruppa et al. 2005).

A workflow of a typical fingerprinting metabolomic experiment, where a rapid snapshot of the global biochemical state is measured (as oppose to metabolomics profiling where only a number of pre-defined metabolites is identified and quantified) and involving the use of the DI FT-ICR MS technology (as well as routinely conducted at the University of Birmingham) is shown on Figure 1.1. Following the experimental design, the extracted from tissue, urine or serum metabolites (samples) are detected and quantified in the FT-ICR MS analyser. The collected series of mass-to-charge spectra are then processed (or pre-processed; e.g. checked for quality, subjected to the non-biological noise removal methods or mathematically transformed) to

improve further data analysis and yield a desirable data format such as a two dimensional matrix containing metabolite relative abundances (concentrations) organised with each variable (ions indicative of metabolites names) measured in unique column and each sample analysed in a unique row. Final step includes data analysis that employs various machine learning approaches to measure the deviation from 'normality', e.g. the differences of drug-treated samples as oppose to control samples for sample classification and hypothesis generation or to develop statistical models enabling a distinction of newly obtained sample distinction, useful in diagnostic (Ellis, Dunn et al. 2007). The fingerprinting approach offers a mechanistic insight into biochemical pathways altered under given perturbation (caused by disease, influenced by the environmental or therapeutic factors) and is known as a hypothesis-generating strategy (Goodacre, Vaidyanathan et al. 2004). Further data analysis include the integration of the experimental results with other metabolomics, 'omics' or other relevant data to gain a full understanding of the processes at the molecular level.

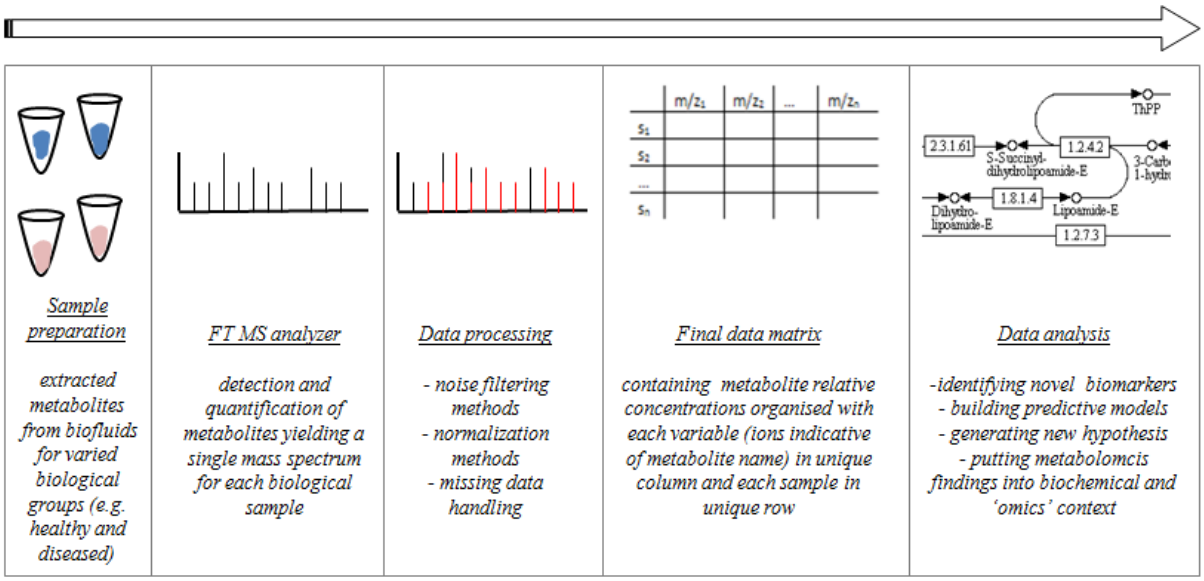

Figure 1.1 A typical workflow of a metabolomics experiment employing DI FT-ICR mass spectrometry

1.2.1 Principle of DI nESI FT-ICR mass spectrometry

DI nESI FT-ICR mass spectrometry is based on detection of coherently excited ions cyclotron motion (Figure 1.2). First, nanoelectrospray ionization (nESI) is used to convert the sample of interest into a fine aerosol containing ions. Briefly, to achieve this homogenized, in a liquid form sample mixed with solvent is spread at the low flow are (ca. <300nL/min) through an emitter (electrospray needle) across a high (typically several kV) potential difference (Gibson, Mugo et al. 2009). Ions already present in the solution and the ones formed due to electrochemical reactions with the solvent are released at the end of the needle as charged droplets. These get smaller as the solvent evaporates and reach the Rayleigh limit, i.e. the point where the surface tension can no longer sustain the excess of ions (positive or negative depending on the ionisation mode) and where they undergo "Coulomb explosions" ripping them apart to form smaller droplets. This process reiterates until the droplets are small enough to produce gas-phase ions (Wilm and Mann 1996, Gibson, Mugo et al. 2009) .

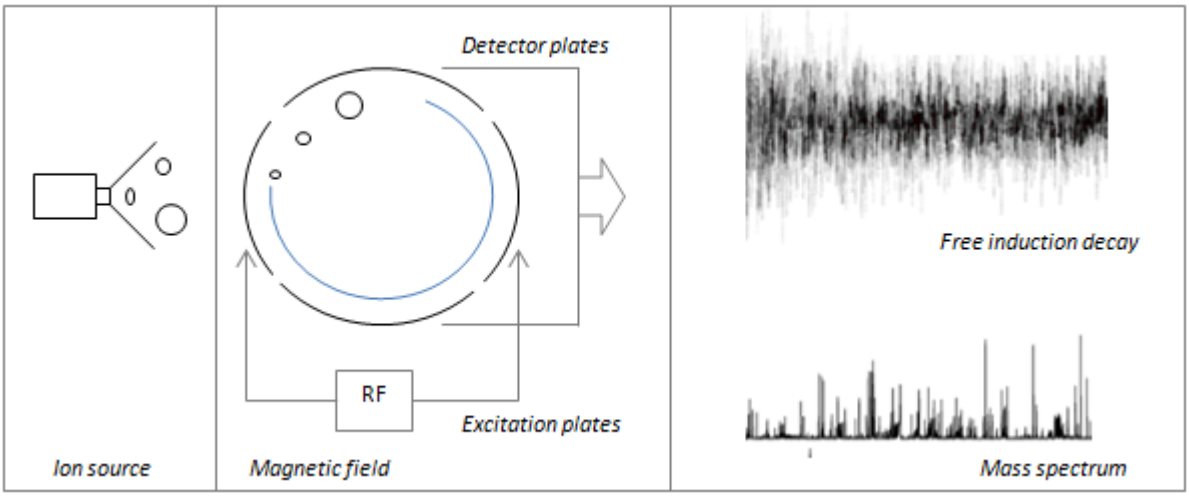

Figure 1.2 Brief principle of ultrahigh resolution mass spectrometry using FT-ICR/MS

Generated during the nanoelectrospray ions pass through a series of pumping stages at increasingly high vacuum to eventually enter the ion trap (with two trapping plates at both ends preventing ions to escape and with pressures ranging from 10^{-10} to 10^{-11} mBar and temperature close to absolute zero) located inside a spatially uniform magnetic field (typically of 4.7 to 13 T). Each of these ions moving in a presence of such field is subject to a Lorentz force given by Eq. 1.1 in which *m, z* and *v* denote mass, charge and velocity (respectively) of the moving ion and ***B*** is the strength of the magnetic field which bend the ion path into a circle of radius *r* (magnetic component of the Lorentz field is perpendicular to the plane determined by vectors ***v*** and ***B***, where $\boldsymbol{B} = -B_0 * k$) (Marshall, Hendrickson et al. 1998, Barrow, Burkitt et al. 2005).

$$F = m\frac{d\boldsymbol{v}}{dt} = z\boldsymbol{v} \times \boldsymbol{B} \qquad \text{Eq. 1.1}$$

For the ion velocity in the *xy* plane (plane perpendicular to ***B***) it becomes Eq. 1.2 showing that the frequency of the ions rotation is dependent on their mass to charge ratio (*m/z),* where ω_c denotes the induced cyclotron frequency.

$$\omega_c = \frac{zB_0}{m} \qquad \text{Eq. 1.2}$$

At this stage, however, the radius *r* of the motion is not big enough for the detector plates to receive the signal and therefore the excitation plates are used to accelerate ions to a larger detectable orbital radius by applying a spatially uniform electric field oscillating near the cyclotron frequency of ions – a frequency sweep pulse is used to excite all the ions. This results in the simultaneous measurement of all the ions and produces a signal of all the ions frequencies in time domain and a resulting signal called free induction decay (FID) composed of interferogram of superimposed sine waves (due to the ions moving in the circle). Here, Fourier

transform (mathematical transformation) is used to decompose the FID signal into its constituent frequencies and convert them into the mass spectrum (Marshall, Hendrickson et al. 1998, Marshall and Hendrickson 2002).

1.2.2 Sample preparation and data acquisition

Metabolites in the samples that are about to be measured via DI nESI FT-ICR mass spectrometry have to first be extracted from the biological organisms. Typically, extracellular metabolites present in humans are obtained non-invasively from urine or invasively in the form of serum, plasma or cerebrospinal fluid (Dunn, Bailey et al. 2005) and the intracellular metabolites can be extracted via a tissue biopsy (Wu, Southam et al. 2008, Hrydziuszko, Silva et al. 2010). The extraction protocol can affect the metabolome's composition and in turn the subsequent data interpretation. Wu et al. have shown that a two-step method with 10 min partitioning provides the most accurate (when compared to stepwise and all-in-one methods) snapshot of metabolome for both NMR and MS metabolomics (Wu, Southam et al. 2008). Briefly, in the first step of this extraction method, sample is being homogenised in the presence of methanol and water, and in the second step the chloroform and water are added, maintaining the methanol/chloroform/water ratio of 2:2:1.8 respectively throughout the whole extraction process. Samples are then centrifuged and polar fractions can be removed and stored (at -80°C) until the FT-ICR mass spectrometry analysis (Wu, Southam et al. 2008, Taylor, Weber et al. 2009).

At the other end of the metabolites quantification process one wishes to detect physico-chemically varied metabolites, occurring both as the low and high abundance ions maintaining a high mass accuracy to enable consequent metabolites identification: unique empirical formulas

are derived from the mass-to-charge values. This means that in a conventional data acquisition mode increased number of ions has to enter the FT-ICR MS analyser with a potential upside of their measured mass accuracy being impaired due to undesirable ions interactions resulting in space-charge effects. The spectral stitching method has been proposed in 2007 to address this limitation (Southam, Payne et al. 2007). Instead of using a standard wide scan range mode, ions in the limited mass-to-charge ratio range (e.g. 70 to 90 *m/z* ratio only) are transmitted and detected by the FT-ICR MS analyser, resulting in the increased sensitivity and overcoming the problem of undesirable ions interactions. To obtain a wide range spectrum (for all the ions present), the data acquisition is repeated for several overlapping *m/z* ranges and the information from the overlapping regions is used to 'stitch' the spectra together. The application of this method has resulted in a reduced mass accuracy error and improved dynamic range (the ratio of the highest to the lowest concentration metabolite detected) when compared to a typical wide scan method, with a final absolute mass error of 0.48 ppm and over 3000 ions detected in flatfish liver extract (4.3-fold increase) (a hybrid 7-T FT-ICR mass spectrometer from Thermo Fisher Scientific and 200nL/min flow rate and 1.65kV nanoelectrospray from NanoMate, Advion Biosciences)(Southam, Payne et al. 2007).

1.2.3 Data processing: measuring the metabolome

Following data acquisition, the obtained mass spectra are subjected to various data processing (or pre-processing) methods, i.e. mathematical operations, in order to facilitate and enhance the subsequent data analysis. These methods typically include, but are not limited to, filtering methods to remove the measurement noise (not representing biological information), peak detection methods to detect their exact position and quantify them, normalization approaches to

reduce the systematic variation both between the samples and across the peaks, transformations to change the scale of the data, method of handling outliers and missing data (Goodacre, Broadhurst et al. 2007, Katajamaa and Orešič 2007). Depending on the experimental design, specific features of the analytical platform and the acquired data and as well as the subsequent data analysis, various data pre-treatment methods may be applicable.

A three-stage filter is an example of the possible measurement noise filtering method that is applied routinely at the metabolomics research group at the University of Birmingham (Payne, Southam et al. 2009). In this approach, each biological sample and the polar fraction of the "extract blank" (prepared in the same way as biological sample but with no biological material added) is measured in triplicate. The frequency centre of each peak is determined by applying KC*e* (*e* = 5.5) interpolation to local maxima (Keefe and Comisarow 1990). The measurement noise is distinguished from real (arising from biological samples not from technical artefacts) signal and removed in three steps in which: i) the all the peaks below (low intensity) a pre-defined (typically 3.5) hard signal-to-noise ratio (SNR) are removed, ii) the three technical replicates are merged together to yield a single mass spectrum that represent each sample that includes peaks present in at least two out of three technical replicates (peaks are considered to arise from the same metabolite when appearing with 1.5 ppm spread) and iii) biological spectra are formed into a rectangular dataset with intensities values for each *m/z* (columns) and for each biological sample (rows) that includes only those peaks that are present across a pre-defined percentage (typically 50%) across all of the biological samples. In addition, following the 'replicate' filter (second step) peaks appearing both in biological sample and the "extract blank" are being removed if they are of higher intensity in the "extract blank" (Payne, Southam et al. 2009, Taylor, Weber et al. 2009). The spectral noise appeared to be white and to determine the

noise threshold, the noise in the spectrum was approximated with Rayleigh distribution model (verified by experimentation) with a single parameter σ. The threshold was then determined using a probabilistic model described by multiple steps including selection of the initial sub-range of the spectrum, fitting the sub-range data to the Rayleigh model using maximum likelihood algorithm yielding a value of σ, solving cumulative distribution function of the Rayleigh distribution model as well as further steps to e.g. ensure that the optimal sub-range of the spectrum is used for the threshold estimation (described in details in (Southam, Payne et al. 2007). This approach enables to eliminate the measurement noise that may arise from technical imperfections of the MS analysed e.g. arising from undesirable ion-ion interactions inside the detector cell, ion suppression or limited electrospray ionization process as well as possible contaminations of the biological samples during their preparation stage ("extract blank").

Normalization and transformation methods may include a vast range of approaches. Some of most common ones include reducing the variance between samples by normalizing to a constant sum, to a constant feature (e.g. to an internal standard) or to a reference sample as in probabilistic quotient normalization where the reference spectrum is derived by calculating the median of the control samples and all the peaks are divided by the median of their quotients obtained when compared to the reference spectrum (Dieterle, Ross et al. 2006). Logarithm transformation is used to reduce the variance between the variables (peaks), especially important prior of the multivariate analysis to reduce the chances of the most intense peaks being the most dominating ones. Generalised logarithm transformation in which data is modified depending both on the original intensity and the transformation parameter, has been proposed as a further improvement to the logarithm transformation and has shown to outperform the autoscaling (within column, subtracting from each value the mean and diving by the standard deviation) and Pareto scaling

(same as autoscaling but diving by the square root of the standard deviation) for the discriminating between classes of the NMR metabolomics datasets (van den Berg, Hoefsloot et al. 2006, Goodacre, Broadhurst et al. 2007, Parsons, Ludwig et al. 2007).

Another key issue is deciding on treatment the missing data in the final dataset. These missing data can arise both from the technical and biological reasons, i.e. metabolites not detected from some of the samples due to technical limitations of the MS analyser (e.g. intensity values below the limit of detection) or metabolites intensity values truly equal to zero due to a genuine biological heterogeneity between the analysed samples (Hrydziuszko and Viant). A common approach for metabolomics study is to impute the missing data with a plausible value to yield a complete datasets for subsequent data analysis. Some of the imputation methods include substitution missing entries with a small arbitrary chosen value (0.001 or half of the minimum value found in the dataset), substitution with mean or median value of the non-missing data across a given peak or more advanced imputation methods such as *k*-nearest imputation, multiple imputation or Bayesian Principal Component Analysis estimation. The imputation of missing values in the metabolomic data processing pipeline is one of the main objectives of this thesis and therefore discussed in considerably greater detail in Chapter 2.

1.2.4 Data analysis: univariate and multivariate statistical methods for high-dimensional metabolome data

The applied data-processing methods depend upon the analysed datasets and may vary between the research groups, different experimental designs and the questions to be addressed. Even more is true of the methods used in the subsequent data analysis, as these may draw upon statistics, data mining, machine learning, artificial intelligence, probability theory etc. and with the existing methods being constantly redefined and with the new methods being developed

(Goodacre, Vaidyanathan et al. 2004, Goodacre, Broadhurst et al. 2007). It is not possible to discuss each of the data analysis used, however some of the most common techniques are outlined below. Considering a fingerprinting metabolomics experiment, in which a snapshot of all metabolites is measured to generate new hypotheses, to identify novel biomarkers or to provide a sample classification tool based on their biological properties (e.g. healthy and diseased), data analysis methods may include univariate and multivariate statistical methods, unsupervised and supervised learning methods (Goodacre, Vaidyanathan et al. 2004).

Typically, statistical hypothesis tests such as t-test, z-test or ANOVA are the univariate methods (taking account one variable i.e. here applied to each peak individually) used to test whether the means of the groups (e.g. healthy and diseased) are equal and thus to identify the potential biomarkers (Kirkwood and Sterne 2003). When the data does not meet the criterion of the normality assumption, the non-parametric univariate methods, such as Mann-Whitney U or Kruskal-Wallis test are used. Along the statistical hypothesis test, the fold change is being calculated to provide the information on the direction of the concentration change (e.g. a given peak concentration increasing or decreasing following an exposure to the tested drug or the treatment). Due to several thousands of peaks being subjected to a statistical hypothesis testing, an important issue of the increased chances of committing a type I error (incorrectly rejecting a null hypothesis) arises, thus leading to wrongly identifying a set of peaks as significantly different between the groups. This is addressed via applying various correction methods, including a simple Bonferroni or a more advanced Benjamini and Hochberg methods (Benjamini and Hochberg 1995).

One of the most commonly used in metabolomics multivariate data analysis method is Principal Component Analysis (PCA) - an unsupervised (using unlabeled data and no prior

information) method applied to gain an overview of the large datasets, identify outliers and trends in the data. PCA is a mathematical transformation that maps high-dimensional data onto a lower-dimensional space the captures most of the information from the original data; or in different terms it converts a large number of intercorrelated variables to a smaller set of uncorrelated variables (principal components) while maintaining the variation of the dataset. Therefore, the

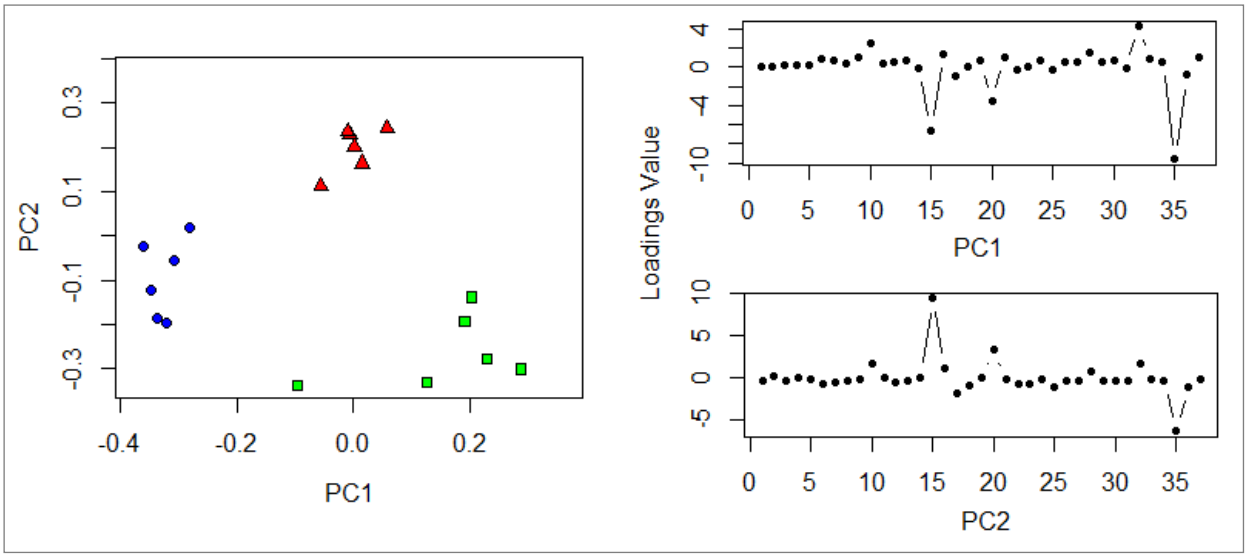

Figure 1.3 Example of a principal component scores plot for three biologically different samples groups (on the left) and corresponding loadings values for the PC1 and PC2 (on the right); Blue circles: cancer leukemia cell line, control group; red triangles: exposed to medroxyprogestorne; green squares: exposed to indomethacin

original samples are represented by coordinates *scores* (***T***) in the new space which dimensions are linear combinations of the original variables, called *loadings* (***P***) Figure 1.3 (Abdi and Williams 2010). The original data ***X*** can be then approximated via

$$\hat{X} = T_a P_a^T \qquad \text{Eq. 1.3}$$

where ^{T} denotes the transpose of the matrix, and the subscript a denotes number of components taken into account, with the

$$a_{max} = \min(n, p) \qquad \text{Eq. 1.4}$$

where *n* and *p* indicating the number of rows and columns in the matrix. For the a_{max}, the approximation of the original data is perfect with

$$\hat{X} = X \qquad \text{Eq. 1.5}$$

The principal components are found in such a way that the first one captures the highest variance (accounts for as much of the variability in the original datasets as possible) and the subsequent principal components capture the remaining variance in the decreasing order. They are orthogonal combinations of variables defined in such a way that the variances of the scores are maximal, the sum of Euclidean distances between the scores is maximal and the reconstruction of *X* is as close as possible to the original.

Principal component analysis is not a classification method, yet is it often a first choice method in a metabolomics experiment to give a quick overview of the data, possible groupings and outliers. Other unsupervised methods include hierarchical clustering methods or Kohonen neural methods. When the labels are available and the sample size of the study justifies the use of the supervised methods (sample size not leading towards obvious over-fitting) methods such as discriminate analysis, partial least squares regression or neural networks may used or the evolutionary-based algorithms such a genetic algorithms to find a 'model' that will correctly link the inputs with the targets (Goodacre, Vaidyanathan et al. 2004).

Finally, placing the statistically significant finding from the data analysis into a biological context requires assigning peaks to their metabolites identities. The ultra-high mass accuracy of the DI FT-ICR mass spectrometer enables the assignment of the empirical formulae to the peaks using an elemental composition calculator and to match the observed *m/z* values of the peak to the metabolite identifies found in the online databases such as Kyoto Encyclopaedia of Genes and

Genomes (KEGG) (Kanehisa, Goto et al. 2010) or Human Metabolome Database (HMDB) (Wishart, Knox et al. 2009, Kanehisa, Goto et al. 2010).

1.2.5 Current limitations

DI FT-ICR mass spectrometer may allow achieving the highest mass resolution and mass accuracy, and therefore enables the analysis of the complex metabolic extracts (Brown, Kruppa et al. 2005, Dunn 2008). The accurate mass measurements provide a rapid way to identify the majority of the metabolites without a need to resort to chromatography methods; making DI FT-ICR based metabolomics an ideal tool to simultaneously acquire all the available information (Brown, Kruppa et al. 2005). However, there are still challenges ahead.

The future progress in various aspects of DI FT-ICR MS based metabolomics will benefit the applicability and the effectiveness of this approach. The study designs and protocols need to be optimized, especially if DI FT-ICR MS based metabolomics is introduced to a new area such as nutrition or oncology research (Scalbert, Brennan et al. 2009). The metabolites detection and quantification should account for any unwanted sources of variations and yet yield a replicable, comparable between laboratories measurements. Finally, to turn the collected raw metabolite data into a biological knowledge, an assortment of data processing, statistical analysis and data storage formats is needed (Fiehn, Robertson et al. 2007, Allwood, Ellis et al. 2008). The limitations central to this thesis include the issue of missing data (data processing) and the metabolites identification (data analysis).

Within data processing stage there is a common problem of missing data occurrence, which can arise from both technical and biological reasons impeding the usage of the majority of the statistical analysis methods (Steuer, Morgenthal et al. 2007, Hrydziuszko and Viant 2011). A

further knowledge of the exact nature of these missing data and the extent of their effect of the data analysis is of importance for a better linkage between raw data and the biological knowledge. Currently, there is limited research on missing data such as their optimal ways of handling prior data analysis, e.g. by applying missing data estimation algorithms. As stated in section *1.2.3 Data processing: measuring metabolome,* addressing the problem of missing data is one of the main objectives of this thesis, therefore this limitation and its implications are discussed at length in Chapters 2 and 3.

Drawing biological conclusions and/or obtaining a novel knowledge remain the hardest parts of the DI FT-ICR MS based metabolomics experiment. The high mass accuracy and resolution may allow resolving all metabolite peaks that represent different metabolites masses as well obtain their elemental compositions. Further heuristic rules are applied to narrow the possible molecular formulae by filtering out the unfeasible ones (e.g. containing unlikely high number of elements). These in combination with using the metabolite databases enable assigning a metabolite name. However, despite these advances, elucidation of the correct compound structure remains a challenge (Kind and Fiehn 2007). The number of potential empirical formulae calculated for a given peak increases with the *m/z* values. Mass steroisomers and isometric molecules cannot be distinguished, e.g. glucose and galactose sharing the same empirical formula of $C_6H_{12}O_6$ (Weber, Southam et al. 2011). And the databases do not yet contain information on all the vast numbers of metabolites (Scalbert, Brennan et al. 2009).

1.3 Metabolomics in surgery

Being the building blocks and substrates of all cellular processes, metabolites reflect the dynamic processes underlying cellular homeostasis. Therefore, metabolomics studies of human

diseases can contribute to the elucidation of their pathophysiological mechanisms, leading to an improved diagnostics, treatment, patients' monitoring and finally, where applicable, surgical interventions. Recent years have brought an initial research into the application of metabolomics mainly, but not limited to organ (e.g. kidney, liver and heart) transplantation.

Adequate monitoring of organ physiology during and following transplantation to detect organ reperfusion injury (tissue damage caused by the returning blood supply following the period of ischemia, lack of oxygen) and to assess organ function or dysfunction can alone decide upon the success of the whole procedure and a patient's full recovery. The metabolomics studies in this area have been predominately focused on kidney transplantation, the most frequent (excluding cornea transplantation and exceeding 2500 procedures a year in UK only) organ transplantation procedure for the patients with end-stage renal diseases (NHS Blood and Transplant). With over 30 metabolomics studies published, the main trends seem to be utilizing metabolomics in assessing ischemia-reperfusion injury, characterizing immunosuppressive drug toxicity and organ (dys)function (Wishart 2006). For example, Wang et al. have used the matrix-assisted laser desorption/ionization Fourier transform mass spectrometry to analyse urinary samples from transplant patients to address the questions whether urinary metabolites can distinguish and predict the patients developing acute clinical rejection following the transplantation (Wang, Zhou et al. 2008). Examination of the metabolic mass spectra from patients with and without the evidence of rejection pinpointed a set of eight small molecules (yet to be identified) that enable patients' distinction. Mao et al. also addressed the same question, however via studying serum metabolites and using gas-chromatograph mass spectrometry (Mao, Bai et al. 2008). They have identified 17 metabolites that were significantly higher in the group of patients developing acute rejection and resulted in 77.3% prediction accuracy; these included, among others, amino acid

(phenylalanine, serine, glycine, threonine, valine), carbohydrates (galactose oxime, glycose, fructose), caroboxylic acid, lactate, urea and myo-inositol. These and similar studies are of great importance, since there is still a need to find an ideal biomarker (ideal set of biomarkers) for kidney function following transplantation; currently used serum creatinine lacks high sensitivity and specificity as it is dependent on muscle mass, hydration status and can be elevated due to multiple cases of graft injury (Sarwal 2009). Other metabolomics studies have proven to be of advantage when applied to assessing the effects of the immunosuppressive therapy: high pressure liquid chromatography mass spectrometry was used to track serum concentrations of cyclosporine (CsA, immunosuppressive drug) and its metabolites (Vollenbroeker, Koch et al. 2005). It has been previously shown that the long-term use of CsA is limited by nephrotoxicity and elevated risks of cardiovascular diseases. Vollenbroeker at al. have shown that some of the CsA metabolites correlate with several inflammatory and artherosclerotic markers, and therefore the use of CsA may be improved (e.g. by choosing the optimal, shorter, postoperative treatment time). Metabolomics applied to heart transplantation studies aims at addressing questions similar to those of kidney transplantation, mainly there has been a considerable interest in developing fast and non-invasive methods to measure cardiac function and/or to detect cardiac rejection following the transplantation. For example, Eugene et al. have used proton magnetic resonance spectroscopy method of plasma samples to try to detect the acute cardiac rejection in patients undergoing heart transplantation, achieving high sensitivity and specificity (>90%) based on the selected methyl and methylene peaks mainly arising from lipoproteins (Eugene, Le Moyec et al. 1991). It is beyond the scope of this section to discuss all the metabolomics studies in organ transplantation, however a solid review was done and presented by Wishart (Wishart 2005).

Although the majority of 'surgical' metabolomics studies focusing on organ transplantation, there are few others showing that metabolomics is applicable to a broader range of clinical studies. Mutch et al. used successfully both gas and liquid chromatography-coupled mass spectrometry to analyse serum metabolites from patients before and following roux-en-Y gastric bypass surgery to understand the numerous metabolic adaptations associated with the procedure such as weight loss, increased insulin sensitivity and glucose homeostasis (Mutch, Fuhrmann et al. 2009). Finally, metabolomics has become of interest to surgical oncology, as a potential new tool improving cancer detection and treatment. There has been much metabolomics research aimed at identifying biomarkers capable of early diagnosis of breast, ovarian, colon and prostate cancer showing that metabolomics holds promise as an non-invasive mean of detecting early-stage malignancy and also monitoring treatment efficacy (Davis, Bathe et al. 2011). Metabolomics tools could especially play an important role when the diagnosis of cancer type with the currently available methods remains a challenge – for number of tumours discriminating between benign and malignant disease is still not straightforward with the potential misdiagnosis leading to life-threatening operations being performed for benign and inadequate treatment being undertaken for the malignant disease.

1.3.1 Metabolomics in orthotopic liver transplantation

Liver transplantation is the currently only known treatment for the end-stage liver diseases, with over 600 liver transplants taking place in United Kingdom each year. In orthotopic liver transplantation (OLT), most commonly used technique, the native liver is removed and replaced by the donor organ (as oppose to split donation). It is estimated that due to the shortage of available organs, approximately 10% of potential recipients die while on the waiting list (2007-

2008 Transplant Activity in UK, National Health Service annual report). Furthermore, the molecular mechanisms undergoing in the graft during the transplantation process (removal from the donor, ice storage and reimplantation) are not fully understood. Therefore OLT could benefit from the metabolomics studies, which providing novel molecular insight into the biochemical pathways altered during OLT could lead to a) increasing donor pools by including marginal donors or those obtained by donation after cardiac death, b) improving the therapeutic interventions to minimize tissue damage and maximize the likelihood of grafts success, c) advancing post-OLT patients' monitoring and identification of graft dysfunction or poor function or finally d) providing further information for a more advanced system of donor-recipient matching. As in the kidney transplantation, several metabolomics studies have been already carried out to assess whether novel metabolomics biomarkers can be found to inform upon liver function following OLT. Majority of these studies employed NMR-based metabolomics methods and have resulted in the promising findings (Wishart 2005)(see Chapter 5 for further details). DI FT-ICR mass spectrometry based metabolomics, with its ability to detect thousands of peaks in the biological sample has a considerable potential for investigating liver metabolism during and following OLT. Studies to verify the applicability of this metabolomics approach to this extremely varied clinical settings (large human-to-human metabolic variations; differences in the OLT procedure e.g. varying graft ice storage times; variations between donor and recipient metabolism; different indications for performing OLT; multiple prior and post OLT treatments) need still to be undertaken.

1.4. Research aims

The overall aim of this work is to advance and to develop the data processing pipeline for the DI FT-ICR mass spectrometry based metabolomics to improve the human liver transplantation surgery. In particular, the below main questions are asked:

i) does the treatment of missing data during the data processing stage influence the subsequent data analysis? If so, what are consequences of choosing different approaches? What are the nature and the reasons for the occurrence of missing data? Can this information be used to identify the optimal missing data estimation algorithm that applied once during data processing stage will yield the correct results of the subsequent univariate and multivariate data analyses (Chapter 2)

ii) can the survival analysis methods for the right censored data be used as an alternative way to handle missing values? Can DI FT-ICR mass spectrometry metabolomic data be represented as right censored data? If so, what are the advantages and limitations of this novel approach? (Chapter 3)

iii) can DI FT-ICR mass spectrometry based metabolomics be applicable to a highly variable clinical study to metabolic processes of liver grafts during human orthotopic liver transplantation (OLT)? Can it inform the OLT by characterizing multiple metabolic changes occurring upon OLT, providing novel insight into the biochemical pathways and/or suggesting novel therapeutic interventions for the improved outcomes? (Chapter 5)

1.5 Thesis structure

Chapter 1 contains a general introduction to the thesis. It briefly outlines, quite extensive now, field of metabolomics, focusing on the human metabolomics, i.e. studies on the molecules found in humans, as oppose to plant or environmental metabolomics. It introduces metabolomics research based on the analytical platform of direct infusion Fourier transform ion cyclotron resonance mass spectrometry, discussing its theoretical background and a typical experimental pipe-line and current limitations. It finishes with a section on applicability of metabolomics in surgery.

Chapter 2 is the first of the research chapters containing published paper on missing data in DI FT-ICR mass spectrometry based metabolomics (Hrydziuszko and Viant). It addresses the question on the nature of missing data and how to choose the optimal imputation method, with the proposed algorithms and findings transferable to metabolomics studies employing other mass spectrometry platforms.

Chapter 3 introduces a branch of statistics called survival analysis, especially the definitions of censored data and their methods of analyses. Further, it introduces the mathematical lemma that the above methods can be used to analyse the left censored data. Further it discusses the applicability to the above survival analysis approach to handle missing data in DI FT-ICR mass spectrometry based metabolomics (novel approach) and presents the results on testing this approach.

Chapter 4 includes the supplementary work on advancing data processing and data analysis methods for the DI FT-ICR mass spectrometry based metabolomics that was carried out as a result of additional questions that arose while addressing the main research aims. Mainly it presents a software developed for the “extract blank” analysis (applied in a published study;

(Taylor, Weber et al. 2009) and the algorithm for comparing the ordered lists, both of which can be applicable to and of potential use during the data processing and analysis stages (Hrydziuszko and Viant 2012).

Chapter 5 discusses the applied part of the research, i.e. to the clinical study of liver transplantation. In particular it presents a pilot study used to verify the applicability of the DI FT-ICR MS to the study of human liver transplantation and it is an edited version of the published article (Hrydziuszko, Silva et al. 2010).

Chapter 6 contains the summary and conclusions of this thesis. It also highlights and discusses the need for further work that has emerged during the presented in this thesis research.

CHAPTER 2

Missing Data – Towards Optimal Imputation Method

2.1 Introduction

2.1.1 Missing data in statistical analysis

Missing data are a common complication of any real-world study, mostly occurring due to study design, due to chance or due to technical reasons (Nicholas and Ken 2008). Where missing data cannot be prevented by design the need for the statistical analysis with missing data arises. A vast range of missing data handling approaches has become available in recent years. These can be grouped into four major categories based on the generic approaches of addressing the problem of missing data occurrence as summarized in Figure 2.1 (García-Laencina, Sancho-Gómez et al. 2010).

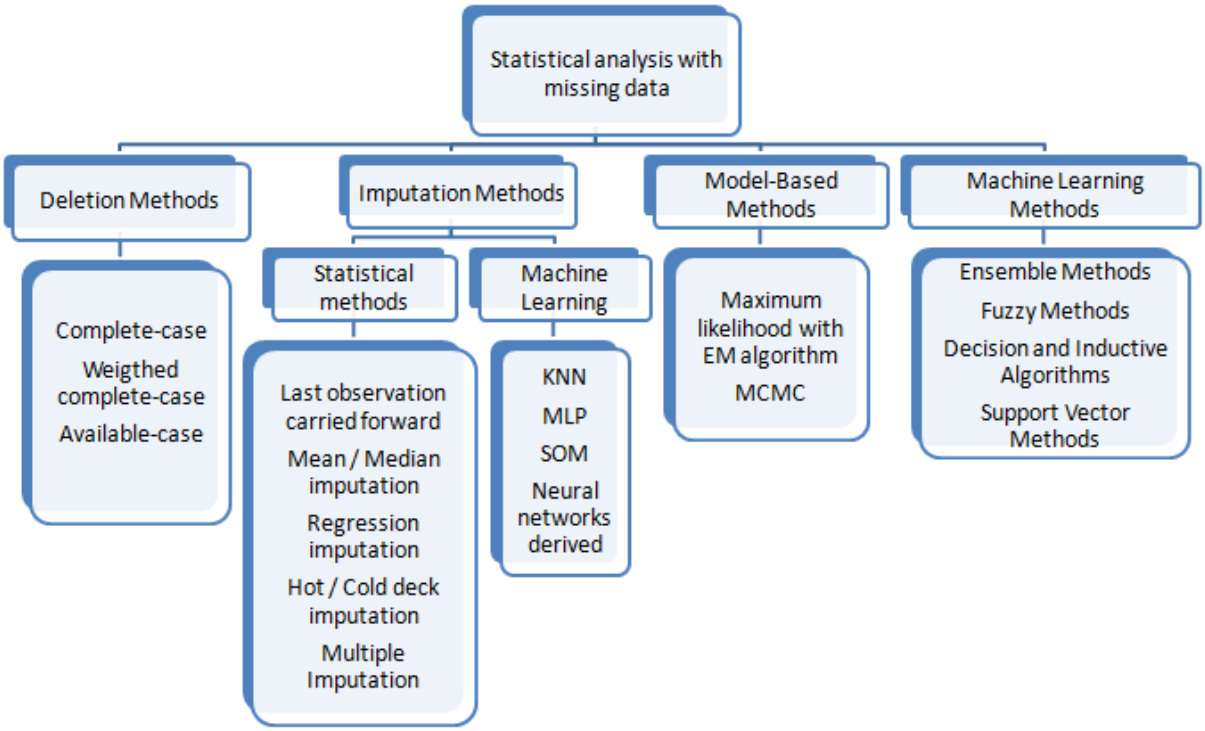

Figure 2.1 Classification of missing data handling methods in statistical analysis with an application to pattern classification.

Historically the oldest and the simplest category comprises deletion methods which focus on the analysis of the observed values, such as complete-case method, weighted complete-case analysis or

available-case analysis (Little and Rubin 2002). In complete-case method only cases (samples) where all the variables are present are considered (list-wise deletion), resulting in a simplistic approach with potential disadvantage of a major information loss and analysis bias when the missing data are not present at random (MCAR, defined below) (Schafer and Graham 2002). To adjust for this bias, Rubin introduced the idea of constructing weights for complete cases, a concept related to weighting in randomization inference for finite population surveys (Little and Rubin 2002). Available-case method discards data at the level of variables (pair-wise deletion), making a better use of data than complete-case analysis (all available values are considered). This in turn leads to a different number of variables for each case and thus making a subsequent data analysis (e.g. classification methods) challenging (García-Laencina, Sancho-Gómez et al. 2010).

The objective of the missing data imputation method is to produce a complete datasets by replacing missing values with plausible estimates to enable statistical analysis using standard data analysis approaches (Myers 2000). The earliest used methods included simple approaches where only one plausible value was assigned to each missing datum. The last observation carried forward was commonly implemented in longitudinal measurements observed for the patients, where the last value was used to replace the subsequent missing data (Myers 2000). Another undemanding method is mean or median imputation, in which missing values for a variable are replaced by the observed mean or median for that particular variable (remaining cases) (McKnight, McKnight et al. 2007). The last observation carried forwards as well as mean and median imputation tend to artificially reduce the variance, biasing the estimates by not taking into account the likely correlations between the various components of the data (García-Laencina, Sancho-Gómez et al. 2010). When the variables of interest are correlated with the data that are present in the complete sample regression imputation may be used. Here, the missing components are replaced by the predicted values from a regression model (linear or non-linear) built using the components of the vector (across samples) that are present.

This enables to preserve the variance and the covariance of variables with missing data, yet the imputed values follow a single regression model and cannot reflect any inherent variation in the data (García-Laencina, Sancho-Gómez et al. 2010). Hot and cold deck procedures replace the missing data from a similar complete data vector from the given the same (hot) or other similar (cold) dataset. Here, the disadvantage is the underestimation of standard errors due to decreased variability since the imputed values are drawn from the data already present in the dataset (McKnight, McKnight et al. 2007). To overcome the above shortcomings of reduced variability and not taking into account covariance structure, multiple imputation methods have been proposed in which missing data is filled *n* times according to an appropriate assumed model (e.g. linear), yielding *n* datasets. Each of these datasets is subjected to a statistical analysis and the results from the *n* datasets are combined for inference, providing valid statistical inferences reflecting the uncertainty due to missing data, e.g. valid confidence intervals for the estimated parameters (Little and Rubin 2002). Second group of missing data imputation methods is based on the machine learning procedures (often complex), in which a predictive model is built to estimate values used for missing data substitution. Example methods comprise k-nearest neighbour (KNN) imputation (details in 2.3 Materials and Methods), self-organizing map (SOM), multi-layer perceptron (MLP) imputation or other neural networks based methods such as auto-associative neural network, recurrent neural network or multi-task networks (Marlin 2008, García-Laencina, Sancho-Gómez et al. 2010, Jerez, Molina et al. 2010).

Model-based procedures allow obtaining parameter estimates given the observed data, the relationships among observed variables and constrains imposed by underlying distribution. The model-based name reflects the fact that the assumption about the joint distribution of all variables in the model has to be made (McKnight, McKnight et al. 2007, García-Laencina, Sancho-Gómez et al. 2010). Unlike the imputation methods, model-based procedures do not assign any values to the missing data, but they treat the data 'as if' they were observed, yielding estimates of parameters and statistics. The most popular model-based

procedure for handling missing data seem to be maximum likelihood (ML) with a basic principle of choosing estimates with values that maximize the probability of obtaining the observed data. This is accomplished by the use of the likelihood function that estimates the probability of the data as a function of the data and the unknown parameters (Enders 2001). The Expectation Maximization (EM) allows obtaining ML estimates when data are missing. EM handles missing data by solving smaller, complete data problems, providing estimates for the entire dataset that is including both the observed and missing values. The procedure is repeated in an iterative process including imputing values for missing data using ML, generating parameters estimates incorporating the imputed data, re-imputing values based on these parameters and re-estimating parameters based on the re-imputed data; all eventually converging on ML estimates (Enders 2001, Little and Rubin 2002, McKnight, McKnight et al. 2007). When the underlying distribution is unknown, Markov chain Monte Carlo (MCMC) procedures offers greater flexibility. Here, briefly the ultimate goal is to obtain a desired posterior distribution, given the observed data and the information gained from the data to update the statistical model that can be used for parameter estimation. MCMC methods generate simulated values in a Markov chain (sequence of random values whose probabilities depend only on the values at the previous step). Similar to EM, the observed data are augmented with simulated values of the missing values in an iterative algorithm (imputation and posterior step) to yield parameter estimates (McKnight, McKnight et al. 2007).

In pattern classification problems several methods have been proposed that does not require a prior missing data imputation; with the missing values being handled in classification problems avoiding the need for the explicit imputations. Some of these methods include neural network ensembles, decision trees, fuzzy approaches or support vector machines, other are multivariate exploratory and predictive approaches such as modification introduced to Principal Component Analysis or Partial Least Squares Discriminant Analysis to handle some amount of missing (Andersson and Bro 1998, Walczak and Massart 2001, García-Laencina, Sancho-Gómez et al. 2010).

2.1.2 Missing data in metabolomics

Many questions addressed using metabolomic approaches are similar to those being asked in transcriptomics and/or proteomic investigations, e.g. which metabolites, genes and/or proteins differ significantly between biological groups under considerations such as healthy *vs*. diseased, or control *vs*. drug-treated samples (Defamie, Cursio et al. 2008). Moreover, for many 'omics experiments the final data formats (after instrument specific processing) are alike, with a rectangular matrix containing gene expression values or metabolite relative abundances, and organised with each variable measured in a unique column and each sample analysed in a unique row. This consistency of data format has facilitated the use of the same or similar univariate and multivariate statistical methods (including computational data analysis or pattern recognition methods) in metabolomics as are used in other 'omics' approaches (Goodacre, Vaidyanathan et al. 2004). However, while in other 'omics' fields there has and continues to be considerable interest in understanding and developing appropriate techniques to handle missing data (prior to statistical analysis), it has received minimal attention in metabolomics. Missing values (also referred to as missing data or missing entries) may arise in metabolomics experiments for a number of reasons. In the case of direct infusion Fourier transform ion cyclotron resonance (DI FT-ICR) mass spectrometry (MS) based metabolomics, they could have a biological and/or technical origin. A metabolite abundance value for a specific sample may not be available when several samples are analysed and then all the measurements are compiled into a data matrix for further comparison or analysis. For some samples a specific peak may not be present for genuine biological reasons, e.g. due to heterogeneity between samples, or in other cases its abundance is below the detection limit of the mass spectrometer, or alternatively it was not measured properly

owing to a technical problem such as a temporary reduction in electrospray performance due to particulate material in the spray nozzle (Payne, Southam et al. 2009).

Metabolomics researches often face such problems with missing data. These problems can be (and have been) addressed by a) simply disregarding all the variables for which missing data are present (Xia, Psychogios et al. 2009), b) using data analysis methods (including univariate and multivariate) that have been shown to be able to handle some proportion of missing data (Kenny, Broadhurst et al. 2010, Blanchet, Smolinska et al. 2011) or c) estimating missing data with various imputation algorithms (Xia, Psychogios et al. 2009, Kenny, Broadhurst et al. 2010) Focusing on the part of the data for which all the measurements are present could be an optimal solution when only a small proportion of variables are affected by missing data, however, this is typically not the case in most metabolomics experiments. Some statistical software packages allow univariate statistical testing (e.g. t test or ANOVA in R or Matlab) on samples with missing data by simply disregarding the missing entries, and some multivariate exploratory or predictive approaches have been developed to handle some amount of missing data (e.g. missing data Principal Component Analysis or Partial Least Squares Discriminant Analysis (Andersson and Bro 1998, Walczak and Massart 2001).

While this strategy may be appropriate when dealing with large sample size studies that contain few missing data, it may be problematic for metabolomics studies in which the sample size is often limited and thus ignoring missing data could diminish the power of the statistical tests. Furthermore, it is not uncommon that missing data occurs predominantly in one biological group (at least for the case of DI FT-ICR MS metabolomics) and when combined with a small sample size this can lead to an insufficient number of metabolites measured for this approach (e.g. at least two detected measurements per biological group are needed to perform a t test in the

R or Matlab environments). Hence, imputing missing data prior to data analysis represents a practical solution in applied metabolomics since i) it yields a simple, consistent, rapid and automated data processing pipeline, ii) the resulting data matrix is compatible with a very wide array of univariate or multivariate analyses, iii) this approach facilitates a comparison of univariate and multivariate statistical results for specific metabolites of interest (e.g. biomarkers), and iv) it provides a complete profile of metabolite concentrations that can be used in a consistent manner in other types of data analysis (e.g. integration with other 'omics datasets). The importance of appropriate handling of missing data has been recognised in the analysis of DNA microarray (Troyanskaya, Cantor et al. 2001) and gel-based proteomics data (Pedreschi, Hertog et al. 2008, Albrecht, Kniemeyer et al. 2010). For example, studies have been reported on how missing values affect statistical parameter estimations (Troyanskaya, Cantor et al. 2001), how they influence the results of univariate (de Brevern, Hazout et al. 2004, Scheel, Aldrin et al. 2005) and multivariate data analysis (Pedreschi, Hertog et al. 2008), what is the optimal method of their imputation (Jörnsten, Wang et al. 2005, Scheel, Aldrin et al. 2005, Kim, Lee et al. 2007, Tuikkala, Elo et al. 2008), and how to develop robust data analysis algorithms for datasets with significant amount of missing entries (Kim, Lee et al. 2007). In metabolomics, none of the above questions has yet been thoroughly addressed. This is particularly surprisingly given that mass spectrometry based metabolomic analyses (e.g., DI FT-ICR MS) typically generate datasets with considerable amounts of missing data (Southam, Payne et al. 2007, Han, Danell et al. 2008, Taylor, Weber et al. 2009). Furthermore, it has been suggested that missing values do not affect the data analysis outcome, but that their treatment (i.e. deletion or estimation) is carried out only for computational convenience (Steuer, Morgenthal et al. 2007). Current methods for handling missing data in metabolomics involve simple methods such as replacing a missing value by the

mean or median of the available measurements for that variable, replacing with some small arbitrary number, or k-nearest neighbour imputation (Steuer, Morgenthal et al. 2007) (GeneSpring MS software (Alignment Technologies). A quite different approach, reported by Sangster et al. (Sangster, Wingate et al. 2007), estimates missing values by returning to the raw spectral data and integrating the areas of the missing peaks which are below the applied signal-to-noise ratio (SNR) threshold, but in close proximity to the peaks' known *m/z* value.

Here, we analyse missing data in the context of DI FT-ICR MS based metabolomics measurements, but with the findings of our analyses potentially transferable and of importance for other metabolomics studies. We investigate not only the nature of the missing data but also their effects on data analysis, both univariate and multivariate. Specifically we addressed the following questions: what are the potential origins of missing data in metabolomic datasets? Do they appear at random, or as a function of peak intensity and/or *m/z* value? Do they affect the outcome of commonly used univariate and multivariate data analyses? And if so, what is the optimal method of replacing their values as part of a consistent and automated data processing pipeline that will provide the metabolomics researchers with a complete data matrix that is compatible with many univariate and multivariate statistical analyses or other data mining algorithms? With more than a dozen imputation methods available and published (not limited to 'omics' studies), we focus our investigations on the eight commonly used and reported methods in applied 'omics studies that are readily implementable (by other researchers) in the R computing environment, ultimately providing a (potentially expandable) benchmark for the questions above. Finally, to maximise the generality of our findings, we have investigated three widely differing biological datasets (including cellular, tissue and whole organism extracts from

in vitro and *in vivo* experimentation) that were measured in positive and/or negative ion mode FT-ICR MS.

2.2 Materials and Methods

2.2.1 Mass spectrometry datasets

The three FT-ICR MS datasets were available for theoretical consideration of missing data problem, experimental design and data collection done by other researches, at the Environmental Metabolomics Research Laboratory, University of Birmingham. These comprised of 1) CCL – cancer cell line, specifically an acute myeloid leukaemia cell line (K562) cultured and treated under hypoxic conditions, comprising of 6 control samples, 6 samples exposed to indomethacin (non-steroidal anti-inflammatory drug) and 6 samples treated with medroxyprogesterone acetate (component of hormonal contraceptives), all measured in positive (CCL_p) and negative ion mode (CCL_n); 2) DM – *Daphnia magna* (a freshwater invertebrate) exposed for 24 h to 1.5 mg/L of 2,4-dinitrophenol, a cellular metabolic toxicant which obstructs oxidative phosphorylation, comprising of 10 control and 10 exposed samples measured in negative ion mode only (Taylor, Weber et al. 2010); 3) HL – human liver biopsies taken throughout orthotopic liver transplantation, comprising 7 biopsies taken soon after organ retrieval and 7 further biopsies taken post-reperfusion of blood circulation in the recipient patient, measured in positive ion mode only (Hrydziuszko, Silva et al. 2010). All cell, tissue or whole organism samples were extracted using a methanol/chloroform/water method (Wu, Southam et al. 2008), and the polar metabolites were analysed using a hybrid 7-T direct infusion nanoelectrospray FT-ICR mass spectrometer (Thermo Fisher Scientific LTQ FT) over the range *m/z* 70 to 500. All CCL and DM samples were analysed in triplicate, and the HL samples in duplicate; these represent technical replicates of each sample for use in the subsequent noise filtering algorithm. Spectra were processed as

described previously (Taylor, Weber et al. 2009), including a 3-step filtering algorithm to eliminate noise peaks (Payne, Southam et al. 2009). Specifically, the first filtering step comprised of a hard SNR threshold, below which peaks were rejected (2.5 for CCL datasets and 3.5 for DM and HL datasets). In the second step, only peaks present in 2 out of 3 technical replicates (for CCL and DM datasets) and 2 out of 2 replicates (HL dataset) were retained, and the intensities averaged to create a single spectrum per biological sample. In the third step, only peaks present in at least 50% of the samples were retained (specifically across all samples for each of the DM and HL datasets, and across samples within each biological group for the CCL datasets; Table 1). Probabilistic quotient normalisation was then performed on all of the datasets (prior to univariate and multivariate analyses) (Dieterle, Ross et al. 2006) followed by the generalised log transformation (prior to multivariate analysis only, in order to stabilise the variance across the peaks and avoid the highest abundance peaks dominating in the multivariate analyses) (Parsons, Ludwig et al. 2007). Individual peak intensities were confirmed to follow normal distributions as tested with the Shapiro-Wilk normality test (for >99% of the peaks that have no missing data and that have been measured in at least three samples; this is in agreement with our previous unpublished observations for other (including larger sample size) metabolomics datasets obtained via DI FT-ICR MS). At this stage of analysis, each dataset contained *m* peaks and *n* samples with multiple missing values (Table 1). The median of the coefficients of variation of the peak intensities, reflecting biological diversity within each dataset (Parsons, Ekman et al. 2009), was 17.21%, 20.24%, 25.59% and 60.99% for CCL_n, CCL_p, DM and HL, respectively (excluding peaks with missing data) confirming, as expected, that the metabolic heterogeneity increased from cell line extracts to laboratory cultured organisms to clinical samples.

Table 2.1 List of the DI FT-ICR MS based metabolomic datasets analysed together with some of their basic properties.

Dataset	Brief Description	Median of coefficient of variation [%]	No .of samples	No. of groups	No. of peaks	Missing values [%]	Peaks with missing values [%]
CCL_n	Human cancer cell line K562, negative ion mode	17.21	18	3	6770	22.01	51.67
CCL_p	Human cancer cell line K562, positive ion mode	20.24	18	3	4426	28.53	64.96
DM	*Daphnia magna* exposed for 24h to dinitrophenol, negative ion mode	25.59	20	2	4196	14.63	55.22
HL	Human liver tissue prior and post liver transplantation, positive ion mode	60.99	14	2	1805	23.66	78.73

2.2.2 Occurrence and distribution patterns of missing data

The properties of the missing values in the FT-ICR MS datasets were examined using two methods. First, the distribution of the missing values across each dataset was determined to be 'missing completely at random' (MCAR) or not. This employed Little's test of MCAR for multivariate data with missing values (Little 1988). Second, their occurrence patterns were assessed using Pearson's correlation between missing data properties and the dataset features, specifically the amount of missing data *vs*. both the abundances and *m/z* values of the non-missing data peaks.

2.2.3 Impact of missing data imputation on statistical analyses

We then compared eight common and/or readily available missing data imputation (or estimation) methods in terms of their impact on univariate and multivariate data analysis as well as in terms of their performance for handling missing values (experimental design summarised in Figure 2.2). This was performed on all three FT-ICR MS datasets, as described below. Specifically, the eight estimation methods comprised: 1) *S* – substitution of missing values with a small predefined value (e.g. 0.01) (as used in GeneSpring MS software; Alignment Technologies); 2) *HM* – substitution with half of the minimum value found in the non-missing data (Xia,

Psychogios et al. 2009); 3) *M* – substitution with the mean of the non-missing values across all samples for that peak (Steuer, Morgenthal et al. 2007); 4) *MED* – substitution with the median of the non-missing values across all samples (Steuer, Morgenthal et al. 2007); 5) *KNN* – weighted k-nearest neighbour algorithm in which *k* (here $k = 5$; different *k* values did not significantly affect our analysis, data not shown) metabolites most similar in terms of their intensity profiles across all samples are identified based on the Euclidean distance similarity measure to the metabolite having a missing datum for a given sample; the missing datum is then estimated as the weighted average of the *k* metabolites for that sample with their contribution weighted by their similarity (Troyanskaya, Cantor et al. 2001, Steuer, Morgenthal et al. 2007); 6) *BPCA* – Bayesian PCA missing value estimation, a three-stage algorithm based on principal component regression, Bayesian estimation and the expectation-maximisation repetitive algorithm; briefly during the principal component regression the missing data of a metabolite's intensity profile are estimated from the observed values using the PCA result, followed by a Bayesian estimation in which residual error and the projection of metabolites on the principal components are considered as normal independent variables with unknown parameters which are inferred in the final expectation-maximization algorithm step (Oba, Sato et al. 2003, Xia, Psychogios et al. 2009); 7) *MI* – multivariate imputation by chained equations (Buuren and Groothuis-Oudshoorn 2010); 8) *REP* – modified version of Sangster's method as used by us previously, for which a missing value is substituted with the average intensity of the nearest (in term of *m/z* value) peaks from the raw measurements of the technical replicates (Sangster, Wingate et al. 2007). Methods *S* and *HM* substitute the missing values with a relatively small value and act on the assumption that missing data do not influence the outcome of the subsequent data analysis due to a low amount of missing data. Methods *M* and *MED* impute missing data using a row mean or median and assume that the

metabolite's intensity is similar across all the experiments (i.e. samples). Furthermore, along with the *S* and *HM* methods, *M* and *MED* do not use the information contained in the structure of the data. *KNN* searches for the *k* metabolites (or more specifically peaks in the mass spectra) that have similar measured signal intensities across the biological samples as compared to the peak for which the missing entry is present. The missing value is then replaced with the weighted average of the corresponding non-missing values from the group of *k* peaks that were identified as most similar. *BPCA* and *MI* methods use the global structure of the dataset, in a way that all the metabolites are taken into consideration to obtain the imputed value. *BPCA* estimates the missing data in a three-stage process starting with principal component regression, followed by Bayesian estimation and finishing with an expectation-maximisation like repetitive algorithm; this approach has been shown to outperform *KNN* for gene expression data (cDNA microarrays) when the number of samples was large (>30) and the missing data occurred randomly (Oba, Sato et al. 2003, Albrecht, Kniemeyer et al. 2010). *MI,* multivariate imputation by chained equations (available in the R environment in a MICE library) is a method of multiple imputation in which each variable is estimated using a regression model conditional on all the other variables iteratively looping through all the variables with missing data; here we used the predictive mean matching implementation (Little and Rubin 2002), similar to the regression method. Method *REP* attempts to utilise peak abundance information captured in the technical replicates that lies beneath the SNR threshold by estimating missing data via the average of the closest (in terms of *m/z*) peaks in each of the three (two for HL dataset) technical replicates.

After imputing the missing values for the three FT-ICR MS datasets, using all eight methods, statistical tests were employed to determine the metabolic differences between the sample classes (e.g. between the control and two drug-treated groups in the CCL study). This allowed us to

examine the impact of each imputation method on finding significantly changed peaks via univariate testing (t-test or ANOVA between groups with Benjamini and Hochberg correction for multiple testing (Benjamini and Hochberg 1995) as well as via multivariate principal component analysis (PCA) scores and loadings values. This approach was chosen as it is routinely used in metabolomics to provide an initial unsupervised explorative analysis and because it is appropriate for the sizes of the datasets investigated here (containing thousands of peaks and only up to ten samples per biological group); supervised methods such as partial least squares discriminate analysis would likely lead to over-fitting in the modelling (Broadhurst and Kell 2006, Westerhuis, Hoefsloot et al. 2008). The eight missing value estimation methods were further compared in terms of the outcome of both univariate and multivariate analysis via hierarchical clustering, in which the Euclidean distance was calculated for the groups of peaks significantly changed between biological groups for the univariate analysis and for the top 5% of peaks contributing to the separation along the first and the second principal components (PC1 and PC2) based on their loading values. For the univariate approach, the Euclidean distance was calculated based on the overlap (percentage) of the number of identified as significantly changed peaks between every two missing data estimation methods. For the PCA, since the order (i.e. ranking based on loading values) of these peaks holds important information about the PCA results, we developed a measure referred to as *ODist* that compares the number of shared peaks between any two imputation methods as well as their rank order; i.e. in addition to calculating the amount of overlap in the top 5% of peaks between two imputation methods, we assigned a higher 'similarity' value for methods for which the top peaks are in the same order (see Supplementary Material, 'Impact of missing data imputation on multivariate data analysis').

In addition, to assess the impact on the final biochemical interpretation of the data, we have compared the eight methods in terms of detecting significantly 'active' KEGG (Kyoto Encyclopaedia of Genes and Genomes) human pathways (Kanehisa, Araki et al. 2008). Here we defined a significantly 'active' pathway as follows: for significantly changed peaks between groups (univariate analysis) or the top 5% of peaks contributing towards separation along PC1 or PC2, we assigned one (or more) putative metabolite names (Sumner, Amberg et al. 2007) to each *m/z* value, based upon accurate mass measurements and the KEGG database, taking into account commonly detected ion forms ($[M-e]^+$, $[M+H]^+$, $[M+Na]^+$, $[M+^{39}K]^+$, $[M+2Na-H]^+$, $[M+2^{39}K-H]^+$ for positive ion mode and $[M+e]^-$, $[M-H]^-$, $[M+^{35}Cl]^-$, $[M+^{37}Cl]^-$, $[M+HAc-H]^-$ (HAc, acetic acid) for negative ion mode). Following that, for each putatively identified metabolite, we listed all KEGG pathways for which it is involved. The probability that a peak *i* belongs to a pathway *j* was calculated via $P_i(pathway_j \mid peak_i)$ = ∑ putative metabolite assignments of peak *i* that are involved in pathway *j* / ∑ putative assignments for peak *i*. We marked pathways as significantly 'active' if for at least one of the peaks they were observed with a probability greater than or equal to 0.75.

2.2.4 Performance of missing data estimation algorithms

To assess the performance of the missing data imputation algorithms, we used 'complete' CCL_n, CCL_p, DM and HL datasets that were created by excluding all peaks that contained missing values. Next, we deliberately introduced missing values, either completely at random (MCAR) or not at random (MNAR) (Little and Rubin 2010), generating datasets with missing data entries but for which we knew the real (original) values. These missing data were again imputed using each of the eight methods described above (for *REP* we estimated the values of the

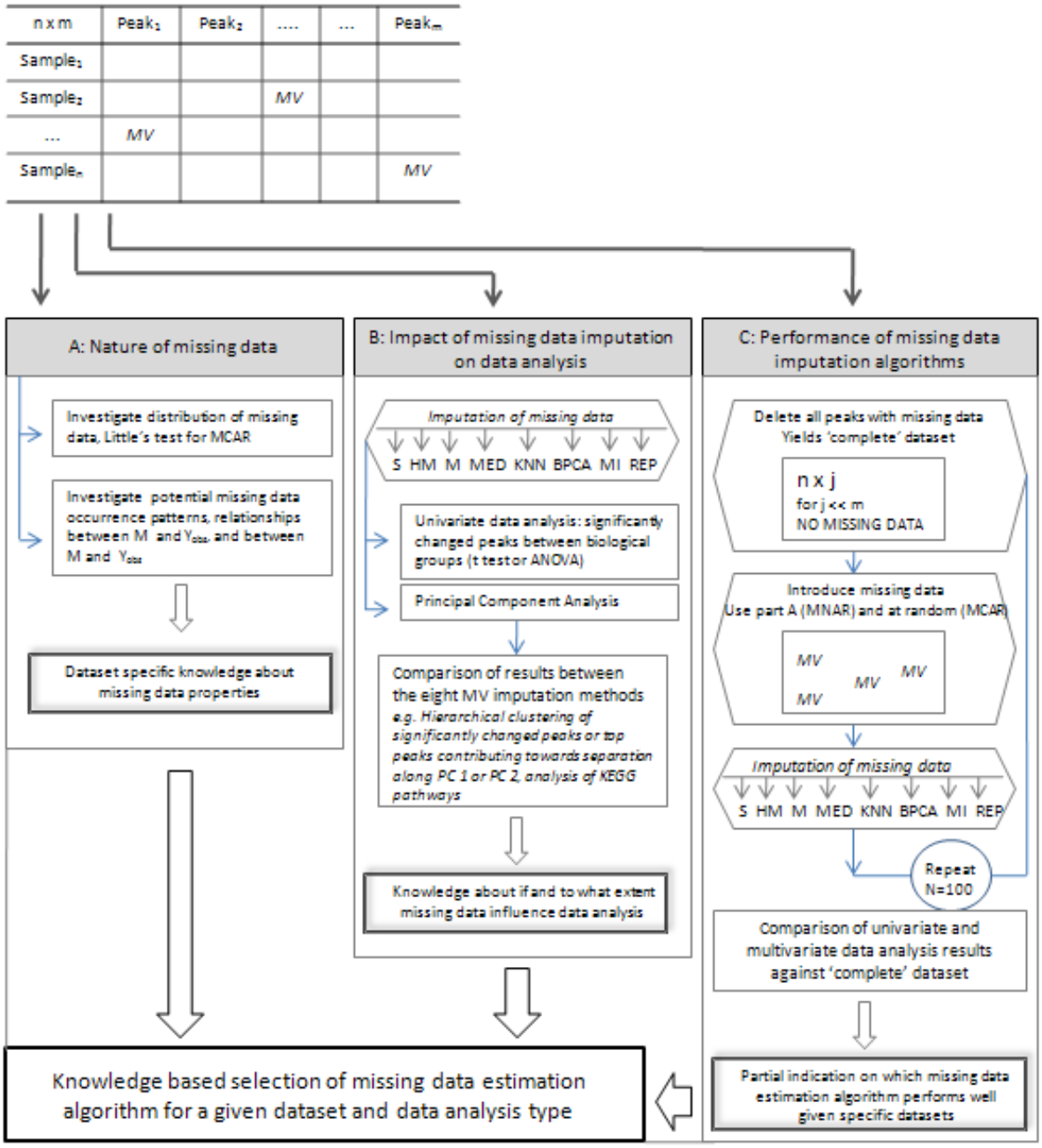

Figure 2.2 Flow chart summarising the approach for the analyses of missing values (MV) performed in this study.

deliberately introduced missing entries using the information from the triplicate technical measurements, explained below). The imputed missing values (for all eight methods) were compared to the original deliberately excluded values in terms of normalised root mean square errors (NRMSE) (Troyanskaya, Cantor et al. 2001) and also the outcome of univariate and multivariate (PCA) data analysis. The whole procedure (both for MCAR and MNAR) was repeated 100 times. Furthermore, the amount of missing data generated was ca. 20% for each 'complete' dataset, similar to the amount of real missing data in the original datasets. For MNAR, missing values were introduced to mirror the missing data properties that we discovered in the original datasets, mainly to capture their relationship with the signal intensity and *m/z* ratio. For each original dataset linear regression model was built to capture the relationship between signal intensity (non-missing data) and *m/z* ratio and the frequencies of missing data occurrence were noted for each SIM window as well as the overall percentage of missing data (ca. 20% as in Table 2.1). The same amount of missing was introduced to the 'complete' datasets by identifying the number of missing data per SIM window to maintain the original datasets proportions and second, based on the regression model, selecting which of the observed data should be labelled as missing.

2.3 Results and Discussion

2.3.1 Occurrence and distribution patterns of missing data in DI FT-ICR MS metabolomics

A typical DI FT-ICR MS based metabolomics dataset measured and processed as described above contains ca. 20% of missing data (Table 2.1), which equates to up to 80% of peaks having at least one missing value across the analysed samples. Little's MCAR test revealed that this missing data does not occur completely at random ($p = 0.029$, 0.032, 0.021 and 0.045 for CCL_n,

CCL_p, DM and HL datasets respectively), i.e. they do not follow a random distribution. If *Y* denotes a rectangular dataset (e.g., with each peak's abundance in a unique column and each

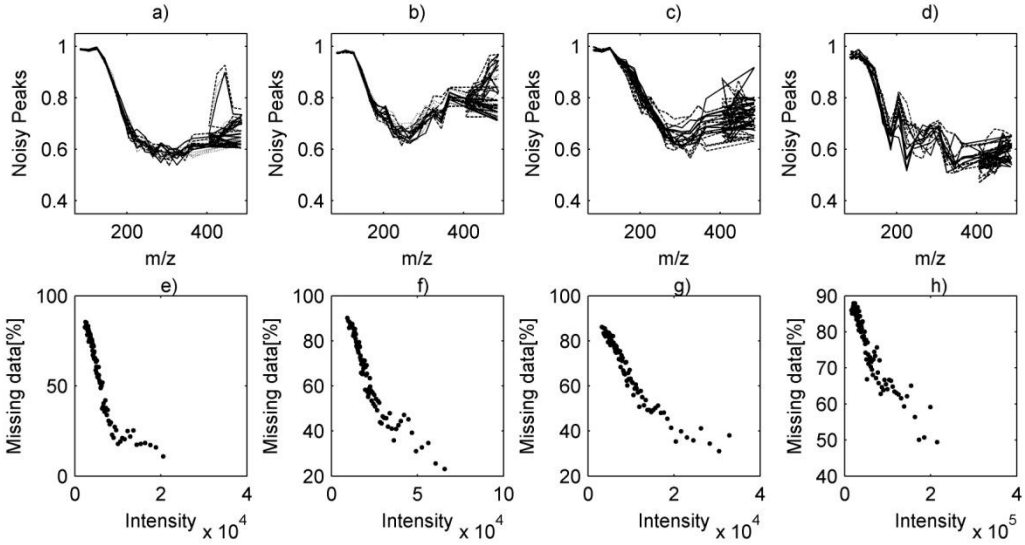

Figure 2.3 Probability of the occurrence of noisy peaks as a function of *m/z* ratio and percentage of missing data vs. mean peaks abundance for the four datasets. Probability of the occurrence of noisy peaks as a function of m/z ratio for a) CCLn, b) CCLp, c) DM and d) HL datasets. Each sample in the CCL and DM datasets was measured as 3 technical replicates, and therefore noisy peaks are defined as occurring in 1 or 2 out of 3 measurements only (prior noise filtering). HL samples were measured as duplicates, with noisy peaks defined as occurring in 1 out of 2 measurements only. Relationships shown for all separately for all available samples in each dataset (marked in multiple lines on the graphs). Percentage of missing data vs. mean peak abundances (binned in 100 intervals with a sample filter ≥50%) for e) CCLn, f) CCLp, g) DM and h) HL datasets. The top 5% of peak abundances have been removed for plotting purposes.

sample in a unique row) containing missing data, *Y* can be split into two subsets Y_o (observed values) and Y_m (missing data). A dummy matrix *D* could then indicate the location of the missing values in *Y*, such that $d_{ij} = 1$ if y_{ij} is missing and $d_{ij} = 0$ if y_{ij} is present. Following Rubin's established categories for the mechanisms of missing data (Rubin 1976) for MCAR (missing completely at random; intuitively perceived as random), by definition, there is no relationship Φ between *D* and either Y_m or $Y_{o,}$ meaning that a pattern (occurrence) of missing data is not dependent on the values that are missing (e.g. all low intensity values in the dataset are missing) nor on the observed data (e.g. missing data occur for those metabolites with measured low

intensity). For non-randomness, missing values may occur as MAR (missing at random, with a relationship Φ between D and Y_o; counter-intuitive since it denotes one type of non-randomness) or MNAR (missing not at random, with a relationship Φ between D and Y_m and possibly Y_o; describing the other type of non-randomness) ((Rubin 1976, Little and Rubin 2002). For both the random and non-random scenarios, missing data may arise due to technical errors, biological factors, or a mixture of the two. The relationship Φ between D and Y_m is theoretical and therefore cannot be assessed due to the lack of information about the missing data. However, the analysis of the relationship between D and Y_o showed that missing data in FT-ICR mass spectra is a function of both the abundances of observed peaks as well as their *m/z* values. For the former, the lower the (mean) peak abundance the greater the amount of missing data that the peak contains (Pearson correlation coefficient of -0.80, -0.85, -0.89 and -0.90 for the CCL_n, CCL_p, DM and HL datasets respectively; Figure 2.3). For the latter, the analysis of the technical replicates showed that the probability of observing noisy peaks (i.e. missing in one or more of the technical replicates) is high for low *m/z* value signals, decreasing in probability for mid-range *m/z* values, and increasing again for high *m/z* signals (Figure 2.3). This trend was quite apparent for both CCL and the DM datasets, with a small exception (i.e. no increase for the high *m/z* peaks) for the most biologically variable HL dataset. This effect was of unknown origin. We have confirmed that these relationships are not due to our data processing, i.e. the three stage noise filtering algorithm. Further investigation of the potential source(s) of this relationship is beyond the scope of our study, possibly arising from a technical peculiarity of the FT-ICR MS instrumentation.

Two relationships were discovered above, i.e. D being correlated significantly with non-missing value peak abundances (Y_o) and also with *m/z* values. These relationships highlight important information about the occurrence and distribution of missing data that needs to be

considered prior to their treatment. The large percentage of peaks containing missing data implies that simply removing them would dramatically reduce the size of the dataset (by ca. 80%). Furthermore the abundance relationship indicates that removing peaks with missing data, or inadequate estimation of missing entries, could result in a substantial bias since only peaks with the highest abundances would be kept, hindering subsequent biomarker discovery. We have demonstrated this latter effect while investigating the influence of the third step of the noise filtering algorithm (retaining only those peaks present in s% or more of the samples, and equivalent to discarding peaks with missing data) on the distribution of missing data across biological groups. Figure A1 (Appendix A) shows the effects on the missing data distributions for s≥0%, s≥25%, s≥50% and s≥75% for the DM and HL datasets. The most interesting result was obtained for the HL dataset, for which there was a significant difference in the number of missing values between the cold-phase and post-reperfusion groups when all peaks were retained (s≥0%), but this difference became non-significant for settings of s≥25% and above. From our biochemical knowledge of the processes occurring during liver transplantation, we hypothesise that this is a case where peaks with missing data that genuinely carried important biological information were mistakenly removed (assumed to represent noise peaks), as it has been shown that during the cold phase liver metabolism ceases and it restarts upon reperfusion with an increased production of bile acids and urea (Hrydziuszko, Silva et al. 2010). Overall, our assessment of the missing data within DI FT-ICR MS datasets reveals that missing values do not occur completely at random but instead as a function of (at least) peak abundance and *m/z* value, and that peaks with missing values potentially carry importance biological information.

2.3.2 Impact of missing data imputation on univariate data analysis

The potential impacts of the eight missing data estimation techniques on the ability to discover significant differences in peak abundances between biological groups were evaluated using univariate statistical tests (either t-tests or ANOVAs). Different imputation methods ultimately yielded quite diverse data analysis outcomes. Specifically, the number of peaks identified as significantly different between groups varied considerably between the eight estimation methods, from 2.65% to 14.70%, from 0.58% to 10.20%, from 7.44% to 14.20%, and from 1.72% to 14.24% for the CCL_n, CCL_p, DM and HL datasets, respectively (Appendix A Table A1). As expected, based on the underlying mechanisms of the missing data estimation algorithms, some methods performed comparably, e.g. *S* and *HM*, and *M* and *MED* (Appendix A Figure A2 and Table A2), while others were strongly dependent on the structure of the dataset, e.g. for the CCL datasets, *BPCA* and *REP* performed similar to *M* and *MED*, while for the DM and HL datasets, *REP* resembled *S* and *HM*.

Further investigation of the peaks detected as significantly different between biological groups showed that substantial proportions of these peaks were comprised of those which initially had missing data. Specifically, for the CCL_n, CCL_p, DM and HL datasets, the minimum percentage occurrence of this type of significant peak (from across all eight estimation methods) was 24.73% (for the *BPCA* method), 15.38% (*BPCA*), 23.72% (*M*) and 32.26% (*MED*), respectively (Appendix A Table A1). The maximum percentages of significant peaks (which originally had missing values) were surprisingly high, at 72.62% (*KNN*), 83.66% (*HM*), 49.66% (*HM*) and 85.99% (*S*) for the same four datasets, respectively. In the worst case scenario of inadequate imputation of missing data entries, these minimum and maximum percentages provide an estimate of the false positive error rate associated with the arguably critical identification of

significantly changing peaks. This error rate would most likely be in the upper range of percentages for methods such as *S* and *HM* if the missing data were to represent high abundance metabolites, and also for *M* and *MED* if the opposite were true with missing entries representing low abundance metabolites. This is further visualised in Figures A3 and A4 (Appendix A) that show the distribution of the number of missing data (prior to imputation) that occur specifically within the significantly changing peaks only.

Having analysed the percentages of significantly changing peaks between biological groups that initially had missing data, we then investigated which of the samples originally had these missing values. Interestingly, the results showed that missing entries tend to be located in one of the biological groups, rather than being spread equally across all the groups (Appendix A Table A3). This is a further important observation that helps to verify our earlier hypothesis that missing data may in fact represent true differences between biological groups, and therefore their accurate imputation is of considerable importance. Also, as above, it demonstrates a potential danger when using substitution (*S* and *HM*) or simple imputation (*M* and *MED*) methods. For the case of low peak abundances, it does not mean that a small arbitrarily chosen value would represent the missing data accurately; also, for high peak abundances, it should not be assumed that the mean or median of the non-missing values would represent an optimally imputed value. Rather, it is quite possible that when an inappropriate missing value estimation method is used we may not only lose the knowledge of which peaks are significant or not, but we may introduce further bias by identifying non-significant peaks as significantly different between groups.

Our results therefore point to the potential bias in the biochemical interpretation of metabolomics data, if missing values are estimated incorrectly. To verify this we have compared what we refer to as ‘active’ human pathways observed following the estimation of missing data,

across all eight algorithms. These again resulted in quite diverse outcomes of 'active' pathways with only 20.0%, 14.4% and 0.0% (for CCL_n, CCL_p and HL datasets, respectively) of pathways observed across all of the missing data algorithms (Table 2.2, Appendix A Tables A4-A5); note that this approach was not applied to the non-human DM dataset. The highest number of 'active' pathways was detected for the *S*, *HM* and *KNN* methods, while the lowest for *M*, *MED* and *BPCA*. Prior biochemical knowledge can also aid the interpretation of these findings. For example, for the HL dataset, arginine and proline metabolism and taurine and hypotaurine metabolism are known to play a substantial role in liver transplantation (Silva, Mirza et al. 2006, Kincius, Liang et al. 2007). Both these pathways were discovered to be 'active' following treatment of missing values by the *S, HM, KNN* and *REP* methods, while *BPCA* treatment did not lead to either being classed as active. Overall, the results presented here provide substantial evidence that the choice of missing value estimation method has a substantial effect on the outcome and interpretation of univariate statistical analysis.

Table 2.2 Summary of which KEGG human pathways are 'active' (i.e. observed with 75% likelihood based on the significantly changing peaks between cold phase and post reperfusion groups) in the human liver (HL) dataset, after estimating the missing values with eight different algorithms.

KEGG pathway	S	HM	M	MED	KNN	BPCA	MI	REP
Purine metabolism	X*	X	X	-	X	X	X	X
ABC transporters	X	X	X	X	X	-	X	X
Neuroactive ligand-receptor interaction, Taurine and hypotaurine metabolism	X	X	-	X	X	-	X	X
Pyrimidine metabolism, Arginine and proline metabolism	X	X	X	-	X	-	-	X
Nicotinate and nicotinamide metabolism, Tyrosine metabolism, Drug metabolism - cytochrome P45	X	X	-	-	X	-	X	X
Glycine, serine and threonine metabolism, Aminoacyl-tRNA biosynthesis, Cysteine and methionine metabolism, Alanine, aspartate and glutamate metabolism	X	X	-	-	-	X	-	X
Galactose metabolism	X	X	-	-	X	-	-	X
Lysine degradation, Histidine metabolism, beta-Alanine metabolism, Phenylalanine metabolism, Tryptophan metabolism	X	X	-	-	-	-	-	X
Ether lipid metabolism, Calcium signalling pathway, Fc gamma R-mediated phagocytosis	X	X	-	-	-	-	-	-
Cyanoamino acid metabolism, Glutathione metabolism, Thiamine metabolism, Nitrogen metabolism	X	-	-	-	-	-	-	-
Taste transduction	-	-	-	-	-	-	-	X

* X indicates that a pathway is 'active' for this particular method.

2.3.3 Impact of missing data imputation on multivariate data analysis

Similar to the results from the univariate analyses, the eight missing data estimation techniques also led to diverse outputs from the multivariate data analysis. Specifically, this was assessed from the clustering (or not) of samples from different biological groups on PCA scores plots (Figure 2.4 and Appendix A Figures A5-A7). The differences between the eight estimation techniques were most evident for the most biologically homogeneous dataset, the cell line extracts. For example, for CCL_n there were clear differences between the control, indomethacin treated and medroxyprogesterone acetate treated groups after estimating the missing entries with *S, HM, KNN* and *MI*, a separation between the control and two drug treated groups (but no separation between drug treatments) after *M, MED* and *REP*, and no separation between any of the groups after *BPCA*. The differential effects of the eight estimation methods on the PCA results were further demonstrated by the large spread of the variances captured by the first two principal components: the relative standard deviations of the variances were 36.10% and 16.58% for PC1 and PC2 respectively for CCL_n, 57.46% and 29.53% for CCL_p, 56.52% and 15.13% for DM, and 38.64% and 11.72% for the HL dataset (Appendix A Table A6).

Comparison of the top 5% of peaks contributing towards the separation along PC1 and along PC2 showed, as for the univariate analyses, that some estimation methods performed quite similarly (Appendix A Figures A8-A9 and Tables A7-A8). For example, the largest overlap in significant peaks for the univariate data analysis was between *S* and *HM* (97.51%, 97.61%, 99.83% and 97.28% overlap for CCL_n, CCL_p, DM and HL, respectively) followed by a slightly smaller overlap between *M* and *MED* (95.33%, 92.86%, 95.38% and 68.29%), while for the top 5% of peaks from the PCAs the similarities (expressed as *ODist*; see Chapter 4, comparing ordered sets) were largest between *M* and *MED* (97.08%, 84.82%, 96.19% and 86.67% for PC1,

and 94.46%, 88.39%, 92.38% and 75.56% for PC2) and followed by *S* and *HM* (76.68%, 70.09%, 82.38% and 61.11% for PC1 and 77.84%, 64.29%, 80.00% and 71.11% for PC2). For the PCA results, the smallest overlap was between the *S* and *BPCA* methods, whereas for the univariate data analysis the smallest overlap was between *S* and *M* or *MED*. An important observation from these findings is that the differences in the statistical results between the eight estimation methods were larger for the PCA than for the univariate analyses, indicating that the multivariate data analysis may be more sensitive to the missing data estimation technique used.

This observation of the higher sensitivity of multivariate analysis to missing data estimation was additionally verified by further examination of the top 5% of peaks contributing towards the separation of biological groups along the principal components (from each PCA). In general, these subsets of *m/z* values contained a larger proportion of peaks that initially contained missing data as well as a larger proportion of missing entries than their univariate equivalents (except for the *BPCA* method applied to the CCL_n, DM and HL datasets). For the *S* and *HM* methods, virtually all the peaks in this top 5% subset contained at least one missing value prior to their estimation (Appendix A Table A6, Figures A10-A11, Tables A9-A10). Furthermore, the considerable differences in the results of the PCAs between the eight estimation methods were further illustrated by substantial heterogeneity in the observed 'active' human pathways. Specifically, there were no common pathways following application of the eight tested estimation methods for CCL_n (for the top 5% of peaks contributing to PC1 and to PC2), CCL_p (for top 5% of peaks contributing to PC2) and HL (for top 5% of peaks contributing to PC1). Virtually none of the remaining dataset and principal component combinations exhibited overlap across all eight estimation methods, except for 2.11% and 3.85% of 'active' pathways for CCL_p (PC1) and HL (PC2), respectively (Appendix A Tables A11-A15). Overall, these analyses provide definitive

evidence that the choice of missing value estimation method has a substantial effect on the results of the multivariate statistical analysis used here.

2.3.4 Performance of missing data estimation algorithms

Assessing the effects of missing data estimation methods on the 'complete' datasets, which had missing values deliberately introduced either at random (MCAR, missing completely at random; mentioned here as a comparative benchmark only and with results reported in the Appendix A) or in a way that mimics the missing data distribution and properties of actual FT-ICR MS metabolomics datasets (MNAR, missing not at random), revealed further interesting findings. Before examining these results, it is worth noting that the 'complete' datasets may not be a perfect representation of the 'original' datasets in terms of the metabolites' intensities as a larger proportion of the peaks that were removed to create the 'complete' datasets were of relatively low intensity (i.e. the actual missing data did not occur at random, but occurred in part as a function of peak intensity). However, once missing data were intentionally introduced to the 'complete' datasets to mimic their distribution in the 'original' data, the 'complete' datasets regain similar properties to the 'original' datasets; in fact this represents the best possible approach to assess the performance of missing values estimation algorithms even when missing data are a function of intensity (Scheel, Aldrin et al. 2005, Albrecht, Kniemeyer et al. 2010) since the estimation methods are compared internally within the 'complete' matrix. Here, an obvious bias could occur for method *S*, with a small predefined value (0.01), while the seven other methods should capture the relationship of the 'original' data. This limitation should, however, be kept in mind and this component of the comparison of imputation algorithms should be combined

with the findings from the missing data distributions and their influence on the data analysis (see above) to select the optimal missing data estimation method.

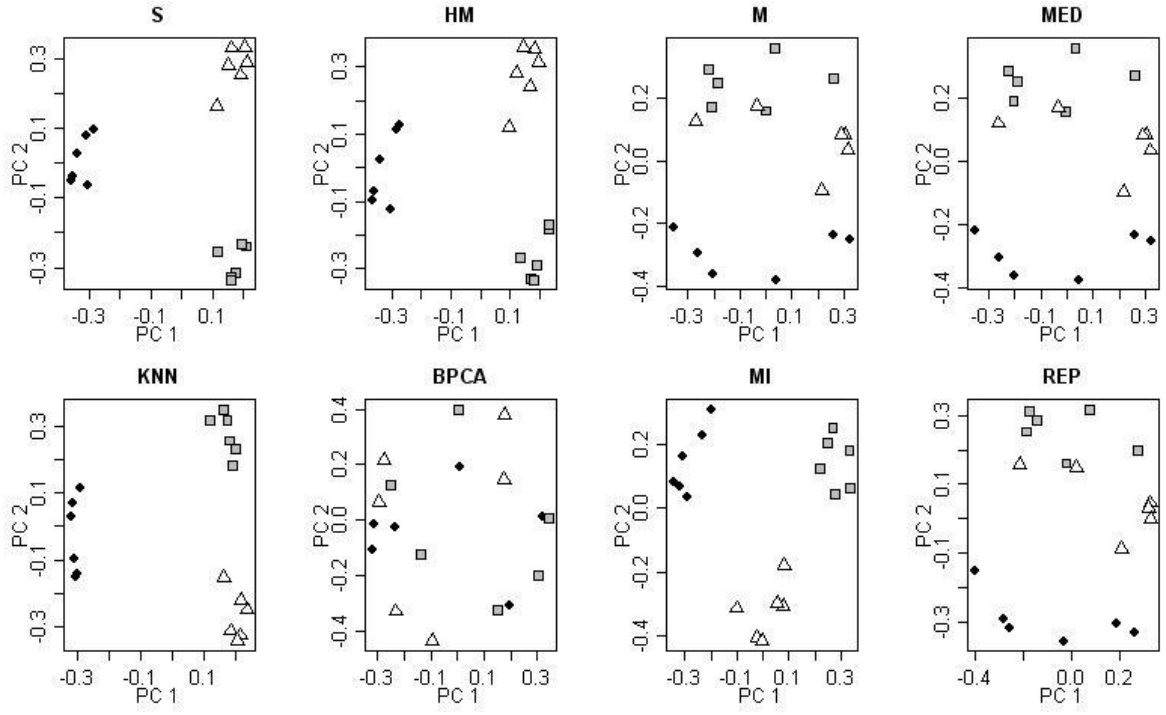

Figure 2.4 Comparison of eight different missing value estimation methods based upon their effects on the PCA scores plots for the CCL_n dataset. Samples labelled as: control cancer cells (diamonds), indomethacin treated (squares) and medroxyprogesterone acetate treated (triangles).

The performances of the eight methods were evaluated in terms of normalised root mean square errors (NRMSE), where the best approach results in the smallest NRMSE between the known, deliberately deleted, 'missing' data and the values that were subsequently estimated for them (averaged across N=100 runs). Five of the eight estimation methods yielded similarly small average NRMSE values, specifically methods *MI*, *BPCA*, *KNN*, *MED* and *M*, while methods *REP*, *S* and *HM* performed poorly (Figure 2.5; Appendix A Figure A12, Table A16). This trend was observed both when the missing data was introduced completely at random (MCAR) as well as for the more realistic case of introduced values not at random (MNAR). For the case of

MCAR, is it expected that *M* and *MED* yield good results since being a least squares estimator method they give the best approximation when no information on the missing data is available (when averaged over many runs). This hints at the challenge of deriving robust conclusions as to which imputation method is optimal for a given nature and distribution of missing values, discussed below.

Next the eight algorithms were assessed in terms of their impact on univariate data analysis, evaluated using the area under the receiver-operating characteristics (ROC) curve. Specifically, for each run (repeated 100 times), data were deliberately removed from the complete dataset (MCAR and MNAR) and then these 'missing values' were estimated via each of the eight methods, and significantly changed peaks at various statistical significance levels were identified and compared with the ones identified for the original complete datasets (with no missing values) using an ROC curve. The area under the ROC curve (AUC) was averaged across 100 runs, with the best method (i.e. closest to the complete dataset) having the highest AUC value (up to 1). The highest performance methods were similar to the best estimation algorithms from the NRMSE assessment, i.e. the largest AUC was observed for the *M*, *MED* and *KNN* methods, and the smallest for *REP*, *S* and *HM* methods (Figure 2.5; Appendix A Figure A12 and Table A17) with no differences being detected between introducing missing data as MCAR or MNAR.

To provide a framework for the interpretation of the multivariate data analysis, we first conducted a PCA on each of the four 'complete' datasets, and then tested the significance of any group separation by calculating p values (t test or ANOVA) for the PC scores along both PC1 and PC2 (Appendix A Figure A13). Similar to the ROC assessment, for missing data introduced more realistically as MNAR, *M*, *MED* and *KNN* outperformed the five other estimation methods by

revealing the known significant separation between the biological groups in PCA scores plots. Specifically a significant separation along PC1 for the DM and HL datasets was revealed after

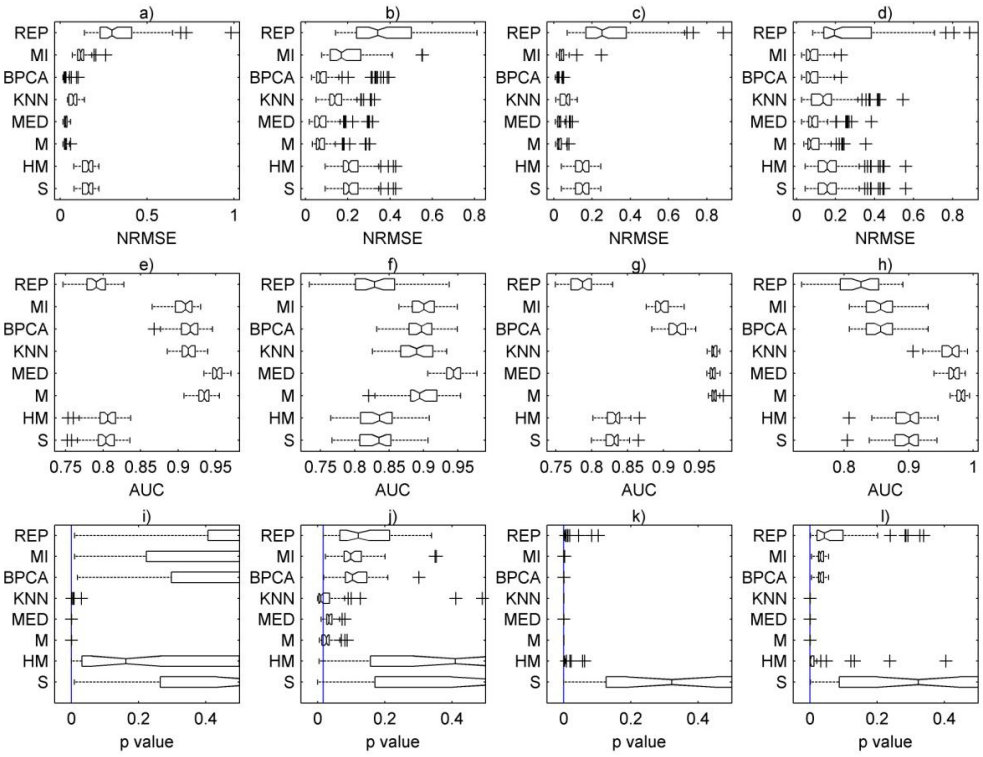

Figure 2.5 Analyses of four DI FT-ICR MS datasets after first introducing and then estimating missing data in the 'complete' datasets as MNAR (average of 100 runs). Boxplots of NRMSE values for the a) CCL_n, b) CCL_p, c) DM and d) HL datasets; boxplots of area under ROC curves (AUC) for e) CCL_n, f) CCL_p, g) DM and h) HL datasets; and distribution of p values (ANOVA or t test on PC scores) for i) CCL_n (PC2 axis), j) CCL_p (PC2 axis), k) DM (PC1 axis) and l) HL (PC1 axis) datasets, where the vertical lines indicate the p values for the complete datasets and therefore represent the ideal result following missing value estimation.

imputing missing entries with the majority of the eight methods (except for method *S* for the DM dataset, and *S* and *REP* for the HL dataset), but there were few false positive errors along PC2 after *M* and *MED* imputation (DM dataset) and *BPCA* and *MI* imputation (HL dataset). Comparison to the MCAR shows that multivariate data analysis is more prone to errors and miscalculations than for the univariate data analysis. This is further supported by the similarities of the top 5% of peaks contributing to the separation of biological groups (along PC1 or PC2), for which the similarity values are quite different between MCAR and MNAR (Appendix A Figures

A14-A15, Table A20). Overall this supports our earlier findings that multivariate data analysis is more sensitive than univariate analysis to the occurrence of missing values. Furthermore we have shown that the *M*, *MED* and *KNN* estimation methods appear to outperform the others, although see the discussion below.

2.4 Missing data estimation: is there an optimal method?

Based upon the studies to date on missing data, described here and conducted in other 'omics' fields, there are currently no grounds to prescribe any one estimation method for dealing with missing entries in metabolomic datasets. However, based upon our findings, it is clearly of considerable importance to address the question of what is the optimal treatment of missing data. For example, simply deleting the variables that contain missing data or, as we have shown, estimating those values with an arbitrarily selected method will likely introduce a large bias to the dataset and significantly affect further data analysis and interpretation. The first step in selecting an appropriate estimation method should be focused on characterising the nature of the missing values within the given dataset. Typically a metabolomics study will have measured thousands of peaks and, as presented here, one should try to infer the relationships between the missing data and the non-missing data. In addition, one should try to establish whether missing data occur as a result of metabolite abundances being below the detection limit of the analytical platform (thought to be the primary case for our FT-ICR MS datasets) or instead if they represent non-detects (i.e. metabolites not measured due to a failure of the analytical platform). With this information, as well as with the assessment of whether and to what extent missing data influence a particular data analysis method, an appropriate missing data estimation technique can be chosen, as discussed below; this three-stage approach is outlined in Figure 2.2.

In the case of the three DI FT-ICR MS based metabolomic datasets investigated here our initial findings suggest that the preferred methods of estimating missing values are *KNN*, *M* and *MED*. These three methods achieved a good balance between enabling the statistical analyses to reveal the expected metabolic differences between biological groups (Figure 2.4 and Appendix A Figures A5-A7) and yet did not identify too many potentially false positive biomarkers, i.e. the significantly changing peaks (univariate; Appendix A Table A1) or those peaks contributing towards separation in PCA space (multivariate; Appendix A Table A6) generally did not contain the highest number of missing data when compared to other estimation methods. The *KNN*, *M* and *MED* also performed the best when assessed using the 'complete' datasets with deliberately introduced missing values that were MNAR. These three methods yielded low NRMSE values, high AUC values associated with the univariate analyses, and impressively low p values associated with the separation of samples in multivariate space (Figure 2.5).

We then sought to further compare these three methods, *M*, *MED* and *KNN*, to find the optimal approach. Although *KNN* was slightly outperformed by *M* and *MED* for the NRMSE and univariate analysis (using the 'complete' datasets with deliberately introduced missing values as MNAR) we hypothesised that the *M* and *MED* methods may introduce a larger bias into the datasets due to the way that they estimate missing values; i.e. it is likely that by calculating the mean or median of the non-missing measured values, estimated peak abundances would take on large values, possibly with the majority of peaks being estimated above the SNR threshold. We therefore evaluated the original datasets (from part B of Figure 1) and confirmed that this is indeed the case. Specifically, for both the *M* and *MED* methods, almost all peaks (ca. 71-95%) were predicted to have intensities above the threshold (for the CCL_n, CCL_p, DM and HL datasets), as opposed to *KNN* for which there was an almost equal split of estimated intensities

below and above the SNR threshold (ca. 38-66% of peaks above the threshold; Appendix A Table A21). Considering the proven relationship in our FT-ICR MS datasets between peak abundances and the number of missing values, it is logical to expect that many, if not the majority, of missing values should lie on or below the SNR threshold. Therefore, based upon our analyses, we conclude that the *KNN* method imputes the most realistic values in these datasets and therefore is the preferred method over *M* and *MED*. This conclusion is strongly backed by existing literature which states that imputing mean values is not a good approach (Little and Rubin 2002, Buuren and Groothuis-Oudshoorn 2010). Furthermore, both the *M* and *MED* methods are disadvantaged by the fact that they can cause an artificial reduction of variance. Overall, we therefore consider *KNN* to be the optimal missing value imputation method for the datasets examined here.

Generalising our findings and drawing upon previously published relevant literature, we recommend some pragmatic guidelines for the applied metabolomics researcher to decide upon which missing data estimation algorithms to use: a) assess whether there is a need to impute missing data (especially if univariate analysis are to be conducted) as any imputation method will potentially bias the data analysis; b) avoid replacing missing data with small arbitrarily chosen numbers (as in *S, HM* and *REP* methods) since this greatly affects the data analysis; only consider these methods when the number of missing data is low and the majority of missing entries are known to result from measurements below the limit of detection; c) if the origin of the missing data is largely unknown or the majority of missing entries are non-detects rather than measurements below the limit of detection, select an estimation algorithm that is based on searching for local or global similarities such as *KNN* (that has been shown to be optimal in this study); d) consider *MED* or *M* imputation only when the missing data represent true non-detects

as opposed to measurements below the limit of detection; and e) ideally evaluate the chosen method against alternative algorithms to avoid obviously biased missing data imputation. This last point need not be limited to the eight imputation approaches presented here, but could possibly include other Multiple Imputation methods or the Expectation-Maximisation algorithm as used in other fields (Schafer 1999, Little and Rubin 2002). Finally, the development of novel imputation methods remains an active field, for example the LinCmb (Jörnsten, Wang et al. 2005) algorithm for microarray expression profiles that adapts to the structure of the data by changing emphasis on the local and global imputation methods, and hence the metabolomics researcher should be aware of on-going progress in this field to help guide their selection in the future.

2.5 Concluding remarks

We have shown that missing values play an important role in DI FT-ICR MS based metabolomics data, and that their estimation is very strongly reflected in the final data analysis outcome, for both univariate and multivariate approaches. Therefore, we conclude that the optimal treatment of missing data is a crucial step in the data processing pipeline to which special attention should be paid. Even though this study is based on three DI FT-ICR MS based metabolomic datasets, our analyses and findings are more generally applicable and of interest to all metabolomics studies. We propose a three step process in order to determine an optimal method for missing value estimation for a given dataset and analytical platform (summarised in Figure 2.2), that includes: assessing the nature of the missing data, analysing the impact of missing data treatments on the final data analysis outcome, and analysing the performance of missing data algorithms on the 'complete' datasets if available. Using this three step approach,

we conclude that the optimal missing data estimation technique for DI FT-ICR MS based metabolomics is the *KNN* method.

CHAPTER 3

Missing Data – Survival Analysis Approach

3.1 Introduction

Chapter 2 discussed the problems of missing data in the DI FT-ICR MS based metabolomics datasets. Various categories of handling missing data (deletion methods, imputation methods, model-based as well as machine learning procedures) were outlined, with the imputation methods being identified as the practical, convenient and potentially robust approaches in the metabolomics data processing pipeline. However, none of the examined in Chapter 4 methods or other known to the author missing data handling approaches was developed to specifically fit the purposes of the DI FT-ICR MS datasets. It appears that not the entire information is being used to infer the properties of the missing data, in particular the knowledge about the noisy peaks seems to be disregarded. Similar to two-colour cDNA microarray data where the analysis involves subtracting background values from foreground (Schützenmeister and Piepho 2010), in DI FT-ICR MS based metabolomics one has an indication of the peaks that fall below the applied signal-to-noise threshold or below the limits of detection of the mass analyzer (noise filtering methods described in Chapter 1) as well as the amount of missing data and the intensity of the non-missing data relationship found and discussed in Chapter 2. Treating missing data without taking into account this information seems like a potential disadvantage and therefore methods of the statistical analysis of the censored data are considered. For these data, it is known that the value of the missing data is beyond/below certain value or in a given range (depending on the censoring data, e.g. patients survived beyond 100 days following surgery or water zinc concentration fall below the limit of detection of a given measurement technology) and appropriate statistical analysis branch (survival analysis) has been developed to take into account such properties to provide unbiased estimates of parameters and statistics. Here, an attempt is being made to bring the methods of survival analysis for the right censored data to the data

analysis (univariate) of the DI FT-ICR MS based metabolomics datasets, as an alternative way to handling missing data that draws upon the available knowledge of the spectra noise.

3.2 Introduction to survival analysis

Survival analysis is a statistical approach designed to analyse the data in the form of times from a well defined *time origin* until the occurrence of some particular event or *time point* (Collett 2003). Originally, the approach was invented to study mortality tables from centuries ago with *death* being the event of focus (interest), hence the term 'survival analysis'. Today, the main use of survival analysis remains in, but is not limited to, the medical and allied areas of research. Other applications include to economics, engineering or geography, with *survival time* being a generic term and possibly referring to time until stockmarket crashes, time until equipment failure or time until an earthquake and so on (Machin, Cheung et al. 2006).

Given an example of a study for 10 patients undergoing a surgery for malignant melanoma, one expects that patients are not all recruited at exactly the same time, but accrue over a period of months or even years (Figure 3.1). After recruitment, patients are followed until they die (event of interest), or until a point in the calendar that marks the end of the study when the data analysis takes place (Collett 2003). If the event (death) occurred in all patients prior the end of the study, the true times to event would be known and therefore, several methods of statistical analysis would be applicable to analyse the data, e.g. to compare the mean survival times between men and women. However, the opposite is usual true, i.e. at the end of the study time the event of interest was not observed for some of the patients, with some of them simply dropping out of the trial, dying from other causes (here patients #1, #3 and #6, Figure 3.1), or still alive at the end of the study (here patients #9 and #10) (Clark, Bradburn et al. 2003).

Survival analysis data are not amendable to standard statistical procedures used in data analysis due to two main reasons. Firstly, there are usually not normally distributed, but are rather positively skewed due to '*many early events and relatively few late ones'* (Clark, Bradburn et al. 2003). Secondly, the survival times are frequently *censored* as the true (exact) time to event is not known. Each patient entering a study at time t_0 dies at time t_0+t, with t being the time to event that ideally would be known. For the censored data (here patients #1, #3, #6, #9 and #10) the t is not available, but the censored survival time is, i.e. the time for which the individual was last known to be alive at time t_0+c, with c denoting maximum observed survival time. For the malignant melanoma example, one can speak of *right censored* data since the censoring occurs after the patient has entered into a study, that is, to the right of the last known (observed) survival

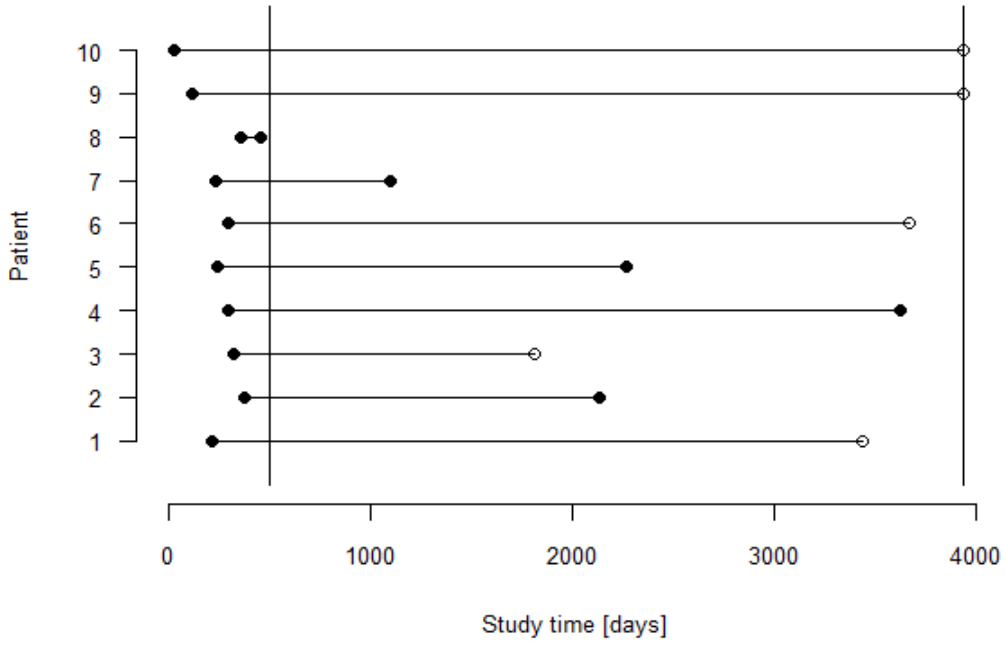

Figure 3.1 Study time (survival time) for ten patients following a surgery for malignant melanoma. The vertical line to the left denotes the end of the recruitment period and the vertical line to the right denotes the end of study. For some of the patients (#1, #3, #6, #9 and #10) the exact survival times (true time to event) are not known, since the patients are lost from the follow-up or have died from a cause other than malignant melanoma (#1, #3 and #6) or are still alive at the end of the study (#9 and #10). Circle markers indicate right censored data, for which the true time to event (i.e. death) is not known.

time with the right-censored time being smaller than the true, not known, survival time (Collett 2003, Dupont 2009). Other types of censoring are also possible, including *left* and *interval censoring*. Considering a study '*investigating the time to recurrence of a cancer following surgical removal of the primary tumour*', the left censoring data could arise in the case for those patients being examined at 3 months and already having a recurrence of cancer (i.e. the event has occurred, but the starting point is not known, here being anytime within the 3 months period). Having a second check-up point, e.g. at 6 months, can yield interval censoring data, for those patients that have had recurrence of cancer during the first and second examination period (Clark, Bradburn et al. 2003). Most survival times are, however, right censored and the majority of statistical methods were developed to take these into account.

3.1.1 Survival function and Kaplan-Meier survival estimate

The survival function plays, next to a related hazard function, an important role in summarizing and modelling survival data. The survival probability (survival function) *S(t)* is the probability that an individual survives from a given time (e.g. surgery for malignant melanoma) to a specified time in the future (Clark, Bradburn et al. 2003). The actual survival time of an individual, *t,* can be regarded as the value of a variable *T*, which can take any non-negative value with different values having a probability distribution and *T* being a random variable associated with the survival time. Given that random variable *T* has a probability distribution with underlying probability density function *f(t),* the distribution function of *T* is as in equation Eq. 3.1 and represents the probability that the survival time is less than some value *t*.

$$F(t) = P(T < t) = \int_0^t f(u)du \qquad \text{Eq. 3.1}$$

The survival function, *S(t)*, can now be defined as a probability that the survival time is greater than or equal to *t* as in Eq. 3.2

$$S(t) = P(T \geq t) = 1 - F(t) \qquad \text{Eq. 3.2}$$

and used to represent the probability that the individual survives from a given time of interest to a time beyond *t*.

Estimating the survival function is '*an initial step in the analysis of survival data*' (Clark, Bradburn et al. 2003), allowing one to present numerical and graphical summaries of the survival times for individuals in different groups (e.g. the difference between survival times between men and women following surgery for malignant melanoma) (Collett 2003). The Kaplan-Meier (K-M, or product-limit) method allows one to obtain this estimation nonparametrically from the observed censored and uncensored survival times (Kaplan and Meier 1958). Assuming that for the *k* out of *n* patients that have entered into the study the events occur independently during the study time at distinct times, such as in Eq. 3.3

$$t_1 < t_2 < t_3 < t_4 < t_5 < \ldots < t_k \qquad \text{Eq. 3.3}$$

The probabilities of surviving from one interval to another may be multiplied together to yield a cumulative survival probability, more precisely the '*probability $S(t_j)$ of being alive at time t_j is calculated from $S(t_{j-1})$ the probability of being alive at t_{j-1}'* (Clark, Bradburn et al. 2003) as in Eq. 3.4

$$S(t_j) = S(t_{j-1})\left(1 - \frac{d_j}{n_j}\right) \qquad \text{Eq. 3.4}$$

where n_j denotes '*the number of patients alive just before t_j and d_j [denotes] the number of events at t_j'* (Clark, Bradburn et al. 2003). K-M estimator '*allows each patient to contribute information*

to the calculations' as long as no event is observed for them, and simply reduces to the ratio of the number of individuals events free (for which the event has not yet happened) at time *t* divided by the number of people in the study when there is no censored data present (Clark, Bradburn et al. 2003). Graphically, the K-M estimate of the survival curve is plotted as a step function, starting from 1 (100% of patients alive at t_0), dropping instantaneously at each time of death (event) to a new level, progressively declining towards 0 as in Figure 3.2 (Machin, Cheung et al. 2006).

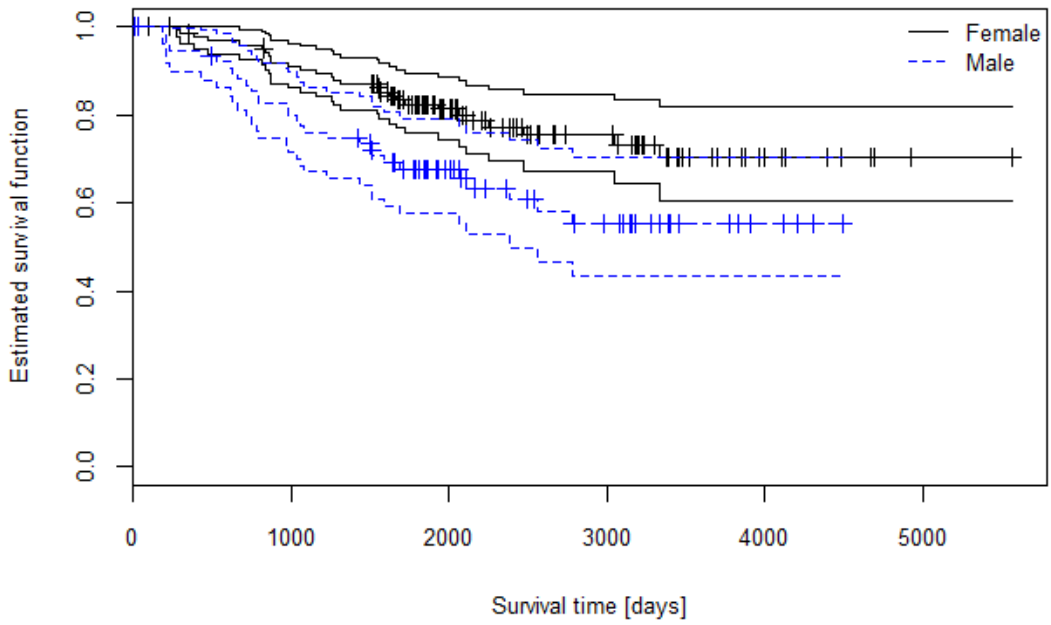

Figure 3.2 Example of the estimated survival function (vertical bars indicate censored data) with confidence intervals for the survival times of women (in black) and men (in blue) following surgery for malignant melanoma

3.1.2 Comparison of survival curves

The estimated survival functions, which allow us to obtain numerical and graphical summaries of the censored data, can be informally used when comparing two groups of patients (e.g. survival times of men and women following the surgery for malignant melanoma). More formal procedures do however exist, e.g. the *log-rank test* which is the most widely used and robust

method of comparing two or more survival curves (Peto, Pike et al. 1977, Clark, Bradburn et al. 2003).

Considering two groups, Group I and Group II having survival functions of $S_1(t)$ and $S_2(t)$ respectively, a log-rank test tests the null hypothesis (Eq. 3.5) of equal survivorship for all *r* distinct death times such as $t_{(1)}<t_{(2)}<\ldots<t_{(r)}$.

$$H_0: S_1(t) = S_2(t) \text{ for all t} \qquad \text{Eq. 3.5}$$

Letting n_{1j} and n_{2j} denote the number of patients at risk of death in the Group I and Group II respectively at the time $t_{(j)}$, there are $d_j = d_{1j} + d_{2j}$ deaths in total out of all $n_j = n_{1j} + n_{2j}$ patients at risk. The observed death rate at the time $t_{(j)}$ is then d_j/n_j and if the null hypothesis is true, the expected number of deaths among patients in Group I given that d_j deaths occurred in both groups is as in Eq. 3.6

$$E[d_{1j} | d_j] = n_{1j}(d_j/n_j) \qquad \text{Eq. 3.6}$$

The greater the difference between d_{1j} and $E[d_{1j} \mid d_j]$, the greater the evidence that the null hypothesis can be rejected (Dupont 2009). To assess the extent of this difference, a 2 x 2 table can be constructed for each *r* distinct death time of the number of patients who die or survive in the two groups (Table 3.1) (Collett 2003) and the overall measure of the deviation of the observed values of d_{1j} from their expected values can be then calculated as a sum of the differences d_{1j}-e_{1j} over the total number of death times *r* for the two groups, where d_{ij} denotes number of deaths in Group I and e_{1j} number of expected deaths in Group I given that d_j deaths occurred in both groups.

Table 3.1 Number of deaths (events) at the j^{th} distinct death time in each of two groups of patients

	Number of deaths at $t_{(j)}$	Number surviving beyond $t_{(j)}$	Number at risk just before $t_{(j)}$
Group I	d_{1j}	n_{1j}-d_{1j}	n_{1j}
Group II	d_{2j}	n_{2j}-d_{2j}	n_{2j}
Total	d_j	n_j-d_j	n_j

The resulting statistic is given by Eq. 3.7 with the corresponding variance shown in Eq. 3.8

$$U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j}) \qquad \text{Eq. 3.7}$$

$$\text{var}(U_L) = \sum_{j=1}^{r} v_{1j} = V_L \qquad \text{Eq. 3.8}$$

where

$$v_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \qquad \text{Eq. 3.9}$$

And finally we can write Eq. 3.10

$$\frac{U_L^2}{V_L} \sim \chi_1^2 \qquad \text{Eq. 3.10}$$

with a log-rank test statistic having a chi-squared distribution with one degree of freedom. Analogously, log-rank test can be extended to enable three or more groups of survival data to be compared (Collett 2003) with U-statistics shown in Eq. 3.11 and for comparing the observed number of deaths in groups 1,2,..., g-1 with their expected values

$$U_{Lk} = \sum_{j=1}^{r} \left(d_{1j} - \frac{n_{kj} d_j}{n_j} \right) \qquad \text{Eq. 3.11}$$

for k=1,2,…,g-1.

3.3 Survival analysis and the univariate analysis of FT-ICR MS based metabolomics spectra

As discussed in previous chapters, an important step in the data processing of the metabolomics data obtained via DI FT-ICR mass spectrometry is the treatment of missing data. In this chapter I address whether a survival analysis approach, i.e. treating the missing peak abundance data as left censored (being below the limit of detection of the MS analyser and/or below the applied signal-to-noise threshold) may offer an alternative strategy for finding statistically changed between biological groups metabolic traits via univariate testing (log-rank test) and without the need of using any of the missing data estimation algorithms. This approach could be of relevance since it was shown in Chapter 2 that even though the missing data may occur for technical and biological reasons, there was a strong association between the amount of missing data and the signal intensity, i.e. strongly suggesting that the majority of missing data are in fact left censored, being below the applied known signal-to-noise threshold. Furthermore, this approach is of interest as it incorporates information about the signal-to-noise threshold and/or the detection limits of the MS analyser, as oppose to treating the missing metabolites as entries with no information available.

Survival analysis methods, described earlier for right censored data and including the K-M estimator of survival function and the log-rank test for comparing the survivorships between groups can easily be applicable to left censored data. Suppose that *n* peaks (metabolic traits) within a given sample (e.g. blood or tissue) are analysed using DI FT-ICR mass spectrometry as described in Chapter 1, then a single observation (peak indicative of metabolite) can be written as

$$X = \max\ (X_0, d_c) \qquad \text{Eq. 3.12}$$

With X_0 denoting the true peak height (indicative of metabolite concentration) and d_c the limit of detection (LOD) of the mass analyzer (or here identified noise threshold), then the mass spectrum containing *n* peaks can be represented as *n* pairs of random variables as in Eq. 3.13

$$\{(X_1, \sigma_1), (X_2\sigma_2), (X_3\sigma_3), \ldots (X_n\sigma_n)\} \quad \text{Eq. 3.13}$$

where X_n *d*enotes observed peak height (either true observed value or the one set at LOD and

$$\sigma_i = \begin{cases} 1\ if\ abundance\ is\ measurable \\ 0\ if\ abundance\ is\ missing \end{cases}, i = 1,2, \ldots n \quad \text{Eq. 3.14}$$

A given peak's abundance (intensity) can be further thought of as being randomly selected from an unknown distribution F(x) as in

$$F(x) = P(X \leq x) = P(-X \geq -x) = P(A - X \geq A - x) = P(Y \geq y) \quad \text{Eq. 3.15}$$

where Y=A-x, y=A-x, and A is a large positive number. Let S(y) = P(Y ≥ y), where S(y) is the survival function described earlier. Therefore it can be seen from equation Eq. 3.15 that F(x) = S(y), meaning that the distribution of the left-censored random variable *X* is equal to the survival function of the right-censored random variable *Y* as shown in Figure 3. 3 (She 1997). Consequently the estimation of the descriptive statistics (e.g. mean, standard deviation) is equivalent to the estimate of the descriptive statistics from the right-censored data by a simple 'flipping' transformation as in Eq. 3.16, in which the initial right-censored data are transformed to left-censored data by subtraction from an arbitrarily chosen large positive value. Similarly, the log-rank test can be employed to compare two or more groups (Helsel 2005).

$$Y = A - X \quad \text{Eq. 3.16}$$

There have been only a few studies in the environmental sciences that have employed this approach. Millard and Deverel studied groundwater chemistry data from two sites in California, with the data obtained characterized with multiple detection limits, showing that only the methods from the survival analysis were applicable to their data (Millard and Deverel 1988). She (She 1997) compared *'the performance of the Kaplan-Meier estimator with maximum likelihood, probability plotting and substitution methods by Monte Carlo simulations to estimate the*

descriptive statistics for the left censored water quality data'. His results showed that the Kaplan-Meier estimator performs as well as or better than the other commonly used methods (She 1997).

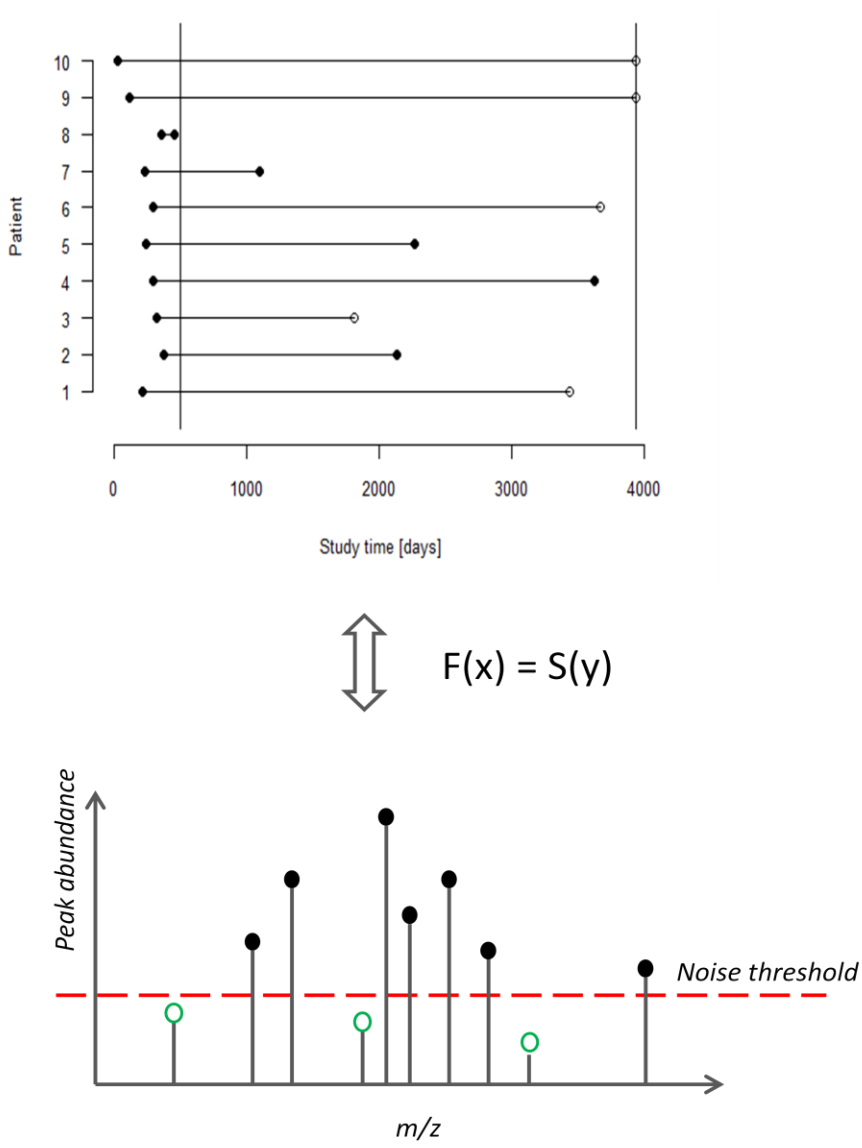

Figure 3. 3 Diagram showing the relationship between right censored medical data and the left censored DI FT-ICR MS based metabolomics data. In both datasets: truly observed data and the censored data can be identified. In medical data unavailable patients' survival data are assigned the last observed data for a given patient and the information that the patient survived beyond it; in DI FT-ICR MS the missing metabolite peak is assigned the noise threshold and the information that its true value is below it. The mathematical transformation can be applied showing the equivalence of the two types of censoring.

Here, the following questions are addressed: i) can survival analysis methods be applicable to the univariate analysis of the DI FT-ICR mass spectrometry based metabolomics data that is characterized by up to 20% amount of missing data affecting ca. 80% of all variables and ii), if so, can these methods improve the univariate analysis in comparison to the missing value estimation methods assessed in Chapter 2, in particular the best performing k-nearest neighbour imputation method?

3.4 Materials and methods

Applicability and performance of the log-rank test to the univariate analysis of the DI FT-ICR mass spectrometry based metabolomics data were tested using the original datasets of *Daphnia magna* (DM) and human liver biopsies (HL) as described in Chapter 2 (Materials and Methods). The cancer cell line (CCL_p and CCL_n) datasets were not analysed in this chapter, as it was shown earlier that due to the noise filtering strategy applied (retaining only peaks present in 50% of the samples within each biological group at the sample filter stage, based on the additional information obtained by a prior analysis of the datasets by a NMR technology) the analyses of these datasets with a consequent interpretation of the results in the context of missing data offers some additional challenges. Here, a single (more typical) DI FT-ICR MS based metabolomics experiment is assumed that cannot draw from any external biological knowledge, thus employing a standard, across all samples, sample filter.

Similarity, not all the missing data estimation methods described and discussed in Chapter 2 were used for testing the applicability and the performance of the new survival approach. For clarity of interpreting the results, the three methods were chosen based upon the results in Chapter 2. These included *S, MED* and *KNN*. Substitution with half of the minimum value found

in the non-missing data (*HM*) and substitution with mean of the non-missing values across all the samples for a given peak (*M*) were disregarded as their performance was shown to yield similar results to *S* and *MED* respectively; Bayesian PCA missing value estimation, multivariate imputation by chained equations and a modified version of Sangster's method (*REP*) were outperformed by other methods, including *M* and *KNN*.

3.4.1 Applicability of the log-rank test to the univariate analysis

As justified above, the applicability of the log-rank test to the univariate analysis of the DI FT-ICR mass spectrometry based metabolomics data was assessed using the 'original' datasets of DM and HL and in the context of the three (*S, MED* and *KNN*) missing data estimation methods. As in Chapter 2, missing data were estimated using the three imputation methods and then univariate testing (t-test or ANOVA) was employed to determine the metabolic differences between sample classes. The log-rank test was applied on a peak-by-peak basis to also test for the occurrence of significantly different peaks between the biological groups. For this approach, the missing values were assumed to be left-censored, i.e. below the applied signal-to-noise threshold (SNR), with the actual threshold values set to 3.5 times the SNR value applied in the first step on the noise filtering algorithm. SNR values were derived from the 'raw' technical replicate spectra (prior to any noise removal) using the SNR definition of '*height of signal peak in magnitude spectrum divided by the standard deviation of the white Gaussian noise in a signal free region of the real and imaginary components of the complex spectrum'* (Payne, Southam et al. 2009). Due to the SIM-stitching data acquisition method, a final mass spectrum (e.g. for a single technical replicate) was assigned a set of 21 threshold values, with a maximum threshold value for a given SIM window carried forward to a single mass spectrum representing a biological sample in the

final datasets (formed from the three technical replicates during the replicate filter stage). With the threshold values defined, the left-censored data were 'flipped' into right-censored data by subtracting them from the maximum value of all of the non-missing data found in the given (DM or HL) dataset. The log-rank test was then employed and the p values were obtained (corrected with Benjamini and Hochberg method for multiple testing (Benjamini and Hochberg 1995)), indicative of the metabolic difference between the sample classes.

The above resulted in a total of four sets of peaks, one for each of the missing data treatment methods (*S, MED* and *KNN* estimation) and one for the log-rank test on the estimated with *K-M* survival curves, with corresponding p values indicative of the significance of the metabolic differences between the biological groups. These four sets were further reduced to include only these peaks that initially contained missing values, resulting in 2317 and 1421 peaks for the DM and HL datasets respectively. Assuming a 0.05 significance level for rejecting the null hypothesis that there are no metabolic differences, the sets were compared in terms of the percentage of peaks indicative of metabolic differences between the biological groups (p value <0.05) and the percentage of missing data points in these peaks (with respect to all data points in the reduced sets). Further set diagrams (Venn diagrams created by VennDiagram R package) were used to show all the possible logical relations between the significantly changed peaks from the four sets (Chen and Boutros 2011). The peaks that were identified as significantly different based solely on the log-rank test and not by any other missing data estimation method were further investigated by comparing to the peaks identified as significantly different following missing data estimation with all three remaining methods. Here, mass-to-charge ratio, percentage of missing data initially present and the abundance of the non-missing data as well as the biological context were examined.

3.4.2 Performance of the log-rank test for the univariate analysis

To assess the performance of the log-rank test, the 'complete' DM and HL datasets were used and missing data were deliberately introduced not at random (MNAR, as described in the previous chapter). The missing data were then imputed with the three missing data estimation methods *S*, *MED* and *KNN* or were assigned a threshold value for the survival analysis approach. These threshold values were selected to reflect their characteristics in the 'original' datasets and derived in a way to retain the similar ratio as in the original datasets between the threshold values and the median values of the non-missing data for a given SIM window, separately for each sample. Results from applying t-tests to the imputed datasets, using *S, MED* and *KNN,* as well as of the log-rank test on the left-censored data were compared against the results obtained for the 'complete' datasets, with the whole procedure repeated 100 times. The percentage of added and lost peaks was calculated for each method, at the 0.05 significance level of rejecting a null hypothesis (as above). The actual errors on the obtained p values were also examined.

3.5 Results and discussion

3.5.1 Applicability of the log-rank test for the univariate analysis

The survival analysis approach yielded the highest percentage of significantly changed peaks across sample classes, resulting in 20.76% and 14.98% for the HL and DM datasets respectively (Table 3.2). These percentages were almost 4 times higher than for the *KNN* and also noticeably higher than for *S*, which yielded the second highest percentage of peaks significantly changed between biological groups. This did not however translate into the highest percentage of data points that initially were missing with approximately 23% and 14% of the data points for HL and DM datasets respectively. The high percentage of significantly changed peaks across sample

classes may indicate that the survival analysis approach does not diminish the power of detecting potential biomarkers. The comparable to other methods percentage of data points that were initially missing may suggest that the identified as significantly changed peaks represent true metabolic difference between biological groups.

Table 3.2 Summary of the significantly changed peaks, between biological groups, for the three datasets treated with missing value estimation methods (S, MED, KNN) and for the untreated dataset that was analysed using the log-rank test

	HL				DM			
	S	MED	KNN	K-M	S	MED	KNN	K-M
SP [%]	15.83	0.14	5.21	20.76	11.83	1.73	4.27	14.98
SDP with MV [%]	22.60	7.14	21.43	23.05	14.16	16.25	14.55	13.92

SP, percentage of significantly changed peaks with respect to all, initially containing missing data peaks (1421 and 2317 peaks for HL and DM datasets respectively); **SDP with MV**, percentage of missing data points in the significantly changed peaks with respect to all data points in the significantly changed peaks

The Venn diagrams of the four sets revealed further two interesting observations, mainly that the peaks identified as significantly different via a survival analysis approach can be split into two subsets (marked in yellow and green on Figure 3.4): one containing peaks that are also identified as significant via other methods (in yellow) and the second one including peaks that are significant only for the survival analysis approach (i.e. not indentified by any of the three remaining missing data estimation methods, in green) (Figure 3.4). The majority of the peaks in the survival analysis set occur within the first subset, with approximately 70% and 75% for the HL and DM datasets respectively. In particular there was a large overlap between survival analysis peaks and the *S* and *KNN* methods for both of the analysed datasets: out of 295 peaks in the survival analysis set, 66.7% and 16.9% overlapped with *S* and *KNN* respectively for the HL dataset. Similarly for the DM dataset: out of 347 peaks 74.90% and 20.17% overlapped with *S* and *KNN*. The second subset (in green) included 98 and 86 peaks for the HL and DM datasets respectively (approx. 30% and 25% with respect to all the peaks in the survival analysis set). The similar classification of the peaks identified as significantly different via the three remaining

methods would not yield equally high percentage of method specific peaks, with only 4.88%, 9.46% and 0% peaks specific for *S*, *KNN* and *MED* respectively for HL dataset and 4.37%, 19.2% and 0% for *S*, *KNN* and *MED* respectively for the DM dataset.

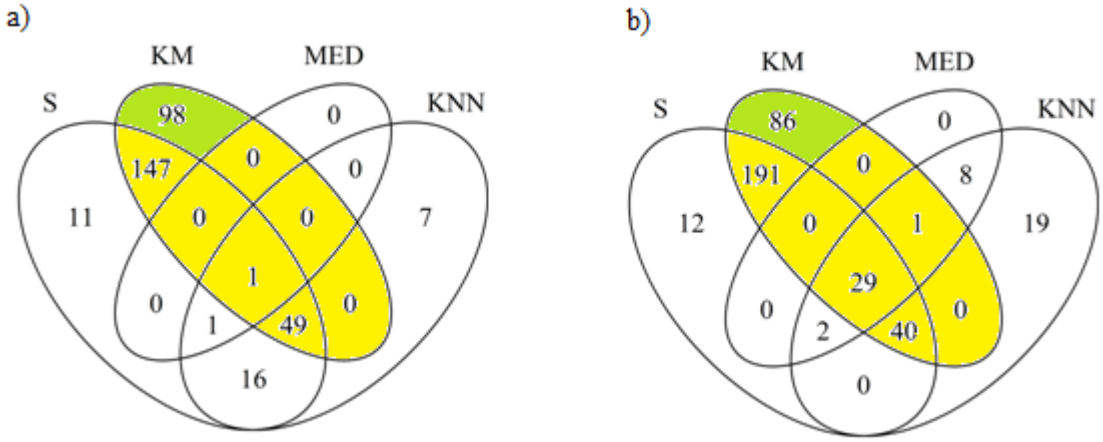

Figure 3.4 Venn diagrams for the a) HL and b) DM datasets, showing the overlap between significantly changed peaks following the missing data estimation with S, MED and KNN and the application of the log-rank test

The above analysis based on the logical relations between the four sets shows that while the survival analysis approach is comparable to the commonly used missing data estimation methods (majority of the peaks identified as significantly different via this method can be also identified via some or all of the remaining methods) and identifies additional significant peaks, informing biomarker discovery. To further understand the latter, those peaks identified solely by survival analysis were investigated by comparing against the peaks that were identified as significantly different following missing data estimation with all three remaining methods (49 and 29 peaks for the HL and DM dataset respectively, excluding *MED* subset for the HL dataset as it contained only two peaks). The identified peaks that are common to the various missing data estimation methods may be regarded as being less sensitive to missing data imputation approaches, having a higher probability of representing true differences in metabolic traits across samples classes and

therefore providing a good reference group of peaks for the validation of new methods (i.e. here, the survival analysis approach); hereafter referred to as "reference peaks".

Analyses of the m/z ratios and intensities of the non-missing data have confirmed that those peaks specific to the survival analysis do not differ in their properties from the reference peaks. The distributions of the mass-to-charge ratios were similar to the corresponding distributions for the reference peaks and the distributions of the number of data initially missing (per peak across all the sample classes) resembled even closer uniform distribution than the one for the reference peaks (Figure 3.5). The former may be interpreted as the specific to the survival analysis approach peaks are not affected (or caused) by the association found and discussed earlier between *m/z* ratio and the amount of missing data (Chapter 2, Occurrence and distribution patterns of missing data). The latter shows (but should be interpreted with caution due to a relatively small group size) that these peaks are not dominated by the 'extreme' cases of the missing data (cases of missing data prevailing in one of the biological groups), but rather they comprise of peaks that initially had missing data spread across the biological groups (data not shown). The interquartile range of the non-missing data was 156 578 and 124 922 for the specific to survival analysis approach peaks and the reference groups respectively for the HL datasets, and similar 21 487 and 24 256 for the DM datasets. These comparable interquartile ranges may indicate that the specific to the survival analysis peaks are also not affected by the second pattern found and discussed in Chapter 2, that is the more missing data present with lower the peaks intensities.

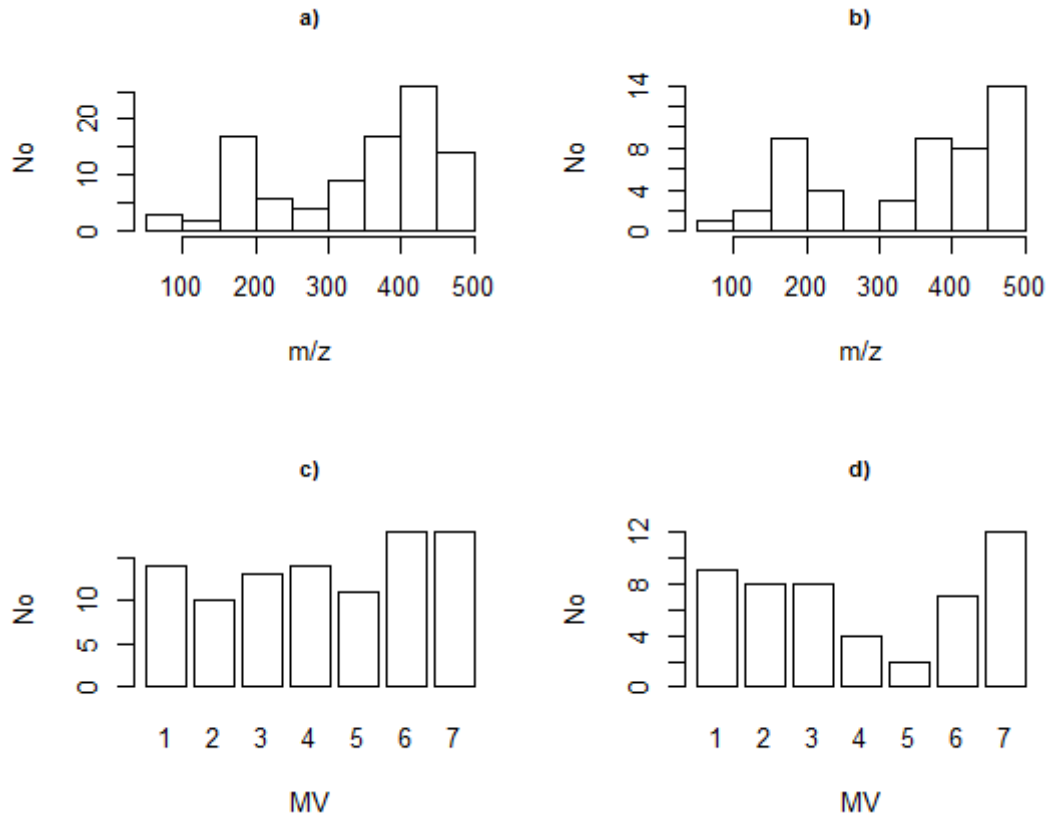

Figure 3.5 Distribution of m/z values and the number of missing data for the significantly changed peaks following the log-rank test and the estimation of the missing data with S, MED and KNN for HL dataset. Distribution of m/z values a) for the 98 peaks specific to log-rank test, b) for the 49 peaks common to the S, MED, KNN and log-rank test. Distribution of peaks having 1 to 7 missing data prior imputation for c) 98 peaks specific to log-rank test and d) 49 peaks common to the S, MED, KNN and log-rank test.

Further encouraging results were obtained when assessing specific to survival analysis peaks following assignment of putative metabolites names based upon accurate mass measurements and the KEGG database for the HL datasets. Out of 98 peaks, 32 were assigned a metabolite name(s) and interpreted within a given biological context of human orthotopic liver transplantation (OLT). This analysis revealed that several metabolites that are expected to change throughout OLT were in fact detected as changing significantly by the survival analysis approach. It is believed (as discussed in Chapter 5) that following reperfusion, liver grafts re-start their metabolic activity that can be observed as an increased urea production (including raised levels of urea cycle intermediates), an increased production of bile acids and the removal of the

compounds found in the preservation solution. The specific to survival analysis approach, in fact included, among others amino acids of alanine, proline and L-citrulline, all increased post reperfusion. The latter is particularly interesting since citrulline is synthesised from ornithine and carbamoyl phosphate in one of the central reactions in the urea cycle (Berg, Tymoczko et al. 2006). Further putative metabolite names included mannitol (part of the preservation solution; decreased post reperfusion) and glycodeoxycholate (a secondary bile acid; increased post reperfusion), all in agreement with the expected biochemical changes. Methionine sulfoxide was another relevant metabolite identified since it is '*an oxidation product of methionine with reactive oxygen species via 2-electron-dependent mechanism*', produced typically by the activated neutrophils and thus reflecting the oxidative stress (Mashima, Nakanishi-Ueda et al. 2003), that occurs in liver grafts during the reperfusion injury (Chapter 5).

Table 3.3 Selected putative metabolite names assigned to the specific to the survival analysis peaks

Metabolite name	Adduct	Fold change
Alanine	H	1.34
Proline	H	3.86
L-Citrulline	H	2.67
Mannitol	2K-H	0.53
Glycodeoxycholate	K^{39}	9.44
Methionine sulfoxide	H	2.58
Succinate	K,H	0.65
sn-Glycerol 3-phosphate	H	0.71
Adenosine 5-diphosphate	2Na-H	10.32
L-Aspartate	K,H	1.91
N-Methyl-L-glutamate	Na	2.45
Uridine monophosphate	2Na-H	2.63

3.5.2 Performance of the log-rank test for the univariate analysis

Consistent with the results above, the survival analysis approach also identified a large proportion of peaks as changing significantly in the 'complete' datasets, following introduction of missing data as MNAR. The averaged percentage (across 100 runs) of significant peaks relative

to all peaks were 19.0%, 13.8%, 13.2%, and 17.4% for *S, MED, KNN* and log-rank test respectively for the HL dataset and 25.75%, 18.15%, 19.95% and 29.9% for the DM dataset (Figure 3.6). This have yielded a low percentage of peaks lost following the survival analysis approach and the missing data estimation with *S,* with averaged across 100 runs values of 22.9%, 32.1%, 27.6% and 21.2% for the *S, MED, KNN* and log-rank test respectively for the HL dataset and 31.1%, 41.5%, 36.5% and 30.4% for the DM dataset. The average error of the added and lost peaks across the runs has indicated that the survival analysis approach yields a comparable error to the *KNN* algorithm with the overall average of 20.95%, 22.95%, 20.4% and 19.3% for the *S, MED, KNN* and log-rank test respectively for the HL dataset and the 26.9%, 29.8%, 26.0% and 25.18% for DM dataset.

Taking into account that the significance level does not always have to be chosen as 5%, the actual errors on the estimated p values (comparing p values from the complete dataset and the p values from each of the four methods) were also assessed. For each run, these were calculated as a percentage error on each peak for each of the methods of handling missing data (three missing data estimation algorithms and survival analysis approach). Pooled across 100 runs, the distributions of these errors were quite comparable between the four methods (Figure 3.7), both for all of the peaks and the peaks that in the 'complete' datasets were significantly different at the 5% significance level. With these errors split into groups based on the number of missing data initially present across samples for a given peak, it was revealed that the errors of p value estimations did not increase linearly with the increasing number of missing data across samples (Figure 3.8) for the survival analysis approach as it did for the missing data estimation method *S.*

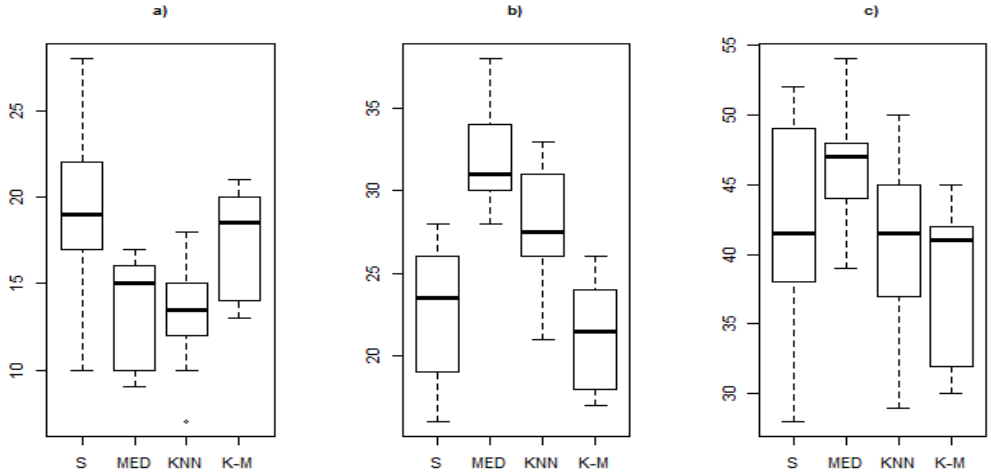

Figure 3.6 Comparison of peaks identified as significantly changed for the HL dataset following missing data estimation with S, MED and KNN and the survival analysis approach. Percentage of a) peaks added and b) peaks lost for the N simulation runs when compared to the reference peak list obtained from the analysis of the 'complete' dataset; c) average of a) and b).

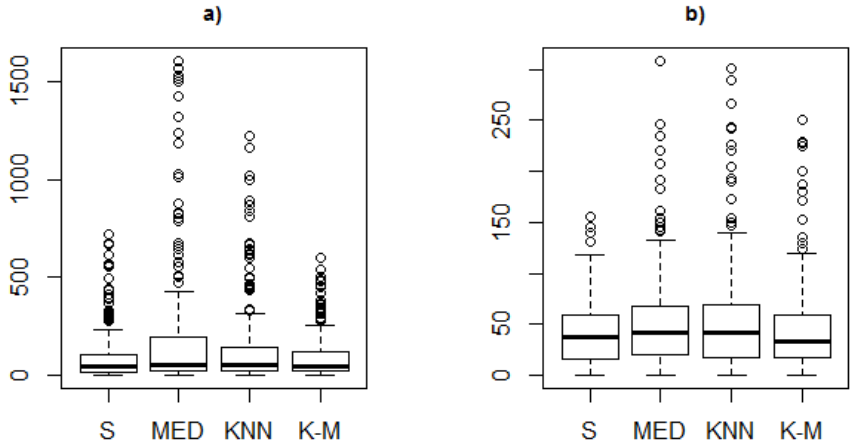

Figure 3.7 Error [%] on the p values obtained following univariate testing on the missing data estimated with S, MED, KNN and the log-rank rank on the left-censored data. HL dataset: a) all the peaks containing missing data (mean percentage of 104.5%, 206.9%, 146.2% and 101.4% for S, MED, KNN and K-M respectively), b) significantly changed peaks as in the complete dataset (mean percentage of 43.3%, 57.4%, 58.5% and 49.0% for S, MED, KNN and K-M respectively). Corresponding median values for the DM dataset (boxplots not shown) are 105.9%, 167.5%, 120.6% and 85.0% for S, MED, KNN and K-M respectively for all the incomplete peaks and 42.5%, 53.4%, 45.05%, 39.1% for the significantly changed peaks.

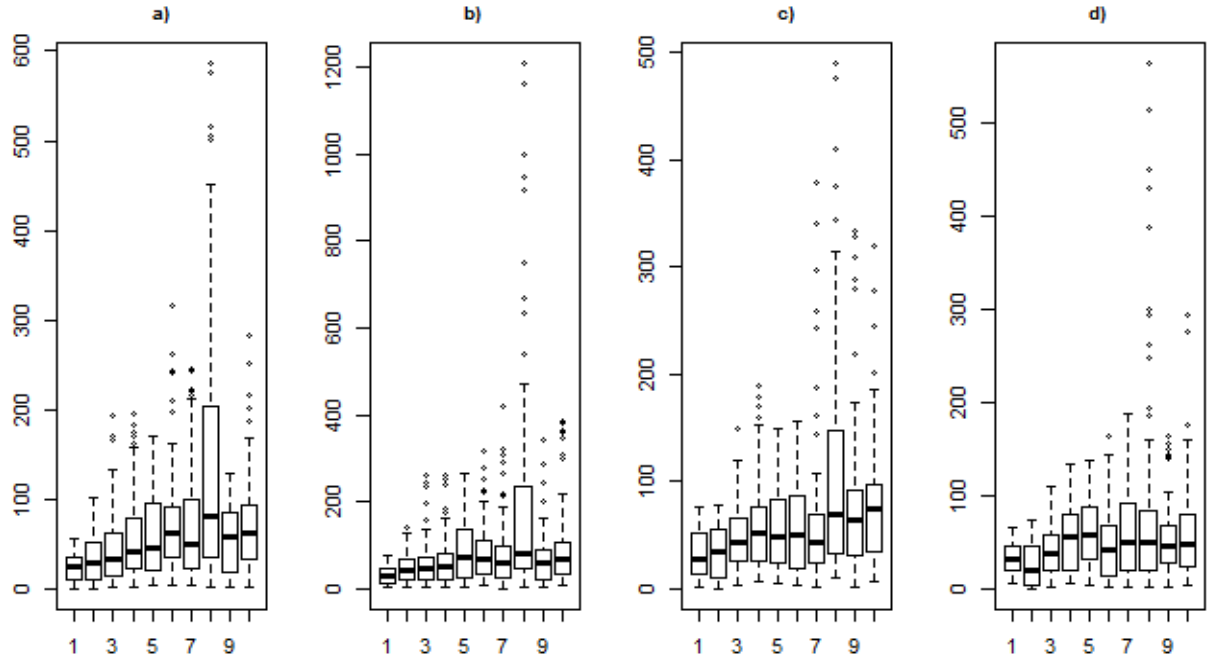

Figure 3.8 Error [%] on the p values of the significantly changed peaks for peaks containing differing amounts of missing data (i.e. from 1 to 10 missing entries). DM dataset, following a) S, b) MED, c) KNN and d) survival analysis approach

3.6 Concluding remarks

Investigations into the applicability and performance of the survival analysis approach to univariate testing of DI FT-ICR MS based metabolomics data have resulted in a set of encouraging observations that suggest that this approach should be considered when dealing with missing data. It has been shown that a simple estimation of noise values (based on the median intensity of the peaks present in the signal free regions of spectrum) used to create left censored data, followed by their transformation into right censored data and identifying peaks indicative of significance metabolic difference between biological groups via a log-rank test was feasible. This approach also enabled detecting a large number of the potentially interesting peaks. These results were obtained for both of the analysed datasets of DM and HL, with the latter one characterized by large metabolic heterogeneity. Further analyses have shown that a sub-set of the peaks

identified as significant using survival analysis approach can also be detected following missing data estimation (*S*, *MED* and *KNN).* The study of the specific to survival analysis approach peaks did not give any grounds to classify the survival analysis approach as not applicable nor as not feasible, either. On the contrary, the characteristics of these peaks such as m/z ratios, distributions of missing data and the intensities of the initially non-missing data were all in line with the selected reference peaks that are less likely to be affected by various missing data estimation approaches, yet still containing missing data. The more 'uniform' distribution of the number of peaks grouped according to the number of missing data present (Figure 3.7) suggests that this approach offers higher sensitivity to detect significantly changed peaks in 'non-extreme' cases (i.e. when missing data are spread across biological groups as oppose to are being clustered in samples from one biological group) but this should be further verified on a bigger datasets. The analyses of these peaks in terms of their biochemical context for the HL datasets also confirmed the applicability of the survival analysis approach. Unfortunately, as discussed in Chapter 2, these results cannot be solely used as an indication whether the tested approach is optimal and/or outperforming other methods (due to the inherit nature of the problem). However, combined with the results of assessing the performance of this method based on the 'complete' datasets and introducing missing data at MNAR, it has been shown that this method should not be excluded when deciding upon the treatment of missing data during the data processing stage. The strength of this method seems to lie in identifying a vast number of peaks that may represent true significant metabolic differences without the actual need to estimate the missing data. This can be especially useful in the hypothesis generating studies for which estimating missing data with the commonly used algorithms fails to provide a large enough subset of peaks indicative to true metabolic difference between biological groups for further verifications and analyses.

CHAPTER 4
Additional Advances In Data Processing And Analysis

4.1 Introduction

Many aspects of the data analysis of the metabolomics datasets can be regarded in terms of finite sets and set theory. A set is a collection of well-defined and distinct objects, which for the purpose of this chapter can be assumed either unordered or ordered. Formally, a set *S* is called finite if for some natural number *n* the one-to-one correspondence function (bijection) exists such as Eq. 4.1 and it can be written as in Eq. 4.2 (Goldrei 1996).

$$f: S \rightarrow \{1,2, \dots, n\} \quad \text{Eq. 4.1}$$

$$\text{S}=\{x_1, x_{2,\dots,} x_n\} \quad \text{Eq. 4.2}$$

A partially ordered set, also known as *poset*, formalizes the intuitive concept of an ordering of the set objects. Let *R* be a random relation on a set *S*, then if *R* satisfies the conditions of reflexivity (Eq. 4.3), antisymmetricity (Eq. 4.4) and transitivity (Eq. 4.5), it is a *partial order* relation and (*S, R)* is called a partially ordered set or a *poset* (Devlin 1993).

$$(x, x) \in R \; for \; all \; x \; \in S \quad \text{Eq. 4.3}$$

$$(x, y) \in R \; and \; (y, x) \in R \Rightarrow x = y \quad \text{Eq. 4.4}$$

$$(x, y) \in R \; and \; (y, z) \in R \; \Rightarrow \; (x, z) \in R \quad \text{Eq. 4.5}$$

Mass spectra obtained via DI FT-ICR MS based metabolomics may be considered as sets containing objects (here peaks indicative of metabolites), therefore the comparison of spectra may be viewed from the set theory perspective. Throughout the metabolomics experiment and the subsequent data analysis, one may be interested in comparing peaks presence (or absence) across two or more biological samples using the binary operations on sets such as union (pooled objects), intersections (common objects), difference (objects specific to one of the sets) and so on (Goldrei 1996). Partially ordered sets can be obtained when comparing two sets of peaks arranged from the most to the least important one according to a specific criterion, e.g. sorted

according to their p values (following univariate testing) or their loadings values (following principal component analysis), for example to compare results between two experiments of interest, e.g. *Daphnia magna* exposed to drugs of different mode-of-action.

The comparison of sets as well as the visualization of these results is a further important part of biology and 'omics' experiments, since often the integration of multiple datasets (sets) is required to gain a full understanding of the underlying biological mechanisms (Chen and Boutros 2011). For the former, there are currently no methods (to the author's knowledge) available that would allow obtaining simple metrics of the (dis)similarity of the two partially ordered sets, here two peak lists that typically would arise when sorting peaks based on their p or loadings values following univariate and multivariate data analysis respectively. Ideally, a simple measure should capture both the amount of the overlapping peaks as well as their position in the two sets. For the latter, visualization methods used include textual tables, network diagrams, heatmaps and the Venn and Euler diagrams, with the Venn diagrams employing circle and ellipses to display all n^2-n+1 possibilities created by the interaction of *n* sets (as on Figure 3.4 in Chapter 3) and being highly popular due to their familiarity and simplicity (Chen and Boutros 2011). Venn himself has shown that his diagrams can be constructed for any number of sets, however the comparison of higher number of sets (n>4) requires adding non-congruent and/or non-convex shapes. A symmetry no longer can be achieved which deprives the Venn diagrams their aesthetics and simplicity of comprehension (Venn 1971). To overcome some of the above issues, Edwards has proposed creating the Venn diagrams (Edward's Venn diagrams) by segmenting the surface of the sphere (Figure 4.1). However, regardless of the method chosen, the more number of sets are being compared, the harder it is to interpret results, especially when different shapes are being introduced to allow for all the interacting possibilities (Edwards 1989).

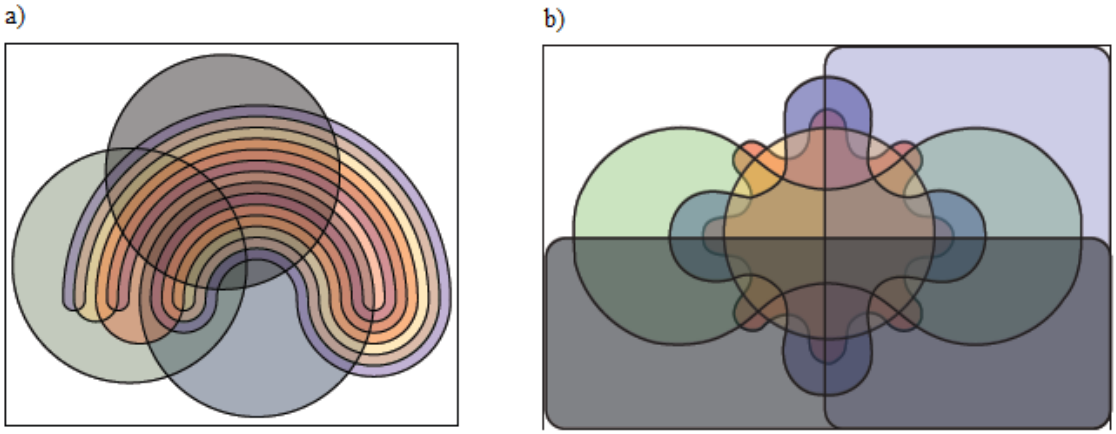

Figure 4.1 Diagrams showing interactions among six sets a) Venn diagram, b) Edward's Venn diagram

While working on the primary objectives of this thesis, two methods (tools) were developed that may be of benefit to 'omics' studies when handling results that may be regarded as sets. The first one is a metric reflecting the similarity of two ordered peak lists (*posets*), taking into account both number of the shared between two lists peaks and their positions that can be assigned based on any of the chosen criteria (e.g. derived from data analysis). The second one is a visualisation tool that identifies and displays the amount of shared peaks between the lists (sets) via a series of colour-coded horizontal bars.

4.2 Comparing ordered sets

Initially, the idea of developing a metric to quantify the (dis)similarity between the two ordered sets arose while assessing the various missing data estimation strategies for DI FT-ICR mass spectrometry data. As discussed in Chapter 2, in order to compare the results of the missing data estimation algorithms in terms of principal component analysis, the lists containing the top 5% peaks contributing towards samples separation along PC 1 or PC 2 (peaks ranked according to their loadings values) had to be considered.

4.2.1 Mathematical representation

The similarity measure, *ordered distance (ODist)* of two ordered lists was developed to capture both the number of common objects among the two sets as well as their positions. Let *A* and *B* denote sets such as Eq. 4.6 and Eq. 4.7

$$A = \{a_1, a_2, \dots a_n\} \quad \text{Eq. 4.6}$$
$$B = \{b_1, b_2, \dots b_m\} \quad \text{Eq. 4.7}$$

where the objects in the sets are ordered according to some external criterion in descending order with the first element of the highest importance (e.g. the smallest p value) and the last element of the least importance. Also, let *C* represent a set containing all the objects which are members of both *A* and *B* as in Eq. 4.8

$$C = A \cap B = \{c_1, c_2, \dots c_k\} \quad \text{Eq. 4.8}$$

where $0 < k \leq n+m$. The similarity measure, $ODist_,$ between the two sets *A* and *B* is expressed by two sub-components. The first one, $ODist_i$, takes into account solely the number of objects common between *A* and *B* as in Eq. 4.9 (also known as Jaccard index)

$$ODist_i = \frac{|A \cap B|}{|A \cup B|} \quad \text{Eq. 4.9}$$

where | | denotes cardinality (number of elements) of the set. The second, $ODist_p$ captures the order of the elements that are common to *A* and *B* as in Eq. 4.10

$$ODist_p = 1 - \frac{S_p}{S_{max}} \quad \text{Eq. 4.10}$$

where S_p is the score measuring, for each element $c \in C$ (elements common to both *A* and *B*) the sum of the difference between their positions in the two sets. Defining functions returning an index (position) of the i^{th} element of the set as in Eq. 4.11 and Eq. 4.12

$$f(a_i) = i \quad \text{Eq. 4.11}$$

$$f(a_i = c_j) = i \quad \text{Eq. 4.12}$$

the S_p can be written as in Eq. 4.13

$$S_p = \sum_{j=1}^{k} |f(a_i = c_j) - f(b_i = c_j)| \quad \text{Eq. 4.13}$$

The S_{max} is the maximum sum of the difference between the positions of the common elements, allowing the elements in A and B to permute (rearrange) to achieve the highest score. Defining functions returning the permuted index (position) of the i^{th} element of the set as in Eq. 4.14 and Eq. 4.15

$$f^*(a_i) = p, where\ \ p \neq i\ and\ p \in (1, k) \quad \text{Eq. 4.14}$$

$$f^*(a_i = c_j) = p, where\ p \neq i\ and\ p \in (1, k) \quad \text{Eq. 4.15}$$

the S_{max} can be written as in Eq. 4.16

$$S_{max} = max \bigvee_{p} \sum_{j=1}^{k} |f^*(a_i = c_j) - f^*(b_i = c_j)| \quad \text{Eq. 4.16}$$

Finally, the overall similarity measure *ODist* is expressed as in Eq. 4.17

$$ODist = c_1 * ODist_i + c_2 * ODist_p, \qquad where\ c_1 + c_2 = 1 \quad \text{Eq. 4.17}$$

and typically $c_1 = c_2 = 0.5$ for equal contributions of the two sub-components being.

The above measure can be easily expanded to measure the similarity between any number of sets with $ODist_i$ as in Eq. 4.18

$$ODist_i = \frac{|A \cap B \cap C \ldots \cap Z|}{|A \cup B \cup C \ldots \cup Z|} \quad \text{Eq. 4.18}$$

and S_{max} being a mean value of all the pair-wise comparison among every two sets as in Eq. 4.19

$$S_{max} = mean\begin{pmatrix} S_{\max AA} & S_{\max AB} & \cdots & S_{\max AZ} \\ S_{\max BA} & S_{\max BB} & \cdots & S_{\max BZ} \\ S_{maxCA} & S_{\max CB} & \cdots & S_{\max CZ} \end{pmatrix} for\ i \neq j \qquad \text{Eq. 4.19}$$

4.2.2 Computational solution and applicability to data analysis in a metabolomic experiment

The similarity measure was implemented in R language as a series of independent functions, including the calculation of $ODist_i$, the calculation of S_p and the S_{max} matrix and based on these, the function calculating of the overall $ODist_p$ and the *ODist* similarity measure. While computational coding of the majority of these calculations was straightforward, obtaining the maximum sum of the difference between the positions of the common elements S_{max} requires finding, for each pair of sets being assessed, factorial *n!* and *m!* number of permutations without replacement, where *n* and *m* are the cardinality of the two sets respectively. For the purpose of this thesis, a basic solution was employed, with a pre-defined number of permutations used and the S_{max} set the maximum value obtained. An accompanying graph was produced to assess whether the increase in the number of permutations yields an increase of the S_{max} value for the specific conditions (in particular number of peaks in the assessed lists) (Figure 4.2).

The developed ordered distance measure was initially developed for and used while addressing the missing data research aims, in particular when assessing the influence of the missing data on the multivariate data analysis. The calculated *ODist* metric for the lists containing the top 5% peaks contributing towards the separation along PC1 and PC2 following imputation of missing data using the eight missing data estimation algorithms (as described in Chapter 2) allowed to identify methods that performed in a similar manner, further strengthening the overall (univariate and multivariate) results (Table 4.1 and Figure 4.3; Appendix A: Figure A8-A9, Table A7-A8). These results are in the published paper: Hrydziuszko, O. and M. Viant

"Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline." Metabolomics: 1-14.

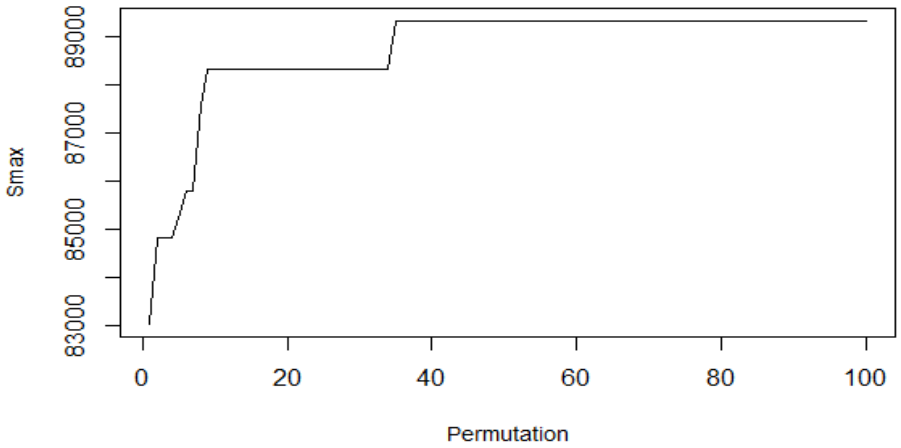

Figure 4.2 An example of the graph showing the increase of the S_{max} value throughout the N permutations; Data generated for two sets of 500 elements each.

Table 4.1 Similarity metric, *ODist*, values between the eight missing value estimation methods based on the 5% top peaks contributing towards the separation along PC1 for the four datasets: CCL_p, CCL_n, DM and HL for $c_1 = c_2 = 0.5$

		S	HM	M	MED	KNN	BPCA	MI	REP
$CCL_p \setminus CCL_n$	S	100.00	76.68	0.87	1.17	27.99	0.58	20.41	6.71
	HM	70.09	100.00	2.62	2.92	14.87	4.08	23.91	9.62
	M	4.46	16.96	100.00	97.08	0.00	21.28	3.50	51.31
	MED	7.56	13.84	84.82	100.00	0.00	20.70	3.79	51.02
	KNN	9.82	8.04	12.50	12.05	100.00	0.00	19.83	3.50
	BPCA	5.80	16.07	25.89	26.34	9.82	100.00	4.96	9.91
	MI	12.05	16.96	13.39	13.39	25.45	14.29	100.00	5.54
	REP	17.86	34.38	37.95	38.84	16.52	24.11	16.96	100.00
		S	HM	M	MED	KNN	BPCA	MI	REP
DM \ HL	S	100.00	82.38	1.90	3.33	4.76	3.33	19.05	22.86
	HM	61.11	100.00	5.24	6.67	11.43	6.19	23.33	30.95
	M	3.33	26.67	100.00	96.19	89.52	34.76	63.33	64.29
	MED	5.56	31.11	86.67	100.00	90.95	35.24	65.24	68.10
	KNN	26.67	47.78	62.22	67.78	100.00	35.24	69.52	70.48
	BPCA	2.22	10.00	23.33	23.33	21.11	100.00	33.81	32.38
	MI	22.22	42.22	52.22	55.56	65.56	21.11	100.00	70.48
	REP	34.44	65.56	50.00	55.56	67.78	21.11	62.22	100.00

Blue: data for CCL_p and DM, yellow: data for CCL_n and HL

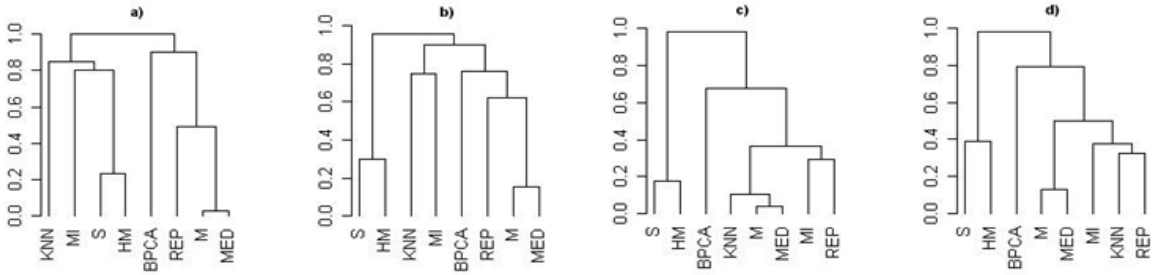

Figure 4.3 Hierarchical clustering (Euclidean distance, agglomeration method: complete) for eight different imputation methods for the top 5% of peaks contributing towards separation along PC1 for a) CCL_n, b) CCL_p, c) DM and d) HL datasets.

4.3 Visualisation tool for sets comparison

The visualisation tool to show the amount of common objects between the peak lists of interest (for all occurring possibilities) was initially developed to investigate the unexpected observation of a very high number of peaks in the blank mass spectrum - an extract blank prepared using identical methods of sample preparation but with no biological material added to the solvents. This number was comparable to the number of peaks detected for the biological samples, therefore a visualisation tool was developed to reveal whether these peaks were common to all the biological samples in this experiment or whether there were specific to blank spectrum, possibly offering some additional insight into this phenomenon.

4.3.1 Realisation

As an alternative to intersecting geometrical shapes as circles or ellipses that are used in Venn Diagrams, it was assumed that it is possible to represent the relationships between the sets via a series of horizontal bars. These bars are colour-coded to denote the amount of objects common between the sets for all of the logical possibilities (Figure 4.4). One of the sets can be marked as a reference one, with its colour-coded bars to be displayed at the very top of the diagram and a bar indicative of the amount of the specific only to this set elements (not present in the remaining

sets) drawn to the very left of the graph (here used for the extract blank) The remaining sets are being added below the reference set, with their corresponding colour-coded bars arranged in a decreasing order, e.g. the number of elements common among the four sets (reference set and the three other sets) followed by the number of elements common among the three other sets, followed by number of elements common among only two sets and finally the number of elements specific to each of the three other sets (towards the right hand side of the diagram). The size of each coloured bar represents the number of elements falling into each specific category (logical possibility, e.g. common among the three other sets) and can be estimated based on the x-axis. The exact numbers of these elements are saved in the output text file.

Representing each set as a horizontal bar, divided into colour-coded regions, enables effortless comparisons of any number of sets since adding a set requires simply appending another horizontal bar below the already present ones, therefore eliminating the necessity to introduce non-congruent shapes. Also, the visual understanding and interpretation of the results is straightforward, even when considering higher number ($n>4$) of sets.

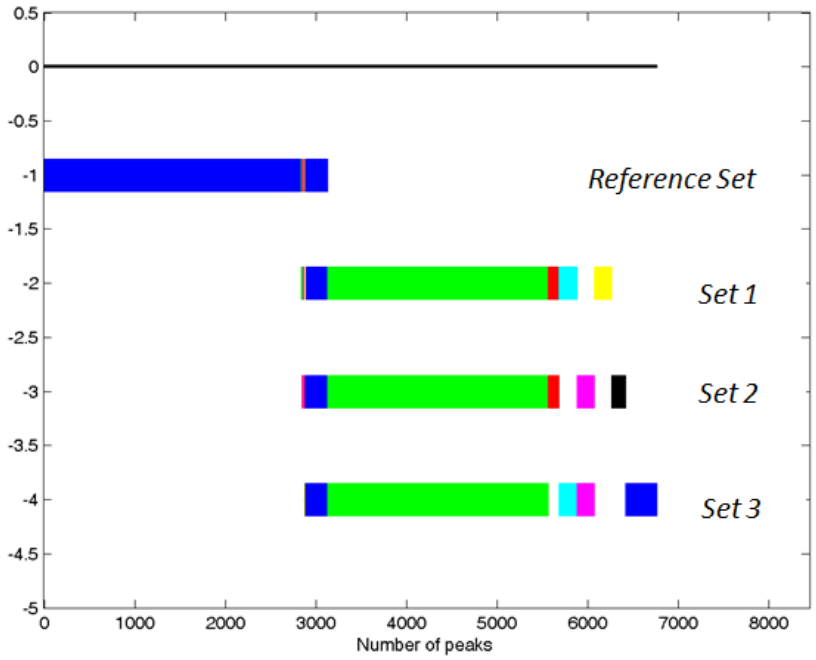

Figure 4.4 Graphical representation of all the interactions between the reference set and three other sets; CCL_p datasets used here for illustration purposes; reference set: blank; set 1: control samples; set 2: samples exposed to indomethacin; set 3: samples exposed to medroxyprogesterone acetate

4.3.2 Applicability to signal processing in a metabolomics experiment

The comparable to biological samples number of peaks detected for the extract blank was initially observed while performing one of the first environmental studies of toxicity testing in *Daphnia magna* employing the DI FT-ICR mass spectrometry metabolomics and the novel SIM-stitching algorithm for the metabolites detection at the Environmental Metabolomics Research Laboratory, University of Birmingham (Taylor, Weber et al. 2009). The extract blank contained several thousand of peaks, similar to the number of peaks in the biological samples of the analysed *Daphnia magna* dataset. This observation was unexpected, since extract blank is prepared following the same protocol as used for the preparation of the biological samples, but with no biological material added (see Chapters 1 and 2). Quite opposite, one would anticipate to detect only a small set of peaks reflecting the presence of methanol, water and formic acid or ammonium acetate for positive and negative ion mode of analysis respectively.

With the aid of the developed visualisation tool, it was shown that the majority of peaks occurring in the extract blank spectrum become suppressed as soon as biological material (metabolites) are added. It was speculated that the peaks measured for the biological samples are the ones of higher concentrations and therefore must have successfully competed for charge during the electrospray ionisation process. The peaks present in extract blank, on the other hand, are likely to include contaminants from the air and/or substances leaking from the plastic sample preparation tubes (plasticisers). Thes results highlight the analytical sensitivity of DI FT-ICR mass spectrometry. This theory is further supported when comparing the extract blank mass spectra obtained throughout various metabolomics studies, e.g. analyzing the cancer cell line, *Daphnia magna* and human liver datasets (as described Chapter 2)(Figure 4.5). Among these three quite different datasets, there were approximately 500 peaks common across the extract blank mass spectra, both for the positive and the negative ion mode of analyses. This phenomenon is a clear demonstration of the finite dynamic range of the DI FT-ICR mass spectrometry detector, which is all now being accounted for during the signal processing stage. In addition to the three stage noise filtering strategy (Chapter 1) peaks present in the extract blank spectrum are being removed from the biological samples spectra based on a user's pre-defined settings, typically these peaks are being removed if they are more intense in the extract blank spectrum than in the biological spectra. The first successful application of this tool was in the same environmental study for which the observation of the high number of peaks for the extract blank was made with the results published in Metabolomics: Taylor, N., R. Weber, et al. (2009). "A new approach to toxicity testing in Daphnia magna: application of high throughput FT-ICR mass spectrometry metabolomics." Metabolomics **5**(1): 44-58.

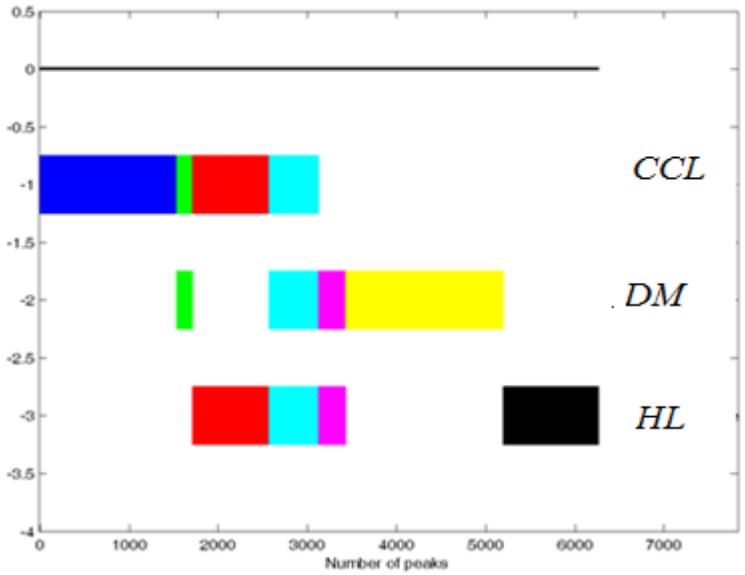

Figure 4.5 The amount of peaks for all the possibilities among the three extract blank spectra obtained for CCL_n, DM and HL datasets, positive ion mode

CHAPTER 5

Metabolomics study of Human Liver Transplantation

5.1 Introduction

The liver is the second most commonly transplanted major organ (after kidney), yet the biochemical mechanisms undergoing in the organ throughout and following surgery are still poorly understood (Seaberg, Belle et al. 1998). Metabolomics, especially mass spectrometry based approach, offers a potentially very powerful tool to gain further insight into graft metabolic activity that could in turn improve the outcome and the survival following the surgery. Here, a proof-of-principle metaboloimics study is presented based on a limited cohort of patients to address the question whether the DI FT-ICR mass spectrometry based metabolomics can be applicable to investigate metabolic process of liver grafts in a highly variable study of clinical liver transplantation. The samples collected and measured (DI FT-ICR MS) resulted in a dataset that was used both to address the outlined question as well as to aid the research on missing data as presented in Chapters 2 and 3. For the latter, it offered a biologically diverse, highly heterogeneous dataset (as measured by the coefficient of variation, Chapter 2) that complemented datasets obtained from more controlled studies (cancer cell line, *Daphnia magna* described in Chapter 2) thus enabling a thorough testing of the missing data estimation methods as well as the novel approach based on survival analysis concept. The proof-of-principle metabolomics study was carried out simultaneously to the research on missing data, and therefore the results presented and published in this chapter (Hrydziuszko, Silva et al. 2010) are based on the method that was being employed at the time at the Environmental Metabolomics Research Laboratory, University of Birmingham, that is the modified version of Sangster's method for which a missing value is substituted with the average intensity of the nearest (in term of *m/z* value) peaks from the raw measurements of the technical replicates (Sangster, Wingate et al. 2007).

5.2 Application of metabolomics to investigate the process of human liver transplantation

Orthotopic liver transplantation (OLT) is the only treatment for end-stage liver disease. In the UK, where ca. 600 liver transplants occur each year, approximately 10% of potential recipients die while on the waiting list (2007-2008 Transplant Activity in UK, National Health Service annual report). The shortage of organs available for OLT has resulted in a drive to increase donor pools by, for example, including marginal organs or those obtained by donation after cardiac death (Reddy, Zilvetti et al. 2004, Attia, Silva et al. 2008). In order to optimize outcomes by improving OLT strategies, knowledge of biochemical and molecular changes in the liver graft during and following transplantation is invaluable (Vascotto, Cesaratto et al. 2006). Of special interest are preservation and ischemia/reperfusion (I/R) injury, multi-factorial processes which affect graft function and recipient survival post OLT (Carini and Albano 2003, Carini and Albano 2003, Fondevila, Busuttil et al. 2003, Fondevila, Busuttil et al. 2003). During the process of liver transplantation, at the donor operation, the liver is retrieved following the arrest of blood circulation, cold-flushed with preservation solution (University of Wisconsin) and stored in ice (commencing cold ischemia; Greek *isch* – restriction, *hema* - blood). In the recipient operation, following explantation of the diseased liver, the graft is taken out of its cold environment for reimplantation. From that time until the reanastomosis of the blood vessels and bile duct and the establishment of recipient blood supply, warm ischemia prevails. Temperature, ischemia and introduction of the oxygenated blood contribute towards the graft injury, and some of the major biochemical processes include influx of sodium and chloride ions into the cell and alteration of calcium homeostasis (Clavien, Harvey et al. 1992, Hansen, Dawson et al. 1994), cease of aerobic glycolysis, which subsequently leads to acidosis, loss of mitochondrial respiration, ATP depletion and deterioration of energy-dependent metabolic pathways and transport processes (ischemia)

(Kang 2002). Additionally, re-oxygenation aggravates ischemic effects (reperfusion injury). These effects could be subdivided into early and late reperfusion injury. Immediately following reperfusion (up to 4 h) a major cause of injury is oxidative stress due to reactive oxygen species (ROS); this is then followed by inflammatory cytokines which results in the late injury (Bilzer and Gerbes 2000, Berrevoet, Schafer et al. 2003). Until now, these molecular mechanisms of I/R injury have been studied at a holistic level using transcriptomics (Conti, Scala et al. 2007, Defamie, Cursio et al. 2008) and proteomics (Vascotto, Cesaratto et al. 2006) approaches. However with the development of technologies for small-molecule analysis (i.e. those typically <1000 Da, arising from carbohydrates, lipids, nucleotides, amino acids, bile acids, other organic acids and bases, etc.), metabolomics could offer a complementary picture to transcriptomics, proteomics and/or histology analyses (Wishart 2005).

Metabolomics (in particular mass spectrometry based approaches) can characterize many hundreds of metabolites simultaneously (Wishart 2005), informing upon multiple metabolic pathways and providing a more comprehensive picture of liver (dys)function. It could provide (i) novel mechanistic insight into the biochemical pathways altered during OLT, (ii) targets for therapeutic interventions to minimize tissue damage and maximize likelihood of graft success, and (iii) molecular biomarker signatures to complement or improve upon existing clinical and histopathological markers of graft dysfunction following liver transplantation (either a reversible initial poor function or irreversible, primary non-function, which results in death of the recipient if retransplantation does not occur (Lemasters and Thurman 1997). This is of importance since metabolic responses are rapid (in seconds or minutes, while other physiological responses are often measured in days and/or weeks)(Wishart 2005). Consequently there is increasing interest in applying metabolomics during/after clinical organ transplantation for monitoring kidney, heart

and liver (Wishart 2005, Sarwal 2009). Examples include a successful attempt to detect acute cardiac rejection by analysing plasma by proton nuclear magnetic resonance (NMR) spectroscopy (Eugene, Le Moyec et al. 1991, Mouly-Bandini, Vion-Dury et al. 2000); profiling acute renal rejection by gas-chromatography mass spectrometry (Mao, Bai et al. 2008); and monitoring kidney-transplant patients' immune responses and drug effects in early recovery using urine samples analysed by NMR (Stenlund, Madsen et al. 2009). In liver transplantation, NMR-based metabolomics studies have showed that individual metabolites may act as indicators of liver function: for example, primary graft dysfunction may be characterized by constant levels of glycerophosphocholine (in the liver tissue) throughout OLT (Duarte, Stanley et al. 2005), by increased glutamine levels (serum and urine), and by decreased urea levels (urine) following OLT (Singh, Yachha et al. 2003). In addition, serum bile acid concentrations were shown to perform better than standard liver biochemical function tests by 48 h (Azer, McCaughan et al. 1994) and AST based evaluation by 1-3 days (Baumgarner, Scholmerich et al. 1995) when assessing liver function. Furthermore, the overall blood metabolite profile can be informative of the successful transplant (Serkova, Zhang et al. 2007).

Direct infusion Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) based metabolomics, in conjunction with selected ion monitoring (SIM) spectral stitching (Southam, Payne et al. 2007), has been shown to detect thousands of peaks (corresponding to many hundreds of metabolites) in biological samples and thus has considerable potential for investigating liver metabolism during OLT at a holistic level. Coulometric electrochemical array detection (CEAD) is a highly sensitive analytical platform specific for the analysis of metabolites involved in reduction and oxidation (redox) reactions and, therefore, highly relevant for studying I/R injury. In addition, it has been shown that microdialysis, coupled to targeted metabolite

analysis, enables continuous monitoring of graft metabolism both during and following OLT (Silva, Richards et al. 2005). Both of these analytical techniques offer considerably higher sensitivity than the already well established NMR metabolomics approach (Viant, Bearden et al. 2008, Taylor, Weber et al. 2009), although current limitations of FT-ICR MS include lower technical reproducibility and less quantitative analysis than NMR spectroscopy (Payne, Southam et al. 2009). However, by avoiding the need for chromatographic separation (i.e. liquid chromatography (LC) or gas chromatography), direct infusion FT-ICR MS benefits from a more rapid analysis than NMR spectroscopy or LC-MS, but subsequent metabolite identification relies solely on the mass-to-charge ratio of each peak with no additional information from LC retention time (Griffiths 2008). The challenges of CEAD metabolomics include difficulties of unambiguous metabolite identification.

Here we report the use of metabolomics in the setting of OLT, applying the complementary techniques of FT-ICR MS (of liver biopsy extracts) and CEAD (of microdialysates). The primary goal of this study is to determine the applicability of these approaches for characterizing I/R injury (and associated metabolic changes during and subsequent to OLT) within only a small cohort of patients; i.e. we address the question as to whether the large biological variation anticipated between patients will mask the metabolic changes induced by OLT. This is anticipated to be much more of a challenge for human samples derived from a clinical setting as compared to inbred animal or plant models derived from a controlled environment (Bijlsma, Bobeldijk et al. 2005). Our longer term goals, using these metabolomics approaches, are to better understand the molecular mechanisms of OLT, generate new testable hypotheses, and discover novel biomarker profiles of diagnostic potential.

5.3 Materials and methods

Eight adult patients undergoing OLT were recruited to study metabolic changes with FT-ICR MS and CEAD in the liver grafts before and after transplantation. The study was approved by the South Birmingham Research Ethics Committee. Patients were consented both for the OLT and for this study.

5.3.1 Clinical data

Standard clinical and biochemical data were collected during organ recovery and OLT. Briefly, median age of recipients was 58 (range 46-62). The median Model for End-Stage Liver Disease score was 20 (range 15-22). The aetiology of liver disease is shown in Table 5.1. Spontaneous intra-cranial bleed was the cause of brain death in all donors except for one donation after cardiac death and the median donor age was 66 (range 40-72). The donors spent a median of 2 (range 1-6) days on the ITU. Seven out of 8 livers were recovered while donors were on mechanical support, one liver was obtained by donation after cardiac death (H7). Standard liver function tests indicated that one recipient (H8) developed features of initial poor function (IPF) in

Table 5.1 Demographic data on recipients and timings of OLT and biopsy samplings (min)

Patient	Age	Sex	Indications for OLT [a]	Time of first biopsy, T_1 [b]	Cold ischemia time	Warm ischemia time	Duration of reperfusion prior to T_2	Time of second biopsy, T_2 [b]	Patient status 2 months after OLT
H1	58	M	A1AT	110	450	40	68	580	Alive
H2	59	M	PSC	100	600	34	76	724	Alive
H3	54	M	Hep C+HCC	110	250	47	86	387	Deceased
H4	62	M	NASH+HCC	120	560	35	83	685	Alive
H5	61	F	PBC	115	410	41	83	541	Alive
H6	51	F	Hep C+HCC	100	410	42	85	542	Alive
H7	46	M	NASH	80	400	32	81	522	Deceased
H8	53	M	ALD+HCC	125	490	27	83	607	Alive

[a] Abbreviations: A1AT, crytogenic cirrhosis; PSC, primary sclerosing cholangitis; Hep C, hepatitis C cirrhosis; HCC, hepatocellular cancer; NASH, non-alcoholic steatohepatitis; PBC primary biliary cirrhosis; ALD, alcoholic liver disease.
[b] After graft first placed on ice.

the graft with AST levels of >1500 IU/L on the second day following OLT (Table B1, Appendix B). Two recipients died within 2 months following OLT (H3 after 6 weeks due to disseminated intravascular coagulation secondary to sepsis and multi organ failure, and H7 after 5 days due to unexplained cardiac arrest).

5.3.2 Liver biopsy and FT-ICR MS metabolomics

Liver tissue samples were obtained by Menghini biopsy for seven liver grafts (not available for H1) at two stages during OLT: T_1, after organ retrieval, perfusion with preservation solution (University of Wisconsin) during the "back-table" preparation, while the liver was maintained at 4°C (during cold ischemia period; cold ischemic injury); T_2, at the end of the recipient procedure before abdominal closure (after warm ischemic period and reperfusion; warm ischemic and reperfusion injury), resulting in a total of 14 samples (Table 5.1). One-half of each biopsy was subject to histological examination and the other half was extracted using a methanol:chloroform:water method (Wu, Southam et al. 2008), and the polar metabolites analysed by ultra-high resolution direct infusion nanoelectrospray FT-ICR mass spectrometry (Thermo Fisher Scientific LTQ FT) from *m/z* 70 to 500. Nanoelectrospray settings included a flow rate of 200nL/min, backing pressure of 0.3 psi, and electrospray voltage of +1.7 and -1.7 kV for positive and negative ion mode respectively. Each sample was analysed in duplicate and spectra were processed as described previously (Taylor, Weber et al. 2009), including a 3-step filtering algorithm (Payne, Southam et al. 2009). Briefly, for each infused sample, molecules are ionised (i.e. by addition of a proton $[M+H]^+$, removal of a proton $[M-H]^-$ or addition of another cation $[M+Na]^+$) due to the high electric field in the electrospray ion source and then analysed based on their mass-to-charge ratio (Kebarle 2000, Griffiths 2008). A single compound can form

multiple ion forms. In this study all commonly detected 'adducts' were taken into account, that comprised of $[M-e]^+$, $[M+H]^+$, $[M+Na]^+$, $[M+^{39}K]^+$, $[M+^{41}K]^+$, $[M+2Na-H]^+$, $[M+2^{39}K-H]^+$, $[M+NH_4]^+$ for positive ion mode and $[M+e]^-$, $[M-H]^-$, $[M+^{35}Cl]^-$, $[M+^{37}Cl]^-$, $[M+HAc-H]^-$, where HAc is acetic acid for negative ion mode (Tong, Bell et al. 1999). Due to the noise filtering process only compounds present in 50% or more of all the samples were retained for further analysis. This exclude from the final peak list all of the known drugs (including all the possible ions forms coming from the drugs) administered to the donors and recipients, since none of them were administered to more than half of the patients.

5.3.3 Extracellular fluid and CEAD metabolomics

At the end of the recipient operation, a microdialysis catheter was inserted into the liver as described previously (Silva, Richards et al. 2005). Serial hourly dialysate samples were collected during the next 48 h (Table B2, Appendix B, not available for H8) and 10 μl of each sample were injected into the HPLC/CEAD system. In this system, a sample is introduced in HPLC and separated on the chromatographic column. Here, 3 electrode elements are present (working, counter and reference) and a fixed potential difference is applied between the working and the reference electrodes. This potential drives an electrochemical reaction at the working electrode's surface, transferring electrons that produce the current, balanced by a current flowing in the opposite direction at the auxiliary electrode. The current from the electrochemical reaction (in pico- or nanoampere) is amplified to a range of ± 1 Volt and when plotted appears as a function of time in a series of peaks (Acworth, Naoi et al. 1997). Coularray 5600A 16-channel metabolomics system; ESA Analytical Ltd, Aylesbury, UK was used. Separation was carried out on a Chromospher ODS column (5 μm, 150x3 mm with guard column; Varian Chromopack,

Walton-on-Thames, UK) using a binary mobile phase gradient which was produced by pumping (1mL/min) 40 mM sodium dihydrogen phosphate buffer, pH 3.2, containing 10^{-4} M sodium heptane sulphonic acid (A) for 3 min, before then introducing methanol:acetonitrile, 9:1 v/v, containing 10^{-4} M sodium heptane sulphonic acid (B). Solvent B was increased linearly from 0 to 5% (v/v) over 7 min, and then from 5 to 30% over the next 25 min. After maintaining these conditions for a further 2 min the system was returned to the original equilibration buffer (A) and allowed to restabilise. Metabolite detection was achieved by incrementing the 16 cells of the array in 60-mV steps from 0 to 900 mV (see Appendix B for details). Chromatographic data were aligned, peak areas integrated, and only reproducible peaks (present in every sample) were kept. The retained peaks in each chromatogram were normalized to unit area, and generalized logarithm transformed (transformation parameter $\lambda=1x10^{-6}$ (Parsons, Ludwig et al. 2007)).

5.3.4 Statistical analyses

Liver biopsy mass spectra were analysed with multivariate (principal components analysis; PCA) and univariate approaches (t-test with Benjamini and Hochberg (BH) correction for multiple testing, after verifying that data follow the normal distribution, Lilliefors test). Grubbs tests (with BH correction) were used to identify outlying peaks in the post reperfusion phase (T_2) for patient H8 (who developed IPF, based on AST levels >1500 IU/L within 2 days after OLT, Table B2, Appendix B). Time course CEAD data were analysed with PCA. This type of analysis (PCA) is widely applied to multivariate metabolomics datasets, i.e. comprising of many samples for which n (n>1) variables (in this case signals in a mass spectrum arising from metabolites) are measured at the same time (Nicholson, K. et al. 1999, Martens and Martens 2001). Each variable can be regarded as a unique dimension, and therefore each sample can be represented in n-

dimensional space. However, such space is difficult to visualise and interpret, thus PCA reduces this dimensionality using a projection technique so that individual samples can be compared in a lower (e.g. two dimensional) space – termed a PCA scores plot. The more clustered the samples are in the scores plot, the more metabolically similar they are and vice versa. Also, the most important variables that explain the variation in the original data set can be identified based upon their contribution (i.e. their loadings) to the sample position (scores) in the new reduced dimension.

5.4 Results

5.4.1 Liver metabolism of cold phase vs. post reperfusion

Liver biopsies had a wide range of micro and macro steatosis (from mild to severe) both in T_1 and T_2 (Table B3, Appendix B). FT-ICR mass spectra of the biopsies contained 1772 and 2437 reproducibly detected peaks for positive and negative ion modes, respectively. Of this total of 4209 peaks detected, 1349 were putatively identified based upon accurate mass measurements and the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa, Araki et al. 2008, Taylor, Weber et al. 2009, Kanehisa, Goto et al. 2010) PCA revealed a clear separation of the biopsies from the cold phase (T_1) and post reperfusion (T_2), along PC1 (Figure 5.1), for all but one patient (H7); this pattern was equally evident in both the positive and negative ion mode datasets, hence verifying the observation. Univariate analysis also identified many significantly changing peaks between T_1 and T_2, specifically 4.6% and 19.8% of all the positive and negative ion mode data, respectively. Based on the putative metabolite assignments, the biggest metabolic changes upon reperfusion (in top 1% of PC1 loadings, with smallest p values and/or largest fold changes; Table 5.1) comprised of an increase of urea production and urea cycle intermediate

levels (e.g. N4-acetylaminobutanal, 5'-Methylthioadenosine), and increased bile acid levels (e.g. chenodeoxyglycocholate, glycodeoxycholate, glycochenodeoxycholate and glycholate). Also, compounds present in the UW preservation solution decreased in relative abundance within the

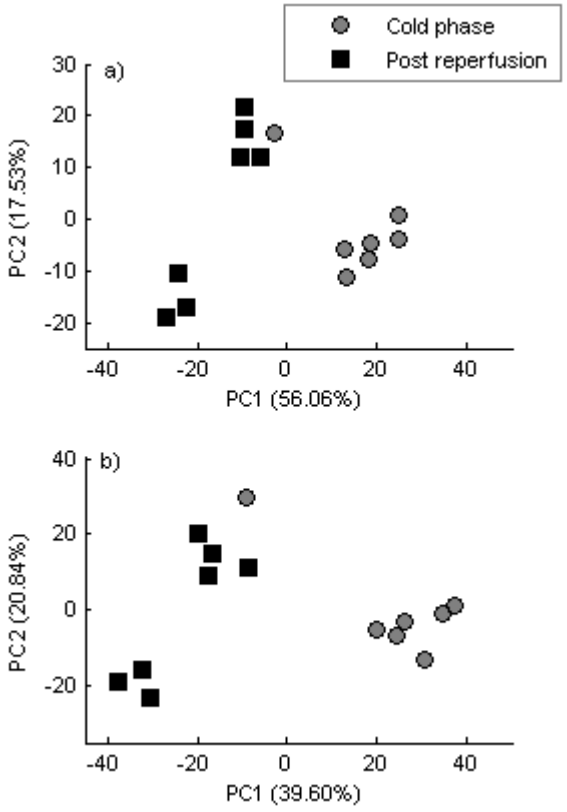

Figure 5.1 Principal component analysis scores plots for (a) positive and (b) negative ion mode FT-ICR mass spectra of liver biopsies, showing separation between the cold phase (T_1, circles) and post reperfusion samples (T_2, squares). A further intriguing separation of the post reperfusion biopsies is evident (liver biopsies H3, H4, H7 and H8 clustered in one group, H2, H5 and H8 in the other).

biopsies following reperfusion (e.g. mannitol, lactabionic acid). The fold change for each of these UW compounds was consistent upon reperfusion, except for those metabolites that occurred both in the preservation solution and endogenously within the liver (e.g. adenosine and glutathione; Table B4 and B5, Appendix B). Further molecular changes included an anticipated disturbance of energy metabolism, with consistent fold increases of several metabolites (e.g. formate,

Table 5.2 Metabolites that changed most significantly between cold phase and post reperfusion [a]

Putative metabolite [b]	Fold change [c]	p value [d]	m/z	Empirical formula	Adduct	Reason for inclusion	Rank
Urea metabolism							
Urea	2.87	0.022	83.02158	CH_4N_2O	Na	t test	9
N4-Acetylaminobutanal	2.76	0.012	130.08627	$C_6H_{11}NO_2$	H	t test	4
5'-Methylthioadenosine	10.8	0.26	338.05072	$C_{11}H_{15}N_5O_3S$	^{41}K	Large increase	13
Bile acid metabolism							
Chenodeoxyglycocholate, Glycodeoxycholate and/or Glycochenodeoxycholate	24.1	0.17	472.30359	$C_{26}H_{43}NO_5$	Na	Large increase, PC1 loadings	2, 17
	9.81	0.13	488.27758		^{39}K	Large increase	15
Glycocholate	14.5	0.061	488.29844	$C_{26}H_{43}NO_6$	Na	Large increase	6
Preservation solution							
Mannitol (5)	0.0877	0.048	205.06824	$C_6H_{14}O_6$	Na	Large decrease	30
Mannitol (5)(3)	0.0637	0.096	221.04218		^{39}K	PC 1 loadings	8
	0.597	0.11	223.04031		^{41}K	PC 1 loadings, large decrease	10, 25
Energy metabolism							
Oxaloacetate (5)	0.0482	0.049	191.01974	$C_4H_4O_5$	HAc-H	PC 1 loadings	17
ADP (3)	14.0	0.21	472.00083	$C_{10}H_{15}N_5O_{10}P_2$	2Na-H	Large increase	7
Other							
Phosphoethanolamine	3.12	0.022	164.00832	$C_2H_8NO_4P$	Na	t test	10
N1-Methyl-2-pyridone-5-carboxamide (2)	2.77	0.022	175.04782	$C_7H_8N_2O_2$	Na	t test	12
L-Histidine	3.72	0.026	178.05874	$C_6H_9N_3O_2$	Na	t test	18
L-Glutamate (8)(5)(5)	4.43	0.0029	182.02257	$C_5H_9NO_4$	^{35}Cl	t test	8
	2.55	0.0029	184.01963		^{37}Cl	t test	9
	8.26	0.071	192.02433		2Na-H	Large increase	20
6-Carboxyhexanoate (3)	12.1	0.42	197.02222	$C_7H_{12}O_4$	^{39}K-2H	Large increase	14
α-Ribazole	0.0433	0.045	315.09320	$C_{14}H_{18}N_2O_4$	^{37}Cl	PC 1 loadings	5
Sphingosine 1-phosphate	9.15	0.19	424.22000	$C_{18}H_{38}NO_5P$	2Na-H	Large increase	18
S-Adenosyl-L-homocysteine	17.5	0.3258	461.04088	$C_{14}H_{20}N_6O_5S$	$2^{39}K$-H	Large increase	4

[a] The selected peaks were within the top 1% of the PC1 loadings, had the smallest p values, or largest up or down fold changes.

[b] Values in parentheses show the number of all possible putative metabolite identities.

[c] Relative metabolite concentration in post reperfusion relative to cold phase sample.

[d] p value – cold phase (T_1) vs. post reperfusion (T_2), p values corrected for multiple testing with Benjamini and Hochberg

orthophosphate, ADP) particularly those involved in oxidative phosphorylation (e.g. fumarate, succinate) in post reperfusion biopsies (Table B6 and B7, Appendix B).

Reconsidering the entire metabolic fingerprints, the PCA scores plots (Figure 5.1) revealed that liver biopsies collected during the cold phase (T_1) were metabolically more similar to each other (tightly clustered) than post reperfusion (except H7, which was identified as an outlier in the cold phase, T_1). In the post reperfusion phase T_2, liver biopsies tended to separate into two groups along PC2 (H3, H4, H6 and H7 in one group and H2, H5 and H8 in the other). The major contributors to this partial separation within the post reperfusion biopsies were, amongst others, putatively identified as L-valine, L-glutamate, L-glutamine, inosine monophosphate (IMP), creatine, taurine, all detected as multiple ionization forms or with a unique putative metabolite assignment (Table B8, Appendix B).

For the one patient that developed IPF (H8), Grubbs tests identified 9 peaks in the biopsy mass spectra that were different (statistical outliers) in the post reperfusion phase (T_2) compared to all other patients. Of these potential indicators of IPF, only creatine and inosine monophosphate (IMP) could be putatively assigned to human metabolites (Table B9, Appendix B).

5.4.2 Redox metabolism in microdialysates post reperfusion

A total of 19 reproducible peaks were detected by CEAD in the microdialysates and subject to PCA (no possible to identify based on the CEAD alone). Time trajectories on the scores plots were quite consistent for all patients (Figure 5.2). The earliest dialysate samples from the patients (5-6h post reperfusion) group together with large positive PC1 scores, as highlighted by the average metabolic trajectory for all patients (Figure 5.2). Samples from subsequent time points were similarly grouped, but towards increasingly more negative PC1 scores. This shift along the PC1 axis was greatest for samples obtained up to 21 h post reperfusion, after which a period of metabolic stability ensued.

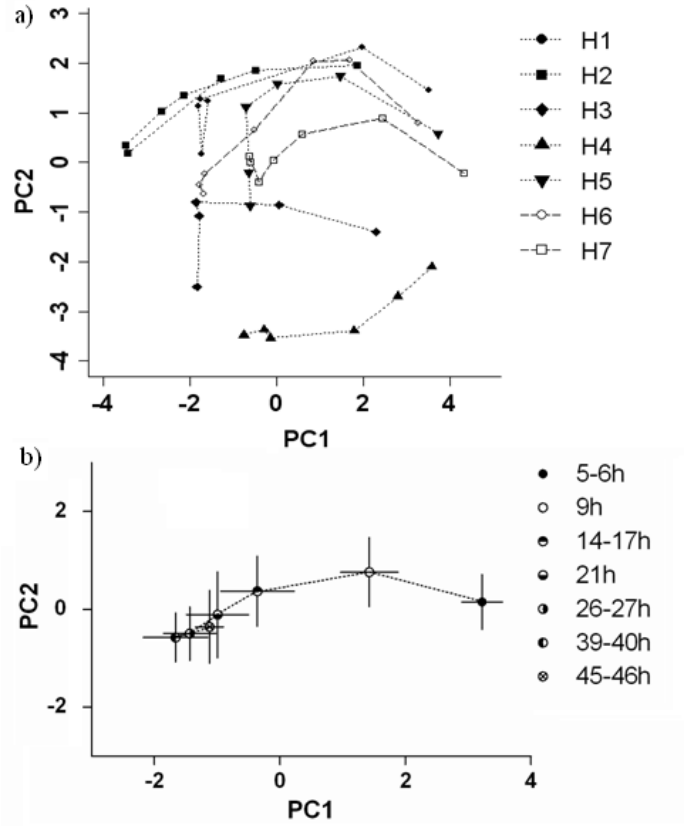

Figure 5.2 Principal components analysis scores plots for CEAD time course data showing that in general redox metabolism following OLT changes rapidly before stabilizing at ca. 21 h post reperfusion. a) PCA scores showing the metabolic trajectory for each patient separately; b) PCA scores showing the average trajectory across all patients (error bars represent SEM). The majority of variance was derived from two patients (H3 and H4).

5.5 Discussion

FT-ICR mass spectrometry of liver biopsy extracts allowed the rapid and reproducible detection of 4209 unique peaks, representing a wealth of metabolic information on the functional biochemical changes associated with liver transplantation. This preliminary study revealed that, from a holistic viewpoint, liver grafts share a similar metabolic profile in the cold phase, suggesting that metabolism is down-regulated in a consistent manner. This is in itself quite remarkable given the relative heterogeneity (in terms of patient-to-patient metabolic variation and the inevitable differences in seven OLT procedures) and small number of donors involved. Furthermore, several anticipated metabolic differences between cold phase and post reperfusion biopsies were identified, which serve to verify the FT-ICR MS approach and its applicability for measuring multiple metabolic pathways simultaneously. Specifically, we documented evidence that reperfused grafts restart their metabolic activity and physiological functions, including synthesis (e.g. bile acid production, urea synthesis) and excretion (clearance of UW solution), with the latter effect also in agreement with a previous NMR study (Singh, Yachha et al. 2003). The metabolic differences between T_1 and T_2 liver biopsies can largely be attributed to temperature changes in the tissue (i.e. low in T_1, affecting enzymatic activity) and/or the presence and absence of blood flow (i.e. no flow in T_1 and in the early T_2 following oxygenated blood flow after reperfusion). Again from a holistic viewpoint, liver biopsies exhibited considerably greater metabolic variability following reperfusion. This is again what would be expected considering the variation in the OLT procedure as well as the impact of the recipient's metabolism on the graft. Considering the second analytical approach, the combination of microdialysis sampling and CEAD time trajectory data allowed the longitudinal analysis of liver metabolism post reperfusion. It detected a series of changes in the redox metabolism of extracellular fluid,

revealing a rapidly changing liver metabolism immediately post reperfusion followed by stabilization after ca. 21 h.

The FT-ICR mass spectra revealed two further intriguing findings, which, due to the small number of patients, must be interpreted with caution. First, the only liver graft obtained by donation after cardiac death corresponded to the only outlier on the PCA scores plot, having a metabolite profile in the cold phase more similar to the other livers' metabolic profiles in the post reperfusion (T_2) stage. This may have resulted from a less effective perfusion of the graft with preservation solution, since it was performed after a period of circulatory arrest (e.g. possibility of micro clot formation), which would have several consequences for graft metabolism. For example, less effective preservation could result in potentially ongoing and injurious metabolic activity within the cold phase graft (T_1) that more closely resembles post reperfusion (T_2) metabolism. The second intriguing finding from the FT-ICR measurements, revealed in both the positive and negative ion PCA scores plots, is the apparent separation of the post reperfusion biopsies into two groups. This may arise from differing rates of metabolic recovery of liver function across the seven patients (1-2 h post reperfusion), which is supported by the CEAD data that shows considerable change in metabolism up to 21 h post reperfusion. It is important to note that it is unlikely that this separation is related to any variation in the OLT procedure (i.e. different hospital units of organ retrieval, different surgical teams etc.). Furthermore, none of the patterns in the PCA scores plots (neither at stage T_1 nor at T_2) could be explained by cold or warm ischemia times, which have been quoted to have a significant effect on the quality of grafts and procedure outcome (Lemasters and Thurman 1997). However, our results are consistent with previous studies, showing that cold and warm ischemia times are not the primary cause of liver dysfunction when they are below 10 h and 60 min for cold and warm ischemia respectively

(Fernández-Merino, Nuño-Garza et al. 2003, Tekin, Imber et al. 2004, Stahl, Kreke et al. 2008). In the current study, the mean cold ischemia time was 446 ± 108 min, and mean warm ischemia time was 37 ± 5 min.

Collectively, this study shows that FT-ICR mass spectrometry and CEAD are useful tools for characterizing multiple metabolic pathways in the liver throughout and following OLT. We have focused our interpretation and discussion on the measurements of known and expected biochemical changes during OLT since this serves to validate the application of these metabolomics methods. However, it is crucial to emphasize that more than 4000 signals were detected in the mass spectra and CEAD chromatograms, which could contain a wealth of novel metabolic information associated with OLT including predictive markers of clinical outcome or IPF. However, extracting such knowledge would require the application of more powerful supervised multivariate statistical methods, which in turn is dependent upon a considerably larger patient cohort. Such studies are now being initiated at Birmingham. In addition, the definitive identification of the metabolites within these metabolomics datasets would further strengthen our interpretation. This awaits the on-going development of automated metabolite identification strategies. However, to our knowledge this study represents the first application of FT-ICR MS based metabolomics to human samples derived from a clinical setting. The initial success of this study, in terms of the ability to identify key metabolic changes within a relatively heterogeneous group of only eight donor livers, is most encouraging.

CHAPTER 6
Final Conclusions And Future Work

The primary objectives of this thesis, as stated in Chapter 1, were to advance the data processing pipeline for a metabolomics experiment that employs the DI FT-ICR mass spectrometry platform and to investigate the applicability of this platform to a clinical study. To address the first of these aims, the issue of missing data occurrence in the final data matrix was investigated. It has been shown that missing data affect majority of the variables (peaks) and their estimation influences the results of the subsequent data analysis, both univariate and multivariate one. Eight missing data estimation algorithms were investigated to identify the optimal imputation approach, by drawing conclusions based upon the analyses of the nature of missing data, results of the specific data analyses (ANOVA and PCA) as well as the analyses of these methods performance assessed on the 'complete' datasets with missing data introduced at MCAR and MNAR (Hrydziuszko and Viant 2011). A novel approach based on the survival analysis was also investigated as an alternative to missing data estimation prior univariate data analysis (manuscript in preparation). To address the second the thesis's aims, consecutive liver biopsies taken throughout human orthotopic liver transplantation were analysed via DI FT-ICR MS based metabolomics with results showing that this approach is feasible and potentially informing upon multiple metabolic changes occurring throughout OLT (Hrydziuszko, Silva et al. 2010). Two supplementary methods of data processing and analysis were developed while addressing the primary objectives of the thesis. These included a single metric defying similarity between two or more ordered sets and a diagram to compare two or more sets. The similarity metric was used in the missing data study (identifying the optimal missing data estimation method) (Hrydziuszko and Viant 2011) and the diagram was introduced to the environmental study of toxicity testing in *Daphnia magna* to investigate the peaks detected in the 'extract blank'(Taylor, Weber et al. 2009).

6.1 Missing data

The occurrence of missing data is not limited to metabolomics, but it is a common complication of any real-world study. The appropriate handling of missing data has become a subject of an extensive research and a field in itself. For instance, various leading commercial statistical software has started providing user with extensions to analyze missing data patterns and methods of their estimation (STATA, SAS, MLwiN) (Horton and Lipsitz 2001, Carlin, Galati et al. 2008). Since multiple imputations have been identified as flexible and powerful approach, many advances were made with examples comprising multilevel multiple imputation with mixed response types (Carpenter, Goldstein et al. 2011), multiple imputation strategies for multiple group structural equation models (Enders and Gottschall 2011) or a parametric fractional imputation for missing data analysis (Kim 2011). On the other hand, survival analysis has been extensively used to analyze the censored data, especially in medical and biological fields where the nature of clinical studies and patients' mortality yields a problem of right censored data (Kirkwood and Sterne 2003). Here, although the majority of methods were developed for such data, the problem of left censored data was recognized and re-definition of left censored into right censored data identified to yield statistics and estimators of interest (Klein and Moeschberger 2003). The research presented in thesis aimed at narrowing the gap between the missing data handling techniques in metabolomics as well as using the 'fit for purposes' methods that would take into account the full available information from the DI FT-ICR MS based metabolomics spectra.

The studies discussed in Chapters 2 and 3 are the first to thoroughly address missing data in the DI FT-ICR mass spectrometry based metabolomics. Here, missing data were investigated with the aim of informing and optimizing data processing pipeline and the subsequent data

analyses, in particular identifying metabolic traits changing significantly between the sample classes. It was shown that missing values constitute up to 20% of all data in the final data matrix, yet they affect up to 80% of all the variables (peaks). The investigation of the properties of these missing entries confirmed that they do not occur at random, but they are i) a function of the peak's intensity (non-missing data observed; the lower the peak's abundance the higher amount of missing data) and ii) a function of peak's mass-to-charge ratio (unexplained increased number of missing data for the lower and upper mass-to-charge ratio peaks). Due to the prevalence of missing data in the final data matrix, any method based on the deletions (i.e. excluding variables with missing data) seems impractical and leads towards a removal of the majority of the dataset and inefficient or biased data analyses (based on the high intensity and from the middle range of the detected mass-to-charge ratios peaks). Further, data analysis methods designed to handle some proportions of missing data may not represent the optimal approach either as it was shown that for the analysed datasets missing data commonly predominate in one of the biological groups. This combined with a typically small sample size (many more variables than samples in the 'omics' experiments) may not be sufficient to provide enough statistical power to discriminate between metabolic changes across sample classes, or may simply fail to provide enough observed measurements to the use of such methods in the first place. From a purely practical point of view, the ideal data processing pipeline would incorporate a missing data estimation procedure that applied once would yield a complete data matrix ready for further data analysis. Therefore, eight common and/or easily accessible missing data estimation algorithms were investigated, taking into account the results of the subsequent data analysis and the biochemical interpretation of the results. It was demonstrated that the choice of the missing data estimation method plays an important role as it largely affects both the univariate and

multivariate data analysis. It was also discussed that out the eight compared methods *k*-nearest neighbour imputation was the optimal choice for the analysed biological datasets and the SIM-stitching method of detecting metabolites.

The above analyses led to development of a three-stage approach that informs data processing pipeline. By i) analyzing the nature of missing data, ii) assessing their influence on the subsequent data analysis (univariate and multivariate) and iii) investigating the performance of the algorithms based on the 'complete' data, the informative decision upon the optimal missing data estimation strategy can be made. This developed approach is not limited to the DI FT-ICR MS, but in fact can be used in the studies employing other metabolomics platforms.

A further work on missing data included introducing a novel approach based on the survival analysis. A theory behind it, in particular the validity of applying survival analysis methods (including log-rank test) designed for the right censored data to the left censored data following the simple transformation of the latter was discussed. It was then demonstrated that missing data in the DI FT-ICR MS based metabolomics can be regarded and represented as left censored with using the spectrum noise levels as censoring information. This approach was further assessed in terms of its applicability (and performance) to the univariate analyses of the DI FT-ICR MS metabolomics datasets. It was shown that to yield plausible results, with a set of identified as significantly changed across biological groups peaks having numerous peaks also identified with other missing data estimation methods as well numerous peaks offering an increased chance of finding predictive biomarkers (not detected with other missing data estimation methods, yet plausible due to their characteristics and biological context interpretation), thus highly relevant to focused on hypotheses generation studies.

Future work on survival analysis approach should address the validity of the assumption that the missing values can be regarded as the left-censored data with the threshold values derived from the applied signal-to-noise ratio. In the datasets discussed in Chapters 2 and 3 this approach seemed to be correct due to the nature of missing data, in particular their increased occurrence for the low intensity peaks (missing data as a function of the peaks intensity). However, this should be further examined as well as other conditions and assumptions under which this approach is feasible and beneficial. These could include the influence of the sample size, number of censored data present, the methods of estimating threshold values or occurrence of missing data (e.g. prevailing in one of or occurring across sample classes). In addition, the multivariate survival analysis methods should be investigated in the light of their applicability to the DI FT-ICR MS metabolomic datasets. Here, the mulitivariate survival data are defined as data for which independence between survival times cannot be assumed, e.g. studies of patients from the same family (common genetic background) or employees from the same company (common environmental exposure) (Hougaard 2000). This in turn could be reflected in the DI FT-ICR MS metabolomics datasets where the peaks (indicative of metabolites) are not independent (e.g. metabolites measured from the same metabolic pathway or the same metabolite measured multiple times due to different ionization forms) and yielding *parallel data* for which the number of times is fixed by design with many early events. To investigate the effects of covariates, to evaluate their dependencies, obtaining estimates and statistics or make predictions shared frailty models or multivariate frailty models can be considered. If feasible, analogical approach to the univariate one, could offer an alternative approach to analyse the multivariate data.

The selection of the optimal missing data estimation method represents an intrinsically insolvable problem since the correct answer can only be known while the true values of the

missing entries are known. Therefore, one can only try inferring the 'right' strategy based on the nature of missing data present, their influence on the subsequent data analysis and incorporating elements of the prior biochemical knowledge, hence the three-step approach was developed in this thesis. To improve the understanding of the missing data, a crucial part of the above approach, additional experiments should be carefully planned and conducted. Based on the resulted presented in this thesis, these would aim at addressing the technical and biological reasons for missing data occurrence to enable easier and robust interpretation of the subsequent data analysis.

6.2 Orthotopic liver transplantation

The analyses of liver biopsies taken throughout the OLT have confirmed that DI FT-ICR mass spectrometry is an adequate tool for characterising multiple metabolic pathways in this clinical setting. The large biological variation anticipated between patients (variability between donors, between recipients and between OLT procedures themselves) did not mask the metabolic changes induced by OLT. Despite the small cohort of patient, the DI FT-ICR MS enabled observing the expected biochemical changes upon OLT, including bile acids and urea synthesis and clearance of UW solution following OLT, all indicative of liver grafts restarting their healthy metabolic activity and physiological functions. In addition to the anticipated changes, DI FT-ICR MS allowed the observation of thousands (>4000) of unique peaks and further interesting patterns (e.g. separation of the biopsies taken post-reperfusion into two groups, albeit on a small number of samples). These initial findings encouraged further studies, which with more biopsies taken to enable more robust statistical analyses are now being analysed in the Environmental Metabolomics Research Laboratory, University of Birmingham. This on-going work will also

incorporate the previous findings on the missing value estimation and handling techniques, hopefully leading to the discovery of novel metabolic mechanisms important for the control of the ischemia/reperfusion injury thus leading towards development of the OLT strategies. Currently, despite hundreds of liver transplants being performed each year in the UK, the knowledge of the biochemical changes in the liver graft during OLT is limited and the overall 1-year survival rate is reported to be around 80%, with 5-year survival rate decreasing to around 60%. These rates depend upon multiple factors, some of which can be addressed with the DI FT-ICR MS based metabolomics, in particular organ preservation routine used, donor and recipient selection (and pairing), surgical and anaesthetic methods and during and post surgery monitoring. With the metabolomic platform verified to be able to detect and characterize many hundreds of metabolites, next research question should address all of the above aspects, aiming to increase the number of immediate successful outcomes as well as survival rates e.g. by providing metabolic markers predictive of graft success or enabling improved donor-recipient matching.

6.3 Additional advances in data processing and analysis

The similarity metric between ordered sets and the diagram to compare different sets were developed while addressing the primary objectives of this thesis. In particular, a similarity metric was developed to compare lists of peaks ranked according to their loading values following principal component analysis when comparing the impact of various missing data estimation algorithms on the multivariate analysis. In addition a visualisation tool was developed to help to understand the observation of the extract blank mass spectrum containing a comparable number of peaks to a biological mass spectrum, when first using a SIM-stitching method of metabolites acquisition for an environmental DI FT-ICR MS based metabolomics study. This tool was

invaluable for reassuring the experimentalists that the peaks measured in biological extracts were in fact specific to those extracts. Although being designed to address very specific issues, both of these methods offer further useful applications to metabolomics data processing and analysis. Hypothesis generating studies, in particular aimed at identifying potential biomarkers of disease, deal with lists of potentially interesting peaks that may be ranked according to some criteria of interest (p values, loadings values etc.). These can be considered in the light of ordered sets and easily compared using the developed similarity metric. Peak lists can also be easily visualised and compared to a reference (e.g. control) peaks list. Both of these tools are now used by researchers at the Environmental Metabolomics Research Laboratory, University of Birmingham. Currently, the similarity metric and the visualisation tool are implemented under R and Matlab respectively. Future work will be focused on creating an R package containing both of these developed tools that can be freely distributed and easily used.

6.4 Concluding remarks

Thesis objectives have been successfully met, both for the optimizing data processing pipeline and contributing towards OLT study. I have developed numerous approaches and algorithms that were reported in publications. These included a research into choosing optimal missing data estimation strategy (Hrydziuszko and Viant 2011) and introducing a novel approach of missing data handling prior univariate data analysis (in preparation). In addition, a similarity metrics was developed to support the studies on missing data (Hrydziuszko and Viant 2011) and a visualisation tool that supported the metabolomics investigation (Taylor, Weber et al. 2009). I have also assessed the applicability of the DI FT-ICR MS based metabolomics for a clinical study of OLT (Hrydziuszko, Silva et al. 2010). These developments and results are now in routine use

Environmental Metabolomics Research Laboratory, University of Birmingham, informing data processing and analysis of new studies.

References

Abdi, H. and L. J. Williams (2010). "Principal component analysis." Wiley Interdisciplinary Reviews: Computational Statistics **2**(4): 433-459.

Acworth, I. N., M. Naoi, H. Parvez and S. Parvez (1997). Coulometric electrode array detectors for HPLC. The Netherlands, VSP.

Albrecht, D., O. Kniemeyer, A. A. Brakhage and R. Guthke (2010). "Missing values in gel-based proteomics." Proteomics **10**(6): 1202-1211.

Allwood, J. W., D. I. Ellis and R. Goodacre (2008). "Metabolomic technologies and their application to the study of plants and plant–host interactions." Physiologia Plantarum **132**(2): 117-135.

Andersson, C. A. and R. Bro (1998). "Improving the speed of multi-way algorithms:: Part I. Tucker3." Chemometrics and Intelligent Laboratory Systems **42**(1-2): 93-103.

Attia, M., M. A. Silva and D. F. Mirza (2008). "The marginal liver donor – an update." Transplant International **21**(8): 713-724.

Azer, S. A., G. W. McCaughan and N. H. Stacey (1994). "Daily determination of individual serum bile acids allows early detection of hepatic allograft dysfunction." Hepatology **20**(6): 1458-1464.

Barrow, M. P., W. I. Burkitt and P. J. Derrick (2005). "Principles of Fourier transform ion cyclotron resonance mass spectrometry and its application in structural biology." Analyst **130**(1): 18-28.

Baumgarner, U., D. Scholmerich, B. Kremer, G. Streckfuss, D. Henne-Bruns, B. L. Mergard, H. Kraemer-Hansesn and E. H. Farthmann (1995). Early detection of graft dysfunction after orthotopic liver transplantation in man by serum and biliary bile acid analysis. Stuttgart, ALLEMAGNE, H.G.E.

Beckwith-Hall, B. M., J. K. Nicholson, A. W. Nicholls, P. J. D. Foxall, J. C. Lindon, S. C. Connor, M. Abdi, J. Connelly and E. Holmes (1998). "Nuclear Magnetic Resonance Spectroscopic and Principal Components Analysis Investigations into Biochemical Effects of Three Model Hepatotoxins." Chem Res Toxicol **11**(4): 260-272.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society. Series B (Methodological) **57**: 289-300.

Berg, J. M., J. L. Tymoczko and L. Stryer (2006). Biochemistry. New York, W.H. Freeman and Company.

Berrevoet, F., T. Schafer, B. Vollmar and M. D. Menger (2003). "Ischemic preconditioning: enough evidence to support clinical application in liver surgery and transplantation?" Acta Chir Belg **103**(5): 485-489.

Bijlsma, S., I. Bobeldijk, E. R. Verheij, R. Ramaker, S. Kochhar, I. A. Macdonald, B. van Ommen and A. K. Smilde (2005). "Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation." Anal Chem **78**(2): 567-574.

Bilzer, M. and A. L. Gerbes (2000). "Preservation injury of the liver: mechanisms and novel therapeutic strategies." Journal of hepatology **32**(3): 508-515.

Blanchet, L., A. Smolinska, A. Attali, M. Stoop, K. Ampt, H. van Aken, E. Suidgeest, T. Tuinstra, S. Wijmenga, T. Luider and L. Buydens (2011). "Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis." BMC Bioinformatics **12**(1): 254.

Broadhurst, D. and D. Kell (2006). "Statistical strategies for avoiding false discoveries in metabolomics and related experiments." Metabolomics **2**(4): 171-196.

Brown, S., C, G. Kruppa and J.-L. Dasseux (2005). "Metabolomics applications of FT-ICR mass spectrometry." Mass spectrometry reviews **24**(2): 223.

Buuren, S. v. and K. Groothuis-Oudshoorn (2010). "MICE: Multivariate Imputation by Chained Equations in R." Journal of statistical software.

Carini, R. and E. Albano (2003). "Recent insights on the mechanisms of liver preconditioning." Gastroenterology **125**(5): 1480-1491.

Carlin, J. B., J. C. Galati and P. Royston (2008). "A new framework for managing and analyzing multiply imputed data in Stata." Stata Journal **8**(1): 49-67.

Carpenter, J. R., H. Goldstein and M. G. Kenward (2011). "REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types." REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types **45**(5).

Chen, H. and P. Boutros (2011). "VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R." BMC Bioinformatics **12**(1): 35.

Clark, T. G., M. J. Bradburn, S. B. Love and D. G. Altman (2003). "Survival analysis part I: basic concepts and first analyses." Br J Cancer **89**(2): 232-238.

Clavien, P. A., P. R. Harvey and S. M. Strasberg (1992). "Preservation and reperfusion injuries in liver alloggrafts. An overview and synthesis of current studies." Transplantation **53**: 957-978.

Collett, D. (2003). Modelling Survival Data in Medical Research, Chapman & Hall/CRC.

Conti, A., S. Scala, P. D'Agostino, E. Alimenti, D. Morelli, B. Andria, A. Tammaro, C. Attanasio, F. D. Ragione, V. Scuderi, F. Fabbrini, M. D'Esposito, E. Di Florio, L. Nitsch, F. Calise and A. Faiella (2007). "Wide gene expression profiling of ischemia-reperfusion injury in human liver transplantation." Liver Transplantation **13**(1): 99-113.

Davis, V. W., O. F. Bathe, D. E. Schiller, C. M. Slupsky and M. B. Sawyer (2011). "Metabolomics and surgical oncology: Potential role for small molecule biomarkers." Journal of Surgical Oncology **103**(5): 451-459.

de Brevern, A. G., S. Hazout and A. Malpertuy (2004). "Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering." BMC Bioinformatics **5**: 114.

Defamie, V., R. Cursio, K. Le Brigand, C. Moreilhon, M. C. Saint-Paul, M. Laurens, D. Crenesse, B. Cardinaud, P. Auberger, J. Gugenheim, P. Barbry and B. Mari (2008). "Gene Expression Profiling of Human Liver Transplants Identifies an Early Transcriptional Signature Associated with Initial Poor Graft Function." American Journal of Transplantation **8**(6): 1221-1236.

Devlin, K. (1993). The joy of sets : fundamentals of contemporary set theory. New York, USA, Springer-Verlag Inc.

Dieterle, F., A. Ross, G. Schlotterbeck and H. Senn (2006). "Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics." Anal Chem **78**(13): 4281-4290.

Duarte, I. F., E. G. Stanley, E. Holmes, J. C. Lindon, A. M. Gil, H. Tang, R. Ferdinand, C. G. McKee, J. K. Nicholson, H. Vilca-Melendez, N. Heaton and G. M. Murphy (2005). "Metabolic Assessment of Human Liver Transplants from Biopsy Samples at the Donor and Recipient Stages Using High-Resolution Magic Angle Spinning 1H NMR Spectroscopy." Anal Chem **77**(17): 5570-5578.

Dunn, W. B. (2008). "Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes." Physical Biology **5**(1): 011001.

Dunn, W. B., N. Bailey and H. Johnson (2005). "Measuring the metabolome: current analytical technologies." Analyst **130**(5): 606-625.

Dunn, W. B., D. I. Broadhurst and S. M. Deepak (2007). "Serum metabolomics reveals many novel metabolic markers of heart failure, including pseudouridine and 2-oxoglutarate." Metabolomics **3**(4): 413-426.

Dupont, W. D. (2009). Statistical Modeling for Biomedical Researchers. New York, Cambridge University Press.

Edwards, A. W. F. (1989). "Venn Diagrams for many sets." New Scientist: 121-126.

Ellis, D. I., W. B. Dunn, J. L. Griffin, J. W. Allwood and R. Goodacre (2007). "Metabolic fingerprinting as a diagnostic tool." Pharmacogenomics **8**(9): 1243-1266.

Enders, C. K. (2001). "A Primer on Maximum Likelihood Algorithms Available for Use With Missing Data." Structural Equation Modeling: A Multidisciplinary Journal **8**(1): 128-141.

Enders, C. K. and A. C. Gottschall (2011). "Multiple Imputation Strategies for Multiple Group Structural Equation Models." Structural Equation Modeling: A Multidisciplinary Journal **18**(1): 35-54.

Eugene, M., L. Le Moyec, J. De Certaines, M. Desruennes, E. Le Rumeur, J. B. Fraysse and C. Cabrol (1991). "Lipoproteins in heart transplantation: Proton magnetic resonance spectroscopy of plasma." Magnetic Resonance in Medicine **18**(1): 93-101.

Fardet, A., C. Canlet, G. Gottardi, B. Lyan, R. Llorach, C. Rémésy, A. Mazur, A. Paris and A. Scalbert (2007). "Whole-Grain and Refined Wheat Flours Show Distinct Metabolic Profiles in Rats as Assessed by a 1H NMR-Based Metabonomic Approach." J Nutr **137**(4): 923-929.

Fernández-Merino, J., J. Nuño-Garza, P. López-Hervás, A. López-Buenadicha, Y. Quijano-Collazo and E. Vicente-López (2003). "Influence of ischemia and surgery times on development of primary dysfunction liver transplant in patients." Transplantation Proceedings **35**(4): 1439-1441.

Fiehn, O. (2002). "Metabolomics – the link between genotypes and phenotypes." Plant Mol Biol **48**(1): 155-171.

Fiehn, O., D. Robertson, J. Griffin, M. van der Werf, B. Nikolau, N. Morrison, L. Sumner, R. Goodacre, N. Hardy, C. Taylor, J. Fostel, B. Kristal, R. Kaddurah-Daouk, P. Mendes, B. van Ommen, J. Lindon and S.-A. Sansone (2007). "The metabolomics standards initiative (MSI)." Metabolomics **3**(3): 175-178.

Fondevila, C., R. W. Busuttil and J. W. Kupiec-Weglinski (2003). "Hepatic ischemia/reperfusion injury—a fresh look." Experimental and Molecular Pathology **74**(2): 86-93.

García-Laencina, P., J.-L. Sancho-Gómez and A. Figueiras-Vidal (2010). "Pattern classification with missing data: a review." Neural Computing & Applications **19**(2): 263-282.

Gibson, G. T. T., S. M. Mugo and R. D. Oleschuk (2009). "Nanoelectrospray emitters: Trends and perspective." Mass spectrometry reviews **28**(6): 918-936.

Goldrei, D. (1996). Classic Set Theory: For Guided Independent Study. Florida, USA, Chapman & Hall.

Goodacre, R., D. Broadhurst, A. Smilde, B. Kristal, J. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, A. Craig, T. Ebbels, D. Kell, C. Manetti, J. Newton, G. Paternostro, R. Somorjai, M. Sjöström, J. Trygg and F. Wulfert (2007). "Proposed minimum reporting standards for data analysis in metabolomics." Metabolomics **3**(3): 231-241.

Goodacre, R., S. Vaidyanathan, W. B. Dunn, G. G. Harrigan and D. B. Kell (2004). "Metabolomics by numbers: acquiring and understanding global metabolite data." Trends Biotechnol **22**(5): 245-252.

Griffiths, W. J. (2008). Metabolomics, metabonomics and metabolite profiling, RSC Publishing.

Han, J., R. Danell, J. Patel, D. Gumerov, C. Scarlett, J. Speir, C. Parker, I. Rusyn, S. Zeisel and C. Borchers (2008). "Towards high-throughput metabolomics using ultrahigh-field Fourier transform ion cyclotron resonance mass spectrometry." Metabolomics **4**(2): 128-140.

Hansen, T. N., P. E. Dawson and K. G. M. Brockbank (1994). "Effects of Hypothermia upon Endothelial Cells: Mechanisms and Clinical Importance." Cryobiology **31**(1): 101-106.

Helsel, D. R. (2005). "More Than Obvious: Better Methods for Interpreting Nondetect Data." Environmental Science & Technology **39**(20): 419A-423A.

Hollywood, K., D. R. Brison and R. Goodacre (2006). "Metabolomics: Current technologies and future trends." Proteomics **6**(17): 4716-4723.

Horton, N. J. and S. R. Lipsitz (2001). "Multiple Imputation in Practice." The American Statistician **55**(3): 244-254.

Hougaard, P. (2000). Analysis of Multivariate Survival Data. United States of America, Springer.

Hrydziuszko, O., M. A. Silva, M. T. P. R. Perera, D. A. Richards, N. Murphy, D. Mirza and M. R. Viant (2010). "Application of Metabolomics to Investigate the Process of Human Orthotopic Liver Transplantation: A Proof-of-Principle Study." OMICS: A Journal of Integrative Biology **14**(2): 143-150.

Hrydziuszko, O. and M. Viant (2012). "Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline." Metabolomics **8**(0): 161-174.

Jerez, J. M., I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín and L. Franco (2010). "Missing data imputation using statistical and machine learning methods in a real breast cancer problem." Artificial intelligence in medicine **50**(2): 105-115.

Jörnsten, R., H.-Y. Wang, W. J. Welsh and M. Ouyang (2005). "DNA microarray data imputation and significance analysis of differential expression." Bioinformatics **21**(22): 4155-4161.

Kaddurah-Daouk, R., B. S. Kristal and R. M. Weinshilboum (2008). "Metabolomics: A Global Biochemical Approach to Drug Response and Disease." Annual Review of Pharmacology and Toxicology **48**(1): 653-683.

Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu and Y. Yamanishi (2008). "KEGG for linking genomes to life and the environment." Nucleic Acids Res **36**(suppl 1): D480-D484.

Kanehisa, M., S. Goto, M. Furumichi, M. Tanabe and M. Hirakawa (2010). "KEGG for representation and analysis of molecular networks involving diseases and drugs." Nucleic Acids Res **38**(suppl 1): D355-D360.

Kang, K. J. (2002). "Mechanism of hepatic ischemia/reperfusion injury and protection against reperfusion injury." Transplantation Proceedings **34**(7): 2659-2661.

Kaplan, E. L. and P. Meier (1958). "Nonparametric Estimation from Incomplete Observations." Journal of the Americal Statistical Association **53**(282): 457-481.

Katajamaa, M. and M. Orešič (2007). "Data processing for mass spectrometry-based metabolomics." Journal of Chromatography A **1158**(1-2): 318-328.

Kebarle, P. (2000). "A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry." Journal of Mass Spectrometry **35**(7): 804-817.

Keefe, C. D. and M. B. Comisarow (1990). "A Family of Highly Accurate Interpolation Functions for Magnitude-Mode Fourier Transform Spectroscopy." Appl. Spectrosc. **44**(4): 600-613.

Kenny, L. C., D. I. Broadhurst, W. Dunn, M. Brown, R. A. North, L. McCowan, C. Roberts, G. J. S. Cooper, D. B. Kell, P. N. Baker and o. b. o. t. S. f. P. E. Consortium (2010). "Robust Early Pregnancy Prediction of Later Preeclampsia Using Metabolomic Biomarkers." Hypertension **56**(4): 741-749.

Kim, D.-W., K.-Y. Lee, K. H. Lee and D. Lee (2007). "Towards clustering of incomplete microarray data without the use of imputation." Bioinformatics **23**(1): 107-113.

Kim, J. K. (2011). "Parametric fractional imputation for missing data analysis." Biometrika **98**(1): 119-132.

Kincius, M., R. Liang, A. Nickkholgh, K. Hoffmann, C. Flechtenmacher, E. Ryschich, C. N. Gutt, M. M. Gebhard, J. Schmidt, M. W. Büchler and P. Schemmer (2007). "Taurine Protects from Liver Injury after Warm Ischemia in Rats: The Role of Kupffer Cells." European Surgical Research **39**(5): 275-283.

Kind, T. and O. Fiehn (2007). "Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry." BMC Bioinformatics **8**(1): 105.

Kirkwood, B. R. and J. A. C. Sterne (2003). Essential medical statistics. Oxford, Blackwell Science.

Klein, J. P. and M. L. Moeschberger (2003). Survival Analysis Techniques for Censored and Truncated Data. United States of America, Springer.

Lemasters, J. J. and R. G. Thurman (1997). "Reperfusion injury after liver preservation for transplatation." Annual Review of Pharmacology and Toxicology **37**(1): 327-338.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. Alexandria, VA, ETATS-UNIS, American Statistical Association.

Little, R. J. A. and D. B. Rubin (2002). Statistical analysis with missing data. New Jersey, John Wiley & Sons.

Machin, D., Y. B. Cheung and M. K. B. Parmar (2006). Survival Analysis. A practical approach. Chichester, John Wiley & Sons, Ltd.

Mao, Y.-y., J.-q. Bai, J.-h. Chen, Z.-f. Shou, Q. He, J.-y. Wu, Y. Chen and Y.-y. Cheng (2008). "A pilot study of GC/MS-based serum metabolic profiling of acute rejection in renal transplantation." Transplant Immunology **19**(1): 74-80.

Marlin, B. M. (2008). Missing Data Problems In Machine Learning, University of Toronto.

Marshall, A. G. and C. L. Hendrickson (2002). "Fourier transform ion cyclotron resonance detection: principles and experimental configurations." International Journal of Mass Spectrometry **215**(1-3): 59-75.

Marshall, A. G., C. L. Hendrickson and G. S. Jackson (1998). "Fourier transform ion cyclotron resonance mass spectrometry: A primer." Mass spectrometry reviews **17**(1): 1-35.

Martens, H. and M. Martens (2001). Analysis of One Data Table X: Principal Component Analysis.

Mashima, R., T. Nakanishi-Ueda and Y. Yamamoto (2003). "Simultaneous determination of methionine sulfoxide and methionine in blood plasma using gas chromatography-mass spectrometry." Anal Biochem **313**(1): 28-33.

Mayr, M. (2008). "Metabolomics: Ready for the Prime Time?" Circulation. Cardiovascular genetics **1**(1): 58-65.

McKnight, P. E., K. M. McKnight, S. Sidani and A. J. Figueredo (2007). Missing Data: a gentle introduction. New York, The Guildford Press.

Millard, S. P. and S. J. Deverel (1988). "Nonparametric statistical methods for comparing two sites based on data with multiple nondetect limits." Water Resour. Res. **24**(12): 2087-2098.

Mouly-Bandini, A., J. Vion-Dury, P. Viout, M. Sciaky, T. Mesana and P. Cozzone (2000). "Detection of acute cardiac rejection by high resolution proton magnetic resonance spectroscopy of plasma." Magnetic Resonance Materials in Physics, Biology and Medicine **11**(1): 27-32.

Mutch, D. M., J. C. Fuhrmann, D. Rein, J. C. Wiemer, J.-L. Bouillot, C. Poitou and K. Clément (2009). "Metabolite Profiling Identifies Candidate Markers Reflecting the Clinical Adaptations Associated with Roux-en-Y Gastric Bypass Surgery." PLoS One **4**(11): e7905.

Myers, W. R. (2000). "Handling Missing Data in Clinical Trials: An Overview." Drug Information Journal **34**(2): 525-533.

Nicholas, J. H. and P. K. Ken (2008). Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models.

Nicholson, J. K., LINDON, J. C., HOLMES and E. (1999). 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. London, ROYAUME-UNI, Taylor & Francis.

Oba, S., M. a. Sato, I. Takemasa, M. Monden, K. i. Matsubara and S. Ishii (2003). "A Bayesian missing value estimation method for gene expression profile data." Bioinformatics **19**(16): 2088-2096.

Ohta, D., D. Shibata and K. Shigehiko (2007). "Metabolic profiling using Fourier-transform ion-cyclotron-resonance mass spectrometry." Anal Bioanal Chem **389**(5): 1469.

Oliver, S. G., M. K. Winson, D. B. Kell and F. Baganz (1998). "Systematic functional analysis of the yeast genome." Trends Biotechnol **16**(9): 373-378.

Parsons, H., C. Ludwig, U. Gunther and M. Viant (2007). "Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation." BMC Bioinformatics **8**(1): 234.

Parsons, H. M., D. R. Ekman, T. W. Collette and M. R. Viant (2009). "Spectral relative standard deviation: a practical benchmark in metabolomics." Analyst **134**(3): 478-485.

Payne, T., A. Southam, T. Arvanitis and M. Viant (2009). "A signal filtering method for improved quantification and noise discrimination in fourier transform ion cyclotron resonance mass spectrometry-based metabolomics data." Journal of The American Society for Mass Spectrometry **20**(6): 1087-1095.

Pedreschi, R., M. L. Hertog, S. C. Carpentier, J. Lammertyn, J. Robben, J. P. Noben, B. Panis, R. Swennen and B. M. Nicolai (2008). "Treatment of missing values for multivariate statistical analysis of gel-based proteomics data." Proteomics **8**(7): 1371-1383.

Peto, R., M. C. Pike, P. Armitage, N. E. Breslow, D. R. Cox, S. V. Horward, N. Mantel, K. McPherson, J. Peto and P. G. Smith (1977). "Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples." Br J Cancer **35**: 1-39.

Raamsdonk, L. M. (2001). "A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations." Nat Biotechnol **19**(1): 45-50.

Reddy, S., M. Zilvetti, J. Brockmann, A. McLaren and P. Friend (2004). "Liver transplantation from non–heart-beating donors: Current status and future prospects." Liver Transplantation **10**(10): 1223-1232.

Rubin, D. B. (1976). "Inference and missing data." Biometrika **63**(3): 581-592.

Sangster, T. P., J. E. Wingate, L. Burton, F. Teichert and I. D. Wilson (2007). "Investigation of analytical variation in metabonomic analysis using liquid chromatography/mass spectrometry." Rapid Commun Mass Spectrom **21**(18): 2965-2970.

Sarwal, M. (2009). "Deconvoluting the 'omics' for organ transplantation." Current opinion in organ transplantation **14**(5): 544-551.

Scalbert, A., L. Brennan, O. Fiehn, T. Hankemeier, B. Kristal, B. van Ommen, E. Pujos-Guillot, E. Verheij, D. Wishart and S. Wopereis (2009). "Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research." Metabolomics **5**(4): 435-458.

Schafer, J. L. (1999). "Multiple imputation: a primer." Statistical Methods in Medical Research **8**(1): 3-15.

Schafer, J. L. and J. W. Graham (2002). "Missing data: our view of the state of the art." Psychological methods **7**(2): 147-177.

Scheel, I., M. Aldrin, I. K. Glad, R. Sorum, H. Lyng and A. Frigessi (2005). "The influence of missing value imputation on detection of differentially expressed genes from microarray data." Bioinformatics **21**(23): 4272-4279.

Schmidt, C. W. (2004). "Metabolomics: What's Happening Downstream of DNA." Environ Health Perspect **112**(7).

Schützenmeister, A. and H.-P. Piepho (2010). "Background correction of two-colour cDNA microarray data using spatial smoothing methods." TAG Theoretical and Applied Genetics **120**(2): 475-490.

Seaberg, E. C., S. H. Belle, K. C. Beringer, J. L. Schivins and K. M. Detre (1998). "Liver transplantation in the United States from 1987-1998: updated results from the Pitt-UNOS Liver Transplant Registry." Clin Transpl: 17-37.

Serkova, N. J., Y. Zhang, J. L. Coatney, L. Hunter, M. E. Wachs, C. U. Niemann and M. S. Mandell (2007). "Early detection of graft failure using the blood metabolic profile of a liver recipient." Transplantation **83**(4): 517.

She, N. (1997). "Aanalyzing censored water quality data using a non-parametric approach." JAWRA Journal of the American Water Resources Association **33**(3): 615-624.

Silva, M. A., D. F. Mirza, J. A. C. Buckels, S. R. Bramhall, D. Mayer, S. J. Wigmore, N. Murphy and D. A. Richards (2006). "Arginine and Urea Metabolism in the Liver Graft: A Study Using Microdialysis in Human Orthotopic Liver Transplantation." Transplantation **82**(10): 1304-1311 1310.1097/1301.tp.0000241099.0000293794.d0000241096.

Silva, M. A., D. A. Richards, S. R. Bramhall, D. H. Adams, D. F. Mirza and N. Murphy (2005). "A Study of the Metabolites of Ischemia-Reperfusion Injury and Selected Amino Acids in the Liver Using Microdialysis during Transplantation." Transplantation **79**(7): 828-835 810.1097/1001.TP.0000153156.0000138617.0000153197.

Singh, H. K., S. K. Yachha, R. Saxena, A. Gupta, G. A. Nagana Gowda, M. Bhandari and C. L. Khetrapal (2003). "A new dimension of 1H-NMR spectroscopy in assessment of liver graft dysfunction." NMR in Biomedicine **16**(4): 185-188.

Southam, A. D., T. G. Payne, H. J. Cooper, T. N. Arvanitis and M. R. Viant (2007). "Dynamic Range and Mass Accuracy of Wide-Scan Direct Infusion Nanoelectrospray Fourier Transform Ion Cyclotron Resonance Mass Spectrometry-Based Metabolomics Increased by the Spectral Stitching Method." Anal Chem **79**(12): 4595-4602.

Stahl, J. E., J. E. Kreke, F. A. A. Malek, A. J. Schaefer and J. Vacanti (2008). "Consequences of Cold-Ischemia Time on Primary Nonfunction and Patient and Graft Survival in Liver Transplantation: A Meta-Analysis." PLoS One **3**(6): e2468.

Stenlund, H., R. Madsen, A. Vivi, M. Calderisi, T. Lundstedt, M. Tassini, M. Carmellini and J. Trygg (2009). "Monitoring kidney-transplant patients using metabolomics and dynamic modeling." Chemometrics and Intelligent Laboratory Systems **98**(1): 45-50.

Steuer, R., K. Morgenthal, W. Weckwerth and J. Selbig (2007). A Gentle Guide to the Analysis of Metabolomics Data. Metabolomics: Methods and Protocols, Humana Press**:** 105-129.

Sumner, L., A. Amberg, D. Barrett, M. Beale, R. Beger, C. Daykin, T. Fan, O. Fiehn, R. Goodacre, J. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. Lane, J. Lindon, P. Marriott, A. Nicholls, M. Reily, J. Thaden and M. Viant (2007). "Proposed minimum reporting standards for chemical analysis." Metabolomics **3**(3): 211-221.

Taylor, N., R. Weber, A. Southam, T. Payne, O. Hrydziuszko, T. Arvanitis and M. Viant (2009). "A new approach to toxicity testing in <i>Daphnia magna : application of high throughput FT-ICR mass spectrometry metabolomics." Metabolomics **5**(1): 44-58.

Taylor, N. S., R. J. M. Weber, T. A. White and M. R. Viant (2010). "Discriminating between Different Acute Chemical Toxicities via Changes in the Daphnid Metabolome." Toxicological Sciences **118**(1): 307-317.

Tekin, K., C. J. Imber, M. Atli, B. K. Gunson, S. R. Bramhall, D. Mayer, J. A. C. Buckels, P. Mcmaster and D. F. Mirza (2004). "A simple scoring system to evaluate the effects of cold ischemia on marginal liver donors1." Transplantation **77**(3): 411-416.

Tong, H., D. Bell, K. Tabei and M. Siegel (1999). "Automated data massaging, interpretation, and e-mailing modules for high throughput open access mass spectrometry." Journal of The American Society for Mass Spectrometry **10**(11): 1174-1187.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. B. Altman (2001). "Missing value estimation methods for DNA microarrays." Bioinformatics **17**(6): 520-525.

Tuikkala, J., L. L. Elo, O. S. Nevalainen and T. Aittokallio (2008). "Missing value imputation improves clustering and interpretation of gene expression microarray data." BMC Bioinformatics **9**: 202.

van den Berg, R., H. Hoefsloot, J. Westerhuis, A. Smilde and M. van der Werf (2006). "Centering, scaling, and transformations: improving the biological information content of metabolomics data." BMC Genomics **7**(1): 142.

Vascotto, C., L. Cesaratto, C. D'Ambrosio, A. Scaloni, C. Avellini, I. Paron, U. Baccarani, G. L. Adani, C. Tiribelli, F. Quadrifoglio and G. Tell (2006). "Proteomic analysis of liver tissues subjected to early ischemia/reperfusion injury during human orthotopic liver transplantation." Proteomics **6**(11): 3455-3465.

Venn, J. (1971). Symbolic Logic. New York, Chelsea Publishing Company.

Viant, M. R., D. W. Bearden, J. G. Bundy, I. W. Burton, T. W. Collette, D. R. Ekman, V. Ezernieks, T. K. Karakach, C. Y. Lin, S. Rochfort, J. S. d. Ropp, Q. Teng, R. S. Tjeerdema, J. A. Walter and H. Wu (2008). "International NMR-Based Environmental Metabolomics Intercomparison Exercise." Environmental Science & Technology **43**(1): 219-225.

Vollenbroeker, B., J. H. Koch, M. Fobker, B. Suwelack, H. Hohage and U. Müller (2005). "Determination of Cyclosporine and Its Metabolites in Blood via HPLC-MS and Correlation to Clinically Important Parameters." Transplantation Proceedings **37**(4): 1741-1744.

Walczak, B. and D. L. Massart (2001). "Dealing with missing data: Part II." Chemometrics and Intelligent Laboratory Systems **58**(1): 29-42.

Wang, J.-N., Y. Zhou, T.-Y. Zhu, X. Wang and Y.-L. Guo (2008). "Prediction of Acute Cellular Renal Allograft Rejection by Urinary Metabolomics Using MALDI-FTMS." Journal of Proteome Research **7**(8): 3597-3601.

Weber, R. J. M., A. D. Southam, U. Sommer and M. R. Viant (2011). "Characterization of Isotopic Abundance Measurements in High Resolution FT-ICR and Orbitrap Mass Spectra for Improved Confidence of Metabolite Identification." Anal Chem **83**(10): 3737-3743.

Westerhuis, J., H. Hoefsloot, S. Smit, D. Vis, A. Smilde, E. van Velzen, J. van Duijnhoven and F. van Dorsten (2008). "Assessment of PLSDA cross validation." Metabolomics **4**(1): 81-89.

Wilm, M. and M. Mann (1996). "Analytical Properties of the Nanoelectrospray Ion Source." Anal Chem **68**(1): 1-8.

Wishart, D. S. (2005). "Metabolomics: The Principles and Potential Applications to Transplantation." American Journal of Transplantation **5**(12): 2814-2820.

Wishart, D. S. (2006). "Metabolomics in monitoring kidney transplants." Current Opinion in Nephrology and Hypertension **15**(6): 637-642

Wishart, D. S. (2008). "Metabolomics: applications to food science and nutrition research." Trends in food science & technology **19**(9): 482-493.

Wishart, D. S., C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, R. Mandal, I. Sinelnikov, J. Xia, L. Jia, J. A. Cruz, E. Lim, C. A. Sobsey, S. Shrivastava, P. Huang, P. Liu, L. Fang, J. Peng, R. Fradette, D. Cheng, D. Tzur, M. Clements, A. Lewis, A. De Souza, A. Zuniga, M. Dawe, Y. Xiong, D. Clive, R. Greiner, A. Nazyrova, R. Shaykhutdinov, L. Li, H. J. Vogel and I. Forsythe (2009). "HMDB: a knowledgebase for the human metabolome." Nucleic Acids Res **37**(suppl 1): D603-D610.

Wu, H., A. D. Southam, A. Hines and M. R. Viant (2008). "High-throughput tissue extraction protocol for NMR- and MS-based metabolomics." Anal Biochem **372**(2): 204-212.

Xia, J., N. Psychogios, N. Young and D. S. Wishart (2009). "MetaboAnalyst: a web server for metabolomic data analysis and interpretation." Nucleic Acids Res **37**(suppl 2): W652-W660.

Appendix A

Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. Supplementary Material

Index of figures

Index of tables

SM: Occurrence and distribution patterns of missing data

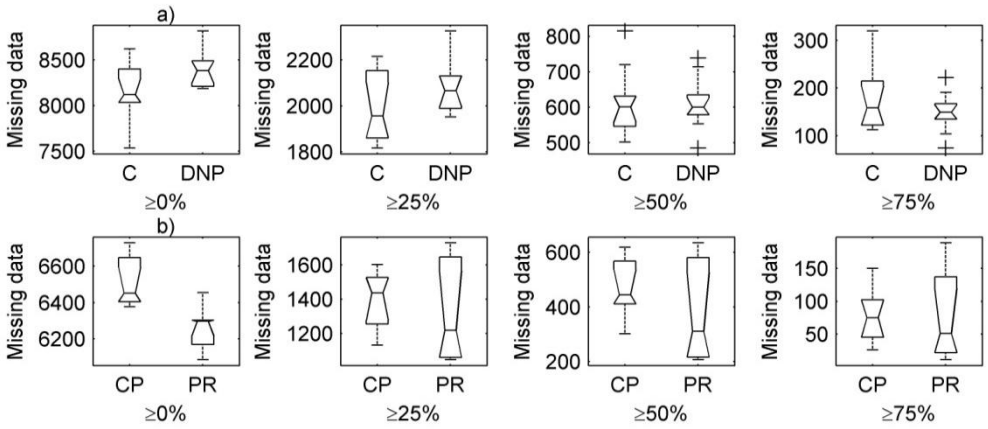

Figure A1 Boxplots of the amount of missing data per biological group for various settings of the sample filter (part of the 3-step noise filtering algorithm) for a) DM (with C: control and DNP: dinitrophenol-treated groups) and b) HL (with CP: cold phase and PR: post reperfusion groups). Noise filtering strategies may hinder or introduce significant differences between biological groups, e.g. reducing the difference between the biological groups for the HL dataset.

SM: Impact of missing data imputation on univariate data analysis

Table A1 Number of significantly changed peaks identified after imputing missing data with eight missing data estimation algorithms.

		S	HM	M	MED	KNN	BPCA	MI	REP	MEAN	RSD
CCL_n	SP [%]	12.83	12.63	3.08	3.02	14.70	2.65	11.05	5.09	8.13	63.21
	SP with MV[%]	70.68	70.32	29.86	28.99	72.62	24.73	68.21	50.72	52.01	40.71
	SDP with MV[%]	41.76	41.30	7.06	6.60	46.26	5.46	42.00	22.68	26.64	68.24
CCL_p	SP [%]	10.20	10.09	0.91	0.89	6.75	0.58	3.14	2.69	4.41	92.01
	SP with MV[%]	83.84	83.66	39.02	37.50	80.53	15.38	70.21	68.60	59.84	43.14
	SDP with MV[%]	51.42	51.05	8.40	6.81	53.89	1.28	42.32	31.86	30.88	71.90
DM	SP [%]	14.18	14.20	7.44	7.70	10.13	9.27	9.03	13.20	10.64	26.44
	SP with MV[%]	49.58	49.66	23.72	25.70	37.65	33.68	32.98	47.11	37.51	27.61
	SDP with MV[%]	14.55	14.54	2.96	3.17	7.76	6.77	6.83	12.36	8.62	54.26
HL	SP [%]	14.24	13.85	2.11	1.72	5.93	4.10	3.10	10.69	6.97	74.70
	SP with MV[%]	85.99	85.60	39.47	32.26	70.09	63.51	55.36	80.83	64.14	32.03
	SDP with MV[%]	28.65	28.49	6.39	3.00	20.16	22.68	11.73	26.28	18.42	54.91

SP [%], number of **S**ignificantly-changed **P**eaks (percentage in respect to the number of all peaks in the dataset). **SP with MV [%],** Number of SP with missing values (MV) [%] (percentage of significantly changed peaks containing at least one missing value across all samples with respect to the number of all significantly changed peaks). **SDP with MV [%],** Number of SDP with missing values [%] (specifically the percentage of **D**ata **P**oints with missing values in the **S**ignificantly-changed peaks, with respect to all data points in the significantly changed peaks).

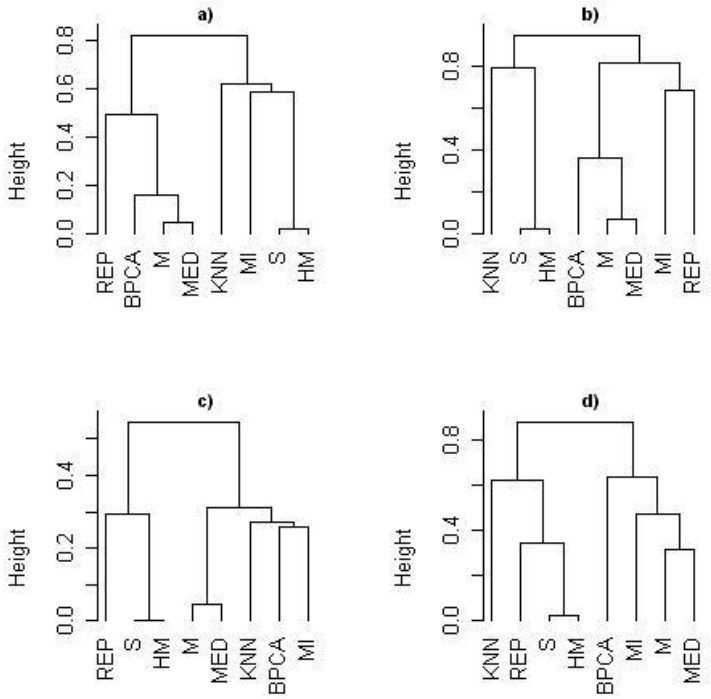

Figure A2 Comparison of the eight missing value estimation methods based upon the hierarchical clustering (Euclidean distance, agglomeration method: complete) of the significantly changed peaks from univariate analysis of the a) CCL_n, b) CCL_p, c) DM and d) HL datasets.

Table A2 Numerical values derived from the hierarchical clustering in Figure A2, showing the similarities between the eight missing value estimation methods in terms of which peaks were found to change significantly using univariate analysis.

		S	HM	M	MED	KNN	BPCA	MI	REP
$CCL_p \setminus CCL_n$	S	100.00	97.51	23.28	22.96	40.27	20.27	41.21	35.65
	HM	97.61	100.00	23.65	23.33	39.85	20.60	41.09	36.52
	M	8.71	8.81	100.00	95.33	20.45	83.64	26.01	57.75
	MED	8.50	8.59	92.86	100.00	20.18	85.24	25.98	57.95
	KNN	20.99	21.35	12.05	11.73	100.00	18.06	37.97	24.61
	BPCA	5.68	5.74	63.41	65.00	8.22	100.00	23.36	50.85
	MI	16.76	16.93	22.15	21.48	19.03	18.44	100.00	31.16
	REP	22.15	22.65	33.88	33.06	21.49	21.49	31.66	100.00
		S	HM	M	MED	KNN	BPCA	MI	REP
HL \ DM	S	100.00	99.83	45.59	47.59	56.44	51.38	55.84	70.73
	HM	97.28	100.00	45.51	47.51	56.36	51.31	56.00	70.88
	M	14.34	14.74	100.00	95.38	69.04	73.95	72.75	52.20
	MED	12.06	12.40	68.29	100.00	71.56	74.51	73.33	53.59
	KNN	37.36	38.37	33.03	28.97	100.00	73.56	72.90	64.26
	BPCA	23.97	23.66	41.77	36.36	36.09	100.00	74.15	57.96
	MI	21.32	21.91	59.32	52.63	48.18	38.30	100.00	59.22
	REP	65.44	66.54	18.46	16.06	45.63	25.35	27.69	100.00

Highlighted in yellow, values for CCL_n and DM, in blue for CCL_p and HL.

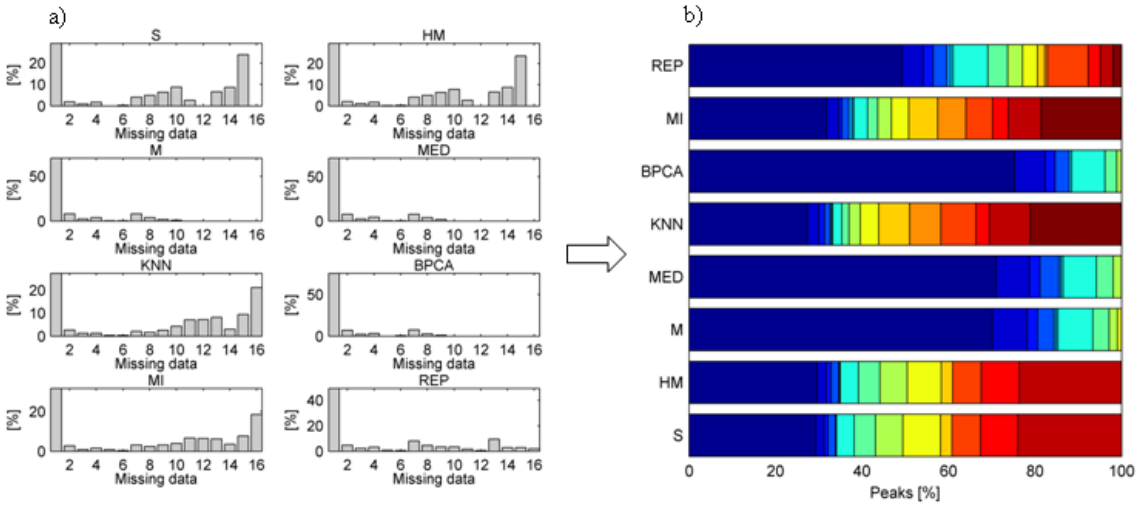

Figure A3 Percentage of peaks having zero up to the maximum number of missing values per sample (i.e. up to 16 for this CCL_n dataset) considering only those peaks that significantly changed (univariate analysis), for each of the eight estimation methods; a) shows distribution plots for each method, b) equivalent information presented as stacked boxplots.

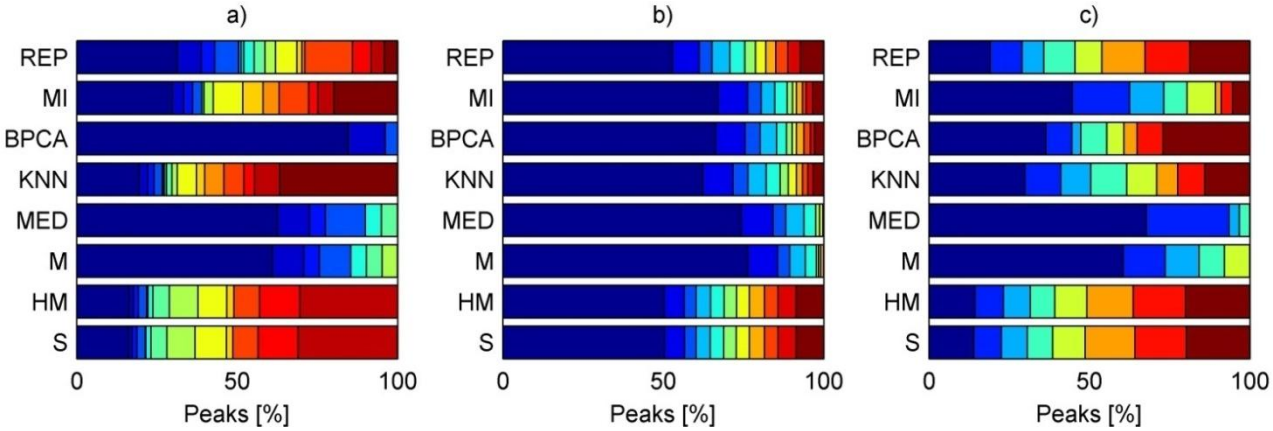

Figure A4 Boxplots (stacked) showing the percentage of significantly changed peaks containing various amounts of missing values per sample, ranging from zero (in blue) up to 15, 10 and 7 (in brown) for the a) CCL_p, b) DM and c) HL datasets respectively.

Table A3 The five most commonly occurring patterns of missing values across samples for the significantly changed peaks identified after missing data estimation with the eight methods.

		Pattern/percentage				
		1	2	3	4	5
CCL_n	S	000 / 29.32*	266 / 23.98	166 / 8.64	066 / 6.59	126 / 5.34
	HM	000 / 29.68	266 / 23.67	166 / 8.78	066 / 6.58	016 / 4.97
	M	000 / 70.14	001 / 8.06	006 / 7.11	016 / 3.79	003 / 2.84
	MED	000 / 71.01	001 / 7.73	006 / 7.25	016 / 3.86	003 / 2.90
	KNN	000 / 27.38	366 / 21.13	266 / 9.42	236 / 7.24	336 / 5.85
	BPCA	000 / 75.27	001 / 7.14	006 / 7.14	003 / 2.75	016 / 2.75
	MI	000 / 31.79	366 / 18.60	266 / 7.65	236 / 6.46	136 / 4.22
	REP	000 / 49.28	066 / 9.46	006 / 7.16	001 / 4.87	016 / 4.58
CCL_p	S	266 / 31.00	000 /16.16	166 / 12.45	066 / 7.86	036 / 7.64
	HM	266 / 30.46	000 /16.34	166 / 12.58	066 / 7.95	036 / 7.51
	M	000 / 60.98	001 /9.76	003 / 7.32	002 / 4.88	006 / 4.88
	MED	000 / 62.50	001 /10.00	003 / 10.00	002 / 5.00	006 / 5.00
	KNN	366 / 36.63	000 / 19.47	266 / 7.92	236 / 5.94	336 / 3.96
	BPCA	000 / 84.62	001 / 11.54	003 / 3.85	002 / 0.00	011 / 0.00
	MI	000 / 29.79	366 / 19.86	066 / 6.38	036 / 4.96	236 / 4.96
	REP	000 / 31.40	066 / 14.88	001 / 7.44	036 / 6.61	003 / 5.79
DM	S	00 / 50.42	01 / 6.22	03 / 4.37	04 / 4.03	02 / 3.53
	HM	00 / 50.34	01 / 6.21	03 / 4.36	04 / 4.03	02 / 3.69
	M	00 / 76.28	01 / 9.29	03 / 2.56	02 / 2.24	12 / 2.24
	MED	00 / 74.30	01 / 9.91	03 / 3.41	02 / 2.48	12 / 2.17
	KNN	00 / 62.35	01 / 9.41	03 / 3.53	02 / 3.06	04 / 2.59
	BPCA	00 / 66.32	01 / 9.25	02 / 3.08	03 / 2.57	12 / 2.57
	MI	00 / 67.02	01 / 9.23	02 / 3.69	03 / 3.43	04 / 2.37
	REP	00 / 52.89	01 / 8.3	03 / 3.79	02 / 3.07	04 / 3.07
HL	S	00 / 14.01	05 / 12.45	16 / 11.67	06 / 9.73	01 /8.56
	HM	00 / 14.40	16 / 11.60	05 / 11.20	06 / 10.00	01 /8.80
	M	00 / 60.53	01 / 13.16	13 / 7.89	02 / 5.26	11 /5.26
	MED	00 / 67.74	01 / 25.81	11 / 3.23	12 / 3.23	02 /0.00
	KNN	00 / 29.91	01 / 11.21	16 / 9.35	02 / 7.48	03 / 7.48
	BPCA	00 / 36.49	07 / 22.97	01 / 8.11	12 / 5.41	13 / 5.41
	MI	00 / 44.64	01 / 17.86	02 / 7.14	03 / 5.36	13 / 5.36
	REP	00 / 19.17	16 / 11.40	01 / 9.84	03 / 7.25	05 / 7.25

*Pattern 000 corresponds to no missing data for none of the 3 groups; pattern 266 corresponds to 2 missing values in one of the groups and 6 missing values in each of the two remaining groups. The number after the slash represents the percent occurrence of the pattern.

Table A4 Summary of which KEGG human pathways are 'active' (i.e. observed with 75% likelihood based on the significantly different peaks between the control and two drug treated groups) in the CCL_n dataset, after estimating the missing values with eight different algorithms.

KEGG pathway	S	HM	M	MED	KNN	BPCA	MI	REP
Drug metabolism	X*	X	X	X	X	-	X	X
Purine metabolism, Glycerophospholipid metabolism	X	X	-	-	X	-	X	X
Sulfur metabolism	X	X	-	-	X	-	-	X
Lysine degradation, Histidine metabolism, Amino sugar and nucleotide sugar metabolism, Phenylalanine metabolism, D-Arginine and D-ornithine metabolism, Tryptophan metabolism, Pantothenate and CoA biosynthesis, Taste transduction	X	X	-	-	X	-	X	-
Cyanoamino acid metabolism, beta-Alanine metabolism	X	X	-	-	-	-	X	-
Ascorbate and aldarate metabolism	X	X	-	-	-	-	-	X
Drug metabolism - cytochrome P450	X	X	-	-	X	-	-	-
Nitrogen metabolism, Aminoacyl-tRNA biosynthesis, Nicotinate and nicotinamide metabolism	X	X	-	-	-	-	-	-
Lysine biosynthesis, Pathways in cancer, Prostate cancer, alpha-Linolenic acid metabolism, Autoimmune thyroid disease, Biosynthesis of unsaturated fatty acids	-	-	-	-	X	-	-	-
Sphingolipid metabolism, Fatty acid biosynthesis, Arachidonic acid metabolism	-	-	-	-	-	-	X	-

* X indicates that a pathway is 'active' for this particular method.

Table A5 Summary of which KEGG human pathways are 'active' (i.e. observed with 75% likelihood based on the significantly different peaks between the control and two drug treated groups) in the CCL_p dataset, after estimating the missing values with eight different algorithms.

KEGG pathway	S	HM	M	MED	KNN	BPCA	MI	REP
Pyrimidine metabolism, Alanine, aspartate and glutamate metabolism	X	X	X	X	X	-	X	X
Thiamine metabolism	X	X	-	-	X	-	X	X
Butanoate metabolism, Drug metabolism - cytochrome P45	X	X	-	-	X	-	X	-
Galactose metabolism, Glycerolipid metabolism, Purine metabolism, Glycine, serine and threonine metabolism, Ubiquinone and other terpenoid-quinone biosynthesis, Steroid hormone biosynthesis, Sphingolipid metabolism, Neuroactive ligand-receptor interaction, Metabolism of xenobiotics by cytochrome P45	X	X	-	-	X	-	-	-
Vitamin B6 metabolism, alpha-Linolenic acid metabolism	X	X	-	-	-	-	X	-
Parkinson's disease	-	-	-	-	X	-	X	X
beta-Alanine metabolism, Pantothenate and CoA biosynthesis, Phenylalanine metabolism, Biotin metabolism, Tyrosine metabolism, Fatty acid metabolism, Nicotinate and nicotinamide metabolism	X	X	-	-	-	-	-	-
Phenylalanine, tyrosine and tryptophan biosynthesis, Taurine and hypotaurine metabolism, Oxidative phosphorylation	-	-	-	-	X	-	-	-
Amino sugar and nucleotide sugar metabolism	-	-	-	-	-	-	X	-

* X indicates that a pathway is 'active' for this particular method.

SM: Impact of missing data imputation on multivariate data analysis

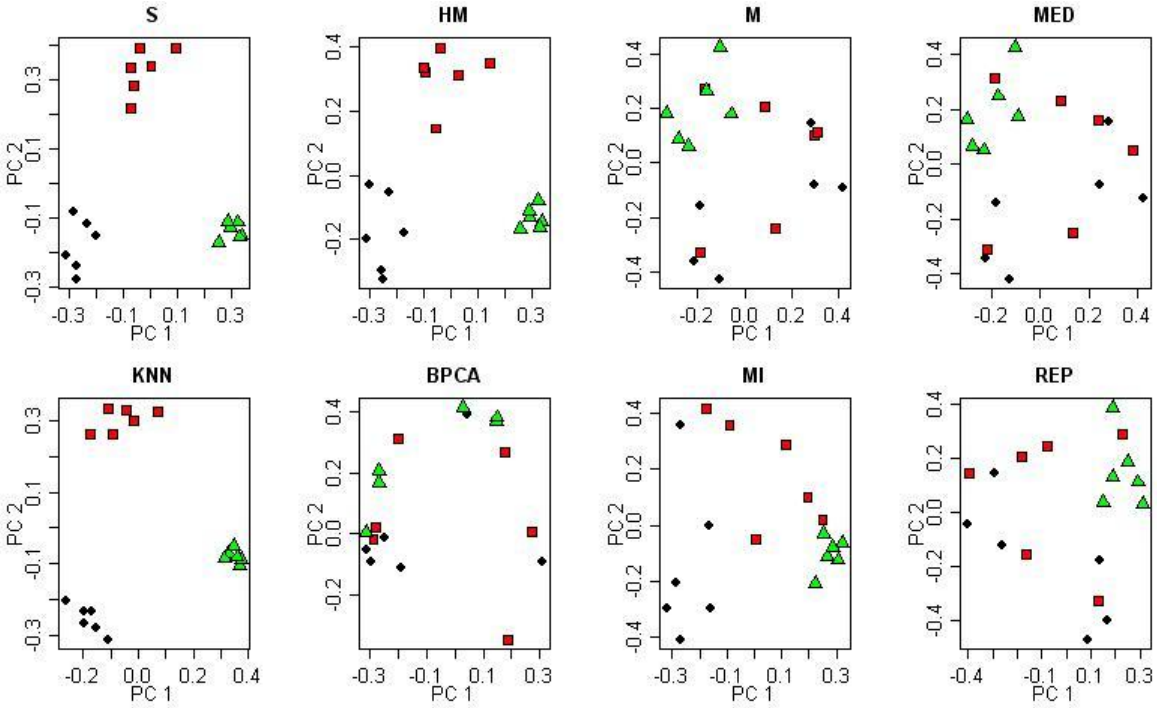

Figure A5 PCA scores plots for the CCL_p datasets obtained after estimating missing data with the eight methods: controls (black diamonds), indomethacin treated (red squares), medroxyprogesterone acetate treated (green triangles) cancer cells.

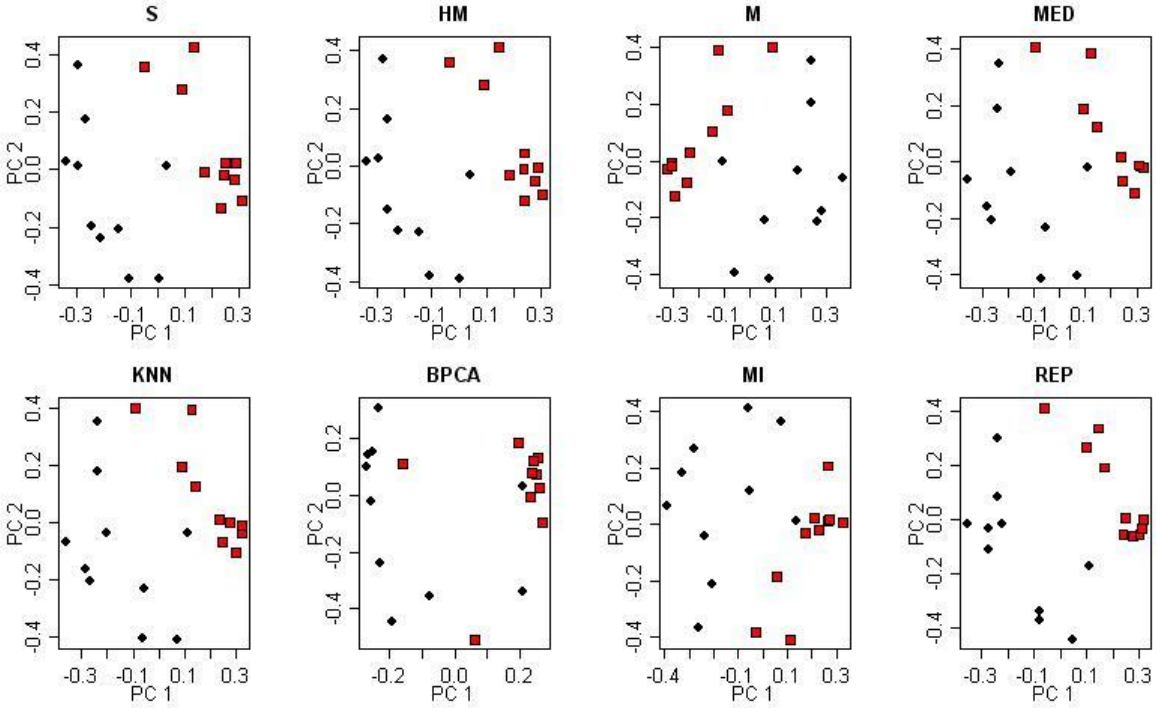

Figure A6 PCA scores plots for the DM datasets obtained after estimating missing data with the eight methods: controls (black diamonds), dinitrophenol exposed (red squares) *Daphnia magna*.

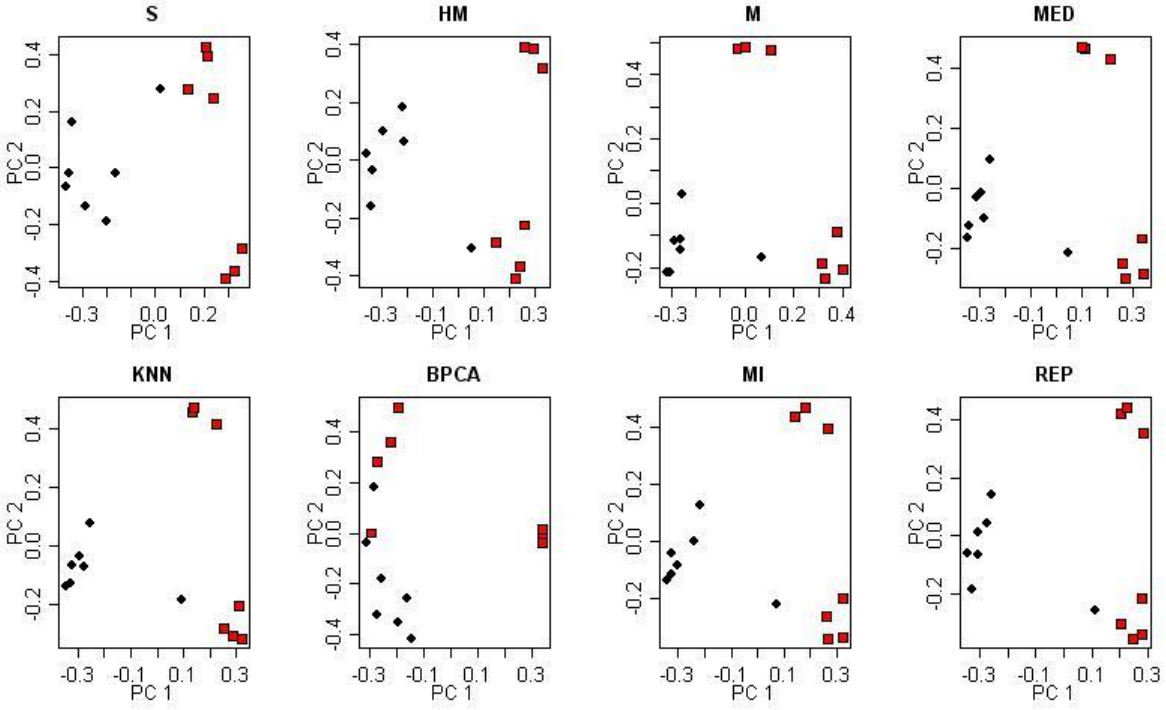

Figure A7 PCA scores plots for the HL datasets obtained after estimating missing data with the eight methods: cold phase (black diamonds), post-reperfusion (red squares) human liver extracts.

Table A6 Influence of missing data estimation algorithms on PCA: variance captured for principal components, peaks with missing data, and percentage of missing data out of the top 5% of peaks contributing towards separation along PC1 and PC2.

Dataset	PC	Parameter	S	HM	M	MED	KNN	BPCA	MI	REP	MEAN	RSD
CCL_n	1	Variance	17.00	17.60	21.12	21.35	32.13	41.29	23.15	17.09	23.84	36.10
		Peaks with MV	100.00	97.08	17.20	17.78	100.00	7.58	95.04	59.48	61.77	67.42
		MV [%]	56.04	51.21	2.24	2.17	67.33	1.75	64.87	24.78	33.80	86.57
	2	Variance	15.43	15.49	14.47	14.81	11.14	17.70	20.14	15.70	15.61	16.58
		Peaks with MV	100.00	100.00	37.61	37.90	91.84	7.29	98.54	79.30	69.06	52.45
		MV [%]	60.50	58.08	8.05	7.66	62.89	1.83	70.08	40.51	38.70	73.65
CCL_p	1	Variance	17.12	17.78	15.71	17.16	18.66	54.75	21.47	19.11	22.72	57.46
		Peaks with MV	100.00	98.66	42.86	43.75	92.86	31.25	91.52	83.93	73.10	39.21
		MV [%]	58.16	57.51	12.80	12.28	66.17	14.04	63.00	47.89	41.48	58.18
	2	Variance	14.18	13.83	15.28	15.83	15.82	28.16	13.60	14.34	16.38	29.53
		Peaks with MV	100.00	100.00	41.96	39.73	97.32	29.02	91.96	74.11	71.76	42.14
		MV [%]	59.35	57.99	10.19	7.42	70.41	13.42	61.78	36.66	39.65	65.79
DM	1	Variance	15.28	17.71	26.21	27.38	26.77	66.49	20.57	27.10	28.44	56.52
		Peaks with MV	100.00	100.00	33.33	34.76	40.00	20.00	57.62	60.00	55.71	54.30
		MV [%]	41.55	39.19	4.31	5.17	6.86	3.40	16.60	17.10	16.77	92.46
	2	Variance	9.51	10.25	12.66	13.23	12.84	8.93	10.90	13.02	11.42	15.13
		Peaks with MV	100.00	100.00	19.05	22.38	31.90	13.33	66.19	61.43	51.79	68.31
		MV [%]	38.12	36.43	2.86	3.33	5.90	1.81	22.10	16.10	15.83	94.91
HL	1	Variance	24.38	28.99	24.49	26.80	30.21	61.82	26.05	36.23	32.37	38.64
		Peaks with MV	100.00	98.89	52.22	54.44	77.78	21.11	84.44	87.78	72.08	37.93
		MV [%]	44.44	37.86	10.40	11.59	23.81	3.97	26.35	28.65	23.38	59.92
	2	Variance	15.38	17.46	19.92	21.35	21.47	16.92	20.01	20.39	19.11	11.72
		Peaks with MV	100.00	100.00	45.56	53.33	62.22	22.22	73.33	76.67	66.67	40.03
		MV [%]	40.08	36.59	9.52	12.78	15.79	4.21	20.40	20.95	20.04	62.79

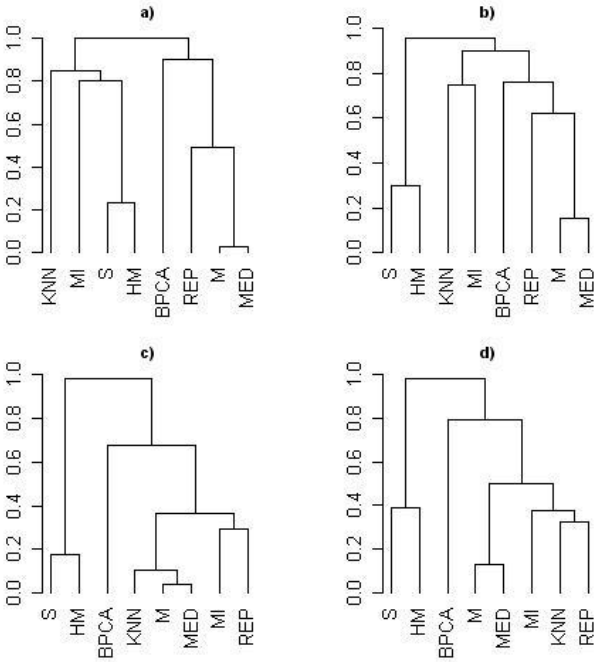

Figure A8 Hierarchical clustering (Euclidean distance, agglomeration method: complete) for eight different imputation methods for the top 5% of peaks contributing towards separation along PC1 for a) CCL_n, b) CCL_p, c) DM and d) HL datasets.

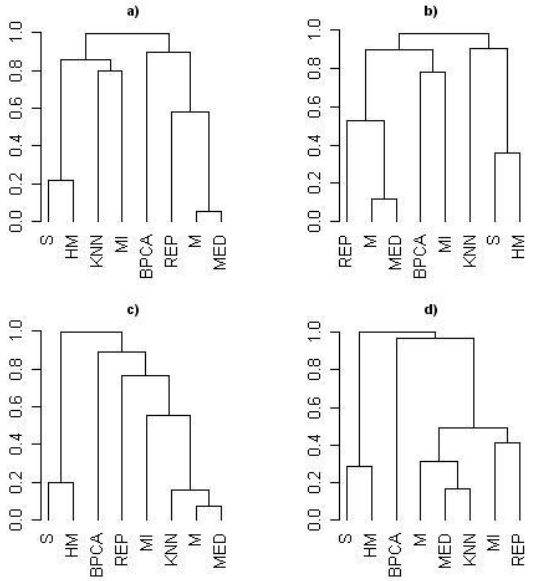

Figure A9 Hierarchical clustering (Euclidean distance, agglomeration method: complete) for eight different imputation methods for the top 5% of peaks contributing towards separation along PC2 for a) CCL_n, b) CCL_p, c) DM and d) HL datasets.

Table A7 Numerical values derived from the hierarchical clustering in Figure A8 showing the similarities between the eight missing value estimation methods in terms of which peaks contribute towards the separation along PC1. Similarities values expressed as R_t and measured between the top 5% peaks contributing towards separation along PC1.

		S	HM	M	MED	KNN	BPCA	MI	REP
CCL_p \ CCL_n	S	100.00	76.68	0.87	1.17	27.99	0.58	20.41	6.71
	HM	70.09	100.00	2.62	2.92	14.87	4.08	23.91	9.62
	M	4.46	16.96	100.00	97.08	0.00	21.28	3.50	51.31
	MED	7.56	13.84	84.82	100.00	0.00	20.70	3.79	51.02
	KNN	9.82	8.04	12.50	12.05	100.00	0.00	19.83	3.50
	BPCA	5.80	16.07	25.89	26.34	9.82	100.00	4.96	9.91
	MI	12.05	16.96	13.39	13.39	25.45	14.29	100.00	5.54
	REP	17.86	34.38	37.95	38.84	16.52	24.11	16.96	100.00
		S	HM	M	MED	KNN	BPCA	MI	REP
DM \ HL	S	100.00	82.38	1.90	3.33	4.76	3.33	19.05	22.86
	HM	61.11	100.00	5.24	6.67	11.43	6.19	23.33	30.95
	M	3.33	26.67	100.00	96.19	89.52	34.76	63.33	64.29
	MED	5.56	31.11	86.67	100.00	90.95	35.24	65.24	68.10
	KNN	26.67	47.78	62.22	67.78	100.00	35.24	69.52	70.48
	BPCA	2.22	10.00	23.33	23.33	21.11	100.00	33.81	32.38
	MI	22.22	42.22	52.22	55.56	65.56	21.11	100.00	70.48
	REP	34.44	65.56	50.00	55.56	67.78	21.11	62.22	100.00

The similarity measure is based on the number of overlapping peaks between the two ranked lists (R_a) and the order of the peaks in the list (R_b), i.e. if the same peak is present in the two ranked lists being compared, R_b is showing the difference in ranks (the positions of the peak) the bigger the value the ranks are more similar (peaks occupy similar position in the two ordered lists) (see Chapter 4 for details)

Table A8 Numerical values derived from the hierarchical clustering in Figure A9 showing the similarities between the eight missing value estimation methods in terms of which peaks contribute towards the separation along PC2. Similarities values expressed as R_t.

		S	HM	M	MED	KNN	BPCA	MI	REP
CCL_p\CCL$_n$	S	100.00	77.84	7.29	7.00	14.29	0.58	21.87	18.66
	HM	64.29	100.00	8.75	8.16	14.58	1.75	19.83	18.95
	M	4.46	9.38	100.00	94.46	11.37	17.78	6.12	41.69
	MED	1.76	4.91	88.39	100.00	11.08	17.78	5.83	43.44
	KNN	13.39	9.38	4.02	4.02	100.00	5.54	20.12	18.95
	BPCA	1.79	12.50	24.55	22.32	2.23	100.00	1.46	10.50
	MI	16.96	18.30	12.50	10.27	16.07	21.88	100.00	12.83
	REP	6.70	14.29	47.32	49.55	8.04	14.29	9.82	100.00
		S	HM	M	MED	KNN	BPCA	MI	REP
DM \ HL	S	100.00	80.00	0.48	1.90	2.38	0.95	20.48	7.14
	HM	71.11	100.00	1.90	4.29	4.76	3.81	20.00	13.33
	M	1.11	4.44	100.00	92.38	83.81	22.86	44.76	40.95
	MED	2.22	8.89	75.56	100.00	88.10	24.76	45.24	47.14
	KNN	4.44	10.00	68.89	83.33	100.00	22.38	46.19	48.10
	BPCA	0.00	1.11	21.11	16.67	13.33	100.00	10.95	15.71
	MI	5.56	8.89	56.67	63.33	65.56	8.89	100.00	23.33
	REP	13.33	21.11	51.11	65.56	71.11	3.33	58.89	100.00

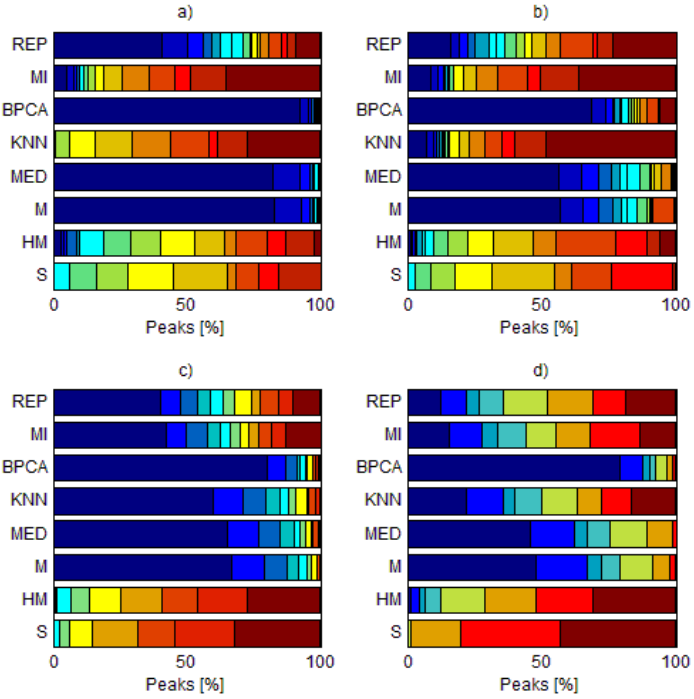

Figure A10 Boxplots (stacked) showing percentage out of top 5% of peaks contributing towards separation along PC1 containing various amounts of missing values per sample, ranging from zero (in blue) up to 15, 15, 10 and 7 (in brown) for the CCL_n, CCL_p, DM and HL datasets respectively.

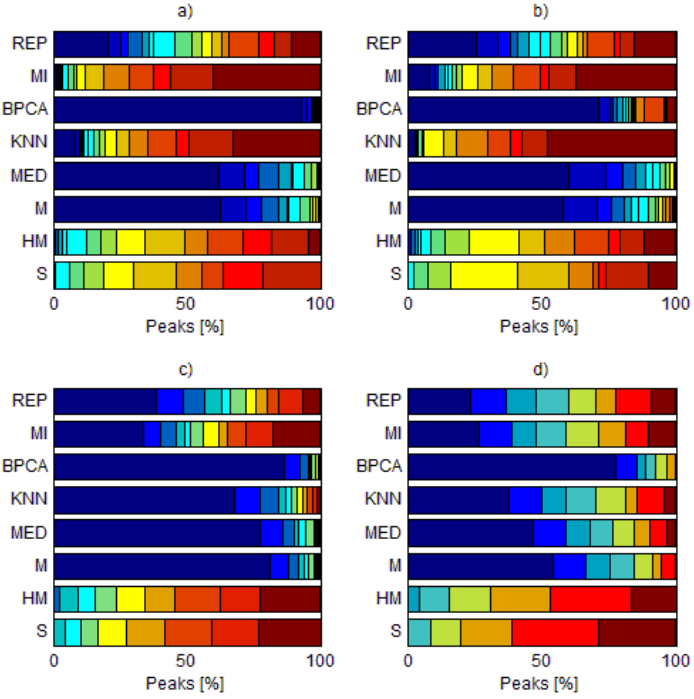

Figure A11 Boxplots (stacked) showing percentage out of top 5% of peaks contributing towards separation along PC2 containing various amounts of missing values per sample, ranging from zero (in blue) up to 15, 15, 10 and 7 (in brown) for the CCL_n, CCL_p, DM and HL datasets respectively. Distributions for PC1 (Figure A10) and for PC2 are shifted towards the middle and to the right, i.e. towards the larger number of missing data across the analysed samples when compared with the univariate equivalents.

Table A9 The five most commonly occurring patterns of missing values across samples for the top 5% of peaks contributing towards separation along PC1 after missing data estimation with the eight methods.

		Pattern / Percentage							
		S	HM	M	MED	KNN	BCPA	MI	REP
CCL_n	1	266 / 15.45	266 / 9.62	000 / 82.8	000 / 82.22	663 / 15.16	000 / 92.42	366 / 13.70	000 /40.52
	2	066 / 8.45	066 / 8.45	100 / 5.83	100 / 6.12	633 / 11.95	100 / 1.46	636 / 13.70	100 /6.12
	3	166 / 7.58	166 / 7.00	001 / 2.33	001 / 2.33	636 / 8.75	001 / 1.17	663 / 8.16	366 /3.21
	4	622 / 7.29	600 / 4.96	010 / 1.75	010 / 1.46	622 / 7.58	110 / 0.87	626 / 5.54	663 /3.21
	5	612 / 6.41	610 / 4.37	101 / 1.46	101 / 1.46	632 / 7.58	010 / 0.58	000 / 4.96	636 /2.62
CCL_p	1	661 / 17.41	336 / 8.48	000 / 57.14	000 / 56.25	663 / 26.34	000/ 68.75	663 / 18.30	000 / 16.07
	2	631 / 10.71	66 / 8.04	336 / 7.59	010 / 6.25	636 / 12.50	336 / 4.02	366 / 14.29	366 / 10.27
	3	066 / 8.04	661 / 8.04	010 / 5.80	326 / 2.68	366 / 9.38	636 / 4.02	662 / 10.27	336 / 9.38
	4	226 / 6.25	631 / 6.7	100 / 1.79	100 / 1.79	662 / 8.93	010 / 3.57	000 / 8.48	636 / 8.93
	5	136 / 5.80	660 / 5.36	110 / 1.79	110 / 1.79	000 / 7.14	366 / 1.79	633 / 6.70	663 / 4.46
DM	1	28 / 5.71	70 / 4.76	00 / 66.67	00 / 65.24	00 / 60.00	00 / 80.00	00 / 42.38	00 /40.00
	2	18 / 5.24	18 / 4.76	10 / 6.67	10 / 6.67	10 / 6.19	01 / 3.81	10 / 4.76	10 / 3.81
	3	70 / 4.76	61 / 4.29	01 / 5.71	01 / 5.24	01 / 4.76	10 / 3.33	20 / 4.29	01 / 3.81
	4	61 / 4.76	91 / 4.29	02 / 2.86	02 / 2.86	20 / 3.33	02 / 1.90	01 / 2.86	11 / 2.38
	5	91 / 4.76	19 / 4.29	11 / 2.86	11 / 2.38	02 / 2.86	11 / 1.43	11 / 2.38	20 / 2.38
HL	1	60 / 21.11	16 / 18.89	00 / 47.78	00 /45.56	00 / 22.22	00 / 78.89	00 / 15.56	16 / 13.33
	2	16 / 18.89	13 / 10.00	01/ 10.00	13/10.00	13/ 10.00	01 / 5.56	13 / 8.89	00 / 12.22
	3	70 / 18.89	15 / 10.00	13/ 10.00	01 /8.89	01 / 7.78	10 / 3.33	01 / 7.78	13 / 12.22
	4	50 / 12.22	50 / 7.78	10 / 8.89	10 /7.78	15 / 7.78	03 / 2.22	14 / 7.78	01 / 7.78
	5	15/ 10.00	70 / 7.78	14 / 5.56	14 /5.56	16 / 7.78	22 / 2.22	60 / 6.67	14 / 6.67

Table A10 The five most commonly occurring patterns of missing values across samples for the top 5% of peaks contributing towards separation along PC2 after missing data estimation with the eight methods.

		Pattern / Percentage							
		S	HM	M	MED	KNN	BCPA	MI	REP
CCL_n	1	626 / 13.12	662 / 7.87	000 / 62.39	000 / 62.10	366 / 16.91	000 / 92.71	663 / 25.66	000 / 20.7
	2	661 / 10.20	661 / 7.00	010 / 5.54	010 / 4.66	663 / 10.79	010 / 0.87	662 / 10.5	066 / 6.41
	3	662 / 8.16	626 / 5.54	060 / 2.92	100 / 2.92	662 / 8.75	100 / 0.87	636 / 7.29	366 / 4.37
	4	361 / 7.00	361 / 5.25	100 / 2.33	060 / 2.92	000 / 8.16	001 / 0.58	366 / 7.00	060 / 3.50
	5	362 / 4.96	606 / 4.37	001 / 2.04	001 / 1.75	266 / 5.25	110 / 0.58	336 / 4.66	166 / 3.50
CCL_p	1	626 / 16.07	636 / 10.71	000 / 58.04	000/ 60.27	366 / 20.09	000 / 70.98	366 / 12.95	000 / 25.89
	2	636 / 10.27	626 / 8.04	010 / 9.82	010 / 10.27	663 / 16.07	336 / 7.59	663 / 12.95	366 /9.38
	3	360 / 7.14	360 / 5.36	100 / 1.79	100 / 2.23	636 / 11.61	010 / 3.13	636 / 11.61	010 / 5.80
	4	261 / 6.70	066 / 5.36	020 / 1.79	020 / 1.79	362 / 7.59	366 / 2.68	000 / 8.04	066 / 5.80
	5	262 / 6.70	261 / 4.91	002 / 1.34	001 / 1.34	662 / 4.91	326 / 1.34	336 / 4.91	663 / 4.02
DM	1	44 / 7.14	44 / 7.62	00 / 80.95	00 / 77.62	00 / 68.1	00 / 86.67	00 / 33.81	00 / 38.57
	2	46 / 7.14	46 / 5.71	01 / 4.76	01 / 4.76	10 / 4.76	10 / 2.86	46 / 5.71	10 / 6.67
	3	34 / 5.24	55 / 4.76	10 / 2.38	10 / 3.81	01 / 4.76	01 / 2.86	01 / 3.81	20 / 3.81
	4	64 / 5.24	12 / 4.29	11 / 1.43	32 / 1.90	11 / 2.86	02 / 1.90	55 / 3.81	01 / 3.33
	5	55 / 5.24	32 / 4.29	20 / 1.43	11 / 1.43	20 / 2.38	11 / 0.95	11 / 3.33	21 / 2.86
HL	1	43 / 23.33	23 / 16.67	00 / 54.44	00 / 46.67	00 / 37.78	00 / 77.78	00 / 26.67	00 / 23.33
	2	23 / 16.67	42 / 13.33	01 / 8.89	01 / 7.78	01 / 8.89	01 / 5.56	01/ 10.00	01/ 10.00
	3	33 / 15.56	43 / 11.11	30 / 6.67	30 / 7.78	30 / 6.67	13 / 3.33	30 / 6.67	42 / 6.67
	4	42 / 10.00	33 / 10.00	20 / 5.56	20 / 5.56	20 / 5.56	14 / 3.33	20 / 4.44	20 / 5.56
	5	12 / 5.56	13 / 8.89	10 / 3.33	10 / 4.44	60 / 4.44	10 / 2.22	03 / 4.44	11 / 5.56

The most commonly occurring missing data sample patterns (in Table A9 and Table A10) included missing entries within a large proportion of the samples per biological group, for the majority of the estimation algorithms. The only exceptions resulted from methods M and MED, for which single missing data entries were dominant.

Table A11 Summary of which KEGG human pathways are 'active' (i.e. observed with 75% likelihood based on the top 5% of peaks contributing towards separation along PC1 between the control and two drug treated groups) in the CCL_n dataset, after estimating the missing values with eight different algorithms.

KEGG pathway	S	HM	M	MED	KNN	BPCA	MI	REP
Sphingolipid metabolism	X	X	X	X	-	X	X	X
Purine metabolism	X	X	-	X	X	X	X	X
Pyrimidine metabolism	X	X	-	-	-	X	X	X
Glycine, serine and threonine metabolism	X	X	-	-	X	X	X	-
Drug metabolism	X	X	-	-	X	-	X	X
Amino sugar and nucleotide sugar metabolism, Histidine metabolism	X	X	-	-	X	-	X	-
Cysteine and methionine metabolism, ABC transporters	X	X	-	-	-	X	X	-
Ascorbate and aldarate metabolism, Nicotinate and nicotinamide metabolism	X	X	-	-	-	X	-	-
Fatty acid biosynthesis	-	-	-	-	-	X	X	X
Lysine degradation	X	-	-	-	X	-	X	-
Arginine and proline metabolism	X	-	-	-	-	X	X	-
Glutathione metabolism	-	X	-	-	-	X	X	-
Drug metabolism - cytochrome P450	X	X	-	-	X	-	-	-
Pyruvate metabolism	X	X	-	-	-	-	-	-
Arachidonic acid metabolism	-	X	-	-	-	-	X	-
Biosynthesis of unsaturated fatty acids	-	-	-	-	-	X	-	X
Phenylalanine metabolism	-	-	-	-	X	-	X	-
Pentose phosphate pathway, Pentose and glucuronate interconversions, Glycerolipid metabolism, Glycerophospholipid metabolism, Tryptophan metabolism, Phenylalanine, tyrosine and tryptophan biosynthesis, beta-Alanine metabolism, Selenoamino acid metabolism, Pantothenate and CoA biosynthesis, Metabolism of xenobiotics by cytochrome P450, Aminoacyl-tRNA biosynthesis, Neuroactive ligand-receptor interaction, Taste transduction	-	-	-	-	-	X	-	-
alpha-Linolenic acid metabolism, Cyanoamino acid metabolism	-	-	-	-	-	-	X	-
Alanine, aspartate and glutamate metabolism	-	-	-	-	-	-	-	X
Lysine biosynthesis	-	-	-	-	X	-	-	-

Table A12 Summary of which KEGG human pathways are 'active' (i.e. observed with 75% likelihood based on the top 5% of peaks contributing towards separation along PC2 between the control and two drug treated groups) in the CCL_n dataset, after estimating the missing values with eight different algorithms.

KEGG pathway	S	HM	M	MED	KNN	BPCA	MI	REP
Pentose and glucuronate interconversions	-	X	X	X	X	X	X	X
Purine metabolism	X	X	X	X	-	X	X	X
Cysteine and methionine metabolism, ABC transporters	X	X	X	X	X	X	-	X
Sphingolipid metabolism	-	X	X	X	-	X	X	X
Pyrimidine metabolism	-	-	X	X	X	-	X	X
Glutathione metabolism	-	-	X	X	X	X	-	X
Pyruvate metabolism, Alanine, aspartate and glutamate metabolism	-	-	X	X	-	-	X	X
Drug metabolism - other enzymes	X	-	X	X	-	-	X	-
Amino sugar and nucleotide sugar metabolism	X	X	-	-	-	X	-	-
Sulfur metabolism	X	X	-	-	-	-	-	X
Biosynthesis of unsaturated fatty acids	-	-	-	X	X	X	-	-
Arginine and proline metabolism, Tryptophan metabolism, Neuroactive ligand-receptor interaction	-	-	X	X	-	X	-	-
Histidine metabolism	-	-	X	X	-	-	X	-
Tyrosine metabolism	-	-	X	X	-	-	-	X
Ascorbate and aldarate metabolism, Glycerophospholipid metabolism, Glycine, serine and threonine metabolism	-	-	-	-	-	X	-	X
Nicotinate and nicotinamide metabolism	-	-	-	-	-	X	X	-
Pantothenate and CoA biosynthesis	-	-	X	-	X	-	-	-
Metabolism of xenobiotics by cytochrome P450	X	X	-	-	-	-	-	-
Fatty acid biosynthesis, Fatty acid metabolism, Steroid biosynthesis, Glycerolipid metabolism, Selenoamino acid metabolism, Terpenoid backbone biosynthesis, Aminoacyl-tRNA biosynthesis	-	-	-	-	-	X	-	-
Arachidonic acid metabolism, Lysine degradation, Phenylalanine metabolism, Autoimmune thyroid disease	-	-	-	-	-	-	X	-
alpha-Linolenic acid metabolism	-	-	-	-	X	-	-	-

Table A13 Summary of which KEGG human pathways are 'active' (i.e. observed with 75% likelihood based on the top 5% of peaks contributing towards separation along PC1 between the control and two drug treated groups) in the CCL_p dataset, after estimating the missing values with eight different algorithms.

KEGG pathway	S	HM	M	MED	KNN	BPCA	MI	REP
Pyrimidine metabolism, Alanine, aspartate and glutamate metabolism	X	X	X	X	X	-	X	X
Drug metabolism - cytochrome P450	-	X	X	X	X	X	X	X
Glycolysis / Gluconeogenesis	X	X	X	X	X	-	X	-
Ether lipid metabolism	X	-	X	X	X	-	X	X
Glycine, serine and threonine metabolism	X	X	X	X	X	X	-	-
Cysteine and methionine metabolism	X	-	X	X	X	X	X	-
beta-Alanine metabolism	X	X	X	X	-	X	X	-
Sphingolipid metabolism	-	-	X	X	X	X	-	X
Linoleic acid metabolism	X	X	X	X	-	-	-	X
Purine metabolism, Glutathione metabolism, ABC transporters	-	-	X	X	-	X	X	X
Taurine and hypotaurine metabolism	-	-	X	X	X	X	X	-
Arginine and proline metabolism	-	-	X	X	-	X	-	X
Butanoate metabolism	-	X	-	-	X	-	X	-
Primary bile acid biosynthesis	-	-	X	-	-	X	-	X
Phenylalanine metabolism, Thiamine metabolism, Metabolism of xenobiotics by cytochrome P450	X	X	-	-	-	X	-	-
D-Arginine and D-ornithine metabolism	-	-	X	X	X	-	-	-
Vitamin B6 metabolism	-	-	X	X	-	-	X	-
Retinol metabolism, Epithelial cell signalling in Helicobacter pylori infection	-	-	X	X	-	-	-	X
Limonene and pinene degradation	-	X	-	-	-	X	X	-
Neuroactive ligand-receptor interaction	-	-	X	-	X	X	-	-
Oocyte meiosis, Progesterone-mediated oocyte maturation, Pathways in cancer, Prostate cancer	X	X	-	-	-	-	X	-
Pentose and glucuronate interconversions, Fatty acid biosynthesis, Fatty acid metabolism	-	-	X	X	-	-	-	-
Galactose metabolism, Glycerolipid metabolism, Pantothenate and CoA biosynthesis	X	X	-	-	-	-	-	-
Steroid biosynthesis	-	-	-	X	-	-	-	X
Arachidonic acid metabolism	-	X	-	-	-	-	-	X
alpha-Linolenic acid metabolism	-	-	-	X	-	-	X	-
Valine, leucine and isoleucine degradation	-	X	-	-	-	X	-	-
Ascorbate and aldarate metabolism, Amino sugar and nucleotide sugar metabolism, Histidine metabolism	-	-	-	-	-	-	X	-
Oxidative phosphorylation, Lysine degradation, Tryptophan metabolism, Parkinson's disease	-	-	-	-	-	X	-	-
Phenylalanine, tyrosine and tryptophan biosynthesis, Ubiquinone and other terpenoid-quinone biosynthesis	-	-	-	-	X	-	-	-

Table A14 Summary of which KEGG human pathways are 'active' (i.e. observed with 75% likelihood based on the top 5% of peaks contributing towards separation along PC2 between the control and two drug treated groups) in the CCL_p dataset, after estimating the missing values with eight different algorithms.

KEGG pathway	S	HM	M	MED	KNN	BPCA	MI	REP
Pentose and glucuronate interconversions	-	X	X	X	X	X	X	X
Purine metabolism	X	X	X	X	-	X	X	X
Cysteine and methionine metabolism, ABC transporters	X	X	X	X	X	X	-	X
Sphingolipid metabolism	-	X	X	X	-	X	X	X
Pyrimidine metabolism	-	-	X	X	X	-	X	X
Glutathione metabolism	-	-	X	X	X	X	-	X
Pyruvate metabolism, Alanine, aspartate and glutamate metabolism	-	-	X	X	-	-	X	X
Drug metabolism - other enzymes	X	-	X	X	-	-	X	-
Amino sugar and nucleotide sugar metabolism	X	X	-	-	-	X	-	-
Sulfur metabolism	X	X	-	-	-	-	-	X
Biosynthesis of unsaturated fatty acids	-	-	-	X	X	X	-	-
Arginine and proline metabolism, Tryptophan metabolism, Neuroactive ligand-receptor interaction	-	-	X	X	-	X	-	-
Histidine metabolism	-	-	X	X	-	-	X	-
Tyrosine metabolism	-	-	X	X	-	-	-	X
Ascorbate and aldarate metabolism, Glycerophospholipid metabolism, Glycine, serine and threonine metabolism	-	-	-	-	-	X	-	X
Nicotinate and nicotinamide metabolism	-	-	-	-	-	X	X	-
Pantothenate and CoA biosynthesis	-	-	X	-	X	-	-	-
Metabolism of xenobiotics by cytochrome P450	X	X	-	-	-	-	-	-
Fatty acid biosynthesis, Fatty acid metabolism, Steroid biosynthesis, Glycerolipid metabolism, Selenoamino acid metabolism, Terpenoid backbone biosynthesis, Aminoacyl-tRNA biosynthesis	-	-	-	-	-	X	-	-
Arachidonic acid metabolism, Lysine degradation, Phenylalanine metabolism, Autoimmune thyroid disease	-	-	-	-	-	-	X	-
alpha-Linolenic acid metabolism	-	-	-	-	X	-	-	-

Table A 15 Summary of which KEGG human pathways are 'active' (i.e. observed with 75% likelihood based on the top 5% of peaks contributing towards separation along PC1 between the cold phase and post-reperfusion groups) in the HL dataset, after estimating the missing values with eight different algorithms.

KEGG pathway	S	HM	M	MED	KNN	BPCA	MI	REP
Sphingolipid metabolism	X	X	X	X	X	X	X	-
Primary bile acid biosynthesis	-	-	X	X	X	X	X	X
Glycerophospholipid metabolism	X	-	X	X	X	X	X	-
Glycine, serine and threonine metabolism, Cysteine and methionine metabolism	X	X	X	X	-	X	X	-
Arginine and proline metabolism	-	X	X	X	X	-	X	X
Aminoacyl-tRNA biosynthesis	X	X	X	X	-	-	X	-
Purine metabolism	-	X	-	-	-	X	X	X
Alanine, aspartate and glutamate metabolism	X	X	-	-	-	-	X	X
Lysine degradation	X	-	-	-	X	X	X	-
ABC transporters	-	-	X	X	-	X	X	-
Neuroactive ligand-receptor interaction	X	X	X	-	-	-	X	-
Metabolism of xenobiotics by cytochrome P450	X	X	-	-	-	-	-	X
Histidine metabolism, beta-Alanine metabolism	-	-	X	X	-	-	-	-
Glutathione metabolism	-	-	-	-	-	X	X	-
Calcium signalling pathway, Fc gamma R-mediated phagocytosis	X	X	-	-	-	-	-	-
Glyoxylate and dicarboxylate metabolism, Nitrogen metabolism, Cyanoamino acid metabolism, Thiamine metabolism, Nicotinate and nicotinamide metabolism	-	-	-	-	-	-	X	-
Oxidative phosphorylation, Taurine and hypotaurine metabolism, Parkinson's disease	-	-	X	-	-	-	-	-
Ether lipid metabolism, Arachidonic acid metabolism, Phenylalanine metabolism	-	-	-	-	-	X	-	-
Valine, leucine and isoleucine degradation	-	X	-	-	-	-	-	-

SM: Missing data imputation algorithms performance

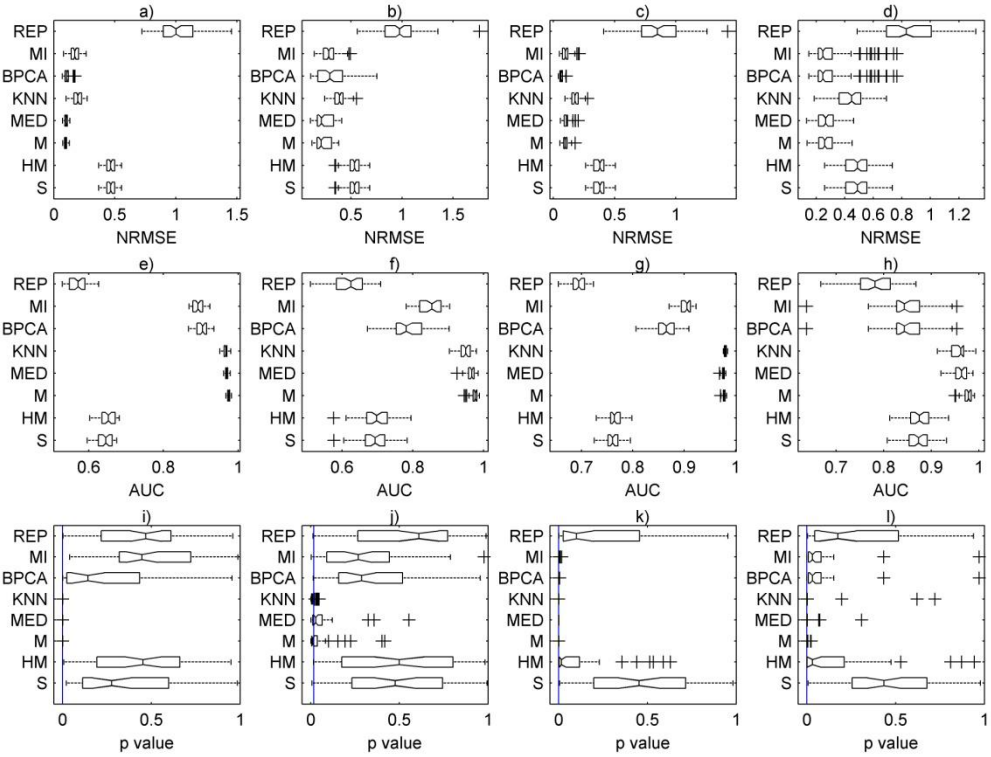

Figure A12 Analyses of four DI FT-ICR MS datasets after first introducing and then estimating missing data in the 'complete' datasets as MCAR (average of 100 runs). Boxplots of NRMSE values for the a) CCL_n, b) CCL_p, c) DM and d) HL datasets; boxplots of area under ROC curves (AUC) for e) CCL_n, f) CCL_p, g) DM and h) HL datasets; and distribution of p values (ANOVA or t test on PC scores) for i) CCL_n (PC2 axis), j) CCL_p (PC2 axis), k) DM (PC1 axis) and l) HL (PC1 axis) datasets, where the vertical lines indicate the p values for the complete datasets and therefore represent the ideal result following missing value estimation.

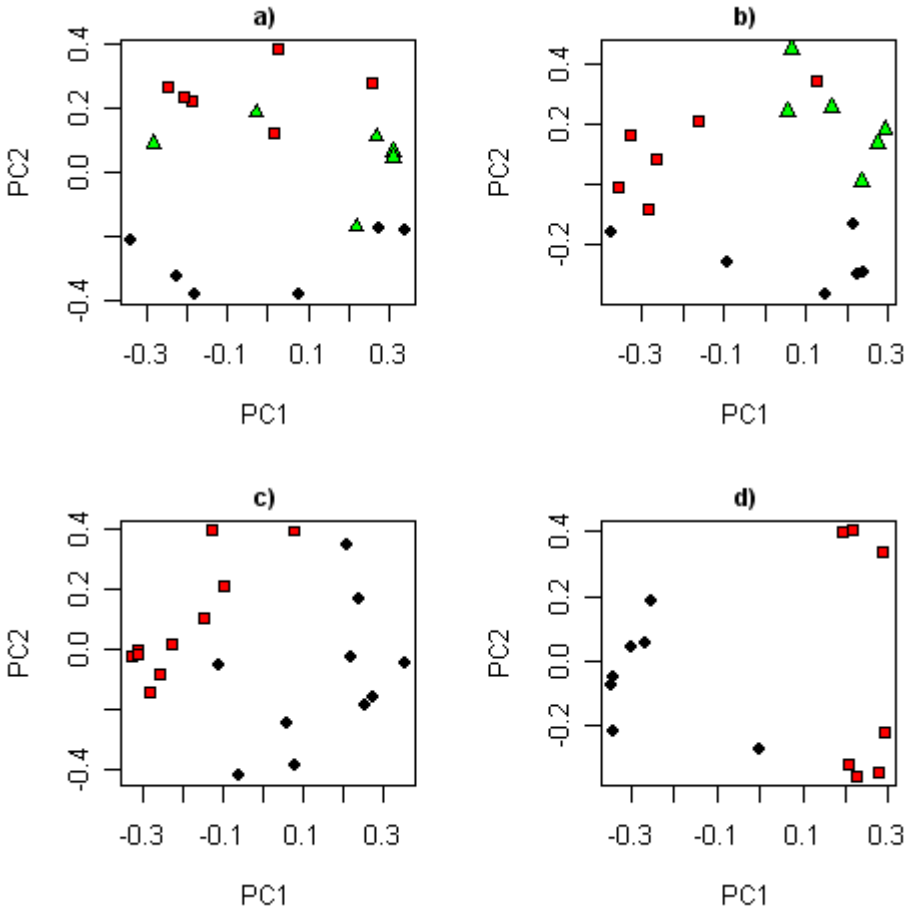

Figure A13 PCA scores plots on the 'complete' datasets (i.e. after excluding any peaks that have missing values) for a) CCL_n, b) CCL_p, c) DM and d) HL datasets. Symbols as defined in Figures A5-A7.

Table A16 Mean NRMSE and corresponding relative standard deviation (RSD) values for N=100 runs when introducing missing data as MCAR and MNAR; complementary to the Figure 5 and Figure A13 boxplots.

MEAN	S	HM	M	MED	KNN	BPCA	MI	REP
CCL_n	0.47, 0.16*	0.47, 0.16	0.10, 0.03	0.10, 0.03	0.19, 0.07	0.11, 0.03	0.17, 0.12	1.02, 0.34
CCL_p	0.53, 0.22	0.53, 0.22	0.23, 0.09	0.23, 0.09	0.39, 0.15	0.31, 0.10	0.29, 0.20	0.97, 0.38
DM	0.38, 0.15	0.38, 0.15	0.10, 0.03	0.11, 0.03	0.18, 0.06	0.06, 0.02	0.10, 0.04	0.87, 0.29
HL	0.48, 0.18	0.48, 0.18	0.27, 0.10	0.27, 0.10	0.44, 0.15	0.29, 0.08	0.29, 0.08	0.86, 0.28
RSD	S	HM	M	MED	KNN	BPCA	MI	REP
CCL_n	9.51, 23.85	9.52, 23.95	12.41, 30.9	14.00, 35.53	19.77, 34.31	21.86, 44.75	23.78, 25.53	16.21, 42.39
CCL_p	13.35, 26.74	13.36, 26.77	33.03, 72.06	36.07, 79.08	15.16, 35.74	49.03, 89.91	26.87, 51.31	19.92, 44.24
DM	15.34, 34.8	15.34, 34.8	22.08, 48.94	25.46, 52.87	22.11, 48.04	21.82, 36.33	33.97, 61.35	22.76, 55.65
HL	22.51, 56.11	22.51, 56.12	25.35, 60.12	29.00, 66.19	25.75, 68.5	47.42, 51.99	47.42, 51.99	24.15, 68.05

* First value corresponds to MCAR, the second to MNAR

Table A17 Mean AUC and corresponding relative standard deviation (RSD) values for N=100 runs when introducing missing data as MCAR and MNAR, complementary to the Figure 5 and Figure A13 boxplots.

MEAN	S	HM	M	MED	KNN	BPCA	MI	REP
CCL_n	0.64, 0.8*	0.65, 0.81	0.98, 0.93	0.97, 0.95	0.97, 0.91	0.90, 0.92	0.89, 0.91	0.57, 0.79
CCL_p	0.69, 0.83	0.7, 0.83	0.97, 0.9	0.96, 0.94	0.95, 0.89	0.79, 0.90	0.85, 0.90	0.62, 0.83
DM	0.76, 0.83	0.76, 0.83	0.98, 0.97	0.98, 0.97	0.98, 0.97	0.86, 0.92	0.9, 0.90	0.69, 0.79
HL	0.87, 0.90	0.87, 0.90	0.98, 0.98	0.96, 0.97	0.96, 0.96	0.85, 0.86	0.85, 0.86	0.78, 0.83
RSD	S	HM	M	MED	KNN	BPCA	MI	REP
CCL_n	3.45, 2.31	3.39, 2.30	0.37, 1.24	0.42, 0.86	0.62, 1.35	1.96, 1.81	1.69, 1.52	4.23, 2.41
CCL_p	6.03, 4.02	6.27, 3.96	0.97, 3.32	1.25, 1.66	1.78, 3.10	6.81, 2.97	4.19, 2.48	7.96, 4.88
DM	2.12, 1.58	2.1, 1.56	0.27, 0.45	0.27, 0.44	0.22, 0.52	2.67, 1.64	1.35, 1.39	2.12, 2.39
HL	3.03, 3.22	3.03, 3.18	0.92, 0.80	1.66, 1.11	2.05, 1.96	5.65, 3.91	5.65, 3.91	5.38, 4.79

* First value corresponds to MCAR, the second to MNAR

Table A18 P values (from t test or ANOVA) with corresponding RSD values for the PC1 and PC2 scores for the four datasets when introducing missing data as MCAR; complementary to the Figure A13 boxplots.

		MEAN				RSD [%]			
		CCL_n	CCL_p	DM	HL	CCL_n	CCL_p	DM	HL
PC1	S	0.51	0.50	0.46	0.47	52.33	66.03	64.43	58.67
	HM	0.42	0.41	0.10	0.15	85.02	74.24	173.03	154.58
	M	0.41	0.20	0.00	0.00	5.09	82.85	35.25	420.16
	MED	0.39	0.12	0.00	0.01	4.74	123.64	31.51	502.77
	KNN	0.41	0.51	0.00	0.03	3.13	37.56	26.20	437.61
	BPCA	0.52	0.61	0.00	0.07	31.39	46.32	137.91	213.33
	MI	0.50	0.60	0.00	0.07	50.53	41.03	192.61	213.33
	REP	0.53	0.46	0.27	0.29	59.13	55.74	110.52	96.36
		CCL_n	CCL_p	DM	HL	CCL_n	CCL_p	DM	HL
PC2	S	0.36	0.47	0.41	0.37	82.73	63.27	69.8	73.8
	HM	0.45	0.49	0.30	0.41	65.30	66.94	94.63	78.03
	M	0.00	0.05	0.07	0.64	151.60	195.22	22.79	44.4
	MED	0.00	0.06	0.06	0.56	143.29	168.45	19.88	54.66
	KNN	0.00	0.01	0.10	0.49	213.50	111.03	13.19	66.43
	BPCA	0.26	0.36	0.67	0.30	110.58	72.57	37.12	101.1
	MI	0.50	0.30	0.57	0.31	56.24	76.83	50.20	101.1
	REP	0.45	0.53	0.26	0.46	59.77	56.14	90.56	65.31

* Highlighted in blue datasets with no missing data introduced for which for the corresponding PC1 there is a significant difference (at 0.05 level) while performing t test or ANOVA on scores values; highlighted in yellow methods for which there is a significant difference after introducing missing data and estimating them with the specific method.

Table A19 P values (from t test or ANOVA) with corresponding RSD values for the PC1 and PC2 scores for the four datasets when introducing missing data as MNAR; complementary to the Figure A5 boxplots.

		MEAN				RSD [%]			
		CCL_n	CCL_p	DM	HL	CCL_n	CCL_p	DM	HL
PC1	S	0.41	0.46	0.43	0.37	74.05	63.98	77.98	79.76
	HM	0.27	0.28	0.02	0.04	95.39	94.68	578.34	344.29
	M	0.35	0.08	0.00	0.00	4.27	52.86	29.2	148.98
	MED	0.35	0.06	0.00	0.00	3.97	38.26	26.84	132.04
	KNN	0.42	0.33	0.00	0.00	27.23	61.52	22.41	370.92
	BPCA	0.50	0.75	0.00	0.03	10.07	9.7	30.24	36.8
	MI	0.49	0.73	0.00	0.03	17.64	15.6	85.75	36.8
	REP	0.46	0.70	0.01	0.10	50.47	27.15	289.95	156.12
		CCL_n	CCL_p	DM	HL	CCL_n	CCL_p	DM	HL
PC2	S	0.52	0.48	0.45	0.46	59.57	64.1	72.35	53.43
	HM	0.28	0.43	0.34	0.27	103.96	72.03	87.31	100.86
	M	0.00	0.03	0.05	0.83	94.68	68.21	15.84	17.19
	MED	0.00	0.04	0.05	0.81	76.94	40.33	14.25	17.05
	KNN	0.00	0.04	0.07	0.72	390.92	222.44	16.86	28.52
	BPCA	0.64	0.12	0.57	0.01	51.01	44.17	22.49	142.09
	MI	0.55	0.11	0.54	0.01	56.86	54.44	38.17	142.09
	REP	0.58	0.18	0.45	0.06	44.44	108.17	57.95	294.54

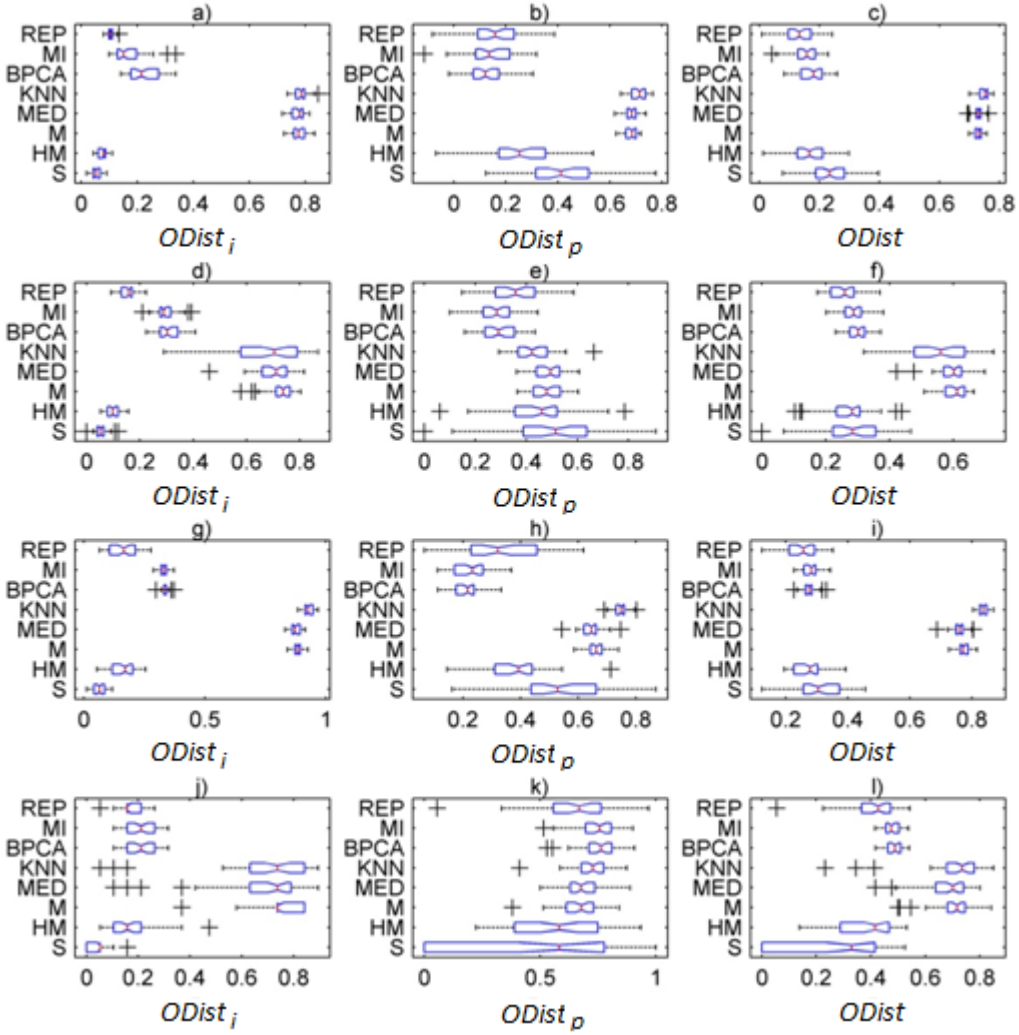

Figure A14 Similarities between the top 5% of peaks contributing towards the separation along PC1 and PC2 expressed as $ODist_i$, $ODist_p$ and *ODist* when introducing missing data as MCAR; a-c) CCL_n for PC2 axis, d-f) CCL_p for PC2 axis, g-i) DM for PC1 axis, and j-l) HL for PC1 axis.

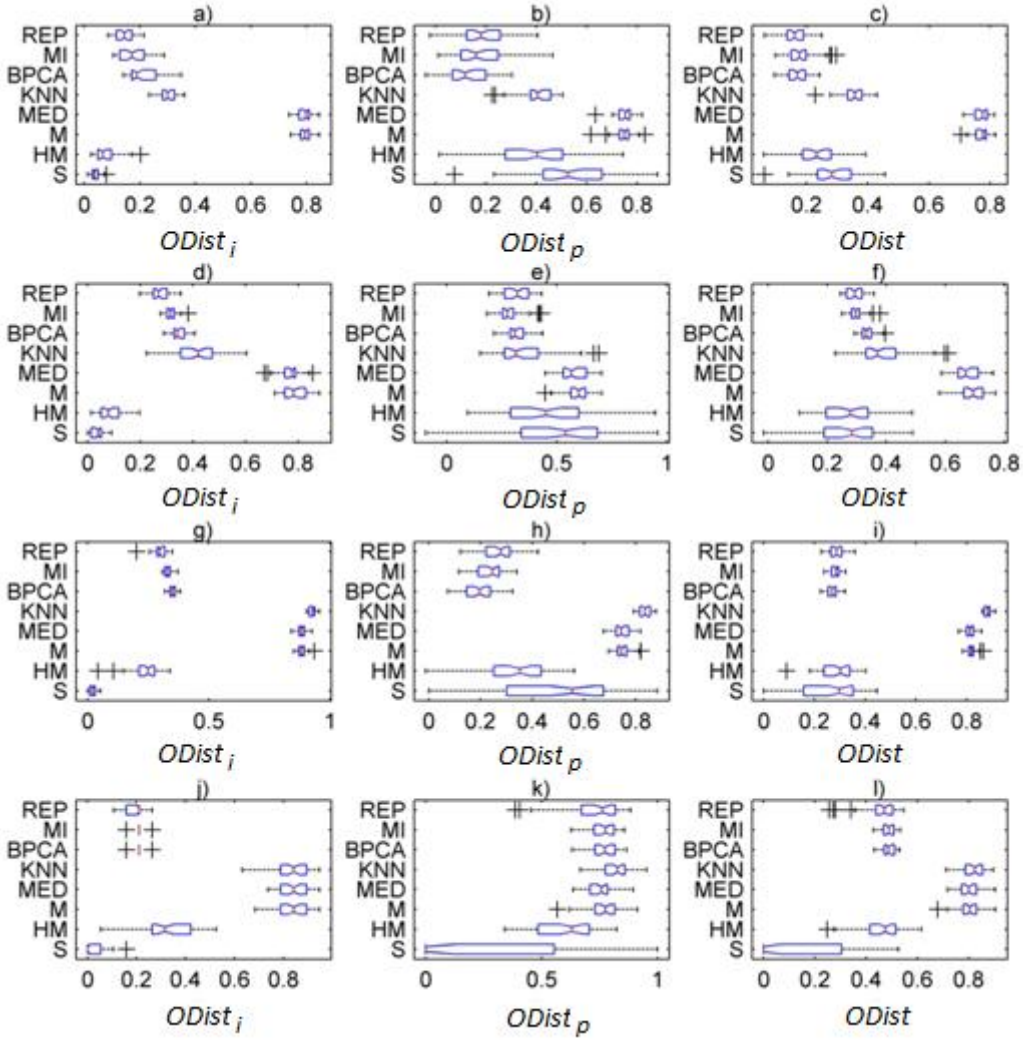

Figure A15 Similarities between the top 5% of peaks contributing towards the separation along PC1 and PC2 expressed as as $ODist_i$, $ODist_p$ and *ODist* when introducing missing data as MNAR; a-c) CCL_n for PC2 axis, d-f) CCL_p for PC2 axis, g-i) DM for PC1 axis, and j-l) HL for PC1 axis.

Table A20 Mean similarity measure expressed as R_a, R_b and R_t across N=100 runs for the top 5% of peaks contributing towards separation along PC1 and PC2 while introducing missing data as MCAR and as MNAR.

			S	HM	M	MED	KNN	BCPA	MI	REP
MCAR	CCL_n (PC2)	$ODist_i$	0.05	0.07	0.78	0.77	0.78	0.23	0.16	0.11
		$ODist_p$	0.41	0.25	0.68	0.68	0.71	0.13	0.14	0.17
		$ODist$	0.23	0.16	0.73	0.73	0.75	0.18	0.15	0.14
	CCL_p (PC2)	$ODist_i$	0.06	0.10	0.73	0.71	0.68	0.31	0.29	0.16
		$ODist_p$	0.52	0.44	0.48	0.49	0.42	0.29	0.28	0.36
		$ODist$	0.29	0.27	0.6	0.60	0.55	0.30	0.29	0.26
	DM (PC1)	$ODist_i$	0.07	0.16	0.88	0.88	0.93	0.34	0.33	0.16
		$ODist_p$	0.56	0.38	0.66	0.64	0.75	0.21	0.23	0.35
		$ODist$	0.31	0.27	0.77	0.76	0.84	0.27	0.28	0.25
	HL (PC1)	$ODist_i$	0.05	0.17	0.75	0.69	0.71	0.21	0.21	0.18
		$ODist_p$	0.48	0.59	0.67	0.68	0.72	0.76	0.75	0.64
		$ODist$	0.26	0.38	0.71	0.68	0.71	0.48	0.48	0.41
MNAR	CCL_n (PC2)	$ODist_i$	0.04	0.07	0.79	0.79	0.30	0.21	0.18	0.14
		$ODist_p$	0.53	0.4	0.75	0.75	0.40	0.13	0.18	0.18
		$ODist$	0.28	0.24	0.77	0.77	0.35	0.17	0.18	0.16
	CCL_p (PC2)	$ODist_i$	0.04	0.09	0.78	0.77	0.42	0.35	0.31	0.27
		$ODist_p$	0.5	0.45	0.60	0.58	0.35	0.32	0.28	0.31
		$ODist$	0.27	0.27	0.69	0.67	0.38	0.33	0.30	0.29
	DM (PC1)	$ODist_i$	0.02	0.24	0.89	0.88	0.93	0.35	0.33	0.3
		$ODist_p$	0.48	0.34	0.75	0.74	0.84	0.19	0.23	0.28
		$ODist$	0.25	0.29	0.82	0.81	0.88	0.27	0.28	0.29
	HL (PC1)	$ODist_i$	0.03	0.33	0.85	0.85	0.82	0.21	0.21	0.19
		$ODist_p$	0.27	0.61	0.77	0.75	0.81	0.76	0.76	0.73
		$ODist$	0.15	0.47	0.81	0.80	0.82	0.49	0.49	0.46

Table A21 Percentage of estimated missing data whose values are above the applied signal-to-noise ratio (SNR) threshold as defined for the original datasets.

	M	MED	KNN
CCL_n	94.56	94.62	65.93
CCL_p	70.58	70.98	37.91
DM	72.29	76.36	57.73
HL	77.06	76.96	58.3

Appendix B

Application of metabolomics to investigate the process of human orthotopic liver transplantation: a proof-of-principle study. Supplementary Material

Contents

Table B1 Liver function tests 48 hours following OLT

Patient	AST [IU/L]	AKP [IU/L]	Bilirubin [µmol/L]	Urea [mg/dL]	Creatine [µmol/L]	INR
H1	387	132	85	14	155	1.8
H2	115	212	122	18.7	203	1
H3	920	101	44	6.3	125	1.5
H4	498	106	76	17.3	2.6	1.9
H5	874	130	103	11.8	161	1.7
H6	255	184	44	11.4	134	1.2
H7	291	156	230	17.1	205	1.4
H8	3939	103	79	16.4	310	1.7

Abbrevations: AST, aspartate aminotransferase, AKP, INR

Table B2 Time of microdialysate collection following OLT

Patient	T_{CEAD1} [h]	T_{CEAD2} [h]	T_{CEAD3} [h]	T_{CEAD4} [h]	T_{CEAD5} [h]	T_{CEAD6} [h]	T_{CEAD7} [h]
H1	5	9	17	21	27	nc	45
H2	5	9	15	21	27	39	45
H3	5	9	15	nc	27	39	nc
H4	6	9	14	21	27	nc	46
H5	5	9	15	nc	26	40	46
H6	5	9	15	21	27	39	45
H7	5	9	15	21	27	39	45

Abbreviations: nc, not collected

Table B3 Histology of liver biopsies: steatosis levels (micro and macro) assessed during cold phase and post reperfusion

	Cold phase		Post reperfusion	
Patient	micro [%]	macro [%]	micro [%]	macro [%]
H1	<5	<5	25	<5
H2	10-15	<5	50-60	<5
H3	20	25-30	50	25
H4	50	5	70	10
H5	10-15	5	20-25	5
H6	30	0	60	<5
H7	35	<5	40	<5
H8	35	70	50-60	25

Table B4 Compounds from UW solution found in the liver biopsy mass spectra with other possible metabolite assignments.

Compound name	Peak index	Adduct	Mass error (ppm)	Other metabolites with same accurate mass
Adenosine	640	[+ H]	0.34	Deoxyguanosine;
Adenosine	717	[+ Na]	0.06	
Adenosine	661	[- H]	0.00	
Adenosine	992	[+ Cl]	-0.21	
Adenosine	1019	[+Cl[37]]	-0.17	
Adenosine	1274	[+ Hac - H]	-0.56	
Allopurinol	64	[+ H]	-0.17	Hypoxanthine
Allopurinol	113	[+ Na]	-0.18	
Allopurinol	191	[+ K]	-0.17	
Allopurinol	37	[- H]	-0.11	
Allopurinol	102	[+ Cl]	-0.04	
Allopurinol	176	[+ Hac - H]	0.08	
Citrate	379	[+ Na]	-0.16	Oxaloacetate;Isocitrate; (1R,2S)-1-Hydroxypropane-1,2,3-tricarboxylate; 5-Dehydro-4-deoxy-D-glucarate; 2,5-Didehydro-D-gluconate; Carboxymethyloxysuccinate; (4R,5S)-4,5,6-Trihydroxy-2,3-dioxohexanoate; (1S,2S)-1-Hydroxypropane-1,2,3-tricarboxylate;
Citrate	464	[+ K]	0.10	
Citrate	483	[2Na-H]	-0.42	
Citrate	645	[2K -H]	0.29	
Citrate	171	[- H]	-0.06	
Citrate	415	[+ Cl]	-0.01	
Glutathione	783	[+ H]	0.24	-
Glutathione	865	[+ Na]	-0.39	
Glutathione	965	[+ K]	-0.24	
Glutathione	1024	[2Na-H]	-0.58	
Glutathione	1178	[2K -H]	0.24	
Glutathione	1041	[- H]	0.04	
Glutathione	1404	[+ Cl]	-0.17	
Lactobionic acid	1055	[+ H]	-0.26	-
Lactobionic acid	1164	[+ Na]	0.24	
Lactobionic acid	1247	[+ K]	-0.03	
Lactobionic acid	1295	[2Na-H]	-0.01	
Lactobionic acid	1472	[2K -H]	0.05	
Lactobionic acid	1537	[- H]	0.40	
Mannitol	352	[+ Na]	0.10	D-Sorbitol; D-Iditol; L-Iditol; Galactitol; L-Glucitol;
Mannitol	406	[+ K]	0.09	
Mannitol	444	[2Na-H]	0.07	
Mannitol	590	[2K -H]	-0.11	
Mannitol	129	[- H]	-0.14	
Mannitol	339	[+ Cl]	0.24	
Mannitol	355	[+Cl[37]]	0.10	
Mannitol	498	[+ Hac - H]	0.14	

Table B5 Fold changes of compounds from UW solution found in the liver biopsy mass spectra, with p values (t test with Benjamini and Hochberg correction for multiple testing).

Compound name	Peak index	Adduct	Fold change	p value
Adenosine	640	[+H]	1.25	0.688
Adenosine	717	[+Na]	2.67	0.081
Adenosine	661	[- H]	1.89	0.128
Adenosine	992	[+ Cl]	3.85	0.044
Adenosine	1019	[+Cl[37]]	4.21	0.042
Adenosine	1274	[+ Hac - H]	1.76	0.360
Allopurinol	64	[+H]	0.96	0.921
Allopurinol	113	[+Na]	0.88	0.685
Allopurinol	191	[+K]	0.36	0.121
Allopurinol	37	[- H]	0.68	0.300
Allopurinol	102	[+ Cl]	1.14	0.801
Allopurinol	176	[+ Hac - H]	1.39	0.351
Citrate	379	[+Na]	0.17	0.182
Citrate	464	[+K]	0.10	0.167
Citrate	475	[+K[41]]	0.11	0.199
Citrate	483	[2Na-H]	0.38	0.178
Citrate	645	[2K-H]	0.19	0.115
Citrate	171	[- H]	0.05	0.049
Citrate	415	[+ Cl]	0.77	0.585
Glutathione	783	[+H]	0.61	0.557
Glutathione	865	[+Na]	0.91	0.865
Glutathione	965	[+K]	0.79	0.779
Glutathione	982	[+K[41]]	0.84	0.818
Glutathione	1024	[2Na-H]	2.58	0.303
Glutathione	1178	[2K-H]	1.32	0.738
Glutathione	1041	[- H]	0.72	0.596
Glutathione	1404	[+ Cl]	1.19	0.768
Lactobionic acid	1055	[+H]	0.35	0.324
Lactobionic acid	1164	[+Na]	0.85	0.726
Lactobionic acid	1247	[+K]	0.52	0.324
Lactobionic acid	1259	[+K[41]]	0.52	0.321
Lactobionic acid	1295	[2Na-H]	1.66	0.513
Lactobionic acid	1472	[2K-H]	0.64	0.500
Lactobionic acid	1537	[- H]	0.41	0.078
Mannitol	352	[+Na]	0.09	0.048
Mannitol	406	[+K]	0.06	0.096
Mannitol	422	[+K[41]]	0.06	0.110
Mannitol	444	[2Na-H]	0.80	0.636
Mannitol	590	[2K-H]	0.45	0.243
Mannitol	129	[- H]	0.09	0.043
Mannitol	339	[+ Cl]	0.08	0.049
Mannitol	355	[+Cl[37]]	0.12	0.044
Mannitol	498	[+ Hac - H]	0.06	0.044

Yellow highlighting indicates those compounds with average abundances that were increased post reperfusion, green highlighting indicates those compounds that change significantly ($p < 0.05$) post reperfusion.

Table B6 Peaks with putative metabolite assignments and fold changes (post reperfusion / cold phase) involved in energy metabolites (based on KEGG database classification).

Peak index	m/z	Fold change	p value	Formula	Putative Metabolite	Adduct	KEGG ID	P [%]*
1	76.03931	1.69	0.3402	C2H5NO2	Glycine	H	C00037	100
21	90.97665	6.42	0.1834	CH2O2	Formate	2*Na-H	C00058	100
46	122.92452	2.47	0.2712	CH2O2	Formate	2*K(39)-H	C00058	100
187	174.89594	2.77	0.2773	H3PO4	Orthophosphate	2*K(39)-H	C00009	100
223	184.05802	2.75	0.3425	C6H11NO4	N-Methyl-L-glutamate	Na	C01046	66.67
1541	450.01887	4.77	0.0606	C10H15N5O10P2	ADP	Na	C00008	66.67
1653	472.00083	14.02	0.2087	C10H15N5O10P2	ADP	2*Na-H	C00008	66.67
5	81.97013	2.07	0.1957	CH3NO	Formamide	K(39)-2*H	C00488	100
7	82.97193	2.89	0.0814	CH2O2	Formate	Cl(37)	C00058	100
19	96.96011	0.17	0.0410	H2SO4	Sulfate	'-H'	C00059	100
21	96.96962	1.35	0.3852	H3PO4	Orthophosphate	'-H'	C00009	100
34	132.94639	1.68	0.0645	H3PO4	Orthophosphate	Cl(35)	C00009	100
49	145.01425	0.15	0.1145	C5H6O5	2-Oxoglutarate	'-H'	C00026	100
187	198.9179	1.24	0.5621	P2H4O7	Pyrophosphate	Na-2*H	C00013	100
344	218.06704	0.77	0.4351	C8H13NO6	O-Succinyl-L-homoserine	'-H'	C01118	100
2183	448.00429	6.14	0.0558	C10H15N5O10P2	ADP	Na-2*H	C00008	66.67
2277	463.97834	3.93	0.2345	C10H15N5O10P2	ADP	K(39)-2*H	C00008	66.67

Yellow highlighting indicates those peaks which increased in concentration post reperfusion. The final column shows the probability (in %) that a given peak is involved in energy metabolism; data filtered to retain peaks with P > 60%; p values corrected for multiple testing with Benjamini-Hochberg

* The probability that a given peak is associated with energy metabolism (with no distinction for any specific energy metabolism pathways, i.e. nitrogen metabolism, sulphur metabolism, oxidative phosphorylation etc.) was calculated as follows: for each $peak_i$ (i=1…n), a putative metabolite identity was assigned based on the accurate mass measurement. Then P_i (energy metabolism | $peak_i$) = ∑ putative assignments for $peak_i$ that are involved in energy metabolism / ∑ putative assignments for $peak_i$.

Table B7 Peaks with putative metabolite assignments and fold changes (post reperfusion / cold phase) involved in oxidative phosphorylation (based on KEGG database classification).

Peak index	m/z	Fold change	p value	Formula	Putative metabolite	Adduct	KEGG ID
62	134.04478	3.33	0.0705	C4H4O4	Fumarate	NH4+	C00122
104	156.98979	0.46	0.0882	C4H6O4	Succinate	K(39)	C00042
187	174.89594	2.77	0.2773	H3PO4	Orthophosphate	2*K(39)-H	C00009
315	194.94563	1.06	0.8983	C4H6O4	Succinate	2*K(39)-H	C00042
1541	450.01887	4.77	0.0606	C10H15N5O10P2	ADP	Na	C00008
1653	472.00083	14.02	0.2087	C10H15N5O10P2	ADP	2*Na-H	C00008
21	96.96962	1.35	0.3853	H3PO4	Orthophosphate	'-H'	C00009
34	132.94639	1.68	0.0645	H3PO4	Orthophosphate	Cl(35)	C00009
112	175.02485	1.53	0.5002	C4H4O4	Fumarate	HAc-H	C00122
114	177.04048	1.45	0.2257	C4H6O4	Succinate	HAc-H	C00042
187	198.9179	1.24	0.5622	P2H4O7	Pyrophosphate	Na-2*H	C00013
2017	426.0223	3.30	0.1179	C10H15N5O10P2	ADP	'-H'	C00008
2183	448.00429	6.14	0.0558	C10H15N5O10P2	ADP	Na-2*H	C00008
2277	463.97834	3.93	0.2346	C10H15N5O10P2	ADP	K(39)-2*H	C00008

Yellow highlighting indicates those peaks which increased in concentration post reperfusion. p values corrected for multiple testing with Benjamini-Hochberg

Table B8 Top 5% of putatively identified peaks that contribute towards an apparent separation of the post reperfusion liver biopsy spectra on the PCA scores plot.

Peak index	m/z	PC1 loadings	Fold change	p value	Empirical formula	Putative metabolite	Adduct
101	156.0421	0.087	4.64	0.208	C5H11NO2	L-Valine;	K(39)
101	156.0421	0.087	4.64	0.208	C5H11NO2	5-Aminopentanoate;	K(39)
101	156.0421	0.087	4.64	0.208	C5H11NO2	Betaine;	K(39)
101	156.0421	0.087	4.64	0.208	C5H11NO2	Amyl nitrite;	K(39)
101	156.0421	0.087	4.64	0.208	C5H11NO2	4-Methylaminobutyrate;	K(39)
108	158.0402	0.079	4.45	0.190	C5H11NO2	L-Valine;	K(41)
108	158.0402	0.079	4.45	0.190	C5H11NO2	5-Aminopentanoate;	K(41)
108	158.0402	0.079	4.45	0.190	C5H11NO2	Betaine;	K(41)
108	158.0402	0.079	4.45	0.190	C5H11NO2	Amyl nitrite;	K(41)
108	158.0402	0.079	4.45	0.190	C5H11NO2	4-Methylaminobutyrate;	K(41)
1600	461.0408	0.078	17.49	0.348	C14H20N6O5S	S-Adenosyl-L-homocysteine;	2*K(39)-H
419	223.9721	0.072	3.97	0.046	C3H8NO6P	O-Phospho-L-serine;	K(39)
419	223.9721	0.072	3.97	0.046	C5H9NO4	L-Glutamate;	2*K(39)-H
419	223.9721	0.072	3.97	0.046	C5H9NO4	D-Glutamate;	2*K(39)-H
419	223.9721	0.072	3.97	0.046	C5H9NO4	Glutamate;	2*K(39)-H
419	223.9721	0.072	3.97	0.046	C5H9NO4	O-Acetyl-L-serine;	2*K(39)-H
419	223.9721	0.072	3.97	0.046	C5H9NO4	L-4-Hydroxyglutamate semialdehyde	2*K(39)-H
419	223.9721	0.072	3.97	0.046	C5H9NO4	2-Oxo-4-hydroxy-5-aminovalerate	2*K(39)-H
419	223.9721	0.072	3.97	0.046	C5H9NO4	N-Methyl-D-aspartic acid;	2*K(39)-H
690	271.1033	0.066	0.22	0.131	C7H16O7	Volemitol;	HAc-H
319	200.0683	0.068	1.42	0.181	C7H15NO3	L-Carnitine;	K(39)
179	197.0222	0.064	12.07	0.033	C7H12O4	6-Carboxyhexanoate;	K(39)-2*H
179	197.0222	0.064	12.07	0.033	C7H12O4	2-Propylsuccinic acid;	K(39)-2*H
179	197.0222	0.064	12.07	0.493	C6H10O5	(R)-3,3-Dimethylmalate	Cl(35)
179	197.0222	0.064	12.07	0.493	C6H10O5	3-Ethylmalate	Cl(35)
179	197.0222	0.064	12.07	0.493	C6H10O5	2-Hydroxyadipate	Cl(35)
179	197.0222	0.064	12.07	0.493	C6H10O5	(R)-2-Ethylmalate	Cl(35)
179	197.0222	0.064	12.07	0.493	C6H10O5	L-Rhamnono-1,4-lactone;	Cl(35)
179	197.0222	0.064	12.07	0.493	C6H10O5	3-Hydroxy-3-methylglutarate;	Cl(35)
179	197.0222	0.064	12.07	0.493	C6H10O5	2-Dehydro-3-deoxy-L-rhamnonate	Cl(35)
179	197.0222	0.064	12.07	0.493	C6H10O5	2-Dehydro-3-deoxy-D-fuconate	Cl(35)
179	197.0222	0.064	12.07	0.493	C6H10O5	(S)-2-(Hydroxymethyl)glutarate	Cl(35)
1145	315.0932	0.061	0.04	0.493	C14H18N2O4	alpha-Ribazole;	Cl(37)
583	261.0112	0.066	8.09	0.304	C5H13O7P	2-C-Methyl-D-erythritol 4-phosphate	2*Na-H
232	185.0323	0.065	2.03	0.035	C5H10N2O3	L-Glutamine;	K(39)
232	185.0323	0.065	2.03	0.035	C5H10N2O3	D-Glutamine;	K(39)
232	185.0323	0.065	2.03	0.035	C5H10N2O3	3-Ureidoisobutyrate	K(39)
1826	384.9957	0.058	3.27	0.165	C10H13N4O8P	IMP;	K(39)-2*H
160	170.0326	0.064	0.47	0.131	C4H9N3O2	Creatine;	K(39)
145	167.0217	0.064	1.76	0.047	C5H8N2O2	5,6-Dihydrothymine;	K(39)
145	167.0217	0.064	1.76	0.047	C5H8N2O2	gamma-Amino-gamma-cyanobutanoate;	K(39)
132	163.9778	0.063	3.02	0.077	C2H7NO3S	Taurine;	K(39)
888	338.0507	0.061	10.81	0.302	C11H15N5O3S	5'-Methylthioadenosine;	K(41)
888	338.0507	0.061	10.81	0.302	C10H15N2O9P	1-(5-Phosphoribosyl)imidazole-4-	'-e'
1413	344.0324	0.056	0.93	0.701	C10H17N3O6S	Glutathione;	K(39)-2*H
242	186.0163	0.06	2.65	0.026	C3H5O6P	Phosphoenolpyruvate;	NH4+
242	186.0163	0.06	2.65	0.026	C3H8NO6P	O-Phospho-L-serine;	H
242	186.0163	0.06	2.65	0.026	C3H5O6P	3-Phosphonopyruvate	NH4+
242	186.0163	0.06	2.65	0.026	C5H9NO4	L-Glutamate;	K(39)

242	186.0163	0.06	2.65	0.026	C5H9NO4	D-Glutamate;	K(39)
242	186.0163	0.06	2.65	0.026	C5H9NO4	Glutamate;	K(39)
242	186.0163	0.06	2.65	0.026	C5H9NO4	O-Acetyl-L-serine;	K(39)
242	186.0163	0.06	2.65	0.026	C5H9NO4	L-4-Hydroxyglutamate semialdehyde	K(39)
242	186.0163	0.06	2.65	0.026	C5H9NO4	2-Oxo-4-hydroxy-5-aminovalerate	K(39)
242	186.0163	0.06	2.65	0.026	C5H9NO4	N-Methyl-D-aspartic acid;	K(39)
187	174.8959	0.06	2.77	0.105	H3PO4	Orthophosphate;	2*K(39)-H
1369	422.3240	0.06	0.34	0.483	C23H45NO4	L-Palmitoylcarnitine	Na
144	166.9507	0.06	2.02	0.027	C3H6O3	Glycerone;	2*K(39)-H
144	166.9507	0.06	2.02	0.027	C3H6O3	(S)-Lactate;	2*K(39)-H
144	166.9507	0.06	2.02	0.027	C3H6O3	(R)-Lactate;	2*K(39)-H
144	166.9507	0.06	2.02	0.027	C3H6O3	D-Glyceraldehyde	2*K(39)-H
144	166.9507	0.06	2.02	0.027	C3H6O3	3-Hydroxypropanoate;	2*K(39)-H
144	166.9507	0.06	2.02	0.027	C3H6O3	Glyceraldehyde;	2*K(39)-H
1259	326.076	0.054	1.69	0.873	C14H15N3O5	Entacapone	Na-2*H
325	203.0079	0.059	3.01	0.089	C10H6O2	1,4-Naphthoquinone;	2*Na-H
325	203.0079	0.059	3.01	0.089	C10H6O2	1,2-Naphthoquinone;	2*Na-H
254	208.9622	0.054	3.39	0.108	C3H9O6P	sn-Glycerol 3-phosphate;	K(39)-2*H
24	98.99552	0.058	2.52	0.199	CH4N2O	Urea;	K(39)
555	254.0190	0.057	0.8	0.036	C5H14NO6P	sn-glycero-3-Phosphoethanolamine;	K(39)
94	154.0264	0.054	3.03	0.068	C5H9NO2	L-Proline;	K(39)
94	154.0264	0.054	3.03	0.068	C5H9NO2	D-Proline	K(39)
257	188.0144	0.054	3.12	0.046	C5H9NO4	L-Glutamate;	K(41)
257	188.0144	0.054	3.12	0.046	C5H9NO4	D-Glutamate;	K(41)
257	188.0144	0.054	3.12	0.046	C5H9NO4	Glutamate;	K(41)
257	188.0144	0.054	3.12	0.046	C5H9NO4	O-Acetyl-L-serine;	K(41)
257	188.0144	0.054	3.12	0.046	C5H9NO4	L-4-Hydroxyglutamate semialdehyde	K(41)
257	188.0144	0.054	3.12	0.046	C5H9NO4	2-Oxo-4-hydroxy-5-aminovalerate	K(41)
257	188.0144	0.054	3.12	0.046	C5H9NO4	N-Methyl-D-aspartic acid;	K(41)

PCA was redone to include only post reperfusion spectra, so that the separation of the post reperfusion biopsies into two groups occurred along PC1. Metabolites were filtered for those present either in the Human Metabolome Database (http://hmdb.ca/) or the KEGG database.

Table B9 Peaks identified by Grubbs test as outliers in the post reperfusion phase for patient H8 that developed IPF.

Peak index	m/z	Fold change	Empirical formula	Metabolite	Adduct	KEGG pathways
178	172.03074	0.95	C4H9N3O2	Creatine	K(41)	ko00260 Glycine, serine and threonine metabolism; ko00330 Arginine and proline metabolism
178	172.03074	0.95	C4H9N3O2	3-Guanidinopropanoate	K(41)	
599	261.14455	1.22				
913	339.07742	1.64	C10H18N2O8	N-Glycosyl-L-asparagine	2*Na-H	
976	346.98844	45.5				
1535	443.34801	2.43				
482	240.05479	21.98				
609	258.9946	0.93				
1842	384.99571	3.51	C10H13N4O8P	IMP	K(39)-2*H	ko00230 Purine metabolism; ko04742 Taste transduction
1857	386.11406	3.39				

Table B10 Retention times, optimum redox potential and direction of concentration change post reperfusion for 19 reproducible peaks detected by CEAD

Peak number	Retention time (min)			Optimum redox potential	Direction of concentration change post reperfusion
	mean	min	max	(mV)	
1	1.680	1.642	1.708	420	unchanged
2	1.696	1.658	1.725	720	unchanged
3	1.768	1.725	1.808	960	increased slightly (~20%) from 21h
4	1.854	1.808	1.883	480	increased slightly (~20%) to 21h, decreasing slightly (~20%) thereafter
5	1.990	1.950	2.025	720	unchanged
6	2.376	2.317	2.417	840	very marked decrease (~95%) to 21h, very low thereafter
7*	2.562	2.492	2.608	720	slight decrease (~30%) to 9h, unchanged thereafter
8	2.645	2.567	2.717	960	slight decrease (~30%) to 9h, unchanged thereafter
9	3.009	2.925	3.075	960	steady but progressive decrease (~40%) over 48h
10*	6.434	6.300	6.542	900	steady but progressive increase (~ x2) over 48h
11	13.089	12.808	13.292	720	decrease (~40%) to 21h, increasing slightly (~20%) thereafter
12*	14.309	14.000	14.533	720	initial decrease (~40%) to 17h and then stable
13	14.373	14.075	14.600	960	initial decrease (~40%) to 9h and then stable
14	16.571	16.358	16.725	960	unchanged
15	17.814	17.558	17.992	720	relatively stable for 27h, then increased (~ x2) to 46h
16	25.754	25.592	25.875	720	very low at 6h, then sharp increase (~ x6) to 21h, then stable
18	26.775	26.658	26.858	360	unchanged
19	26.812	26.692	26.892	540	unchanged

*On the basis of co-elution with authentic standards, and on comparable electrochemical characteristics, the following peaks were ascribed a provisional identity: 7 – tyrosine, 10 – kynurenine and 12 - tryptoph