NMR-based Metabolomics: Global Analysis of Metabolites to Address Problems in Prostate Cancer

Matthew J. Roberts
Centre for Clinical Research / Centre for Advanced Imaging
The University of Queensland, Australia

Horst Joachim Schirra
Centre for Advanced Imaging
The University of Queensland, Australia

Martin F. Lavin
Centre for Clinical Research
The University of Queensland, Australia
Radiation Biology and Oncology
Queensland Institute of Medical Research, Australia

Robert A. Gardiner
Centre for Clinical Research
The University of Queensland, Australia
Department of Urology
Royal Brisbane and Women’s Hospital, Queensland, Australia
1 Introduction

Cancer significantly contributes to the worldwide burden of disease and premature death across many countries. Consequently, current oncology research focuses on discovering and validating new biomarkers to improve early detection. These efforts are of worldwide importance in detecting significant cancer while it is still localised and in lessening associated morbidities and death.

Biomarkers have mostly been sourced from non- or minimally invasive biofluids, such as blood, urine, and biopsy tissue. Traditionally, biomarkers were limited to circulating end-products of altered cellular function in cancer. However, technology advances and emergence of the –omics sciences have improved analysis of genes, gene expression, proteins and metabolites alike – on both an individual and system-wide scale. This field of research, termed “systems biology”, has allowed for molecules at all levels of the cellular hierarchy to be considered as biomarkers. Continuous improvements in sensitivity, resolution and precision of these analytical techniques produces large datasets, allowing for simultaneous characterisation of, ideally all, compounds in a single sample. Subsequent statistical analysis of these datasets and their interpretation with respect to cellular function is the basis of the different -omics technologies, such as genomics, transcriptomics, proteomics and metabolomics.

In this chapter, we will describe principles and processes that are involved in investigating biological or clinical problems with nuclear magnetic resonance (NMR)-based metabolomics - an approach that involves the global analysis of metabolites. In writing for the scope of this book, we have broken this chapter into three sections: (1) First we will describe and illustrate the methods commonly used in NMR-based metabolomics, including spectral processing, data treatment and subsequent statistical analysis. (2) Secondly, we will use prostate cancer (PCa) as a case study to illustrate how NMR-based metabolomics can be applied to a clinical problem. PCa is the second most common type of cancer and the sixth leading cause of cancer-related death worldwide (Center et al., 2012; Ferlay et al., 2010; Siegel et al., 2012). The diagnosis of prostate cancer is currently problematic for a number of reasons that include lack of sensitive and specific tumour markers as well as limitations due to morbidity inherent with the biopsy diagnosis process. Furthermore, many patients harbour early prostate cancer with insignificant tumours that may not progress to produce clinical problems. (3) Lastly, we will briefly outline the future directions for the role of NMR-based metabolomics, including personalized medicine and integration with other -omics datasets, in order to create a holistic, systems biology approach to solving clinical problems.

Outlining the processes, applications and potential of metabolomics will be of assistance to bio-statisticians and bioinformaticians who may be interested in expanding into this area of research. Similarly, we aim to inspire scientists and clinicians who are interested in applying this approach to a scientific or clinical problem.

2 Metabolomics: History and Methods

2.1 What is Metabolomics?

Metabolomics has been highlighted as a technique that is unique and exciting in biomarker discovery (Abate-Shen & Shen, 2009; Bino et al., 2004; Nicholson & Lindon, 2008). It is the quantification of all small molecular weight metabolites to accurately define the metabolite composition of a biological sample (Fiehn, 2002). The term “metabolomics” is often used interchangeably with “metabonomics”, which is defined as ‘the quantitative measurement of the dynamic multiparametric metabolic response of living
systems to pathophysiological stimuli’ (Nicholson et al., 1999). The technical approach to both metabonomics and metabolomics is similar, involving measurement and analysis of metabolite data for a given sample. Conceptually, however, the objective and application of these techniques is slightly different: Metabolomics seeks to describe the composition of complex biological samples, while metabonomics aims to map and understand the change of a biological system in response to external or artificial stimuli. In the current chapter, we will use the term “metabolomics”, while both metabolomics and metabonomics will be discussed. The advent and development of metabolomics/metabonomics has largely been possible due to advances in analytical techniques, such as nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS), as well as chromatographic separation techniques. As the authors have expertise in NMR spectroscopy, the focus of this chapter will be on NMR-based techniques within the field of metabolomics. We refer the reader to several excellent reviews that detail the application of MS in metabolomics and systems biology (Lu et al., 2008; Pasikanti et al., 2008; Theodoridis et al., 2008).

2.2 Historical Perspective

While “metabolomics” is a recently coined term, the analysis of metabolic end products is a long-practiced ancient scientific process. About 2000-1500 BCE, analysis of urine by human taste or animal behaviour (due to the high urinary glucose concentration) helped diagnose patients with diabetes mellitus (van der Greef & Smilde, 2005). In 1506, a “urine wheel” by Ulrich Pinder linked crude changes detected by human senses (colour, smell, taste) with various medical conditions (Pinder, 1506; Weiss & Kim, 2012). Although qualitative measures of metabolism have been performed for centuries, the origins of quantitative studies stem from the measurement of insensible perspiration and other hydration losses in medieval Italy (Eknoyan, 1999). Technological limitations in analytical chemistry hindered further advances until the early 20th century. At this time, the development of sensitive analytical methods allowed quantification of key compounds/metabolites in urine and other sample types (Simoni et al., 2002). This quantitative approach continued to develop with the introduction of various analytical techniques, such as mass spectrometry (Griffiths, 2008; Thomson, 1912) and NMR (Bloch et al., 1946; Freeman, 1995; Purcell et al., 1946), and with the application of these techniques to metabolic research (Gates & Sweeley, 1978; Hoult et al., 1974). The integration of medical science and analytical chemistry at this time led to a greater understanding of metabolic perturbations in medical conditions, e.g. kidney stones among many others (Reginato & Kurnik, 1989). Further improvements in metabolite profiling in the latter half of the 20th century were aided by advances in chemometrics, the foundation behind data analysis in analytical chemistry (Geladi & Esbensen, 1990). Appropriate data processing and interpretation was achieved by multivariate statistical methods, which will be outlined in more detail below.

2.3 Modern Metabolomics and its Varieties/Applications

Metabolomics continues to evolve as a field and is increasingly used in a variety of applications. Initially, biofluids were analyzed specifically to quantify metabolic perturbations due to drug toxicity, disease, and other internal and external influences. For example, perturbations in steroid metabolism were used for anti-doping testing during the Los Angeles Summer Olympics in 1984 (Fitch, 2008). In the following section, the diverse applications of metabolomics are briefly described; more comprehensive accounts are available in other sources (Duarte & Gil, 2012; Ma et al., 2012b; Ng et al., 2011; Rhee & Gerszten, 2012; Spratlin et al., 2009).
2.3.1 Biofluid and Excretion Analysis

Metabolomics has been used to quantify endogenous metabolites in many human biofluids, with those most commonly analyzed being urine and blood (serum, plasma). Analysis of urine metabolites has shown early promise in diagnosing kidney (Ganti & Weiss, 2011) and bladder tumors (Hyndman et al., 2011), as well as more systemic conditions such as type 2 diabetes mellitus (Salek et al., 2007). Indeed, distinct serum metabolite patterns have been characterized for abnormal clinical states including breast cancer (Oakman et al., 2011), leukemia (MacIntyre et al., 2010), sepsis and acute lung injury (Serkova et al., 2011), coronary artery disease (Brindle et al., 2002) and obesity (Oberbach et al., 2011; Xie et al., 2012). Furthermore, cardiovascular health has been assessed by metabolomic analysis of feces, linking perturbations in the metabolite profiles of gut flora to the metabolic syndrome and dyslipidemia (Wang et al., 2011). Studies on cerebrospinal fluid have associated metabolite changes with brain tumors and neurodegenerative disorders, such as multiple sclerosis and Alzheimer’s disease (Han et al., 2011). Salivary metabolomics has been used to investigate oral cancer and pre-malignant changes, such as leukoplakia (Wei et al., 2011). Thus, this minimally invasive approach has enormous potential in providing valuable scientific and clinical information for medical professionals and researchers.

2.3.2 NMR Spectroscopy of Tissues and in vivo Imaging Techniques

High-resolution magic-angle spinning (HR-MAS) NMR spectroscopy can be used to perform non-destructive metabolite profiling of tissue or other solid samples (Keifer, 2007). That means that after HR-MAS NMR spectroscopy, further testing of tissue samples can be performed such as histopathological evaluation, the current gold standard in disease diagnosis, or genome and protein sequencing. As a result, HR-MAS NMR has been used to investigate a number of disease states (Brown et al., 2012; Dittrich et al., 2012; Kurhanewicz et al., 2002; Martínez-Bisbal et al., 2004; Maxeiner et al., 2010; Millis et al., 1997; Sitter et al., 2002).

Metabolomic analysis can also be performed directly in vivo, largely owing to advances in magnetic resonance spectroscopic imaging and positron emission tomography (PET). A standard clinical magnetic resonance imaging (MRI) scan uses similar physical concepts to NMR, but takes many scans across a section of living tissue. This data is processed to produce an anatomically correct image based on physical properties of the tissue. A magnetic resonance spectroscopy imaging sequence is able to produce NMR-like spectra for a targeted volume segment in the body, allowing for visualization of metabolite content in that anatomical location. Metabolic alterations measured in vivo have been shown to correlate with histopathology (Delongchamps et al., 2011; Kwok et al., 2006). Furthermore, in vivo metabolomics is being used to monitor the response to various therapies, such as radiotherapy and chemotherapy (Coy et al., 2011; Lodi & Ronen, 2011).

Different MRI techniques allow the investigation of different phenomena. Dynamic contrast-enhanced MRI uses the uptake and elimination of contrast agents, such as gadolinium, to distinguish between different tissues. The use of dynamic contrast-enhanced MRI in oncology is based on the premise that cancer cells have a higher metabolism, and thus a higher uptake and elimination of gadolinium contrast (Vargas et al., 2012). Diffusion weighted imaging, initially used to investigate connectivity between different brain regions (Behrens et al., 2003), uses slower water diffusion in cancerous tissues compared
with surrounding healthy cells due to a higher nuclear content and cellular density coupled with extracellular changes (Lim & Tan, 2012). Recently, these MRI techniques have been combined into multiparametric MRI, which has been shown to increase accuracy in cancer detection (Hoeks et al., 2011). The value of these in vivo applications and their role in oncology is commonly described and reviewed in radiology literature (Hoeks et al., 2011; Hoh et al., 1997; Koh & Collins, 2007; Padhani, 2002).

Further understanding of altered metabolism in cancer and identification of abnormal pathways facilitates imaging using PET in combination with computed tomography (CT) via PET/CT (Basu et al., 2011; Jones & Price, 2012). After identifying metabolites, that are either preferentially used or upregulated within particular pathways, nuclear isotopes can be chemically attached either to these metabolites or to metabolite analogues. The emission of positrons from these isotopes can then be measured as gamma rays and superimposed on a CT scan during PET/CT scanning. For example, most cancer cells display heightened glycolysis. Thus, fluoro-deoxy-glucose, containing a radiolabeled positron emitter such as $^{18}$F, can be administered and taken up by cancer cells, which are highlighted (Costello & Franklin, 2005). In addition, PET is able to distinguish specific cancers, e.g. $^{11}$C-choline PET is used to detect prostate cancer (Evangelista et al., 2012; Reske et al., 2006). The application of PET in other clinical scenarios is diverse, but widespread use is limited by logistical and financial constraints (Fletcher et al., 2008; Kelloff et al., 2005; Rohren et al., 2004).

Recently, a novel method has been proposed to perform in vivo metabolomics during surgery by using real-time MS analysis of the smoke produced from electric cautery to biochemically recognize malignant/diseased tissue in which macroscopic changes are not present (Balog et al., 2010; Kinross et al., 2011). Although major development is required before clinical use, the initial concept is intriguing in its potential to improve surgical accuracy and treatment outcomes following cancer surgery.

### 2.4 Integration with other -Omics Sciences

Metabolites are part of the complex and interconnected cellular hierarchy involving DNA, RNA, proteins and metabolites, as outlined in Figure 1. Consequently, providing an understanding of the mutual relationships between genomic, transcriptomic, proteomic and metabolomic data is a major aim of systems biology.

Integration of multi-omics data sets is already providing insight into biological processes. This integration is enabled by the availability of new statistical methods to correlate information contained in multiple large datasets (Rantalainen et al., 2006). Furthermore, ever increasing genome-wide association studies (GWAS) are identifying multiple risk loci associated with various disease states. However, the penetrance of these loci, and therefore their relevance, remains unclear. By integrating metabolomics and other -omics sciences with GWAS, it is anticipated that identification of loci with a high penetrance/phenotypic manifestations will be unveiled (Suhre et al., 2011; Weckwerth, 2011). Integration of -omics data sets has major potential in oncology (Casado-Vela et al., 2011) and some studies have used a targeted approach in relating datasets obtained by different analytical methods (Rantalainen et al., 2006). A recently published study related metabolomic changes to genomic disturbances in PCa tissue to demonstrate alterations in m-aconitate and acetyl citrate lyase. Phospholipase A2 group VII and choline kinase $\alpha$ were responsible for altered citrate and choline levels, respectively (Bertilsson et al., 2012). Other integrative works investigated colorectal cancer (Ma et al., 2012a), heart failure (Lin et al., 2011) and other diseases (Adamski, 2012; Chen et al., 2012). These studies show impressive proof of concept of this new approach. As a result, multivariate statistical analysis and integration of large and multi-omics data sets are a valuable strategy for further investigation.
Another key frontier in systems biology is the creation of genome-wide in silico models of cellular metabolism that are able to incorporate and integrate multi-omic data (Bordbar & Palsson, 2012). Such reconstructed metabolic networks have already suggested improvements for targeted treatment strategies of cholesterol homeostasis in human cellular models (Mo et al., 2007).

The first genome-scale model of human metabolism published in 2007 was a promising milestone (Duarte et al., 2007). The next step in metabolic modeling in oncology will be to construct cancer-specific metabolic models that will incorporate –omics data. This daunting task requires improved curation and annotation of genome databases, as well as integration of high quality –omics datasets from studies of specific cancer types, or of other disease states to create disease-specific reconstructions. Public accessibility and maintenance of data from publicly funded research is critical to these efforts (Field et al., 2009), as is data management (Sansone et al., 2012) and interpretation (Leader et al., 2011).

2.5 Analytical Techniques

Of the many analytical techniques that are used in metabolomics to investigate physiological and pathological states, NMR spectroscopy and gas chromatography (GC) or liquid chromatography (LC) combined with MS are the two most commonly used methods. Both have low running costs, are diverse in sample type and allow for accurate metabolite identification. Other techniques in use are ultra-performance LC-MS, inductively coupled plasma MS, Fourier-transform MS, Fourier-transform infrared spectrometry and thin layer chromatography (Zhang et al., 2012). We will briefly describe the processes involved in MS-based metabolomics, with more extensive reviews available elsewhere (Dunn et al., 2005; Dunn & Ellis, 2005; Lu et al., 2008; Mishur & Rea, 2012; Pasikanti et al., 2008; Theodoridis et al., 2008). Subsequently, we will focus on NMR-based metabolomics, describing the basic principles and statistical approaches that are currently in use.
2.5.1 GC-/LC-MS

Mass spectrometry detects ionized compounds in biological samples according to their mass/charge \((m/z)\) ratio following chromatographic (e.g. GC or LC) separation, and metabolites can be identified in the resulting mass spectrum with reference to internal standards (Wilson et al., 2005). LC- and GC-MS are well used techniques in metabolite analysis and have similar sensitivity, with the major difference being that GC requires more sample preparation (derivatization) and higher analysis temperatures, thus LC may be preferred for this reason (Issaq et al., 2009; Mishur & Rea, 2012). We will briefly outline the basic processes in MS-based metabolomics, which are sample preparation, separation (via liquid-/gas-chromatography), ionization, mass analysis and detection, and finally, data processing.

Sample preparation for MS is dependent on the type of sample. Simple biofluid preparation often involves removing macromolecules through protein precipitation and centrifugation or filtration. Similarly, sampling of the exometabolome (metabolites secreted by cells or organisms into the growth medium) is straightforward. In contrast, to obtain intracellular metabolites, tissues or cells need to be extracted in an appropriate solvent system (Mishur & Rea, 2012). As different solvent systems are biased towards particular classes of metabolites (e.g. polar extraction systems yielding predominantly polar metabolites), the exact choice of solvent depends upon the metabolites of interest. To achieve consistency between samples, internal standards are typically added during/after extraction (Mishur & Rea, 2012).

In GC-MS, derivatization is a further necessary preparation step, which is applied to non-volatile metabolite classes, such as amino and organic acids, sugars, amines and lipids, to render them volatile and thermally stable for GC (Dunn et al., 2005; Dunn & Ellis, 2005; Mishur & Rea, 2012). Derivatization can introduce bias towards individual metabolites if the derivatizing agents are not provided in excess, as the derivatizing reactions have different efficiencies with different metabolites (Mishur & Rea 2012). In addition, metabolites with multiple exchangeable protons will create multiple derivatization products that will show up as separate peaks, thus complicating the final mass spectrum. For GC-MS, electron impact ionization is almost exclusively used (Dunn et al., 2005).

LC-MS is rapidly replacing GC-MS as method of choice in metabolomics, as both methods are similarly sensitive, but sample preparation for LC-MS is simpler, because derivatization is not required (Griffin & Shockcor, 2004). LC typically runs as reverse-phase high-performance LC (HPLC), or recently even as ultra-performance LC (UPLC), and electrospray ionization is typically used in LC-MS systems (Mishur & Rea, 2012). Electrospray ionization MS can run in positive or negative ionization mode, and because individual metabolites are generally only detected in one of those two modes, both ionization modes need to be run to improve coverage of the metabolome (Dunn & Ellis, 2005).

In the resulting mass spectrum, metabolites are quantified by external calibration or by comparison with internal standards (Dunn & Ellis, 2005). GC-MS experiments may also require the use of deconvolution software to adequately analyse overlapping chromatographic peaks (Dunn & Ellis, 2005). To allow comparable results between experiments, data may also undergo further pre-treatment steps, including spectral alignment and automated picking of metabolite peaks (Dettmer et al., 2007). The subsequent multivariate analysis of processed MS data is similar to data obtained by NMR-spectroscopy.

2.5.2 NMR Spectroscopy

NMR spectroscopy is a quantitative technique used to accurately determine metabolite concentrations in samples. Chemical compounds in biological samples are identified by their characteristic peak patterns and signal positions in the NMR spectrum with the aid of online databases (Wishart et al., 2009)
More comprehensive accounts of NMR theory and application in metabolomics are available in dedicated texts (Ross et al., 2007). A comparison of the strengths and limitations of NMR and MS is provided in Table 1.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
</table>
| NMR       | - high reproducibility  
- high resolution  
- non-destructive  
- quantitative  
- low running costs  
- minimal sample preparation/no derivatization  
- unbiased metabolite profile  
- analysis of tissue (HR-MAS)  
- translation to in vivo (MRI)  
- rapid analysis  
- ability for automation  
- structural identification (2D, 3D) | - low sensitivity  
- peak overlap  
- libraries of limited use due to complex matrix  
- long acquisition times for heteronuclear techniques, e.g. $^{13}$C  
- high initial capital cost - reduced availability |
| GC-MS     | - high sensitivity  
- large linear range  
- robust  
- identification of wide range of metabolites  
(wider range with LC-MS)  
- analysis of complex biofluids  
- non-targeted  
- established databases  
- widely available and comparably low capital cost  
- preferred for targeted analysis | - slow  
- sample unable to be re-used  
- requires chemical derivatization  
- potentially multiple derivatization products for metabolites  
- many analytes thermally unstable  
- metabolite weight limitation (<1400 Da) |
| LC-MS     | - high sensitivity  
- high reproducibility  
- large linear range  
- no chemical derivatization needed | - slow  
- limited commercial libraries  
- sample unable to be re-used  
- generation of adducts  
- higher capital cost (HPLC-MS) |

Table 1: Comparison of the advantages and disadvantages of NMR, GC-MS, and LC-MS (Issaq et al., 2009; Shepherd et al., 2011; Shulaev, 2006; Zhang et al., 2012).

2.6 Processing of NMR Data

Processing of NMR data comprises four steps: Fourier transformation, phase correction, baseline correction and calibration. Fourier transformation transforms the raw real-time data into the frequency domain, and phase correction corrects the phase of the resulting NMR spectrum. Baseline correction ensures a constant zero baseline across an NMR spectrum, and calibration is needed to ensure a consistent chemical shift scale/axis across all spectra. Inadequate processing introduces artifacts that confound statistical analysis and jeopardize data integrity (see Figure 2). Minimum standards for reporting and processing
have been outlined (Goodacre et al., 2007), and continue to be a good guide for authors publishing metabolomics research.

Usually, NMR spectra require a phase correction following Fourier transformation in order to achieve pure absorptive line shapes for all peaks in a NMR spectrum. Where possible, phasing should be performed automatically or by the same operator across all samples to ensure consistency. Incorrect phasing can distort peak integrals and thus, the subsequent multivariate statistical analysis (MVSA; see Figure 2A/2B).

Baseline correction is the third critical processing step in producing consistent, comparable and reliable data in NMR spectroscopy. At a minimum, the y-offset of the entire spectrum is corrected to be zero. However, baseline corrections are often more complex, using spline, polynomial or other mathematical functions to accomplish a zero baseline over the whole spectrum. As signal intensities are calculated with reference to zero, inadequate baseline correction will distort spectral peak intensities (Figure 2B/2C), and compromise the subsequent MVSA.

Figure 2: Sequential steps of processing NMR spectra. Shown is a spectrum of human ejaculate in the region of citrate (2.45 – 2.65 ppm) as example. Correct baseline position is illustrated with dotted lines. a - Spectrum after Fourier transformation. The phases of the NMR signals are partly dispersive and in need of phase correction. b - Spectrum after phase correction. Globally reduced metabolite peak intensities and negative values for the baseline occur across the entire spectrum. c - Spectrum after baseline correction, but requiring calibration to a chemical shift standard. Chemical shift values are incorrect across the entire spectrum, resulting in incorrect metabolite identification. d - Correctly processed spectrum suitable for data reduction.

Importantly, the chemical shift axis of each NMR spectrum must be adequately calibrated using a chemical shift standard, such as (deuterated) 4,4-dimethyl-4-silapentane-1-sulfonic acid-d₆ (DSS). In samples containing a high protein content (such as plasma), DSS cannot be used as internal standard. Thus, other endogenous metabolites that are present across all samples, such as lactate, glucose or formate, are used as internal reference. The alternative is to use DSS as an external standard by either inserting a capillary with DSS in deuterium oxide (D₂O) into the NMR tube, or inserting the sample in a capillary into a tube containing DSS in D₂O. The use of 4,4-dimethyl-4-silapentane-1-ammonium trifluoroace-
DSS (Alum et al., 2008). Chemical shift calibration ensures consistent global alignment of spectra in a metabolomics data set which is critical for statistical analysis. In addition, correct spectral alignment is required for reliable metabolite identification. However, even in correctly calibrated spectra, individual peaks can still exhibit differences in chemical shift between individual spectra due to differences in sample pH and ionic strength. These can be corrected post-processing by various automatic peak alignment procedures (Anderson et al., 2011; Giskeødegård et al., 2010; MacKinnon et al., 2012; Savorani et al., 2010; Staab et al., 2010; Wright et al., 2012).

2.7 Statistical Pre-processing

2.7.1 Data Reduction

After processing of raw NMR data, further processing steps are needed to prepare data for MVSA, which are usually termed “statistical pre-processing”. Reducing the full resolution data into small segments of equal width, called bins, or “buckets” (Figure 3), is the most widespread method of data reduction in chemometrics (Wishart, 2008). Compared with analysis at full resolution, this method considerably reduces the size of the data matrix in MVSA, and is particularly helpful when peak positions or widths vary slightly due to changes in pH, ionic strength or other factors. However, due to decreasing data resolution, bucketing can complicate metabolite identification following data analysis. Other pre-processing methods that can be used, particularly in targeted metabolomics, include deconvolution, peak-picking, and weighting factors (Goodacre et al., 2007).

**Figure 3:** Illustration of data reduction by “bucketing”. a - NMR spectrum of human ejaculate with suitable pre-processing (segment shown). b – Same spectrum segmented/data reduced into buckets of 0.04 ppm width across the region 8 – 0 ppm. The region around the water signal from 5.08 – 4.52 ppm was excluded due to artifacts from imperfect water suppression. Note that the area in each individual bucket is integrated across the spectral width of 0.04 ppm and then normalized, yielding intensities similar to a histogram.

2.7.2 Normalization

After data reduction, data need to be normalized to produce data that are comparable between samples (Craig et al., 2006). Normalization is a row operation in the data matrix, and different normalization methods are used to obtain the best representation of the data. In total integral normalization, or normaliz-
ing to total intensity, the spectral intensity in each bucket is divided by the total intensity of each spectrum. This procedure normalizes differences between spectra due to sample concentration/dilution, e.g. due to different water content between samples. However, total integral normalization is vulnerable to distortions when one or a few intense signals change considerably between spectra.

Another method involves normalization to an internal reference compound. For metabolomics analysis of urine, normalization to creatinine has been widely used (Akira et al., 2008; Jentzmik et al., 2010). For physiological reasons, urine creatinine is believed to be a suitable indicator of urine concentration as creatinine excretion is constant. However, creatinine normalization has limitations because it will be confounded by any background pathophysiology that alters serum creatinine levels or creatinine excretion, such as in kidney disease. In these cases, creatinine normalization is not suitable. Furthermore, variations in chemical properties within the sample can distort creatinine alignment, so that other metabolites, especially creatine, will overlap with the creatinine signal and thus impede the proper measurement of creatinine concentration (Ross et al., 2007).

Probabilistic quotient normalization (PQN) is a method that reduces variation caused by large changes in the intensity of one or a few signals across samples, as shown in Figure 4. Thus, PQN can overcome the main weakness of total integral normalization. In PQN, which is usually performed after total integral normalization, each variable (bucket) in a spectrum is first divided by the intensity of the same variable in a reference spectrum. Afterwards the full spectrum is divided by the median of these quotients. This procedure is repeated for all spectra in a data set, using the same reference spectrum (Dieterle et al., 2006).

Recently, Kohl et al. compared many normalization methods, with some derived from genomic data analysis, and recommended more advanced methods, such as quantile normalization for datasets of \( n \geq 50 \), as well as Cubic Spline Normalization and Variance Stabilisation Normalization (Kohl et al., 2012).

![Figure 4: Schematic depiction of spectra containing four NMR signals A-D, following two different normalization methods. a – raw spectra with marked intensity variation present in the first peak A and identical intensities of remaining peaks between all spectra. b - Integral normalization (normalization to total intensity) reduces variation, and therefore influence of the dominant signal A, but also alters relative intensities of the smaller, previously identical signals. c - Probabilistic quotient normalization (PQN) partially reduces variation and influence of the larger signal A, while maintaining the original relationships between smaller peaks to allow optimal comparison during MVSA.](image)

### 2.7.3 Scaling Effects

Following normalization, metabolomic data must be appropriately scaled, or transformed in a column operation in a way that changes how much signals of large and small intensity, respectively, influence the data analysis (Craig et al., 2006; Goodacre et al., 2007). The objective is to reduce noise and maximize
information content in the data. Inappropriate scaling may lead to results that highlight parts of the data unrelated to a biological factor, thus compromising the analysis and biological interpretation of the data. In metabolomics statistical analysis, three scaling methods are largely used.

Centre scaling, or mean centering, subtracts the mean value of each variable/bucket from the original data of that bucket (Craig et al., 2006). This method is the least manipulative, and is also best at minimizing background noise, but large relative variations in small signals may not be detected. Mean centering is usually performed mandatorily. Thus, this method is also sometimes referred to as “no scaling”, as no further scaling is performed after mean centering.

Univariate, or unit variance, scaling divides the raw data obtained after mean centering by the standard deviation of each variable. Univariate scaling gives each variable equal weighting, such that variables with small absolute but large relative variation are highlighted, but this also means that background noise and other unrelated data variation may be overemphasized and thus confound the analysis (Craig et al., 2006).

Pareto scaling is performed by dividing each variable by the square root of its standard deviation (Erikson et al., 1999). This is the recommended scaling method for NMR-based metabolomic data, as it is able to increase the weighting on metabolites with smaller amplitudes, but does not overemphasize the influence of background noise.

2.8 Statistical Analysis

Statistical analysis of metabolomic data depends on the biological question studied and the design of the particular project, thus, the choice of data analysis methods varies between different projects. In addition, the methods for data analysis are continually evolving. Nevertheless, there is a core set of methods of univariate and multivariate statistical analysis that is in use for metabolomics, and minimum reporting standards for data analysis have been established (Goodacre et al., 2007; Liberati et al., 2009; Schulz et al., 2010).

2.8.1 Univariate Statistical Analysis

The role of univariate analysis in metabolomics is largely of a targeted nature. An example would be where metabolites of interest have been identified by MVSA, and detailed analysis of statistical significance of the individual metabolites is desired. Basic univariate methods can be used to analyze whether or not individual metabolites are significantly different between two classes. However, as with any statistical analysis, the distribution of the data determines the type of analysis used. If the data are normally distributed, t-tests, z-tests, and analysis of variance (ANOVA) may be used. In cases where the distribution is not normal, non-parametric methods such as the Kruskal-Wallis test are used (Goodacre et al., 2007). However, given the high number of variables within a metabolomics dataset, use of multiple hypothesis testing corrections, such as Bonferroni correction or false discovery rate/Benjamini-Hochberg are absolutely imperative (Noble, 2009). This means that, to be significant, p-values need to be much smaller (e.g. \( p \leq 5 \times 10^{-5} \)) after correcting for multiple hypothesis testing compared to standard univariate statistical analysis (Broadhurst & Kell, 2006).

2.8.2 Multivariate Statistical Analysis (MVSA)

Modern day metabolomics is largely based on data sets incorporating many variables, between several hundred in the case of bucketed data and up to 65,536 if 1D-NMR spectra are used at full resolution.
Thus, MVSA methods which simultaneously analyze all these variables are preferred (Broadhurst & Kell, 2006). MVSA determines whether there are inherent patterns or groupings within the data that correspond to biological states and also which variables are important in discriminating between the different groupings. Thus, this approach is well suited to analyzing metabolomics datasets, where the aim is to correlate multiple metabolite changes with alterations in biology.

There are two general classes of MVSA methods: unsupervised methods, which analyze patterns within a data matrix \( X \), and supervised methods in which the patterns in \( X \) are correlated with other external data (e.g. clinical data) contained in a \( Y \) matrix or \( Y \) table. The advantages and disadvantages of the most common MVSA methods are summarized in Table 2, and will be discussed in the following sections. Some examples of software programs, both commercial and free/open-source, that are available to perform MVSA are given in Table 3.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td>Principal components analysis (PCA)</td>
<td>• Simplifies data&lt;br&gt;• Describes variation in original data without bias&lt;br&gt;• Groups samples with similar metabolite profiles</td>
<td>• Variation might be unrelated to biological question&lt;br&gt;• Influenced by any confounders</td>
</tr>
<tr>
<td>Supervised</td>
<td>Partial least squares (PLS)</td>
<td>• Simplifies data&lt;br&gt;• Groups samples with similar metabolite profiles&lt;br&gt;• Extracts variation that is correlated with external data/identifiers&lt;br&gt;• more directed to the biological question</td>
<td>• Possibility of introducing bias&lt;br&gt;• Require rigorous validation</td>
</tr>
<tr>
<td></td>
<td>Orthogonal projections to latent structures (OPLS)</td>
<td>• As PLS&lt;br&gt;• Removes orthogonal (unrelated) variation&lt;br&gt;• Improved knowledge extraction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2-way OPLS (O2PLS)</td>
<td>• As OPLS&lt;br&gt;• Two-way data correlation between ( X ) and ( Y )&lt;br&gt;• Potential for unsupervised analysis (when analyzing two large datasets without external ( Y ) table)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kernel OPLS</td>
<td>• As OPLS&lt;br&gt;• Improved model prediction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OnPLS</td>
<td>• As O2PLS&lt;br&gt;• Simultaneous data correlation from multiple (( n )) matrices</td>
<td>• Possibility of introducing bias&lt;br&gt;• Require rigorous validation&lt;br&gt;• Not widely available (commercial/open source)</td>
</tr>
<tr>
<td></td>
<td>Bi-modal OnPLS</td>
<td>• As OnPLS&lt;br&gt;• Data correlation between variables (columns) and samples (rows)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OnPLS path modelling</td>
<td>• As OnPLS&lt;br&gt;• Linkage of matrices along statistically related paths</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Summary of multivariate statistical analysis methods used in metabolomics for information recovery. The advantages and disadvantages of unsupervised and supervised methods are outlined.
Table 3: Examples of software packages used to perform multivariate statistical analysis in metabolomics.

<table>
<thead>
<tr>
<th>Software Package</th>
<th>Use in NMR/MS</th>
<th>Reference/Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIMCA</td>
<td>Both</td>
<td><a href="http://www.umetrics.com">http://www.umetrics.com</a></td>
</tr>
<tr>
<td>MATLAB</td>
<td>Both</td>
<td><a href="http://www.mathworks.com">http://www.mathworks.com</a></td>
</tr>
<tr>
<td>MarkerLynx/MassLynx</td>
<td>MS</td>
<td><a href="http://www.waters.com">http://www.waters.com</a></td>
</tr>
<tr>
<td>STATISTICA Data Miner</td>
<td>Both</td>
<td><a href="http://www.statsoft.com">http://www.statsoft.com</a></td>
</tr>
<tr>
<td>AMIX</td>
<td>Both</td>
<td><a href="http://www.bruker.com">http://www.bruker.com</a></td>
</tr>
<tr>
<td>Agilent Mass Profiler Professional</td>
<td>MS</td>
<td><a href="http://metabolomics.chem.agilent.com">http://metabolomics.chem.agilent.com</a></td>
</tr>
<tr>
<td>Progenesis CoMet</td>
<td>MS</td>
<td><a href="http://www.nonlinear.com">http://www.nonlinear.com</a></td>
</tr>
<tr>
<td>R</td>
<td>Both</td>
<td><a href="http://www.r-project.org">http://www.r-project.org</a></td>
</tr>
<tr>
<td>Metaboanalyst</td>
<td>Both</td>
<td><a href="http://www.metaboanalyst.ca">http://www.metaboanalyst.ca</a> (Xia et al., 2012)</td>
</tr>
<tr>
<td>MAVEN</td>
<td>MS</td>
<td><a href="http://maven.princeton.edu">http://maven.princeton.edu</a> (Melamud et al., 2010)</td>
</tr>
<tr>
<td>MZmine</td>
<td>MS</td>
<td><a href="http://mzmine.sourceforge.net">http://mzmine.sourceforge.net</a> (Pluskal et al., 2010)</td>
</tr>
<tr>
<td>MetDB</td>
<td>MS</td>
<td><a href="https://meltdb.cebitec.uni-bielefeld.de/">https://meltdb.cebitec.uni-bielefeld.de/</a> (Neuweger et al., 2008)</td>
</tr>
<tr>
<td>MetabolomeExpress</td>
<td>MS</td>
<td><a href="https://www.metabolome-express.org">https://www.metabolome-express.org</a> (Carroll et al., 2010)</td>
</tr>
</tbody>
</table>

2.8.3 Unsupervised Methods in Multivariate Statistical Analysis

Often, unsupervised analysis methods are initially used in MVSA as they are excellent data exploration tools that can be either used to simplify the data (dimensionality reduction – principal- or independent-components analysis) or to group samples with similar metabolite patterns (clustering – hierarchal, partitional). Although many different methods exist, principal components analysis (PCA) is the most commonly used method of unsupervised analysis in metabolomics.

PCA simplifies the original multivariate data which form a swarm of data points in a high-dimensional statistical space, by projecting them down into a new space with comparatively few dimensions called principal components (PCs). These PCs are latent variables that describe the variation in the original data. The first PC indicates the direction in which most variation occurs in the data. Subsequent PCs are all orthogonal to each other and sorted in order of descending amount of variation. This arrangement describes the majority of variation in the data within the first few PCs. PCA is visualized by two types of plots, the scores and loadings plots (Figure 5).

The scores plot illustrates the relationship and similarity of samples to each other, and allows for inspection of groupings and outliers. Outliers can be visually identified on the scores plot, or statistically defined as being outside the Hotelling’s 95% confidence range across all components. Further statistical validation can be obtained using the residual variance of the model, known as distance to model plot (Erikson et al., 1999; Trygg et al., 2007). The variables (buckets, ultimately metabolites) attributable to each component in the model are illustrated in the loadings plot (see Figure 5; Denkert et al., 2006; Pan et al., 2007).

The inherent advantage of unsupervised methods is that they are unbiased, i.e. they detect any statistical variation in the data, whether or not it is related to the underlying biological effect (e.g. differ-
ences between cancer and non-cancer samples). This property is also their most noticeable limitation, because when confounding effects are larger than the biological effect, unsupervised analyses will pre

Figure 5: Visualization of a typical multivariate analysis (e.g. PCA, PLS etc.). Sheep urine samples before and after road transport of 12 and 48 hours are shown as an example (Li et al., 2010). a – Scores plot – shows any relationships between samples, such as the presence of separate groups or outliers. This scores plot shows similarity of the pre-transport groups for both transport durations (open inverted triangle = 12 h, black squares = 48 h), and differences between both post-transport groups (48 hours = black diamonds, 12 hours = open triangles). b – Loadings plot – displays the relationship of influential variables that are responsible for the position of outliers or groups seen in the scores plot. Note that the positions/directions of groupings/outliers in the scores plot and responsible variables in the loadings plot correspond to each other (Trygg et al., 2007).

dominantly show the effects of these confounding factors. Distorting variation may also come from uncorrelated background variation, or from noise. For this reason, robust experimental design that limits confounding factors and appropriate pre-processing prior to data analysis are vitally important to ensure meaningful results.

2.8.4 Supervised Methods in Multivariate Statistical Analysis (MVSA)

MVSA can be improved by including external data, such as clinical data, in a Y table or Y matrix. This data inclusion then makes it possible to use a different class of MVSA methods, which are called supervised analysis methods. The biggest advantage of supervised methods is that they can identify the variation (and associated variables) in the biological data that is correlated (or co-varies) with the external data, i.e. they improve information recovery and thus interpretation of the biological data. The main supervised methods used in metabolomics-based biomedical research are partial least squares (PLS) and orthogonal projections to latent structures (OPLS) (Gu et al., 2011; Pan et al., 2007; Trygg & Lundstedt, 2007).

PLS is a method that seeks to identify correlation between the dataset matrix (X) and one or multiple variables (contained in Y). Y data may be categorical (class identities, e.g. healthy vs. disease) or continuous (blood pressure, height etc.). If the external variable(s) are qualitative, then the method will discriminate between the corresponding classes and is known as PLS discriminant analysis (PLS-DA) (Trygg et al., 2007). Figure 6 illustrates this distinction. Supervised analyses in metabolomics can be
affected by systematic variation that is unrelated to the class, as this affects any correlation found by the analysis method.

![Figure 6: Data included in a PLS analysis.](image)

OPLS is an improvement on PLS that separates variation in the data into two parts: one that is correlated with the biological factor(s) and one that is unrelated/orthogonal, i.e. OPLS separates $X$ into variation that is predictive of $Y$ and variation that is orthogonal to $Y$ (Brindle et al., 2002; Trygg et al., 2002, 2003, 2007; Wagner et al., 2005). A further development of OPLS is 2-way OPLS (O2PLS). While OPLS only correlates data in $X$ with $Y$, O2PLS is able to correlate $X$ and $Y$ with each other in both directions (Trygg & Wold, 2003). In addition, individual variables from an O2PLS analysis can be visualized as a bivariate 1D loadings plot facilitating identification of potential metabolites (Cloarec et al., 2005). Both, OPLS and O2PLS have recently been preferred to PLS, as separation and correlation of predictive variation to the $Y$ table has been shown to optimize discriminant analysis, improving overall knowledge extraction (Bylesjö et al., 2006; Pinto et al., 2012; Wiklund et al., 2008). This is because both predictive and orthogonal variation can be examined, which may provide more detailed insight into the factors influencing the biological system (Kirwan et al., 2012). Furthermore, O2PLS can be used to correlate two different data sets with each other, e.g. metabolomic and proteomic datasets in an animal model of prostate cancer (Rantalainen et al., 2006). If applied in this way, O2PLS is essentially an unsupervised analysis that is able to correlate variables of different datasets, providing further insight into related structures and pathways in altered metabolic states (Bylesjö et al., 2007).

Different extensions of OPLS or O2PLS have been published, including kernel-OPLS, which improves model prediction (Rantalainen et al., 2007). OnPLS is an extension of O2PLS which determines correlation not only between two, but multiple ($n$) matrices, allowing for integration of any number of
datasets for a given study (Löfstedt & Trygg, 2011). Bi-modal OnPLS is an extension to OnPLS that is not only able to analyze orthogonal variation in variables (columns), but also in samples (rows) (Löfstedt et al., 2012a). This bi-modal approach should provide more informed data analysis, of both, the variables associated to the biological question, and of confounding factors associated to particular samples. Finally, OnPLS path modeling is a method of linking multiple matrices along a set of paths that flow between data blocks. These paths are assumed to be due to a specific causative mechanism, e.g. changes over time, and are able to extract the minimum number of predictive components that have maximum covariance and correlation (Löfstedt, et al., 2012b). Use of these recent extensions of O2PLS is not yet widespread, but highly promising in improving metabolomic data analysis.

One inherent problem of any supervised MVSA method is that, because they attempt to correlate the experimental data (X) with external data (Y), they are prone to introducing bias in the analysis. This can happen due to overemphasis of spurious correlations in the data that are only coincidental and not caused by biology. As a result, MVSA models have to be rigorously validated when compared to unsupervised analysis methods. The value of a validated supervised model may be higher than that of a model originating from unsupervised methods because supervised methods are more directed toward the biological question. There are several methods of validation. The gold standard is the use of an independent set of data to test the predictive power of the original (training) set of data, combined with external cross-validation of the training data set. An established alternative is permutation analysis in which the data in the Y table are repeatedly permuted at random and the model recalculated with the permuted Y data (Westerhuis et al., 2008). If the model is stable and correlations to Y are only of biological origin, randomization and permutation of Y data will reduce the fit and predictability of the model (see Figure 7A). Weak models in which correlations to Y are due to chance, rather than biology, will produce permuted models that may provide similar or superior prediction than the original model and are thus invalid (see Figure 7B).

![Figure 7: Validation by permutation analysis. a – Example of a valid model (original R^2 and Q^2 plotted on right side of panel), with permutations resulting in models that are less predictive (plotted on left side of panel). The x-axis indicates the distance of the permuted model to the original model, and the y-axis indicates R^2 and Q^2. b – Example of an invalid model, with permutations resulting in models with similar or improved predictability.](attachment:figure7.png)
Measures of validity in this context are the $R^2$ value, which measures goodness of fit, and $Q^2$, which measures model prediction ability. However – similar to cross-validation – permutation analysis becomes less trustworthy the lower the ratio between number of samples ($n$) and variables ($k$) is. The turning point may be near a ratio of $n/k$ of < 0.02 – 0.04, although this is not applicable for all data sets, and each study has to be evaluated on its own merits (Rubingh et al., 2006). In situations where validation via permutation analysis is not easily accessible, cross-validated ANOVA can be used which uses cross-validated predictive residuals using two degrees of freedom for each component, and is more reliable than ANOVA which uses fitted residuals (Eriksson et al., 2008).

3 Current Evidence: Metabolomics in Prostate Cancer

3.1 Prostate Cancer Pathophysiology

Prostate cancer (PCa) is the most common internal cancer in men worldwide and is more prevalent and lethal in Western countries (Siegel et al., 2012). Continually evolving methods for early PCa detection have improved outcomes due to earlier treatment and a better prognosis for patients. Current methods of detection (serum prostate-specific antigen (PSA) and/or digital rectal examination) leading to diagnosis (via trans-rectal ultrasound (TRUS) guided biopsy) require improvement due to limited diagnostic sensitivity and specificity. Improved methods will help to avoid morbidity in men for whom a diagnosis of PCa remains elusive due to limitations and problems associated with TRUS-guided biopsy, as is the current situation. Thus, PCa pathogenesis has been extensively studied to facilitate the discovery of new methods for determining the presence of PCa.

The prostate gland sits in the pelvis below the bladder and in front of the rectum. It is a secretory gland that contributes to the seminal fluid component of ejaculate/semen to facilitate sperm motility and egg fertilization in utero. The secretory portion of the gland is called the peripheral zone (PZ), and constitutes 70% of the gland volume. The epithelium within the PZ secretes prostatic fluid, which contains proteins, such as prostatic acid phosphatase and prostatic specific antigen (PSA), and metabolites, such as citrate and polyamines (e.g. spermine) (Costello & Franklin, 2009). Furthermore, prostatic cells have been shown to be present in EPS and ejaculate, which makes both biofluids suitable media for molecular analysis (Gardiner et al., 1996).

Citrate production, after sequestration, by PZ epithelium results in a higher citrate concentration in EPS when compared with blood plasma (Costello & Franklin, 2009). This process is facilitated by zinc-dependent truncation of the tricarboxylic acid cycle by inhibiting the enzyme m-aconitase, as shown in Figure 8.

ZIP1 is the primary transporter for zinc ions in PZ epithelium, and is expressed by the ZIP1 gene, which has consequently been described as a tumor suppressor gene in PCa (Costello & Franklin, 2006). The expression of ZIP1 and other zinc transporters recently has been described as being regulated by the micro-RNA cluster miR-183-96-182 (Mihelich et al., 2011). Zinc ions inhibit m-aconitase, which converts citrate to isocitrate, the first step of the citric acid cycle. As a result, the preferential sequestration of zinc ions in the PZ epithelium causes citric acid cycle truncation, producing an increased glucose requirement within the PZ epithelium.

The resulting high citrate and Zn$^{2+}$ concentrations in PZ epithelium are reflected in EPS. Citrate is important in seminal ion homeostasis, and is the predominant regulator of calcium ions, which are important in the motility, metabolism and fertilization functions of sperm (Owen & Katz, 2005). Levels of
zinc ions are correlated with those of other cations, such as calcium and magnesium, but are considerably higher in concentration. In semen, these cations are largely redistributed in binding to negatively charged seminal vesicle proteins, such as seminogelins, which are vital in regulating sperm function (de Lamirande, 2007). In seminal fluid, zinc ions are bound mostly to metallothionein, with changes in levels of zinc being paralleled by those of this protein which is mostly derived from the prostate itself (Suzuki et al., 1994).

3.1.2 The Malignant Prostate

Malignant transformation of cells is the result of irreversible genetic alterations, most commonly due to mutations. Specific to PCa, malignant transformation impairs Zn$^{2+}$ accumulation, removing zinc-mediated inhibition of m-aconitase. The result is completion of the citric acid cycle and increased ATP production via oxidative phosphorylation. This is reflected by low zinc and citrate concentrations present in PZ epithelium and prostatic fluid, which have been investigated as potential biomarkers (Costello & Franklin, 2009). Further alterations in gene expression impair normal mitochondrial functioning. Coupled with the relatively rapid division and increased basal metabolic rate in cancer cells, increased glycolysis and lactate fermentation in the presence of oxygen occurring in the malignant state increases glucose uptake, as well as proteolysis and subsequent alanine production. Pyruvate is produced in excess of what can be processed by the tricarboxylic acid cycle, and is converted to lactate. This is known as the Warburg effect, and is seen as a marker of advanced disease in prostate and other cancers (Bayley & Devilee, 2012; Warburg, 1956). This process is outlined in Figure 9. Furthermore, increased membraneogenesis

Figure 8: Pertinent physiology of the healthy PZ epithelium. Biochemical reactions are shown by solid/open arrows and regulatory interactions by dashed arrows. When healthy, ZIP1 mediated uptake of zinc inhibits isomerization of citrate to isocitrate by m-aconitase. The result is high intracellular concentrations of zinc ions and citrate, which are secreted to aid in fertilization. Adapted from (Costello & Franklin, 2009; Roberts et al., 2011).
accompanying increased cellular proliferation adds to the changes in the metabolite profile with malignant transformation, and requires synthesis of choline and creatine, which have been shown to be elevated in malignant prostate tissues (Noworolski et al., 2008).

Figure 9: Pathophysiology of the PZ epithelium after malignant transformation. Impaired zinc uptake reduces inhibition of \( m\)-aconitase, resulting in citrate isomerization and completion of the TCA cycle. Alanine is produced secondary to proteolysis and lactate as a consequence of the Warburg effect. Adapted from (Costello & Franklin, 2009; Israël & Schwartz, 2011; Roberts et al., 2011).

3.2 Individual Biomarkers

A single biomarker that is able to confirm the presence of an altered biological process or indicates progression of a disease is a valuable asset in prompting appropriate management for any medical condition to improve the outcome for a particular patient. For instance, extremely high serum levels of the human hormone \( \beta \)-chorionic gonadotropin (\( \beta \)-hCG, a marker normally used in pregnancy) in a male patient with a small testicular mass strongly indicate the presence of choriocarcinoma. While this is an example of an ideal biomarker, such biomarkers do not currently exist for most scenarios in oncology, particularly in PCa (Cole, 2009).

3.2.1 Serum

The most widely used biomarker for PCa is serum human kallikrein 3, also known as prostatic-specific antigen (PSA). PSA is a serine protease that is normally secreted in seminal fluid to catalyze proteolysis of seminal proteins, such as seminogelin (Lilja, 1985). PSA is elevated in blood in the presence of PCa as well as with other prostatic conditions, such as bacterial prostatitis and benign prostatic hyperplasia (BPH). Despite not being specific for cancer, PSA is clinically valued and widely used (Clarke et al., 2010).

The normal range of serum PSA, based on population studies, is defined as <1.0 ng/ml. Serum PSA is not only frequently elevated in other disease states, but can also change in the absence of pathology due to other confounders, such as racial and environmental variables (Henderson et al., 1997; Marks et al., 2006; Vollmer, 2004). In biomarker research, often a cut-off point is derived from studies to deter-
mine the optimal sensitivity (i.e. the ability of the test to accurately predict true negative patients) and specificity (i.e. the ability of the test to predict accurate true positive patients). A vast body of evidence has shown that a safe cut-off value for PSA does not exist (Schröder et al., 2000; Thompson et al., 2004). Safety in this context refers to a level that is low enough to detect the majority of men with cancer, but not so low as to cause extensive and unnecessary investigation of men without cancer. This finding of the absence of a safe cut-off value for PSA has been used as one of the major arguments against population screening for PCa using PSA (Catalona et al., 2012; Moyer, 2012).

Serum PSA is a clinically valued test when used with discrimination, and various adaptations have been discovered and trialed with varying success, though none has been considered superior to total serum PSA itself (Auprich et al., 2012; Roobol et al., 2009). Examples include free to total PSA levels, PSA velocity and doubling time (time course of an increase by a factor of two), PSA density (serum PSA in relation to the prostate volume determined by TRUS) and, most recently, the prostate health index which incorporates serum PSA, pro-PSA and percentage free PSA (Catalona et al., 2011; Hori et al., 2012; Lughezzani et al., 2012).

Other serum biomarkers in PCa diagnosis vary in type and size, from circulating tumor cells (Danila et al., 2011; Doyen et al., 2012), to microRNAs (Catto et al., 2011; Sevli et al., 2010), with small molecules and ions, such as sarcosine and zinc, having yielded inconsistent results as markers (Daragó et al., 2011; Li et al., 2005; Lucarelli et al., 2012; Struys et al., 2010).

3.2.2 Urine

Ideally, the perfect marker of PCa is sourced from a non-invasive sample/procedure and indicates both the presence and nature of the disease. Currently, the best urinary marker for PCa is PCA3, formerly known as differential display clone 3 (DD3) (Salagierski & Schalken, 2012). The PCA3 test relies on a patient having had a firm digital rectal examination or prostatic massage just before micturition with the flow of urine flushing dislodged prostatic cells in the prostatic urethra to beyond the external meatus with the void for collection, so there is some licence involved in calling this a urine test. PCA3 is a non-coding RNA which has been shown to be highly expressed in and specific for prostatic tissue (Bussemakers et al., 1999; de Kok et al., 2002; Landers et al., 2005). PCA3 in urine is expressed as a ratio to PSA RNA, and improves detection compared with serum PSA. Use of recently described PCA3 isoforms may further improve results (Clarke et al. 2009; Haese et al., 2008; Hessels et al., 2003). PCA3 also contributes to and has been recommended for clinical decision making for men with previous negative biopsies but in whom clinical suspicion is high (Tombal et al., 2012), however its role in PCa detection has yet to be established clinically. Inclusion of the TMPRSS2: ERG fusion gene also has been reported to improve detection of PCa with reference to biopsy (Tomlins et al., 2011).

3.2.3 Seminal Fluid

The main concern with tests using EPS is that both firm DRE / prostatic massage and TRUS biopsy target the posterior part of the prostate, and neglect anterior and anterolateral aspects of the gland in which up to 30% of PCa’s are sited (Quann et al., 2010; Samaratunga et al., 2007). In contrast, seminal fluid contains a prostatic component, which is the result of global smooth muscle contraction, and thus theoretically reflects the pathological status of the whole gland. Furthermore, the ability to produce seminal fluid via ejaculation is an indicator of cardiovascular status, as erectile dysfunction is a known event in deteriorating cardiovascular status (Chew et al., 2011; Schouten et al., 2008). Thus, men who produce seminal flu-
id are expected to have a more favorable mortality outcome following intervention with curative intent for PCa than men who are impotent since cardiovascular disease is the commonest cause of patient demise in this population. Prostatic tissue and prostatic fluid show similar levels of citrate and Zn\textsuperscript{2+}, further suggesting that prostatic fluid reflects intraprostatic pathophysiological status (Zaichick et al., 1996).

### 3.2.4 Metabolite Changes in Seminal Fluid

Historically, changes in citrate and Zn\textsuperscript{2+} in PCa have been the most pronounced and easily detectable in prostatic and seminal fluid (Cooper & Imfeld, 1959). Metabolite profiling and recent metabolomic analysis of seminal fluid and EPS have discovered alterations in other metabolites, summarized in Table 3. Disturbed zinc homeostasis removes the inhibition of m-aconitase, resulting in citrate oxidation in the citric acid cycle. This causes luminal Zn\textsuperscript{2+} and citrate depletion. Zinc depletion only occurs in PCa, and has been shown to be a stable indicator of PCa status and progression (Zaichick et al., 1996).

<table>
<thead>
<tr>
<th>PCa-induced change</th>
<th>Metabolite</th>
<th>Role (normal)</th>
<th>Alteration hypothesis</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Increase</strong></td>
<td>Choline</td>
<td>Membrane phospholipid precursor</td>
<td>Increased membraneogenesis</td>
<td>(DeFeo &amp; Cheng, 2010; Swanson et al., 2003, 2006, 2008)</td>
</tr>
<tr>
<td></td>
<td>Lactate</td>
<td>End product of anaerobic glycolysis</td>
<td>Warburg effect</td>
<td>(DeFeo &amp; Cheng, 2010; Swanson et al., 2006; Tessem et al., 2008)</td>
</tr>
<tr>
<td></td>
<td>Alanine</td>
<td>End product of anaerobic glycolysis</td>
<td>Warburg effect</td>
<td>(DeFeo &amp; Cheng, 2010; Swanson et al., 2006; Tessem et al., 2008)</td>
</tr>
<tr>
<td></td>
<td>Omega-6 fatty acids</td>
<td>Cell membrane biosynthesis, fatty acid oxidation</td>
<td>Altered gene Expression</td>
<td>(Stenman et al., 2009)</td>
</tr>
<tr>
<td></td>
<td>Cholesterol</td>
<td>Membrane biosynthesis, androgen regulated</td>
<td>Increased cell Turnover</td>
<td>(Thysell et al., 2010)</td>
</tr>
<tr>
<td></td>
<td>Sarcosine</td>
<td>Glycine metabolism, purine synthesis</td>
<td>Cell invasion</td>
<td>(Sreekumar et al., 2009)</td>
</tr>
<tr>
<td></td>
<td>(Choline + creatine) / citrate</td>
<td>Metabolite ratio</td>
<td>Increased ratio</td>
<td>(Kurhanewicz et al., 1995; van Asten et al., 2008)</td>
</tr>
<tr>
<td></td>
<td>Choline / citrate</td>
<td>Metabolite ratio</td>
<td>Increased ratio</td>
<td>(van Asten et al., 2008)</td>
</tr>
<tr>
<td></td>
<td>Choline / creatine</td>
<td>Metabolite ratio</td>
<td>Increased ratio</td>
<td>(van Asten et al., 2008)</td>
</tr>
<tr>
<td><strong>Decrease</strong></td>
<td>Citrate</td>
<td>Ion homeostasis, pH buffer</td>
<td>m-aconitase activation</td>
<td>(Serkova et al., 2008)</td>
</tr>
<tr>
<td></td>
<td>Spermine</td>
<td>Polyamine synthesis</td>
<td>Oxidative stress, enzyme alteration</td>
<td>(Serkova et al., 2008; van der Graaf et al., 2000)</td>
</tr>
<tr>
<td></td>
<td>Myo-inositol</td>
<td>Membrane biosynthesis</td>
<td></td>
<td>(Serkova et al., 2008)</td>
</tr>
<tr>
<td></td>
<td>Citrate / spermine</td>
<td>Metabolite ratio</td>
<td>Decreased ratio</td>
<td>(Lynch &amp; Nicholson, 1997)</td>
</tr>
<tr>
<td></td>
<td>Citrate / creatine</td>
<td>Metabolite ratio</td>
<td>Decreased ratio</td>
<td>(van Asten et al., 2008)</td>
</tr>
</tbody>
</table>

**Table 3**: Summary of metabolite changes in prostate cancer. Changes generic to cancer, such as lactate and alanine, are listed together with changes specific to prostate physiology, such as citrate, sarcosine and spermine. Adapted from (Roberts et al., 2011).
Although citrate levels are altered in other pathophysiological states, such as BPH and prostatitis (Cooper & Farid, 1964), reduced citrate concentrations in histologically benign prostatic tissue is considered to precede microscopic evidence of PCa (Dittrich et al., 2012). In poorly differentiated tumors, these normally abundant metabolites are present in very low concentrations (Kurhanewicz et al., 1993). This metabolite relationship in PCa has also been correlated with the Gleason histological scoring system and is more accurate than serum PSA (Kline et al., 2006). Such biochemical changes reflect early neoplastic processes that may not be histologically identifiable, a concept familiar in oncology as the “field effect” (Costello & Franklin, 2009). This further supports the role of metabolomics in identifying significant metabolic alterations in pre-malignant tissue.

Other metabolite changes seen in oncology that are not prostate-specific are also present in seminal fluid. Disturbed synthesis and intracellular depletion of polyamines, such as spermine, are reflected in prostatic fluid (Cheng et al., 2001; Kline et al., 2006; Serkova et al., 2008; van der Graaf et al., 2000). The prostate contains the highest levels of spermine in the body, and disturbances in ornithine-decarboxylase in polyamine metabolism have been a hypothesized mechanism for spermine depletion (Mohan et al., 1999; Simoneau et al., 2008; van der Graaf et al., 2000). A role of increased reactive oxygen species production by increased expression of spermine oxidase in PCa has linked inflammation with PCa carcinogenesis (Goodwin et al., 2008). Levels of myo-inositol, a molecule involved in membrane biosynthesis, have also been shown to be reduced in prostatic fluid (Lynch & Nicholson, 1997; Serkova et al., 2008).

Some changes in metabolite levels in prostatic tissue are not reflected in prostatic fluid. Choline is upregulated in PCa tissue, both in vitro and in vivo, being hypothesized as another metabolite involved in membrane biosynthesis. The use of choline as a marker in prostatic fluid is compromised by the endogenous conversion of phosphocholine (from the seminal vesicles) to choline catalyzed by prostatic acid phosphatase (from the prostate) shortly following ejaculation. This produces a biological artefact in choline concentration. Lactate and alanine are also increased in PCa tissue as part of the Warburg effect. However, spermatozoa utilize fructose from the seminal vesicles and glucose via glycolysis to produce ATP to fuel flagellal movement in utero to aid fertilization, resulting in varying levels of lactate and alanine as metabolic by-products. This illustrates a confounding factor between external cellular components and intraprostatic metabolites.

### 3.3 Metabolomics in Prostate Cancer Diagnosis: Finding the Best Combination

Metabolite concentrations and ratios have aided in distinguishing PCa from benign prostates (Swanson et al., 2003, 2008; Tessem et al., 2008; van Asten et al., 2008). Yet, despite promising preliminary results, there is currently no test available that is accepted as an accurate, stand-alone diagnostic or screening test. With improved data acquisition and processing technology, the concept of using entire metabolic profiles as a large-scale combination of biomarkers has become feasible. Furthermore, metabolite profiles have been shown to be more sensitive as predictors of PCa, and in predicting metastatic potential (Hricak et al., 2007) (Cheng et al., 2005; Mazaheri et al., 2008; Wu et al., 2010). This concept has been demonstrated by metabolic imaging, in which multivoxel MR spectra of intact prostates were analysed with MVSA. This was able to detect highly significant changes between the global metabolite profiles of benign and malignant prostate tissue without the need to identify specific metabolites (Wu et al., 2010). Similar relationships were demonstrated using freshly frozen PCa tissue when microarray gene expression data and metabolomic data were combined using PLS, providing further insight into mechanisms of metabolite alterations in PCa (Bertilsson et al., 2012). In another study, metabolomic profiling provided
an accurate prediction of biochemical recurrence of PCa, that is a rise in serum PSA, following intervention (Maxeiner et al., 2010). This illustrates the potential of metabolomics as a suitable method for monitoring PCa behavior following clinical interventions (Roberts et al., 2011).

The concept of multiple markers in cancer diagnosis has been examined for some time. Specific to PCa, panels of molecular and protein-based markers have been used to improve serum PSA-based PCa detection (Cao et al., 2011; Cuperlovic-Culf et al., 2010; Talesa et al., 2009). The most widely publicized and promising appear to be the combination of PCA3 and TMPRSS2-ERG fusion transcripts in post-massage urine as previously described. Multiple studies have shown improved sensitivity when combining PCA3 and TMPRSS2-ERG compared with PCA3 or serum PSA alone (Hessels et al., 2003; Tomlins et al., 2011). Multiple mRNA markers (GalNAc-T3, PSMA, Hepsin and PCA3) in malignant prostate tissue have been able to provide optimal detection rates (Clarke et al., 2010; Landers et al., 2005, 2008). Other researchers have attempted to combine single markers of different origins to improve PCa diagnosis. For example, a multiplex model utilizing gene-, protein- and metabolite-based targets for PCa outperformed any single biomarker (Cao et al., 2011). However, although these studies are promising in improving PCa diagnosis, many are impractical for use in a clinical setting, mostly due to financial and logistical constraints. Thus, in addition to improved accuracy, a further potential benefit of using metabolomics in PCa diagnosis is a reduction in cost and logistical requirements for each sample, although a high initial capital equipment financial outlay is required. Further limitations of this approach are outlined below.

3.4 Limiting/Confounding Factors

As previously discussed, detecting PCa is difficult. This is due to many confounding factors relating to the pathophysiology of PCa, but also due to concomitant prostatic disease mimicking PCa. In the majority of studies to date, the greatest interference arises from concomitant pathophysiology of the prostate, such as BPH and prostatitis. Serum PSA is known to be elevated in BPH: androgens contributing to BPH development drive PSA synthesis which is mirrored in serum levels. Prostatitis associated with inflammation and prostatic cell lysis, results in increased release of intracellular PSA into the bloodstream, elevating serum PSA. It is for these reasons that serum PSA lacks sensitivity (detecting many false positives) and that there is sustained criticism directed toward serum PSA testing.

Tissue and prostatic fluid levels of citrate were reported to be initially promising in PCa diagnosis compared with serum PSA, but were observed in the past to be depleted in prostatitis (Averna et al., 2005; Kavanagh et al., 1982; Kline et al., 2006). Notwithstanding similar potential dilemmas as those experienced with serum PSA, diminished levels of both zinc and citrate in these samples may provide improved sensitivity in PCa detection, although conclusive evidence in conjunction with PSA elevations is lacking to date (Daragó et al., 2011; Kavanagh et al., 1982; Zaichick et al., 1996). The relationship of citrate depletion in tissues following radiotherapy or hormonal therapy also relates to biochemical recurrence, defined as a rising serum PSA following intervention with curative intent (Menard et al., 2001; Mueller-Lisse et al., 2001). Furthermore, reduced specificity of serum PSA when compared with metabolite diagnosis may be due to pre-malignant disturbances in metabolic homeostasis that are not histologically visible (Costello & Franklin, 2009). The underlying issue of biochemical characterisation preceding histopathology creates ongoing uncertainty, as tissue histopathology is the current gold standard for PCa diagnosis.

Intra- and extracellular citrate levels are also known to increase in BPH, thus citrate estimation in biopsy tissue may be unreliable. In these circumstances, prostatic fluid may be a more appropriate sample
to use since prostatic fluid is produced mostly in the peripheral zone, which is also where most PCa is located (Costello & Franklin, 2009), whereas BPH develops in the transition (central) zone of the prostate. A small proportion of transition zone tumors may be missed, but in the large majority of cases these are less aggressive and therefore of less significance (Grignon & Sakr, 1994).

4 Future Directions

4.1 Pharmacometabolomics and Theranosis: Towards Personalized Medicine

Pharmacometabolomics seeks to predict the metabolic response to exogenous therapeutic agents prior to or during drug administration, and theranosis is the identification and monitoring of optimal treatments for patients as guided by diagnostic tests (Clayton et al., 2006; DeNardo & DeNardo, 2012; Nicholson et al., 2012). Both are important aspects, given the recently emerging evidence of inter-individual differences in drug pharmacokinetics, being the ability of an individual to absorb, distribute, metabolize and excrete an administered drug (Suhre et al., 2011). The result is reduced therapeutic efficacy, but may also be responsible for toxicity and adverse drug effects. The etiology of these differences in drug metabolism is diverse and is well understood in only a limited number of circumstances (e.g. Cytochrome p450 enzyme family (De Gregori et al., 2010)). Gender, age, race and concomitant diseases have been suggested as inherent factors, but have yet to be substantiated (Nebert et al., 2003). External factors such as dietary and lifestyle habits, as well as toxin exposure, may also have a large influence on therapeutic efficacy. Furthermore, less obvious but important factors may contribute, such as altered gut flora in various circumstances (Dumas et al., 2006; Ley et al., 2006; Wang et al., 2011). As such, metabolic phenotypes are diverse and complex due to these many influencing factors (Holmes et al., 2008). Although genetic profiling across different disorders is important, metabolite profiling promises to better reflect the phenotype of disease states, and advanced analysis between both methods may help to identify genes with significant penetrance.

As has been illustrated, the carcinogenic changes in various cancers will cause common changes to individual metabolic profiles that can be investigated with metabolomics. In contrast, each individual patient will exhibit inherently different metabolic profiles, while also responding differently to therapeutic interventions (Wilson, 2009). Thus, the concept of personalized medicine, where treatments are tailored to an individual’s personal metabolic or genetic phenotype, is one that is exciting, and important in advancing medical treatments (Loscalzo et al., 2007). Using metabolite profiles as a representation of metabolic phenotype promises to enable theranosis by providing the most useful information for predicting inter-individual variation that will guide and assess efficacy of treatment outcomes.

Early research focused on genetic predisposition to cancer or alterations in drug-metabolizing enzymes (Yong et al., 2006). Recent research has focused on identifying varying metabolic profiles that indicate significantly affected drug metabolism, with links to altered gut flora homeostasis (Clayton et al., 2006, 2009). This research related individual background urinary metabolic phenotype to biological and therapeutic outcomes of drug metabolism. Other research has identified alterations in metabolite profiles to illustrate pharmacokinetics and early toxicity, important in preventing adverse outcomes from drug toxicity (Winnike et al., 2010). Recently, this approach has been applied to surgery, and allows for per-
sonalized pre-, intra- and post-operative care to improve patient outcomes (Kinross et al., 2011; Mirnezami et al., 2012). The concept of pharmacometabolomics can also be applied to outcome prediction, similar to that used in GWAS, with evidence of serum metabolite levels to be predictive of body mass following chemotherapy for breast cancer (Keun et al., 2009).

Theoretically, pharmacometabolomics has advantages over pharmacogenomics in representing the phenotype resulting from multiple genetic effects. However, it is believed that a combination of approaches will provide best prediction and outcomes (Nicholson et al., 2011). Given current variations in efficacy, toxicity and adverse outcomes of treatments, developing personalized medicine is imperative to provide better medical care to patients while also reducing health budget costs. Thus, the pharmacometabolomic approach is one that has potential to change the therapeutic landscape not only in oncology, but across all fields of medicine.

4.2 Metabolomics to Elucidate Biological Mechanisms

As outlined, metabolomics has been useful in displaying changes in metabolites in various healthy and pathological states. The analysis of metabolites illustrates the end product of normally functioning or disturbed cellular processes and mechanisms. Thus, analysis of changes in metabolite profiles can lead to insights about the underlying biochemical or biological mechanisms. This has been demonstrated in different areas, including but not limited to, drug toxicity, cancer and plant studies (Bylesjö et al., 2007, 2008; Klenø et al., 2004; Rantalainen et al., 2006).

For example, metabolomics could explain how altered STAT5 signaling as a result of truncated intracellular domains of growth hormone receptor in liver tissue leads to late-onset obesity, as systemic metabolite changes were consistent with globally altered metabolism contributing to obesity (Schirra et al., 2008). A similar approach was used for data obtained from a human prostate cancer xenograft model in mice measured by NMR-based metabonomics and proteomics (two dimensional difference gel electrophoresis). Pathway analysis was used to link altered protein expression to changes in amino acids, which contributed to the metabolic phenotype (Rantalainen et al., 2006).

4.3 Integration with other -omics

As previously outlined, the -omics approach to sample analysis provides data sets that require complex statistical analysis to extract meaningful information. In isolation, each -omics field provides insight into that particular level of cell function, and interactions and influences causing the results are hypothesized based on previous research or logical thinking. Thus, appropriate integration of -omics datasets has become an important step in providing meaningful information in systems biology (Crockford et al., 2006; Bylesjö et al., 2008). This approach was e.g. used in insulin resistant mice using NMR-based metabonomics and genomics (quantitative trait locus mapping), and showed altered gut metabolites that were linked with genomic alterations (Dumas et al., 2007). In both studies, large datasets were used to determine which metabolites were similarly affected by alterations in precursor compounds.

A suggested method that uses O2PLS for integration of large datasets for optimal information recovery is outlined in Figure 10 (Bylesjö et al., 2008), using the example of a study that has data from transcriptomics, proteomics and metabolomics experiments. As O2PLS is able to extract information from two datasets at a time, the method for correlating multiple datasets is naturally a multi-step procedure. In the first step the joint variation between two of the three datasets (e.g. transcript and metabolite data) is extracted by O2PLS. This means using one of the datasets as $X$ matrix in O2PLS and the other
one as $Y$ matrix. Note that in this case, O2PLS is effectively run as unsupervised analysis, because only two experimental datasets are correlated against each other, without including a set of external metadata. In the second step, the joint variation between transcript and metabolite data obtained in step 1 is then correlated with the third dataset (proteomics) in a second O2PLS, which will yield the variation common to all three datasets. In the final step, the joint variation that is common to all three datasets is deflated from the original datasets in a series of three parallel O2PLS analyses, to produce variation that is specific to each dataset.

**Figure 10:** Graphic representation of stepwise data integration of multiple –omics datasets with O2PLS. In the first step, joint variation between two –omics datasets (e.g. transcript and metabolite data) is determined. Using O2PLS in a second step, this joint variation is then correlated with the third –omics dataset (e.g. protein data) to determine variation that is joint to all three sets of data. The third step removes the joint variation between all three individual datasets to produce variation that is specific to each dataset. This dataset-specific variation may be important in helping to address the biological question. Adapted from (Bylesjö et al., 2008).

It is trivial to extend this scheme to more than three datasets by adding on further O2PLS steps between steps 2 and 3 that each time introduce a further set of experimental data into the analysis. It should also be noted that it is prudent to repeat the first two steps with different orders of combining the three datasets – e.g. transcript and metabolite data first, then including protein data, versus metabolite and protein data first, then including transcript data, etc. – in order to rule out that potential slight imperfections in the symmetry behavior of O2PLS might cause secondary effects on the data analysis.

### 4.4 Use of Computational Modeling

As is widely highlighted, the current approach to metabolomics including valid statistical analysis, metabolite identification, and biological interpretation is highly time-consuming. As such, the quest to develop computerized methods of metabolite analysis and identification is underway (Aggio et al., 2011; Tulpan et al., 2011). Following metabolite identification, the next step is to determine the relationship and similarities, if any, of the identified metabolites to metabolic pathways, of which some preliminary
Programming applications have been released to address this issue (Aggio et al., 2010; Bebek & Yang, 2007; Leader et al., 2011).

Even more promising is the development of genome-scale computer models of metabolic networks. Extensive work has been completed on bacteria such as *Escherichia coli*, with *in silico* simulation reported to mimic experimental changes (Edwards et al., 2001). Application of these reconstructed networks is more complicated in eukaryotes, such as human cells, due to the complex cellular and organismic organisation, including intra- and extra-cellular regulation and interactions. Despite these, an initial model human cell was constructed to provide a general baseline in expression and response to biological variables (Bordbar & Palsson, 2012; Mo et al., 2007). Furthermore, a reconstruction of healthy liver cells was combined with whole-body pharmacokinetics to investigate multiple levels in biological organisation and provide mechanistic insights into for various drug-induced scenarios (Krauss et al., 2012). Alterations to such models to accurately reflect cancer and other pathophysiological states by incorporating known and emerging evidence will better describe the response of these cells (Jerby & Ruppin, 2012). Depending on the type of model, spatiotemporal processes and interactions within cells that may be undefinable or difficult to quantify are currently difficult to incorporate and apply to an artificial model (Materi & Wishart, 2007). Further development of these reconstructed networks may occur via integration of –omics data sets, and research in this field is continuing (Joyce & Palsson, 2006). Computational modeling by incorporating multiple data sets represents a logical and informative, yet challenging, approach to oncology research to guide pharmaceutical development strategies. The result will be better informed treatment approaches and improved treatment outcomes for these patients.

5 Conclusions

Metabolomics is a novel, modern and robust scientific approach that has shown great advances across many fields in biomedical research. The application of metabolomics to differing fields in medical science, including pathophysiology insight, drug development and *in vivo* imaging make it unique from all other approaches. Further research and collaboration to develop reconstructed networks, via integration of many terabytes of –omics data, is the next frontier in providing valuable insights to advance medical research and treatments in various human disease states.

References


