A metabolomics study of selected perturbations of normal human metabolism.

Elmarie Davoren

BScHons Biochemistry

13106597

Dissertation submitted in partial fulfilment of the requirements for the degree

Master of Science in Biochemistry

at the Potchefstroom campus of the North-West University

Supervisor: Prof C Reinecke

Co-supervisor: Dr G Koekemoer

June 2010

ABSTRACT

Metabolism is an integrated network of biochemical pathways involving a series of enzyme-catalysed anabolic or catabolic reactions in cells. Metabolites are chemical compounds that are involved in or are products of metabolic pathways, and the metabolome is defined as the total complement of all the low molecular weight metabolites present in a cell at any given time. Metabolomics is a relatively new research technology utilised for the global investigation, identification and quantification of the metabolome. Three aims were defined for the metabolomics study presented here:

- The use of metabolomics technology to generate new biological information;
- Application of the metabolomics technology to gain information on the three natural perturbations, namely the menstrual cycle, pregnancy and aging; and
- Reflection on metabolomic studies as a hypothesis-generating approach.

I obtained three sets of urine samples from women during their menstrual cycle, samples from sixteen pregnant and eleven non-pregnant women for the second natural perturbation, and data sets from previous investigations on infant and child groups, as well as thirty-two urine samples from adults for the study of the metabolomic profiles due to age. These urine samples were analysed to determine the organic acid metabolite profiles. The metabolites were identified by means of AMDIS and were manually quantified. Data matrixes were compiled, which underwent certain data reduction steps, prior to statistical analysis. Different statistical approaches were used to generate information on these three natural perturbations due to the clear differences between the three experimental groups used. The investigation of the menstrual cycle did not show a distinct difference between the three phases involved in the cycle, whereas the pregnancy perturbation showed a difference between pregnant groups and non-pregnant groups. The most pronounced difference in metabolite profiles were found when the different age groups were compared to one another. Finally a hypothesis on the effect of age on metabolism was defined and an experimental approach was proposed to evaluate this hypothesis.

In conclusion three proposals were formulated from this investigation:

1. If it appears that an insufficient number of participants can be generated for a metabolomics study, such a study should be discarded in the interest of a more feasible investigation.
2. It is advisable that a number of appropriate analytical validation parameters should be incorporated in the early stages of a metabolomics study, specifically linked to the context of the perturbation chosen for the investigation.

3. The control and experimental groups should be homogenous that is to say as comparable as possible with regard to age, ethnicity, diet, and gender, lifestyle habits and other possible confounding influences, except for the specific perturbation being studied. In a perfect world this would be possible, specifically when hypothesis formulation, testing and finally the expansion of scientific knowledge is a desired outcome of the investigation.

Key words: Metabolomics, natural perturbations, menstrual cycle, pregnancy, infants, children, adults, organic acids, multivariate analysis, biomarkers

OPSOMMING

Metabolisme is ‘n geïntegreerde netwerk van biochemiese weë van ‘n reeks ensiem-gekataliseerde anaboliese en kataboliese reaksies in selle. Metaboliete is die chemiese verbindings wat betrokke is by of produkte is van metaboliese weë, en die metaboloom word gedefinieer as die totale komplement van al die laemolekulêregewig-metaboliete teenwoordig in ‘n sel op ‘n gegewe tyd. Metabolomika is ‘n relatiewe nuwe navorsingstegnologie wat gebruik word vir die indentifisering en kwantifisering van die totale metaboloom. Drie doelwitte is gedefinieer vir diè metabolomikastudie van hierdie verhandeling:

- ervaring in die gebruik van metabolomikategnologie om nuwe biologiese inligting te genereer;
- die gebruik van die metabolomikategnologie om inligting te kry oor drie natuurlike perturbasies, naamlik die menstruale siklus, swangerskap en veroudering; en
- evaluering op hierdie metabolomikastudie kan lei to ‘n moontlike hipotese vir verdere navorsing.

Ek het drie stelle urienmonsters ontvang vanaf vroue gedurende hulle menstruele siklus, vanaf sestien swanger en elf nie-swanger vrouens vir die tweede natuurlike perturbasie, en datastelle van vorige studies op baba- en kindergroepe, asook twee-en-dertig urienmonsters vanaf volwassenes vir die bestudering van die metabolomikaprofiele op grond van ouderdom. Hierdie urienmonsters is geanaliseer om die organiesesuurmetaboliet-profiel te bepaal. Die metaboliete is geïdentifiseer d.m.v. AMDIS en is met die hand gekwantifiseer. Datamatrikse is saamgestel, deur ‘n aantal datareduksiestappe toe te pas, waarna statistiese analise gedoen is. Verskillende statistiese benaderinge is gebruik om inligting te genereer vir die drie natuurlike perturbasies op grond van duidelike verskille tussen die drie eksperimentele groepe wat gebruik is. Die bestudering van die menstruele siklus het nie ‘n duidelike verskil tussen die drie fases aangetoon nie, terwyl die swangerskapperturbasie ‘n verskil tussen die swanger en nie-swanger groepe uitgewys het. Die duidelikste verskil in metabolietprofiele is gevind by die verskillende ouderdomsgroepe wat met mekaar vergelyk is. ‘n Hipotese oor die effek van ouderdom op die metabolisme is geformuleer en ‘n eksperimentele benadering is voorgestel om die hipotese te evalueer:

Ter afsluiting is drie benaderings geformuleer wat uit hierdie ondersoek afgelei kan word:

1. Wanneer ‘n onvoldoende aantal deelnemers gegenereer word vir ‘n metabolomika-studie, dan moet ‘n alternatiewe en meer uitvoerbare studie eerder gedoen word, eerder as om met onvoldoende eksperimentele deelnemers voort te gaan.

2. Dit is raadsaam dat ‘n aantal gepaste analitiese validasieparameters, wat beïnvloed word deur die spesifieke perturbasie wat ondersoek word, geïnkorporeer word in die vroeë stadium van ‘n metabolomikastudie.

3. Die kontrole en eksperimentele groepe moet so homogeen as moontlik wees ten opsigte van hulle ouderdom, etnisiteit, dieet, geslag, leefstyl en ander moontlike invloede, behalwe vir die spesifieke perturbasie wat ondersoek word. In ideale omstandighede behoort dit moontlik te wees, veral wanneer hipoteseformulering, toetsing en uitbreiding van wetenskaplike kennis ‘n verlangde uitkoms van die ondersoek is.

Sleutelwoorde: metabolisme, natuurlike perturbasies, menstruale siklus, swangerskap, baba, kinders, volwassenes, organiese sure, meerveranderlike analise, biomerkers

ACKNOWLEDGEMENTS

I would like to thank the following people for their contribution to my study:

- Prof Carools Reinecke, my supervisor for his guidance and support.
- Dr Gerhard Koekemoer, my co-supervisor for his assistance with regard to the statistical aspects of this study.
- Dr M Nelson for the language editing of my dissertation.
- The personnel of the Laboratory for Inherited Metabolic Defects and Mr Peet Jansen van Rensburg for their assistance during the experimental component of this study.
- Dr Hlengiwe Mbongwa, for her help, support, and guidance during my study.

This study is dedicated to the following people who mean the world to me and without their support and love I would not have achieved my Masters degree: my parents, William and Suset, my sisters, Elandrie and Sanel and my best friends, Carien Mulder and Amaria van Huyssteen.

“To strive, to seek, to find, and not to yield.”

- *(Ulysses, Lord Alfred Tennyson)*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

3-HIA	3-hydroxyisovaleric acid
A	
A	Adults
AMDIS	Automated mass spectral deconvolution and identification system
B	
B C/B L	Black children
BMI	Body mass index
BSTFA	O-bis(trimethylsislyl)-trifluoracetamide
C	
C	Child
Ca C	Caucasian children
CE-MS	Capillary electrophoresis mass spectrometry
Co/Co C	Coloured children
D	
DNA	Deoxyribonucleotide acid
F	
FSH	Follicle-stimulating hormone
G	
GC-MS	Gas chromatography mass spectrometry
H	
hCG	Human chorionic gonadotropin
HCl	Hydrochloric acid

HDLs	High-density lipoproteins
hPL	Human placental lactogen
I	
IEM	Inborn Errors of Metabolism
IM	Infants younger than a year
L	
LC-MS	Liquid chromatography mass spectrometry
LDLs	Low-density lipoproteins
LH	Luteinizing hormone
N	
NMR	Nuclear magnetic resonance
NIST	National Institute of Standards and Technology
P	
PC1	Principal component one
PC2	Principal component two
PCA	Principal component analysis
PCR-RFLP	Polymerase chain reaction-restriction fragment length polymorphism
PKU	Phenylketonuria
PLS-DA	Partial least squares-discriminant analysis
R	
ROS	Reactive oxygen species
T	
TCA	Tricarboxylic acid cycle

TMCS	Trimethylchlorosilane
TMS	Trimethylsilyl
V	
VLDLs	Very low-density lipoproteins

Chapter 1 – Introduction

The metabolome is the total complement of all low molecular weight molecules in a cell at any given physiological or developmental state (Goodacre *et al.*, 2004). These molecules exhibit a high diversity of chemical structures and abundances which require different analytical platforms complementary to each other as well as multivariate statistical analyses (MVA) in order to determine its extensive coverage (Werner *et al.*, 2008). For this reason metabolomics is viewed as "*a discipline dedicated to the global study of metabolites, their dynamics, composition, interactions, and responses to interventions or to changes in their environment, in cells, tissues, and biofluids*" (Katajamaa and Oresic, 2007).

Metabolomics is a relatively new research field and therefore not many studies have been done on normal human metabolism, since the main focus has been on major metabolic perturbations which have a detrimental effect on those individuals suffering from the perturbation. Natural perturbations of human metabolism may also influence the normal metabolome profile. Consequently the results obtained from metabolomics studies should take variation in the normal metabolome profile into account. Three such natural perturbations (the menstrual cycle, pregnancy and age) were chosen to investigate possible differences in the excretion pattern of metabolites found in the control and experimental groups. This could lead to a better understanding of the human metabolome with respect to these natural perturbations. The metabolomics approach chosen for all three these cases was an untargeted analysis of the metabolites of an applicable subsection of the metabolome, namely a targeted measurement of the urinary organic acids.

The organic acid metabolism, i.e. measurement of the organic acids, was selected for the untargeted metabolomics approach, since the profiling of these metabolites may lead to information about the pathophysiological and physiological condition of various metabolic pathways and their interdependant metabolites. Organic acids also include important components related to normal detoxification pathways, like hippuric acid.

Scope of the dissertation

The material presented in this dissertation is covered in four chapters, each designated to a specific aspect of this investigation.

Chapter 2 - Literature Review

This chapter will review current literature about metabolomics (Section 2.1), the organic acid metabolism (Section 2.2), the three perturbations (Section 2.4) as well as a description of the different steps of a metabolomics approach i.e. workflow (Section 2.5). From this overview three specific aims for this investigation were defined (Section 2.6), related to

- the use of metabolomics technology;
- application of the metabolomics technology to gain information on the three natural perturbations studies; and
- reflection on metabolomics studies as a hypothesis generating approach.

Chapter 3 - Materials, methods and experimental subjects

The aim of this chapter is to discuss the experimental aspects involved in an untargeted metabolomics approach. Section 3.2 describes the experimental subjects for each of the perturbations; Sections 3.3 to 3.5.1 describe the experimental design with regard to sample storage, and the determination of urinary organic acids. In Sections 3.5.2 and 3.6 the analytical method used as well as the protocol followed to identify and quantify the metabolites are discussed. Finally the statistical methods utilised are discussed in Section 3.7.

Chapter 4 - Results

This chapter will discuss the results obtained for each of the three perturbations with regard to the statistical analyses used. Sections 4.2, 4.3 and 4.4 include a short discussion of the respective perturbations namely the menstrual cycle, pregnancy and age. This does not include a comprehensive discussion of the results, as this is presented in the next Chapter.

Chapter 5 - Discussion

This is the final chapter of this Master's study and will focus on the discussion of results and key aspects obtained in relation to the three aims defined in Chapter 2, presented in Sections 5.2, 5.3, and 5.4 respectively. At the end of the chapter (Section 5.5) I will attempt to make a clear recommendation and motivation linked to each of the three aims and the findings in this dissertation, taking future studies into consideration.

Chapter 2 - Literature Review

2.1 Metabolomics investigations

2.1.1 Orientation

Contemporary undergraduate textbooks in human physiology and biochemistry typically defines metabolism as all the chemical reactions that convert nutrient biomolecules like lipids, carbohydrates and proteins to release energy and in addition synthesise or break down molecules present in all organisms (e.g., in Silverthorn (2010) and Garrett and Grisham (2005)). Metabolic reactions thus involve hundreds of enzymatic reactions that are organised into discrete pathways which proceed in a step-wise manner, i.e. the products of one reaction become the substrate for the following reaction or reactions. The molecules that participate in a pathway are called intermediates with key intermediates present in more than one pathway. These intermediates act as branch points for several metabolic pathways as they direct substrates into one or more directions.

The living organism thus contains a large number of metabolites involved in its biological processes on the level of low molecular weight substances, like amino acids, lipids and carbohydrates and their derivatives, as well as macromolecules such as proteins and nucleic acids. Most of these metabolites are internally produced as intermediates and end-products of biosynthetic and catabolism pathways while some are obtained externally. All these low molecular weight metabolites are grouped together to form the **metabolome** of an organism or cell.

Oliver first used the term metabolome, and defined it as a measure of the concentration of as many metabolites as possible (Oliver *et al.,* 1998), which was redefined by Goodacre as "*the quantitative complement of all of the low molecular weight molecules present in cells in a particular physiological or developmental state*" (Goodacre *et al.,* 2004). Harrigan and Goodacre (as referred to by Dunn and Ellis, 2005), define it as "*the qualitative and quantitative collection of all low molecular weight molecules (metabolites) present in a cell that are participants in general metabolic reactions and that are required for the maintenance, growth and normal function of a cell"*. This discrepancy in definitions has given rise to an active debate about the most accurate definition of the "metabolome" (Goodacre *et al.,* 2004). The above-mentioned definitions are, however, all valid as they take into account that organisms are susceptible to even minor changes, i.e. **perturbations**, in their external and internal cellular environment, which leads to variations in their metabolic profiles, however, still within the **homeostatic** state of the organism.

Homeostasis is a well-known phenomenon that refers to the stable, internal state of an organism. This steady state is achieved by numerous complex metabolic reactions and regulatory mechanisms of the organism. Thus it is a dynamic, ever-changing state given that it is constantly adapting to changes in its internal and external environment. When the homeostatic state cannot be maintained i.e. due to an imbalance, this can lead to certain metabolic aberrations that underlie various diseases and may even lead to death.

According to Steuer (2006) three different scenarios can be distinguished on concomitant changes in metabolite concentrations, namely specific perturbations, global perturbations and intrinsic variability. Intrinsic variability and specific perturbations, as well as global perturbations will be discussed in 2.1.1 and 2.1.2 respectively. Traditionally **biomarkers** of perturbations were defined by a change in one or more metabolites e.g. phenylpyruvic acid and phenylalanine which are indicative of phenylketonuria (PKU). With the advent of the "omics-revolution" and its subsequent technologies (e.g. genomics, proteomics, transcriptomics, and metabolomics) it became possible to get a more holistic view of the metabolome as well as the perturbations that influence it.

Fiehn introduced the term '**metabolomics**' in 2002 as "*a comprehensive analysis in which all the metabolites of a biological system are identified and quantified*". Jeremy Nicholson (2003) subsequently defined it as follows "*Metabolomics involves the study of multivariate metabolic response of complex multicellular organisms to pathological stressors and the consequent disruption of system regulation*" and according to Harrigan *et al.,* (2005) "*Metabolic profiling involves the acquisition of metabolome data sets of sufficient spectral and/or chromatographic richness and resolution for multivariate statistical analyses and for metabolite identification and quantification*". Katajamaa and Oresic (2007) recently defined it as "*a discipline dedicated to the global study of metabolites, their dynamics, composition, interactions, and responses to interventions or to changes in their environment, in cells, tissues, and biofluids*".

Three important key concepts that arise from these definitions are the following, which will be discussed briefly:

1. Interventions or changes (see 2.1.1)
2. The metabolite profile (see 2.1.2)
3. Multivariate metabolic responses and multivariate statistical analysis (see 2.1.3).

2.1 2 Interventions or changes

Metabolite levels can change because of their interrelated cellular metabolism and not because of deliberate experimental perturbations or changes in their physiological state. This is referred to as **intrinsic variability** (Steuer, 2006). It is especially evident when there is a variation in the experimental data which is not caused by the experimental design but by the intrinsic variability of cellular metabolism. Intrinsic variability also occurs because organisms differ from one another according to their enzyme concentrations which in turn affect concentrations of metabolites and lead to an interdependency between the metabolites. This phenomenon can be seen in a study done by Martins *et al.,* (2004). They compared the exponential and post-diauxic growth phases of *Saccharomyces cerevisiae* and found that there was a stronger correlation between two of the metabolites (Fumaric acid and alpha-ketoglutaric acid) in the post-diauxic phase than in the exponential phase even though the concentration of alpha-ketoglutaric acid did not change significantly between the two phases. This study was done on plants but illustrates the importance of studying metabolite correlations in parallel with concentration changes i.e. inherent biological variation.

Cellular metabolism is also influenced by environmental factors such as temperature, altitude and exposure to exogenous compounds. **Specific perturbations** differ from intrinsic variability as the changes in metabolite levels resulting from specific localised interference or fluctuations within the underlying network of biochemical reactions for example the knockout or over-expression of a gene coding for an enzyme (Steuer, 2006).

Global perturbations are induced changes within the metabolic network at multiple sites or are caused by external factors that influence different metabolites at the same time, such as environmental changes, transient or diurnal measurements in a time series (Steuer, 2006). In a study performed by Roessner *et al.,* (2001) the effect of environmental manipulation on wild-type potato tissue demonstrates the outcome of global perturbations. They incubated potato tissue in various concentrations of glucose and only the potato tissue incubated at the highest concentrations (200 and 500 mM) of glucose exhibited significant differences. One of the potato tissues had a genetically determined change. This was clearly seen when the metabolite profile of the specific potato tissue was compared to the metabolite profile of a transgenic potato plant. Even though this example is directed at plant metabolism it shows that global perturbations can influence the metabolic network of an organism, which would equally well apply to other eukaryotic systems, like the one studied in this investigation.

2.1.3 The metabolite profile

It is difficult to define “normal human metabolism” since there is such a great difference between and within individuals as well as population groups based on variations due to the diet, environment, genotypes and enzyme concentrations. As a result metabolomic studies are dependent on the experimental condition under which the metabolite profiles of participants were obtained. A metabolite profile contains the estimated quantity of a set of metabolically or analytically related metabolites and their derivatives, detected in biological samples via specialised analytical techniques (Gates and Sweeley, 1978; Villas-Boas *et al.,* 2005). The global metabolite profile may then be defined as the comprehensive and integrated metabolic profiles of the metabolome as it gives a biochemical characterisation of the organisms’ metabolic response and interrelationships and as such is the hardest to interpret. As a result not all research involving metabolite profiling will form part of a metabolomics study, but a lot of the current metabolomic studies make use of metabolic profiling (Villas-Boas *et al.,* 2005).

2.1.4 Multivariate metabolic responses and multivariate statistical analysis

As mentioned in 2.1 the living organism contains a large number of metabolites i.e. the end products of cellular regulatory processes. According to Fiehn (2002) the levels of metabolites can be regarded as the ultimate response of biological systems to genetic or environmental changes. This metabolic response then leads to the excretion of hundreds of metabolites in a single biological sample, such as in urine. A metabolomics investigation comprises of numerous samples differing from one another according to diet, gender, ethnicity etc.. Moreover each sample contains hundreds of metabolites which might even differ from one another according to their retention times, concentrations and mass spectra. Hence a metabolomics investigation produces copious amounts of data and in order to extract relevant biological information from the datasets multivariate analysis is applied. This statistical technique makes it possible to study two or more dependent observations or variables at the same time. Algorithms do not drive a metabolomics investigation, the research question does. However it is a necessary part of any metabolomics investigation as it defines and/or determines possible correlations between and within the groups or variables under investigation via supervised or unsupervised methods (Goodacre).

2.2 The selection of a subsection of the metabolome for this study

Metabolites differ from one another when their chemical (polarity, solubility etc.) and physical properties are compared. This diversity occurs because of the difference in atomic arrangements of these metabolites. These differences affect the separation as well as identification of these substances as done in metabolomics research (Dunn and Ellis*.,* 2005). For the purposes of this study we focused on the organic acid metabolism given that these low molecular weight metabolites i.e. organic acids are involved in many of the pathways of intermediary metabolism, as well as in the metabolism of exogenous compounds (Hoffmann and Feyh, 2003; Figure 2.1). If these metabolites are analysed comprehensively for example by means of metabolomic profiling they may possibly lead to information about the pathophysiological and physiological condition of various metabolic pathways and their interdependant metabolites (Hoffmann and Feyh, 2003).

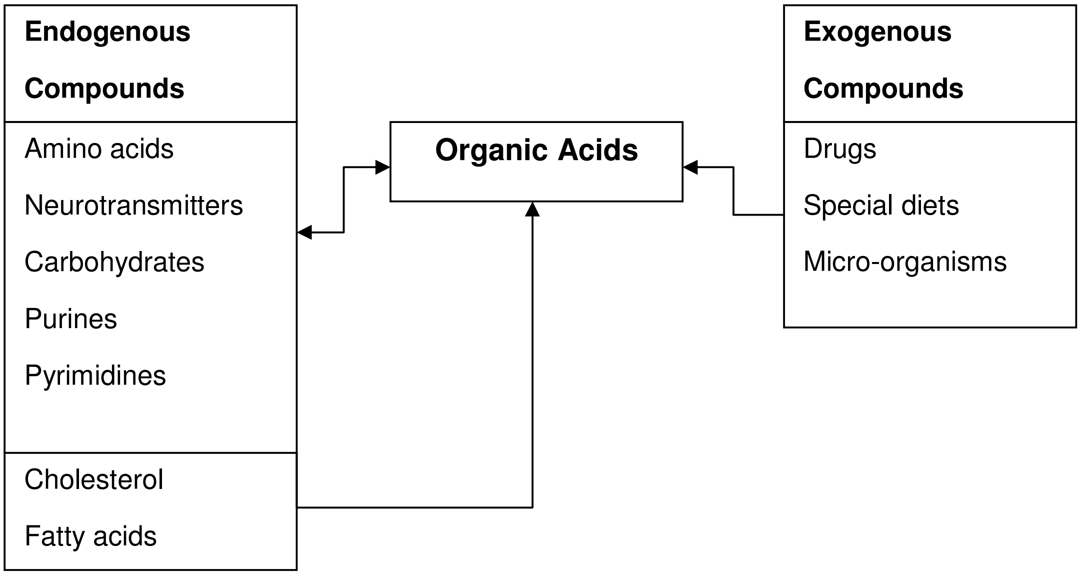


Figure 2.1: Endogenous and exogenous compounds involved in intermediary metabolism (adapted from Hoffmann and Feyh, 2003).

The intermediary metabolic pathways where organic acids play a role include pathways associated with fatty acid metabolism, ketogenesis, the tricarboxylic acid cycle (TCA), pathways of carbohydrate and pyruvate metabolism as well as the amino acid metabolism (Seymour *et al.,* 1997). Consequently organic acids are complex and diverse but these characteristics have also hampered the development of quantitative methods to facilitate comprehensive analysis specifically where all the organic acids can be extracted, analysed and indentified in a single run (Lehotay *et al.,* 1995).

Organic acids are analysed in body fluids (urine or serum) by means of methods generally based upon mass spectrometry or gas chromatography. It involves the extraction of the organic acids as well as their conversion into thermally and chemically stable derivatives for chromatography (Seymour *et al.,* 1997). Organic acids are usually analysed to determine if someone has a disorder or defect in their organic acid metabolism. Organic acidurias or organic acideamias are thus examples of permanent perturbations which influence the metabolome. This analytical technique can also be used to determine possible differences within and between population groups according to their inherent variability or how they react to induced specific perturbations such as alcohol consumption, medication or an infectious disorder.

In a study done by Witten *et al*., (1973) they compared the urinary organic acid profile of 21 healthy young adults. This was done in order to determine the excretion rates of specific organic acids in normal adults, while they were on a controlled diet for three days. This study found that even though the subjects were on a controlled diet there were individual metabolic variations which in turn influenced the excretion of organic acids. This serves as an example of intrinsic variability influencing organic acid profiles.

Organic acidurias involve or occur in the cytosol and in certain organelles such as the peroxisomes, microsomes and mitochondria. They give rise to an accumulation of organic acids, their esters and conjugates in body fluids, body tissues and are mainly excreted in the urine. These disorders may differ from one another but share similarities in biochemistry and chemistry which lead to certain common clinical characteristics that include acute presentation in early life with ketosis, hyperammonaemia, vomiting, convulsions etc. In the newborn infant or young child these disorders are most often fatal and survivors can be mentally or physically handicapped and other patients who present later in childhood do so with neurological deterioration, established failure to thrive etc. (Seymour *et al.,* 1997).

Apart from these biological aspects, analysis of the organic acids are also important from an analytical point of view since organic acids are readily isolated, the derivatisation techniques are well described, the mass spectra of most organic acids are well described and is identified and quantified by means of the AMDIS (Automated Mass Spectral Deconvolution and Identification System) methodology (Lehotay *et al.,* 1995).

2.3 Motivation for the title

During the BScHons project which was assigned to me, we found a significant week-to-week variance in the metabolite profile of single individuals. This is shown in Figure 2.2 for two such experimental subjects. The experiment was set up to determine if the week or even the day of sample collection may influence the urinary organic acid profile. Early morning urine samples were provided by a male and female participant collected over a period of three to four weeks. All the samples were analysed by two independent analysts, in triplicate. Table 2.1 gives the schedule for sample collection from the female and male participant. It shows the number of samples collected in each week and for each day.

Table 2.1: Cross-section of specific days between the weeks for the two participants.

FEMALE PARTICIPANT						MALE PARTICIPANT							
Week	**Day**	**Week**	**Day**	**Week**	**Day**	**Week**	**Day**	**Week**	**Day**	**Week**	**Day**	**Week**	**Day**
1	-	**2**	1	**3**	1	**1**	-	**2**	1	**3**	1	**4**	1
	2		2		2		-		2		2		2
	3		3		3		3		3		3		3
	4		4		4		4		4		4		-
	5		5		-		-		-		-		-

Table 2.1 describes the schedule of urine collection for the female and male participant as well as the number of samples collected in each week and each day. For the statistical analysis we compared the different weeks of each participant with one another. The results of this analysis for the respective participants are shown in Figure 2.2 A and B. Secondly we compared the different days with one another i.e. the day 2 samples were compared to one another for the three or four weeks of urine collection. The results of the day-comparison are shown in Figure 2.3 A and B.

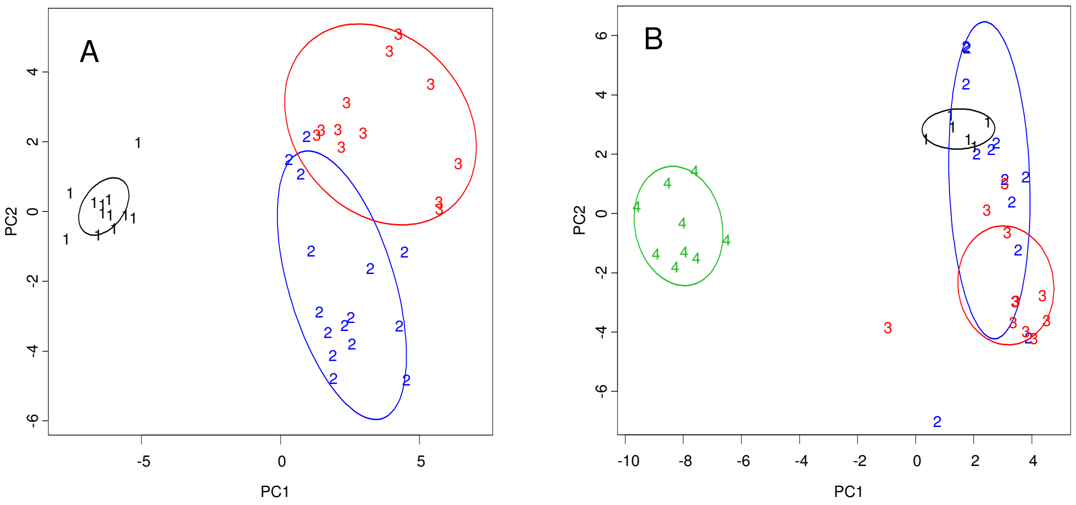


Figure 2.2: A and B. PCA scores plots (see Section 3.7.6.1) give an indication of the week effect in participant A and participant B during the course of three to four weeks of urine samples collected (Week 1: black; week 2: blue; week 3: red and week 4: green)[1].

At first the week-to-week samples were compared and the results obtained showed a distinct difference (Figure 2.2), which led to the following research question that is to say, if a cross-section of specific days between the weeks shows the same tendency? When the day-to-day samples of different weeks were compared i.e. the day 3 to day 3 samples, there was an overlap in data as depicted in Figure 2.3, meaning that there was not a distinct difference between the day-to-day samples. For the week-to-week comparison the female, participant A, showed a more pronounced separation than participant B, the male.

[1] Note: The experimental details for such a metabolomics experiment will be presented later in this dissertation. These results only serve the purpose to introduce an example of a normal perturbation as shown by a metabolomics experiment.

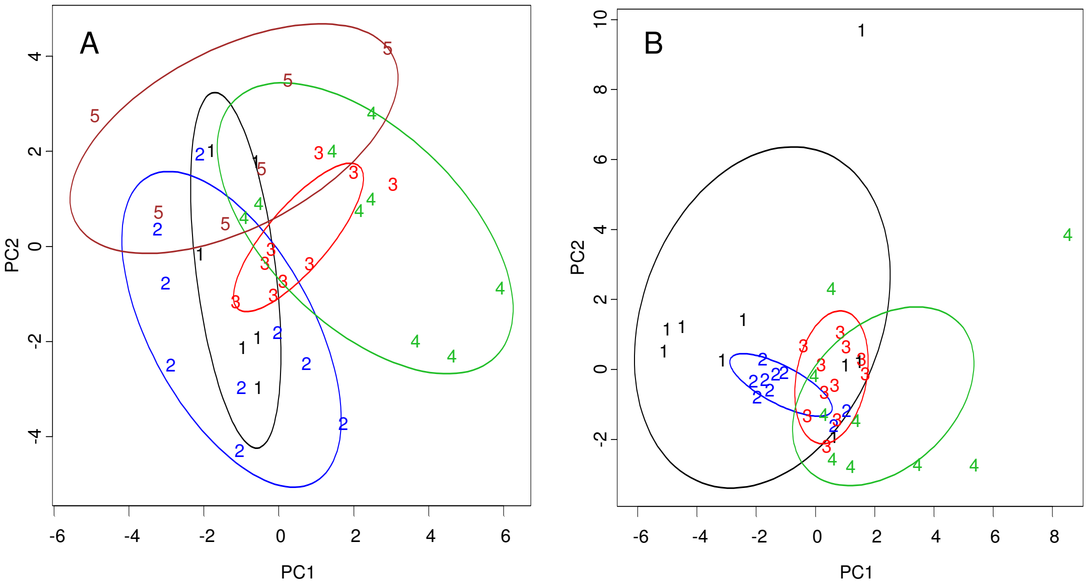


Figure 2.3: A and B. PCA scores plots (see Section 3.7.6.1) give an indication of the day to day effect in participant A and participant B during the course of three to four weeks of urine samples collected (Day 1: black; day 2: blue; day 3: red; day 4: green and day 5: maroon).

This is an important observation with regard to any metabolomics study, as it indicates that a measurable perturbation already occurs on a week-to-week basis under normal physiological conditions. It is very important to design experiments as specific as possible for the biological questions they pose as not all metabolites, in the case of metabolomics, will be relevant to the research question under investigation. Hence the physiological or developmental state of the organism, cells or metabolites should be defined as accurately as possible (Oliver *et al.,* 1998).

We therefore decided to study this phenomenon of perturbations associated with normal conditions in more detail, which thus forms the basis for this MSc investigation, as indicated in the title of the dissertation: **A metabolomics study of selected perturbations of normal human metabolism**. The important aspects in the dissertation title are thus:

- Selected perturbations of normal human metabolism (see 2.3)
- A metabolomics study (see 2.4)
- Bioinformatic analysis (see 2.5)

The final aims of this study are presented at the end of this literature review (see 2.7).

2.4 Selected perturbations of normal human metabolism

Normal perturbations, as defined for this study, may also be defined differently, for example as “a response to a physiological challenge” (Kochar *et al.,* 2006). It was thus emphasised by those authors that the development of a “lifestyle database” of a normal healthy control population group is essential to interpret responses to stimuli like disease, environment or nutrition. For such a control database, they propose that aspects such as gender, age, body mass index and lifestyle should be considered. Kochar *et al.,* (2006) chose NMR (nuclear magnetic resonance) spectroscopy for the generation of their metabonomics data, as metabonomics data, coupled to the necessary multivariate statistical tools “allows the better visualisation of the changing endogenous biological profile in response to physiological challenge” or stimuli than those mentioned above. We have decided to investigate three perturbations that occur during normal physiological conditions, but which are of a progressive complex nature, namely (1) the monthly menstrual cycle in females (measured as perturbation due to the physiological action of specific hormones) (2) pregnancy (seen as the perturbation due to the prenatal foetal development in the mother) and (3) aging (time-dependent metabolic changes). We thus studied normal perturbations with a more intrinsic variability (menstrual cycle), a specific perturbation (pregnancy) and global perturbations (aging).

2.4.1 Menstrual cycle

2.4.1.1 General overview

The first perturbation focused on the menstrual cycle. Female reproduction is characterised by a physiological process that is cyclic since females produce gametes in monthly cycles of twenty five to thirty five days, called the menstrual cycle (e.g. Silverthorn, 2010[2]). These cycles of gamete production, hormone interaction and feedback pathways form part of a complex control system in the body. The menstrual cycle is divided into three phases namely the follicular phase, ovulatory phase and luteal phase. Throughout the cycle four reproductive hormones, namely; estrogen, progesterone, luteinizing hormone (LH) and follicle-stimulating hormone (FSH) undergo cyclical changes, apparently with a circadian rhythm superimposed on the menstrual-associated rhythm, and *vice versa* (Baker *et al.,* 2007). The concentrations of these hormones in blood differ from phase to phase, somewhat between woman to woman and even between multiple cycles of an individual woman (Gandara *et al.,* 2007).

[2] The recent edition of the textbook on Human Physiology by Silverthorn (2010) will be used as the basis for this brief overview on the well-known characteristics of the menstrual cycle.

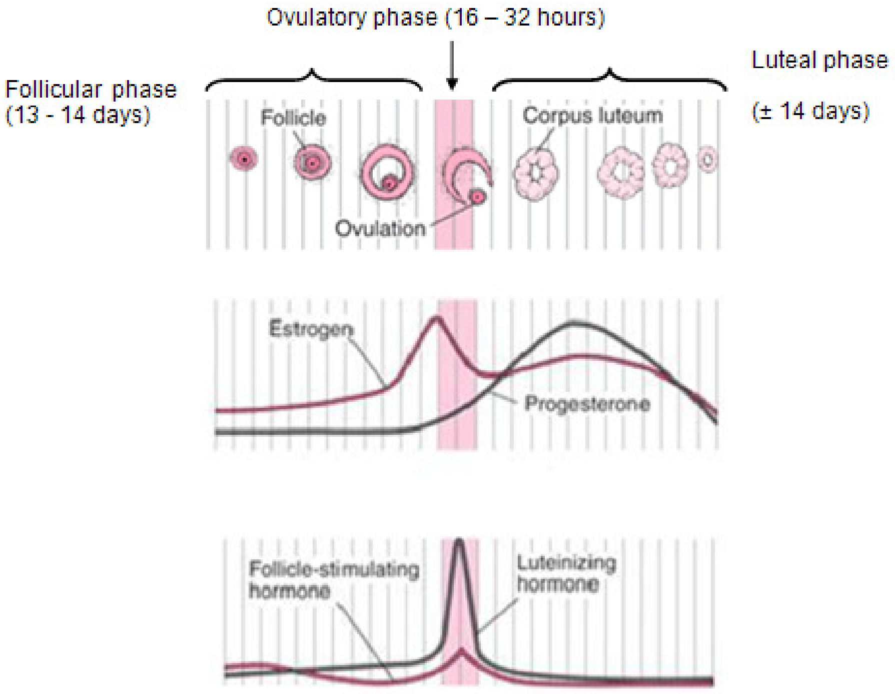


Figure 2.4: Hormonal changes during the menstrual cycle (adapted from Rosenblatt, 2007).

Figure 2.4 gives a visual representation of the increase and/or decrease of the four hormones in each phase. The days of the cycle are not indicated since it varies between women; however the follicular phase usually lasts about thirteen to fourteen days, the ovulatory phase about sixteen to thirty-two hours and the luteal phase about fourteen days (Rosenblatt, 2007). The menstrual cycle can be influenced by external stimuli like circadian rhythms (Baker *et al.,* 2007); high altitude (Escudero *et al.,* 1996), exogenous compounds for example oral contraceptives (Baker *et al.,* 2007) and special diets.

The follicular phase is the first part of the ovarian cycle and is characterised by the maturation of ovarian follicles with estrogen being the dominant steroid hormone. During the last couple of days of the previous cycle there is an increase in the secretion of gonadotropin by the anterior pituitary gland (Silverthorn, 2010). Gonadotropins are hormones that stimulate gonadal function for example FSH and LH (Lawrence, 2005). The increase in FSH secretion leads to the maturation of several follicles in the ovaries, each containing an egg. The granulosa and thecal cells of the follicles start producing estrogen as the follicles grow. These cells are under the control of FSH and LH respectively (Silverthorn, 2010). As estrogen levels increase there is a decrease in the levels of FSH and LH. Thus additional follicles are prevented from being developed given that estrogen exerts a negative feedback control on the secretion of FSH and

LH. However estrogen is still being produced via the granulosa cells as these cells are stimulated by the estrogen present in the circulation (Silverthorn, 2010).

In the early follicular phase menstruation ends and the endometrium proliferates by means of cells, glands and blood vessels. This proliferation is under the influence of the estrogen produced by the developing follicles. Estrogen levels reach their peak as the follicular phase comes to an end. Furthermore only one follicle is still being developed as the other follicles have undergone cell death. The remaining follicle secretes inhibin, progesterone and estrogen but the increased level of estrogen leads to a surge in the secretion of the gonadotropin-releasing hormones (GnRH) before ovulation in addition to preparing the uterus for possible pregnancy. There is a dramatic increase in the LH levels which is crucial for ovulation seeing that maturation of the oocyte would otherwise not be possible (Silverthorn, 2010).

Within twenty-four hours after the increase in LH ovulation takes place whereby the ovum is released into the fallopian tube for either fertilization or death. Thus LH promotes follicular rupture and leads to the transformation of the follicular thecal and granulosa cells to luteal cells of the corpus luteum and stimulates the luteal body to secrete progesterone. For the period of ovulation there is a decrease in the synthesis of estrogen (Silverthorn, 2010).

During the luteal phase the corpus luteum gradually produces increasing amounts of estrogen and in particular progesterone. Estrogen and progesterone exert a negative feedback on the secretion of GnRH. This decrease in FSH and LH secretion is further aggravated by the production of luteal inhibin. Hence the dominant hormone in the luteal phase is progesterone and it is responsible for preparing the endometrium for possible implantation of the fertilised ovum. As the concentration of the progesterone increases, LH production is inhibited. This leads to the apoptosis of the luteal body as it depends on LH. The luteal body degenerates and estrogen and progesterone levels decrease. The decrease in progesterone and estrogen leads to an increase in the secretion of FSH and LH. A new menstrual cycle begins seeing that the endometrium is dependent on progesterone levels for its maintenance. The blood vessels in the surface layer of the endometrium contract, oxygen and nutrient levels decrease and cell death occurs (Silverthorn, 2010).

2.4.1.2 Perturbations of organic acids during the menstrual cycle

No specific study on the urinary organic acid profile as a function of the menstrual cycle has yet been reported. A few important studies have been reported, however, on related aspects which are of significance to my investigation as they included also aspects of the organic acid profile in relation to their investigations.

More recently, Kochhar, *et al.,* (2006) focused on gender-specific differences in the metabolism of humans, using NMR based metabonomics. In that investigation the role of estrogen with regard to the metabolic profile was also included, linked to a specific phase in the menstrual cycle. The main aim of the study was, however, to try and understand the effect nutritional or environmental changes can have on the metabolism of a healthy human control population.

The study group comprised 66 men and 84 women. The participants completed confidential questionnaires about their health (sport or exercise activities), lifestyle (alcohol, dietary regimes and coffee consumption), age, body mass index (BMI) and gender. Participants were excluded from the study if they were acutely ill or pregnant. They collected blood samples from fasted individuals and second morning urine samples. Blood and urine samples were only collected from the women if they fell within the 10th to 15th day of the menstrual cycle, when estrogen is at its highest excretion peak (Figure 2.4). The samples were analysed via NMR analysis and the recorded profiles were investigated by means of multivariate statistical methods such as principal component analysis (PCA) and partial least squares-discriminant analysis (PLS-DA). See section 3.7.6.1 and 3.7.6.2 for a discussion of these methods.

The study detected lactate, glucose, lipids and amino acids as the major compounds in the samples and noticed that urine was a more complex biological fluid than plasma because of its varying metabolite composition and concentration. There was a distinct difference in lipid composition between men and women, given that the concentrations of very low-density lipoproteins (VLDLs) in plasma were greater in men, whereas the concentrations of low-density lipoproteins (LDLs) and high-density lipoproteins (HDLs) were greater in women.

When the data was analysed according to age i.e. participants younger than 30 years and participants older than 46 years the amounts of valine, alanine, tyrosine and isoleucine in plasma were statistically greater in older women than in younger women, whilst there was no significant difference between young and old men. The BMI was also taken into account as a possible cause of differentiation between genders. Tyrosine, glycoprotein and isoleucine concentrations were elevated in the plasma of participants with high BMIs and citrate and choline concentrations were higher in participants with low BMIs. The total amount of lipoproteins was constant for men at all BMIs, whereas it tended to be less in women with a higher BMI.

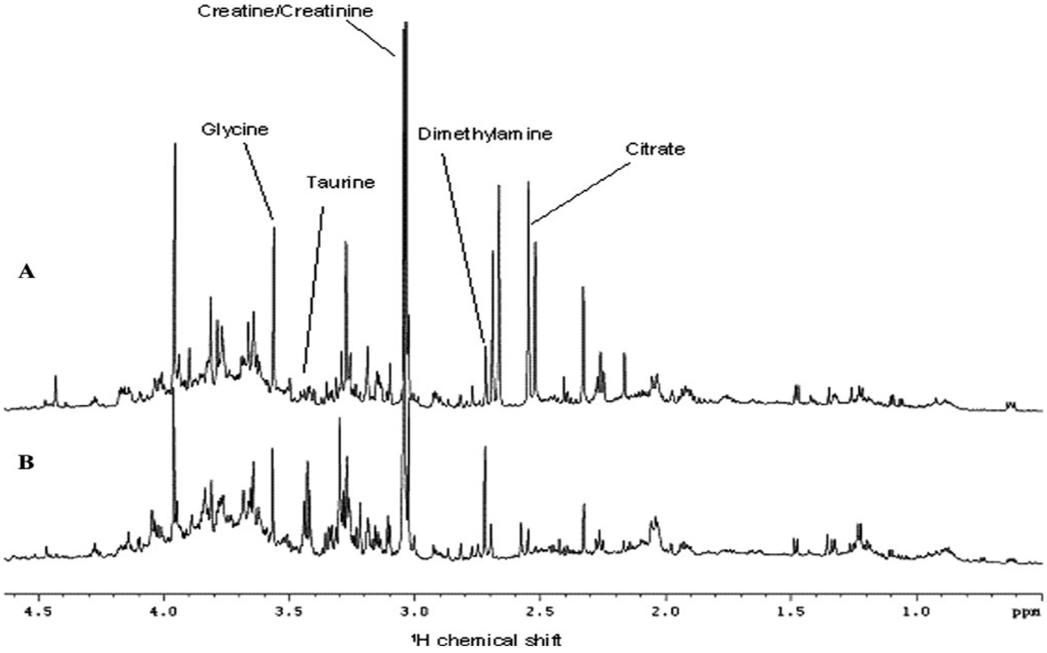


Figure 2.5: H nuclear magnetic resonance (NMR) 600-MHz spectra of second morning urine samples collected from two healthy human participants: (A) female and (B) male with differing ages and BMIs (Kochhar *et al.*, 2006).

The NMR data of the urine samples showed a good separation based on age, BMI and gender. Gender differences were caused by the following metabolites i.e. creatinine, taurine and citrate. The citrate levels were higher in the urine from women and were unrelated to BMI, whereas the urine of men had higher levels of taurine and creatinine and their citrate levels were related to BMI. Four metabolites were responsible for the age differences namely citrate, creatinine, dimethylamine and glycine. Dimethylamine decreased with the increasing age of the men in the experimental group, whilst citrate increased. For the women dimethylamine and citrate remained constant with increasing age, even though glycine showed an increase. Figure 2.5 depicts the NMR spectra of urine samples for a man and a woman compared to each other at different ages and BMIs.

Kochhar *et al*., (2006) concluded that there was a definite difference in the NMR spectra of men and women caused by varying metabolite concentrations in urine and plasma, which was influenced by age and BMI. Finally they believe that estrogen plays a very important role in the communication and regulation between and within lipid and protein biosynthesis, which leads to the subsequent difference in gender profiles.

An additional conclusion from this review is the importance of gender specificity in metabolomics study. USBioTek International, Inc. for example, reported about the necessity of gender specific reference ranges for organic acid results seeing that organic acids are reported relative to creatinine and the rate of creatinine excretion is 25% higher for males than for females. This was particularly seen when they compared the reference range concentration of alpha-ketoglutaric acid for firstly males and females and secondly an only male group. It showed a dramatic difference in reference values, specifically 4 - 18 µg/mg creatinine when the male and female results were combined and 4 - 8.5 µg/mg for the male results alone.

2.4.2 Pregnancy

<u>2.4.2.1 General overview</u>

The second perturbation that we studied, was pregnancy. Metabolic processes undergo major alterations during pregnancy and these processes in the pregnant woman are mediated by the endocrine system (Blackburn and Loper, 1992). At first it was believed that the maternal metabolism adapted with the introduction of the foetus i.e. seen as a 'parasite' on normal metabolism. This concept is, however, no longer held, since numerous metabolic adjustments occur in the early stages of pregnancy when the foetus is still small. As the pregnancy progresses this two-way interaction between mother and foetus will result in more complex metabolic adjustments in the carbohydrate, lipid and amino acid metabolism (Hadden *et al.,* 2008).

The maternal metabolism adapts by continuously adjusting various metabolic pathways like the protein, lipid, fatty acid and carbohydrate metabolism. It is resposible for (1) the adequate growth and development of the foetus, (2) the provision of enough energy stores for the foetus needed after delivery, (3) the management of the increased physiological demands of the pregnant state on the mother and (4) to provide the mother with sufficient enerygy stores for pregnancy, labour and lactation (Blackburn and Loper, 1992). There are also significant physiological and anatomical changes in the pregnant woman as can be seen in the endocrine, cardiovascular, renal, immune and respiratory systems (Oats and Abraham, 2005). These adjustments vary between women since they depend on prepregnancy nutrition, lifestyle of the mother, genetics and foetal size (King, 2000). The following paragraphs will discuss physiological and metabolic adaptations specific to maternal metabolism.

When the ovum is fertilised and implanted on the endometrium, the embryo produces human chorionic gonadotropin (hCG) so as to maintain the corpus luteum and progesterone secretion. As a result estrogen and progesterone are secreted by the corpus luteum until it degenerates and the placenta is responsible for the production of the hormones. The placenta also secretes human placental lactogen (hPL), a pregnancy specific hormone that mobilises free fatty acids from maternal body stores which leads to a reduction in the utilisation of maternal glucose (Oats and Abraham, 2005).

Carbohydrate metabolism has been the key focal point of physiological and pathophysiological research of the maternal-foetal system as glucose is the most important substrate and source of energy for the foetus. Maternal glucose levels are higher when compared to the levels in the foetus, but are lower when compared to non-pregnant women (Hadden *et al.,* 2008; Blackburn and Loper, 1992). These levels are regulated by insulin production which depends on the balance between insulin secretion and insulin clearance and the effect of insulin on maternal muscle, fat and liver. As the insulin action increases plasma glucose levels decrease by means of the inhibition of hepatic glucose release. The increased insulin action reduces the levels of free fatty acids and plasma amino acids in circulation, whilst reduced insulin action increases ketone production, lipolysis and fatty acid oxidation. Maternal adaptations in the carbohydrate metabolism are made possible by an increase in insulin production and a decrease in its sensitivity to the normal hormonal control system (Hadden *et al.,* 2008).

The placenta secretes hormones that influence the nutrient metabolism and depending on the nutrient, certain adjustments can occur e.g. the accumulation of nutrients in new tissue, an increased rate of metabolism or redistribution among tissues. These adjustments are complex, change continuously throughout pregnancy and are determined by foetal demands, maternal nutrient supply and hormonal changes. The demand for sufficient nutrients can double during pregnancy and in order to conserve energy for foetal development the following can occur: the intensity of physical activity can be altered, the rate of lipid synthesis can be reduced and additional food can be consumed. In order to support these adjustments ingested nutrients may be altered through increased intestinal absorption or by minimising the excretion via the gastrointestinal tract or kidney. The nutritional status before pregnancy as well as maternal living conditions influence fetal growth and energy metabolism (King, 2000).

During pregnancy there is a decrease in the concentrations of serum amino acids and proteins even though it is needed by both the mother and foetus. This decrease is related to an increase in placental uptake, hepatic diversion of amino acids for gluconeogenesis, increase in insulin levels and a transfer of amino acids to the foetus. The foetus utilises the amino acids, especially alanine for glucose formation. Protein metabolism adjusts via a biphasic pattern where there is an increase in maternal protein storage during the first half of gestation and a decrease during the second half of gestation, given that there is a decrease in urinary nitrogen excretion (Blackburn and Loper, 1992).

During pregnancy there is an increase in circulating lipids namely phospholipids, cholesterol and especially triglycerides. This increase is accompanied by morphological and functional changes in the adipocytes. Hypertrophy of these cells leads to increased fat storage during the first two trimesters. The number of insulin receptors on these cells increase, which leads to an increased responsiveness to insulin. In the first trimester there is an increase in maternal fat deposition which is used as the energy source for the mother as glucose is used by the foetus. In the third trimester there is a decrease in glucose transport and oxidation as well as lipogenesis within the adipocytes (Blackburn and Loper, 1992).

For the purposes of this study we will be focusing on an untargeted analysis of a subsection of the metabolome (organic acids) as a function of the early, middle and end phases of pregnancy, as compared with controls.

2.4.2.2 Organic acid perturbations during pregnancy

No systematic metabolomics study on the time-dependent changes in metabolite profiles during pregnancy has yet been reported. Baggot *et al*., (2008) did, however, examine organic acids from the amniotic fluid of mothers carrying 41 normal and 22 Down syndrome foetuses obtained via amniocenteses. The primary goal of the study was to determine if there were possible foetal biochemical differences in the metabolite profiles of foetuses diagnosed with Down syndrome, when compared to normal foetuses.

I compiled a table, based on the results of Baggot *et al.,* to illustrate the salient aspects of their investigation (Table 2.2). The data is in a non-parametric form obtained via the Mann-Whitney rank sum tests. The statistical analysis showed certain metabolites that were statistically non-significant i.e. a p-value higher than 0.05, for example glyceric acid, 3-hydroxyisovaleric acid, fumaric acid and suberic acid. Metabolites with a p-value lower than 0.05, were significant and included the following; methylsuccinic acid, 5-hydroxycaproic acid, adipic acid, alpha-ketoglutaric acid and phenylpyruvic acid. These metabolites were significant in view of the fact that they were highly elevated in the amniotic fluid of the Down syndrome foetuses and not in the normal foetuses. Most of these metabolites are collectively associated with riboflavin deficiency, except for phenylpyruvic acid which relates to the metabolism of phenylalanine and neurotransmitters.

Table 2.2: Organic acid metabolites in Down syndrome and normal amniotic fluid.

Metabolite	**Down median**	**Down 5-95% CI**	**Normal median**	**Normal 5-95% CI**	**P**
No significance (1)					
Glyceric acid	26.500	0.000-56.500	25.000	0.000-60.000	0.943
3-Hydroxyisovaleric acid	1.325	0.000-8.650	2.350	0.000-10.400	0.840
No significance (2)					
Fumaric acid	0.050	0.000-0.450	0.000	0.000-0.550	0.186
Suberic acid	0.125	0.000-0.750	0.050	0.000-0.800	0.147
Statistical significance (p < 0.05)					
Phenylpyruvic acid	0.075	0.000-0.550	0.000	0.000-0.250	0.045
α-Hydroxybutyric acid	0.750	0.00-42.000	15.500	0.000-61.000	0.028
5-Hydroxycaproic acid	0.100	0.000-1.600	0.000	0.000-0.550	0.010
α-Ketoglutaric acid	6.250	0.000-45.000	0.000	0.000-23.000	0.019
Adipic acid	0.050	0.000-0.550	0.000	0.000-0.300	0.012
Methylsuccinic acid	0.715	0.000-3.100	0.000	0.000-0.350	0.004

Baggot *et al*., (2006) concluded that even though there was a metabolic difference between normal and Down syndrome foetuses, this type of analysis should not replace the current analysis of chromosomal diagnosis since a foetus with riboflavin deficiency could be misdiagnosed as having Down syndrome. Finally this study may be more useful in the understanding of the physiology and biochemistry of Down syndrome than as a method of diagnosis.

A study conducted by Mock *et al*. (2002) tried to determine if (1) an increased excretion of 3-hydroxyisovaleric acid (3-HIA) served as an indicator of biotin deficiency in pregnant women and if (2) biotin supplementation led to a decrease in 3-HIA excretion. 3-HIA increases as the activity of methylcrotonyl-CoA carboxylase decreases. This enzyme is biotin-dependent and catalyses a vital step in leucine degradation. The study obtained samples from 26 pregnant women as well as 5 non-pregnant women who served as a control group. Ten of the pregnant women were in their early pregnancy period and 16 in their late pregnancy period. Pregnant women were included in the trial if they had highly elevated 3-HIA levels, under the care of a physician, drinking the recommended vitamin intake which did not contain biotin.

This study was a randomised, placebo-controlled trial. Five of the women in the early pregnancy group and five in the late pregnancy group took the placebo, whereas five women in the early pregnancy group and eight women in the late pregnancy group took the biotin supplementation. All of the non-pregnant women received the capsule containing biotin. Urine was collected before supplementation, after which they ingested the biotin or placebo capsule for 14 days. A urine sample was then collected after 14 days and the samples were analysed.

The results that Mock *et al*., (2002) obtained showed that there was an overall decrease in 3-HIA excretion for the women taking the biotin supplement, whereas the women in the early and late pregnancy group, taking the placebo, showed an increase in 3-HIA excretion. Consequently there was a difference between the biotin supplement and placebo groups, which was not caused by a varying biotin status prior to treatment since the 3-HIA excretion levels of the women in the study were equal. They concluded that a reduced biotin status is related to an increase in the level of 3-HIA excretion and that marginal biotin deficiency occurs frequently in the first trimester of pregnancy.

A master’s study completed in 1982 (Christie) attempted to profile urinary metabolites, specifically organic acids and steroids in human pregnancy. The objectives of the study were as follows: the standardisation of suitable steroid and organic acid profiling methods in biological fluids, and utilisation of these methods with regard to a 24-hour urine sample from non-pregnant and pregnant women, the identification of metabolites responsible for any alteration in the course of a normal pregnancy and finally to compare these metabolites identified in normal pregnancy with metabolites in high-risk pregnancies. She obtained 25 urine samples from pregnant women at two time intervals during pregnancy i.e. weeks 12 to 15 and weeks 24 to 27. Her control group included samples from 15 non-pregnant women. The organic acid profiles contained 50 identifiable peaks with 10 of these peaks being unequivocally identified, namely threonic acid, erythronic acid, hippuric acid, citric acid, lactic acid, glycolic acid, glucuronic acid, sulfate, phosphate and uric acid. The study focused on only three (glycolic acid, erythronic acid, and lactic acid) of the ten metabolites since the excretion of these metabolites increased with the progression of pregnancy.

2.4.3 Age

2.4.3.1 General overview

Aging is an extensively studied phenomenon, and was the third aspect included in the present metabolomics study. According to Ashok and Ali (1999) aging is the accumulation of changes responsible for the sequential alterations that accompany advancing age and its associated progressive increases in the chance of disease and death. Allen and Balin (2003) stated that aging is a progressive, time-dependent deterioration of the capacity of an organism to respond adaptively to environmental change. This process results in an increased and irreversible vulnerability to certain diseases and death. It affects all members of a species or population group and the aging process contributes very little to the changes that occur early in life but this contribution increases with age since the process is exponential by nature (Allen & Balin., 2003).

There are many theories that have been put forward to account for the possible causes of aging. Many of these theories originated from studies that investigated changes that accumulated over time. Unfortunately there has of yet not been a single theory that is generally acceptable and the viewpoint has been expressed that it is doubtful that the mechanisms involved in the aging process will be explained by a single theory in the near future (Ashok and Ali., 1999).

Rubner presented one of the earliest theories in 1908 to explain aging via a single underlying mechanism, called “Rubner’s metabolic potential”. Rubner observed that the total energy expenditure in five different domesticated animal species for their lifetime, per unit weight was similar even though their lifespans differed. He hypothesised that there was a limited number of molecular rearrangements in living material. Hence the total amount of metabolic work any organism can perform during its lifetime is governed by a fixed constant. This theory has been challenged by many other studies as well as by Rubner's own work since he demonstrated that human beings exert a higher metabolic potential than the other animals he examined (cited by Allen and Balin., 2003).

There is strong evidence that metabolic processes influence longevity, particularly in cold-blooded animals. Unfortunately there is no universal constant that can be used to define this relationship. Pearl summarised this relationship in 1928 with his "Rate of Living Theory" which stated that all the species exhibit a metabolic potential determined by genetics and that the lifespan of a species as a result is dependent on the rate of metabolism. Many studies have supported this theory with regard to cold-blooded animals as they exhibit a species specific maximum energy expenditure. It is also possible to alter their lifespan experimentally when the metabolic rate is changed. It is difficult to establish the validity of this theory where mammals are concerned as they exhibit hypertrophy when there is an increase in activity or atrophy when there is a sufficient decrease in activity (cited by Allen and Balin, 2003).

One of the most established theories regarding the aging process is the free radical theory of aging. This theory was proposed by Harman in 1956 who postulated that the damage to cellular macromolecules, in aerobic organisms is caused by the production of free radicals which in turn can influence the life span of an organism (as referred to by Kregel and Zhang, 2007). Free radicals are molecules that contain one or more unpaired electrons which leads to an increase in reactivity (Lawrence, 2005). Organisms are exposed to free radical-containing reactive oxygen species (ROS) in the environment or produce them intracelullarly from different sources, in particular via the mitochondria. ROS include metabolites with a variety of diverse chemical species e.g. unstable oxygen radicals (superoxide), hydroxyl radicals and nonradical molecules (hydrogen peroxide) (Finkel and Holbrook, 2000; Kregel and Zhang, 2007).

ROS are responsible for some of the changes observed in cells during differentiation, transformation and aging as they influence biochemical and molecular processes. It has been demonstrated that free radicals damage cells intracellularly which limits the cells' ability to adapt to environmental change and thereby makes them more vulnerable to death. Free radicals modify proteins, damage DNA (deoxyribonucleotide acid), inactivate enzymes and initiate chain reactions responsible for peroxidizing lipids (Allen and Balin., 2003). When cells are placed under stress because of DNA or protein damage caused by the production of large amounts of free radicals this is known as oxidative stress (Lawrence, 2005). When DNA is damaged an organism's lifespan can be decreased as the agents or processes that cause DNA damage diminish DNA repair systems. Alternatively, the rate of damage can be decreased by inhibiting the agents or processes responsible for damage which in turn increases the fidelity of the DNA repair systems (Ashok and Ali., 1999).

The aging process, therefore, cannot be explained by a single theory since it is a multifaceted process which is not entirely understood. The lifespan of an organism is shaped by both genetic and environmental factors since the aging phenotype is heterogenous among individuals of the same species (Weinert and Timiras, 2003). When the aging process is finally understood ROS will definitely be a key factor in the determination of longevity.

2.4.3.2 Organic acid perturbations during aging

In contrast to the normal perturbations associated with the menstrual cycle or pregnancy, the influence of age on the metabolite profile is very well studied. It is anticipated that the age of the group or organism will influence the experimental and/or data analysis methods employed in a study. It is a particularly important aspect of any experimental design when perturbations or pathological conditions are evaluated and compared to reference ranges obtained from healthy subjects. This is clearly demonstrated in a study performed by Guneral and Bachmann (1994) where they found that the concentrations of most of the urinary organic acids change with age and that there was a considerable difference between the age groups under investigation. This study will be discussed in detail to give a clear example of the influence of age on metabolite profiles, being the urinary organic acids in this case.

Guneral and Bachmann (1994) obtained 161 urine samples from a healthy Turkish paediatric population ranging in age from 2 days to 16 years. These samples were grouped according to 5 defined age groups, namely: newborns between 2 and 28 days (n = 57), infants between 1 and 6 months (n = 8), children between 2 and 6 years (n = 66), children between 6 and 10 years (n = 14) and children older than 10 years (n = 16). The participants were clinically in good health, were on a free and unrestricted diet and took no medication. In their analysis these investigators quantified 69 organic acids. Of these organic acids, 32 were found in more than 95% of the samples whilst 37 were encountered in only some of the samples. The median values as well as the 2.5 and 97.5 percentiles of each organic acid for each of the 5 age groups were reported.

When the values of the individual organic acids were compared to each other for the 5 age groups, different trends were observed in relation to age. Guneral and Bachmann (1994) defined four different trends in this study, namely metabolites that decreased with age, metabolites increasing with age, metabolites with an irregular pattern and metabolites with an unchanged pattern. The metabolites that decreased with age were classified into three additional groups i.e. metabolites that decreased with a regular pattern, decreasing pattern with non-significant fluctuations as well as a decreasing pattern with an increase at ages 1 to 6 months.

Guneral and Bachmann thus included the following metabolites in their list of metabolites which adhered to the particular trends:

- Metabolites decreasing with age:
    - Metabolites regularly decreasing with age:
        - Citric acid, **2-hydroxyglutaric acid**[3], quinolinic, citramalic acid, acetylaspartic acid and 4-hydroxy-3-methoxybenzoic acid.
    - Metabolites decreasing with age with a pattern of non-significant fluctuations:
        - Lactic acid, succinic acid, pyruvic acid, glyceric acid, glutaric acid, fumaric acid, malic acid, 2- and 3-hydroxyadipic acid, adipic acid, suberic acid, 2-oxoglutaric acid, p-hydroxyphenyllactic acid, p-hydroxyphenylacetic acid, p-hydroxypyruvic acid, cis-aconitic acid, vanillylmandelic acid, methylcitric acid, ethylhydracrylic acid, 5-hydroxycaproic, ethylmalonic acid, indoleacetic acid, hexadecanoic acid, 2- and 3-deoxytetronic acid, 3-hydroxysebacinate, erythronic acid and threonic acid.

[3] Some figures based on the data from Guneral and Bachmann are included in this review. The organic acids chosen for the development of these figures are shown in bold letters.

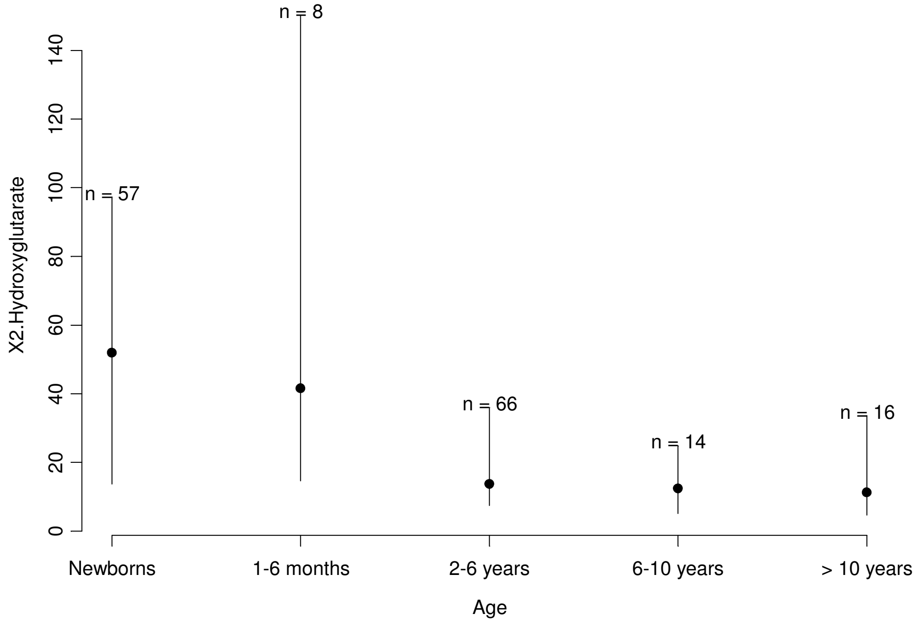


Figure 2.6: Example of a metabolite (2-hydroxyglutaric acid) decreasing with age.

- Metabolites decreasing with age but with a significant increase at ages 1-6 months:
  - 3-hydroxyisobutyric acid, **3-hydroxyisovaleric acid**, isocitric acid, homovanillic acid, methylglutaric acid, pyroglutamic acid, methylsuccinic acid and 3-hydroxy-3-methylglutaric acid.

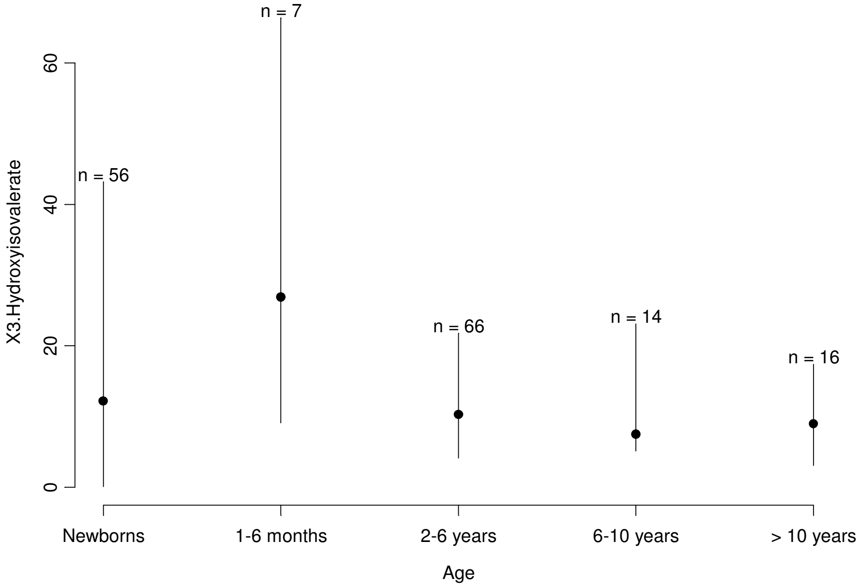


Figure 2.7: Example of a metabolite (3-hydroxyisovaleric acid) decreasing in age but with a significant increase at ages 1 to 6 months.

- Metabolites increasing with age:
  - **2-hydroxyisobutyric acid**, 3-methylglutaric acid, 3-hydroxy-2-methylbutyric acid, glycolic acid and tiglylglycine.

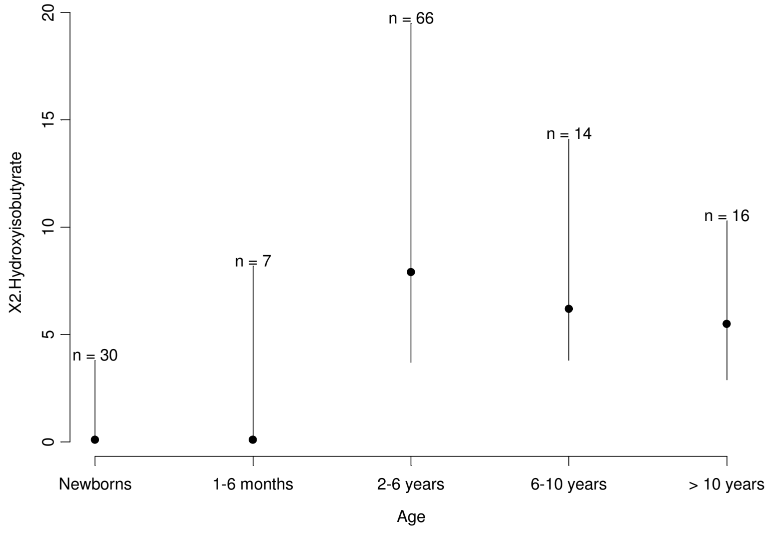


Figure 2.8: Example of a metabolite (2-hydroxyisobutyric acid) increasing with age.

- Metabolites with an irregular pattern:
    - 3-hydroxypropionic acid, methylmalonic acid, 3-hydroxybutyric acid, **pimelic acid**, erythro-4-deoxytetronic acid and threo-4-deoxytetronic acid.

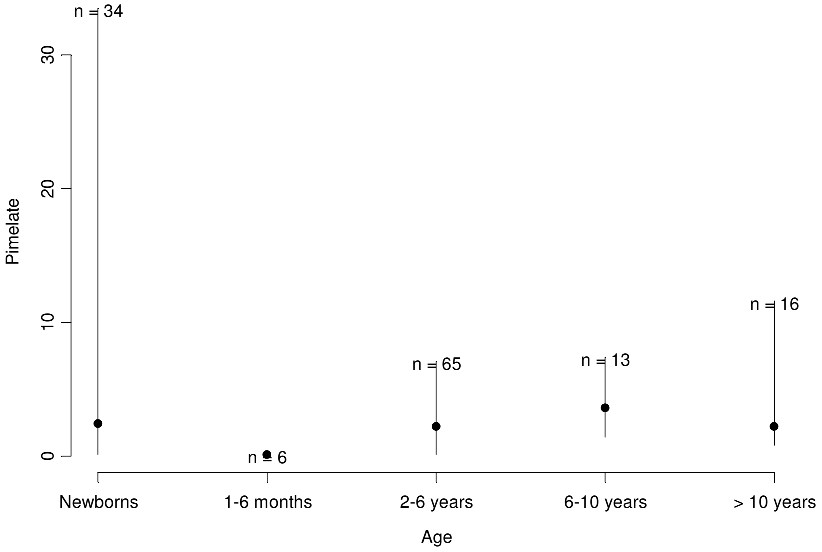


Figure 2.9: Example of a metabolite (pimelic acid) with an irregular pattern.

- Metabolites with an unchanged pattern
    - **Azelaic acid** and 4-hydroxybenzoic acid.

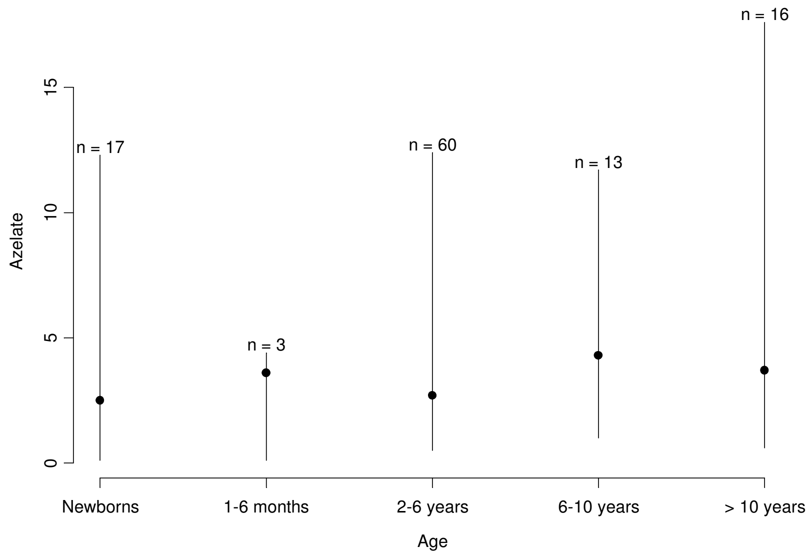


Figure 2.10: Example of a metabolite (azelaic acid) with an unchanged pattern.

Figures 2.6 to 2.10 depict examples of the metabolites identified by the study that fit the main trends. The median values of the metabolites are indicated as black dots and the line through the median indicates the 2.5 and 97.5 percentiles of the metabolites respectively. The sample size for each encountered metabolite is indicated as an “n” above the black lines.

Guneral and Bachmann concluded that it is necessary to use age-related reference ranges for the diagnosis of metabolic diseases, since there was a significant difference in metabolite concentrations between the different age groups (as can be seen in Figures 2.6 to 2.10). This study was not repeated and subsequently the results obtained from it cannot be applied to other pediatric groups as normal reference ranges for the quantified metabolites. It is furthermore not a metabolomics study but it indicates the influence that variation in age can have on the metabolism of an individual.

There are several publications on age specific reference ranges for the different organic acids within and between different population groups. An example of one such publication can be seen in Table 2.3. Table 2.3 contains examples of a few reference values of organic acids in 5 different age groups as well as possible disorders in which they are elevated. These reference values are presented in the “*Physician’s guide to the laboratory diagnosis of metabolic diseases*”, but as stated by Hoffman & Feyh (2003), they should be used as a guideline since reference values may vary between laboratories. Thus in conclusion when conducting any metabolomics investigation it is important to take into account the age of the control and experimental group.

Table 2.3: Published reference values for different age groups (Hoffmann and Feyh, 2003).

Compound		Premature infants ≤ 36 weeks (mmol/mol creatinine)	Term newborns > 36 weeks (mmol/mol creatinine)	Children ≤ 5 years (mmol/mol creatinine)	Children > 5 years (mmol/mol creatinine)	Adults (mmol/mol creatinine)	Disorder to be considered
Lactic acid	Mean	49	51	86	76	25	Mitochondrial and biotin disorders; dihydrolipoyl (E_3) dehydrogenase deficiency; circulatory failure
	Min	1	0.5	33	35	13	
	Max	927	156	285	131	46	
Benzoic acid	Mean	2.5	0.2	2.2	1.9	4.2	Benzoate treatment; can also originate from gut bacteria
	Min	n.d.	n.d.	0.6	n.d.	1.9	
	Max	31	7	7.7	4.3	6.5	
4-Hydroxyphenylacetic acid	Mean	17.9	33.3	37	19.4	11	All forms of tyrosinemias; can also originate from gut bacteria
	Min	3	3	12.3	7.4	3.5	
	Max	78	240	174	30.1	22	
Citric acid	Mean	401	480	385	386	155	-
	Min	93	117	75	120	70	
	Max	1022	1422	667	582	226	

2.5 The metabolomics workflow

2.5.1 General outline

Metabolomics aims to improve the current status of information related to the metabolome by analysing all the metabolites present in an organism, cell or biofluid by means of a single analysis. Unfortunately this is presently technically not possible given that metabolites are diverse and complex in their chemical and physical properties which hampers any complete analysis from a technological point of view, as discussed above. Nevertheless, any metabolomics experiment will generally follow a similar experimental design i.e. the analytic metabolomics workflow, as depicted in Figure 2.11.

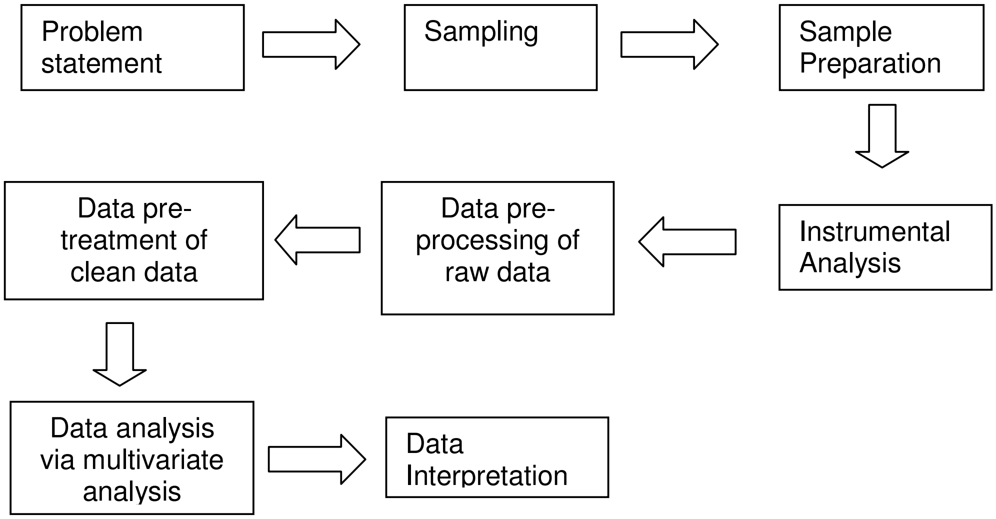

Problem statement
Sampling
Sample Preparation
Instrumental Analysis
Data pre-processing of raw data
Data pre-treatment of clean data
Data analysis via multivariate analysis
Data Interpretation

Figure 2.11: Workflow of a metabolomics experiment.

2.5.2 Problem statement

Any research investigation begins with an initial problem statement that determines the type of samples needed, collection protocol, sample preparation, analytical methods and data analysing techniques required. This is especially relevant with regard to a metabolomics analysis seeing that the metabolome comprises hundreds to thousands of metabolites each with their own unique characteristic, which in turn influences the type of samples needed, collection, preparation and analytical approach. Finally metabolomics as with all research aims to translate the obtained data into biological relevant information which either confirms or negates the research question (van der Werf *et al.*, 2007).

2.5.3 Sampling and sample preparation

A typical metabolomics study can fall into two broad groups namely experiments conducted in a controlled or uncontrolled environment. Studies conducted in a controlled environment include animals housed within the laboratory or microbial experiments, whereas an uncontrolled environment usually refers to human studies. When animal studies are carried out it is possible to determine environmental variables that may influence the metabolome of the animal i.e. the type of housing, light cycle, food intake and water consumption. In animal studies it is also possible to determine the frequency, duration, time and method of sampling as well as the storage conditions prior to analysis (Griffin *et al*., 2007).

The disadvantage of human studies is that it is not always possible to determine the environmental factors that may influence the metabolome of the study group. It is also necessary to obtain a detailed medical history of the participant as this can influence sample collection as well as the interpretation of the data; for example in the case of a participant diagnosed as HIV+ or any other infectious diease. This leads to an important observation i.e. the selection of participants, which is determined by certain criteria for inclusion and exclusion, that is established and influenced by the problem statement (Griffin *et al*., 2007). Hence sampling and sample preparation is influenced by the type of organism under investigation and the problem statement.

When samples are collected they provide a ‘snapshot’ of the organisms’ metabolome at the time of sampling. For example if samples are collected before and after a specific perturbation is induced it will give an indication of how the metabolome may react to this specific perturbation. Sampling and sample storage are influenced by many factors, like the time of sampling, the method of sample collection, storing conditions and continued freeze/thawing (Dettmer *et al*., 2007). The composition of the metabolome is influenced by variations like age, gender, dietary and health status of the organism or study population under investigation. These factors can influence the accuracy and reproducibility of the results obtained (Bollard *et al*., 2005). According to Roessner *et al.,* (2000) the biological variability observed is generally greater than the analytical variability, even when controlled sampling and sample preperation are employed.

The type of sample required for any metabolomics investigation influences and/or is influenced by the metabolomics strategy or type of experiment being employed. Since metabolic processes are so rapid any possible enzymatic reaction or oxidation process that leads to metabolite formation or degradation has to be inhibited by means of freeze clamping, freezing in liquid nitrogen and acid treatment (Dettmer *et al.,* 2007). In the case of urine, the samples are non-invasive and easily obtainable and contain extra-cellular metabolites that give a representation of the metabolic activity over a period of change e.g. of amino acids, purines, pyrimindines, amines and organic acids. These compounds vary in concentration and when measured can lead to a better understanding of the gene and metabolome function. Thus urine provides substantial information on biological systems as the levels of metabolites reflect the response of the system to endogenous or exogenous compounds or to influences which are under investigation (Kuhara., 2005).

Sample preparation is an important component of the metabolomics approach as it is responsible for extracting analytes from complex biological matrices in order to bring them into a compatible format for the analytical method being used. Samples are usually prepared before instrumental analysis via polar or non-polar extraction methods that disrupt the physical and chemical properties of the cells, remove the cell pellet by means of centrifugation and distribute the metabolites to polar and non-polar solvents, such as in the case of liquid-liquid extraction, solid phase extraction, protein precipitation etc. The method used for sample preparation depends on the type of sample, for example urine or blood samples required for investigation. The selection of the method is influenced by the problem statement and in order for it to be a true metabolomics analysis it should be as universal and simple as possible, since specific metabolites are not being targeted. In the case of a targeted analysis, however, the method can be tailored to the metabolites under investigation as the analytes are known (Dettmer *et al.,* 2007).

2.5.4 Data generation

Data generation is an important step in any metabolomics study as it lays the foundation for data interpretation. It is greatly dependent on the most suitable separation method (LC, GC or CE) chosen, which is influenced by certain criteria, for example the detection limit, precision, coverage, dynamic range, selectivity and accuracy provided by a specific method (Scalbert *et al*., 2009). MS is a vital part of any metabolomics study as it provides a fast, sensitive, quantitative and/or qualitative analysis for measuring the molecular weights and quantities of different metabolites (Shulaev., 2006).

Mass spectrometry alone cannot distinguish between two or more substances of the same molecular weight, therefore substances are separated by means of different types of chromatography columns, before analysis on the mass spectrometer. The different types of columns responsible for separation, coupled to mass spectrometers include gas chromatography (GC), liquid chromatography (LC) and capillary electrophoresis (CE) (Tomita, 2005). The type of mass spectrometry system utilised will depend on the nature of the sample being analysed.

Each technique has its associated advantages and disadvantages (Table 2.4) and a single technique cannot be ideally used to examine all the metabolites within a cell since they have different polarities and molecular weights. Subsequently a combination of different analytical techniques has to be used in order to gain a comprehensive view of the metabolome.

Table 2.4: Advantages and disadvantanges of some analytical techniques used in metabolomics (adapted from Shulaev, 2006).

Analytical method	Advantage	Disadvantage
GC-MS	• Robust • Sensitive • Large linear range • Large commercial and public libraries	• Slow • Often requires derivatisation • Many analytes too large for analysis • Many analytes thermally unstable
LC-MS	• Usually no derivatisation required • Large sample capacity • Many modes of separation	• Slow • Limited commercial libraries
CE-MS	• Small sample requirements • Usually no derivatisation • High separation power	• Poor reproducibility of retention time • Limited commercial libraries
NMR	• High resolution • Non-destructive • Rapid analysis • No derivatisation needed	• More than one peak per component • Low sensitivity • Libraries of limited use due to complex matrix

GC-MS is regarded at present as the gold standard of metabolomic analyses as it provides high-chromatographic metabolite resolution, metabolite quantification, analyte-specific detection and supports the indentification of metabolites through the use of well-developed data-bases (Harrigan and Goodacre, 2003). A significant advantage of GC-MS is the availibility of many searchable mass spectral libraries as well as a number of public libraries. Most mass spectral libraries do not include a large number of naturally occurring metabolites and their intermediates, since these libraries are predominantly aimed at the chemical industry and drug studies. Subsequently this limits their use in metabolomics studies (Shulaev, 2006).

Most metabolites have to be chemically derivatised before GC-MS analysis as it provides thermal stability and volatility to the sample. During analysis small amounts of the derivatised samples are analysed on GC columns of differing polarity that facilitate high sensitivity and high chromatographic resolution of the compounds. Hence it analyses volatile metabolites which are of a low molecular weight. The result of the GC-MS analysis is given in chromatogram format but it is very complex as it contains hundreds of metabolite peaks as well as multiple derivatisation products. The peaks of the metabolites are defined via deconvolution software for example AMDIS (Dunn and Ellis*.,* 2005).

AMDIS extracts the spectrum of each compound from a sample mixture and compares it to the spectra of a pure compound represented in a reference library. The overall process involves four sequential steps namely noise analysis, component perception, spectrum deconvolution and compound identification. Metabolites are identified when their retention time is matched to the retention time of a pure compound that was previously analysed under the same instrumental conditions. The reference library used for compound identification can either be the one provided with the software or can be a custom library built with the experiment in mind (Kind, T., 2003). AMDIS also exports a feature list of the identified and/or unidentified compounds with corresponding information such as the retention time.

LC-MS is being increasingly used in metabolomics applications due to its high sensitivity, high resolution and analytical flexibility. It can be used for targeted analysis of for example specific metabolites, class of compounds or a broad range of compound classes. Unfortunately LC-MS lacks a transferable mass spectral library which makes untargeted analysis difficult. Nevertheless it can be used to elucidate the structure of unknown compounds. Sample preparation is simplified seeing that samples are analysed at lower temperatures than GC-MS and hence sample volatility i.e. derivatisation is not required (Shulaev, 2006).

2.5.5 Data analysis and interpretation

Metabolomics studies produce very large data sets since it is a complete biochemical characterisation of an organism (Mendes., 2002). There are several steps involved in any metabolomics investigation after the relevant biological sample is obtained and before the results are interpreted, which influence the outcome of the study. These steps are depicted in Figure 2.11 and include data preprocessing, pretreatment and data analysis which represent a bioinformatics workflow.

Data preprocessing involves methods that convert the raw data obtained from the instrumental part of the experimental analysis into clean data that can be used in data analysis, specifically univariate or multivariate techniques. There are many methods that are responsible for preprocessing data and most of them are order-dependent. Examples of these methods include peak-picking, target analysis and alignment (Goodacre *et al.,* 2007).

Data pretreatment methods correct any aspects of the data that may hinder the interpretation of the data sets from a biological standpoint. The factors that hinder data interpretation include differences in concentrations between and within metabolites, technical variation and values below the detection limit of the instrument (Van den Berg *et al.,* 2006). The method used will depend on the data analysis method chosen and the research question. Data can be pre-treated via normalisation, centering, scaling, and transformation, as will be discussed in more detail in the experimental sections of the dissertation. Each pretreatment method emphasises a different aspect of the data and has its own advantage or disadvantage which means that it is crucial to select the proper method as it will affect the metabolites identified as the most important variable responsible for group separation (Katajamaa and Oresic., 2007).

The final step in any metabolomics study before biological interpretation is the analysis of the pretreated data with a type of multivariate analysis. Section 3.7 will describe the different pre-treatment and data analysis methods used in the specific investigation of the selected perturbations. Section 3.7 will discuss the following statistical methods, namely: descriptive and inferential statistics, data smoothing, effect sizes, data transformation and imputation and multivariate methods such as PCA and PLS-DA.

Any pattern or correlation observed in the data should primarily provide information about the physiological state of the metabolic system. Metabolites can then either show a significant variation between the experimental conditions or groups, or they can show a pair-wise relationship with other metabolites. Any observed correlation in metabolite profiles can only be validated if there are repeated measurements of the samples under identical experimental conditions (Steuer, 2006) .

One possible approach to coping with the torrents of data is to create metabolomics specific databases that contain metadata, raw and processed experimental data. Metadata include the experimental design, nature of the samples, sample storage, sample preparation, analytical techniques and data processing details. It will make it possible for other laboratories to reproduce the experimental conditions and compare the results stored on the databases (Shulaev, 2006). Metabolomics databases can be specific or comprehensive, for example single species-based databases, databases listing all known metabolites for each biological species, databases compiling established biochemical facts, databases that integrate genome and metabolome data and databases storing metadata, raw data and processed data of metabolomics experiments (Goodacre *et al*., 2004).

2.6 Conclusions and aims

1. The field of metabolomics is relatively new, but great strides have been made in determining its advantages and disadvantages as a research tool. Critical steps have been defined as being vital in a successful study, with regard to the characterisation of the metabolome in different biological tissues, or extracts or biofluids, including the steps comprising sample preparation, sample storage and analytical analysis.

> **The first aim of this study was therefore to become acquainted with the technology of metabolomics to generate analytical data with sufficient chromatographic richness and resolution for multivariate statistical analysis and for metabolite identification and quantification (Harrigan *et al.,* (2005).**

The “workflow” needed to complete the first aim is shown in Figure 2.11 by means of the first five text boxes.

2. Metabolite identification, data analysis via multivariate techniques and the interpretation of the metabolite profiles is the next aim of a metabolomics study (Moco *et al*., 2007), which is depicted in Figure 2.11 via the last three text boxes.

> **The second aim of this study was to investigate three natural perturbations not formerly subjected to a detailed metabolomics study and of sufficient complexity**

to become acquainted with the biostatistical research methods to generate new biological information on these perturbations.

3. Metabolomics studies are not designed for hypothesis testing but for hypothesis generation (Kell, 2004). Metabolomics is thus directed towards the simultaneous analysis of multiple metabolites, thereby capturing the status of diverse biochemical pathways at a particular moment in time (i.e. a metabolic snapshot), defining all or any specific metabolic perturbation(s).

Although it may not be expected that a study of normal perturbations would generate a hypothesis, this aspect of metabolomics investigations was nevertheless included in this study.

The third aim was thus an attempt to formulate a hypothesis, based on the metabolomics of natural perturbations, and to propose an approach to test such a hypothesis.

Chapter 3 - Materials and Methods

3.1 Introduction

The aim of this chapter is to discuss the important experimental aspects of an untargeted metabolomics study. The experimental subjects for the menstrual cycle, pregnancy and aging studies are described in Section 3.2. Section 3.3 describes the experimental design implemented for each perturbation focusing on sample storage and Sections 3.4 and 3.5.1 explain the materials and method utilised for the determination of the organic acids in urine. Section 3.5.2 discusses the analytical method employed, its specifications and protocol. The data matrix generated via deconvolution, peak identification and quantification as well as the statistical methods employed for the three perturbations is discussed in Sections 3.6 and 3.7.

3.2 Experimental Subjects

The following section will discuss pertinent information about the experimental subjects who participated in the respective investigations of the menstrual cycle, pregnancy and aging. The information will comprise age, gender, and ethnicity and in some cases any known medical condition. The selection of the participants in each study is vital since unwanted sources of variation in the group can negatively influence the statistical results obtained, for example by means of a “false positive result”.

3.2.1 Menstrual Cycle Participants

The participants in this study signed informed consent forms (Appendix A) before any samples were collected. Two participants provided the necessary urine samples for this study, however one of the participants provided an extra month of samples. The time that elapsed between these two cycles was three months (twelve weeks). From results presented in Section 2.3 (concerning the week effect), we also included the extra sample in subsequent statistical analysis and treated this sample as independent of the other two samples. Table 3.1 lists relevant information on the participants such as their age, ethnicity, known medical condition as well as the month the samples were taken and analysed. Participant 1 and 2 each provided urine samples, but participant 1 provided in addition to the urine samples, blood and saliva samples (see Section 3.3.1). Only two participants were included in the study given that it was an exploratory study the purpose of which was to determine whether the excretion of organic acids is influenced by the menstrual cycle. This was especially relevant for the pregnancy study, since urine samples were obtained randomly from the participants in the control group. Hence it was not known at what stage in the menstrual cycle the samples were collected for the participants and if a specific phase would have an influence on the organic acid profile of the control group.

Table 3.1: Information on participants in the menstrual cycle study.

Participant no.	Age	Ethnicity	Months samples taken	Any known Disease	Medication or any other drugs being taken
Participant 1a	25	Caucasian	March – April	No	Vitamin supplement
Participant 1b			July - Augustus		
Participant 2	23	Caucasian	March – April	No	Vitamin supplement

Ideally, a large experimental group would be advisable for a study as envisioned. However, it is also evident from Table 3.1 that the study group is homogenous and an acceptable group for the purpose of this part of the study.

3.2.2 Pregnancy Participants

Informed consent (Appendix B) was obtained from the voluntary participants before any samples were collected. The participants filled in a questionnaire in order to obtain relevant information (Table 3.2). Twenty-eight participants were initially included in the pregnancy study, twelve formed part of the control group and four were in their first trimester of pregnancy, seven in their second trimester and five in the final trimester. The blood and urine samples used to distinguish between the successive phases of pregnancy was obtained from the patients of Dr. Thomas, a local gynaecologist with consulting rooms in Potchefstroom, whereas the samples of the control (non-pregnant) group were mostly obtained from female individuals from North-West University. All the participants were on a free and unrestricted diet.

Table 3.2: Information on participants in the pregnancy study.

Participant no.	Age	Ethnicity	Weeks, days pregnant	Date the blood sample was received	Date the urine sample was received	Any known medical condition
C1[4]	41	Caucasian	None	25 Feb ’09	25 Feb ’09	No
C2	25	Caucasian	None	7 Apr ’09	3 Apr ’09	No
C3	25	Caucasian	None	7 Apr ’09	16 March ’09	No
C4	23	Caucasian	None	7 Apr ’09	8 Apr ’09	No
C5	25	Caucasian	None	7 Apr ’09	19 March ’09	No
C6	24	Caucasian	None	7 Apr ’09	18 March ’09	No
C7	25	Caucasian	None	20 March ’09	20 March ’09	No
C8	28	Caucasian	None	18 March ’09	18 March ’09	No
C9	23	Caucasian	None	20 Apr ’09	20 Apr ’09	Porphyria
C10	26	Caucasian	None	23 March ’09	23 March ’09	No
C11	21	Caucasian	None	7 May ’09	26 March ’09	No
C12	24	Caucasian	None	7 May ’09	24 March ’09	No
P.1.1[5]	27	Caucasian	7,6	30 Oct ’08	30 Oct ’08	No
P.1.2.	29	Caucasian	5,2	7 Nov ’08	7 Nov ’08	No
P.1.3	27	Caucasian	11,3	29 Oct ’08	31 Oct ’08	No
P.1.4	30	Caucasian	12	7 Nov ’08	7 Nov ’08	No
P.2.1[6]	31	Caucasian	13	22 Oct ’08	22 Oct ’08	No
P.2.2	26	Caucasian	16	6 Nov ’08	6 Nov ’08	No
P.2.3	27	Caucasian	15	29 Oct ’08	27 Nov ’08	No
P.2.4	22	Caucasian	12,6	31 Oct ’08	31 Oct ’08	No
P.2.5	36	Caucasian	20,5	19 Nov ’08	19 Nov ’08	No
P.2.6	30	Caucasian	22	-	17 Feb ’09	No
P.2.7	24	Caucasian	28	8 Oct ’08	10 Oct ’08	No
P.3.1[7]	27	Caucasian	29	29 Oct ’08	2 March ’09	No
P.3.2	31	Caucasian	35,6	17 Feb ’09	17 Feb ’09	No
P.3.3	28	Caucasian	33	8 Apr ’09	6 Apr ’09	No
P.3.4	31	Caucasian	35	-	7 Apr ’09	No
P.3.5	32	Caucasian	39,2	26 Aug ’08	26 Aug ’08	No

As can be seen in Table 3.2 all of the women in the study were Caucasian, they were mostly in their twenties, some in their thirties and one participant was 41 years old. Twenty-seven of the participants did not have any known medical condition except for C9 who has Porphyria, an

[4] C indicates the participants in the control group
[5] P.1 indicates the participants in the first trimester
[6] P.2 indicates the participants in the second trimester
[7] P.3 indicates the participants in the third trimester

inherited disorder of certain enzymes in the heme biosynthetic pathway. Therefore her data was not included in the statistical analysis of this perturbation. This study group was as homogenous as possible, seeing that the samples were obtained from voluntary participants over an eight month period.

3.2.3 Age Participants

For this perturbation I defined four experimental groups; (1) a young infant group (3 days – 7 months old), (2) an older infant group (1 year - 2.5 years), (3) a child group (11 – 13 years old) and (4) a young adult group (20 – 27 years old). The ten infants younger than a year old are henceforth labelled as IM and the remaining infants who were older than a year by IY. Due to the practical difficulty to obtain urine samples from infants and children, I obtained two previously processed data sets containing the organic acid results of these groups. One of the datasets was obtained from Prof. Izelle Smuts, a Pediatrician at the Steve Biko Academic Hospital at the University of Pretoria, and contained the organic acid results of eight older infants (IY) and twenty-five children. The second dataset was obtained from the Laboratory for Inherited Metabolic Defects at North-West University and included the organic acid results of ten young infants (IM). The data sets for the adult group was generated by myself, using the exact methodology as for the first two groups as mentioned above. From the data set I obtained pertinent information about the participants in these studies which are listed in Tables 3.3, 3.4 and 3.5.

Table 3.3: Information on the young infant group (IM).

Case no.	**Age at sample collection**	**Ethnicity**	**Gender**
I4	1 month	Black	Male
I5	2 months	Black	Male
I6	1 month	Black	Male
I9	4 days	Caucasian	Male
I10	1 month	Indian	Male
I11	1 month	Caucasian	Male
I13	3 days	Caucasian	Female
I14	5 months	Indian	Female
I16	7 months	Indian	Male
I18	1 month	Caucasian	Male

Table 3.4: Information on the older infant group (IY).

Case no.	Age (in years) at sample collection	Ethnicity	Gender
I63	1	Black	Male
I64	1	Black	Male
I65	1	Black	Male
I47	1	Caucasian	Male
I62	1	Indian	Male
I61	1	Indian	Female
I93	2	Caucasian	Male
I100	2.5	Caucasian	Male

Table 3.5: Information on the child group.

Case	Age (in years)	Ethnicity	Gender	Case	Age (in years)	Ethnicity	Gender
C60	11	Black	Male	C112	12	Black	Male
C04	11	Caucasian	Male	C111	12	Caucasian	Male
C06	11	Caucasian	Male	C34	12	Caucasian	Male
C81	11	Caucasian	Male	C119	12	Coloured	Male
C96	11	Caucasian	Male	C76	12	Coloured	Male
C118	11	Coloured	Male	C77	12	Coloured	Male
C122	11	Coloured	Male	C78	12	Coloured	Male
C135	11	Coloured	Male	C30	12	Caucasian	Female
C48	11	Coloured	Male	C79	12	Coloured	Female
C67	11	Black	Female	C80	12	Coloured	Female
C12	11	Caucasian	Female	C54	13	Black	Male
C49	11	Coloured	Female	C74	13	Coloured	Male
C99	11	Black	Male				

All the participants in the adult group were voluntary participants and signed informed consent forms (Appendix C) before any samples were taken. The samples of the adult group were obtained from individuals from the North-West University, which met the criteria for individuals between the ages of 20 and 29. These participants were on a free and unrestricted diet. Information about the adult group such as age and gender is listed in Table 3.6.

Table 3.6: Information on the adult group.

Case	Age (in years)	Ethnicity	Gender	Case	Age (in years)	Ethnicity	Gender
A1	23	Caucasian	Female	A18	23	Caucasian	Female
A2	27	Caucasian	Female	A19	22	Caucasian	Female
A3	24	Caucasian	Female	A20	22	Caucasian	Female
A4	23	Caucasian	Male	A21	22	Caucasian	Female
A5	23	Caucasian	Female	A22	26	Caucasian	Male
A6	22	Caucasian	Female	A23	22	Caucasian	Female
A7	23	Caucasian	Female	A24	24	Caucasian	Female
A8	23	Caucasian	Male	A25	20	Caucasian	Female
A9	23	Caucasian	Male	A26	22	Caucasian	Female
A10	24	Caucasian	Female	A27	27	Caucasian	Male
A11	21	Caucasian	Male	A28	22	Caucasian	Male
A12	21	Caucasian	Male	A29	22	Caucasian	Female
A13	25	Caucasian	Female	A30	21	Caucasian	Female
A14	27	Caucasian	Male	A31	23	Caucasian	Female
A15	27	Caucasian	Female	A32	24	Caucasian	Female
A16	26	Caucasian	Female	-	-	-	-
A17	23	Caucasian	Female	-	-	-	-

The information listed in Tables 3.3, 3.4, 3.5 and 3.6 are visualised in Figures 3.1, 3.2 and 3.3 according to gender, ethnicity and age, respectively.

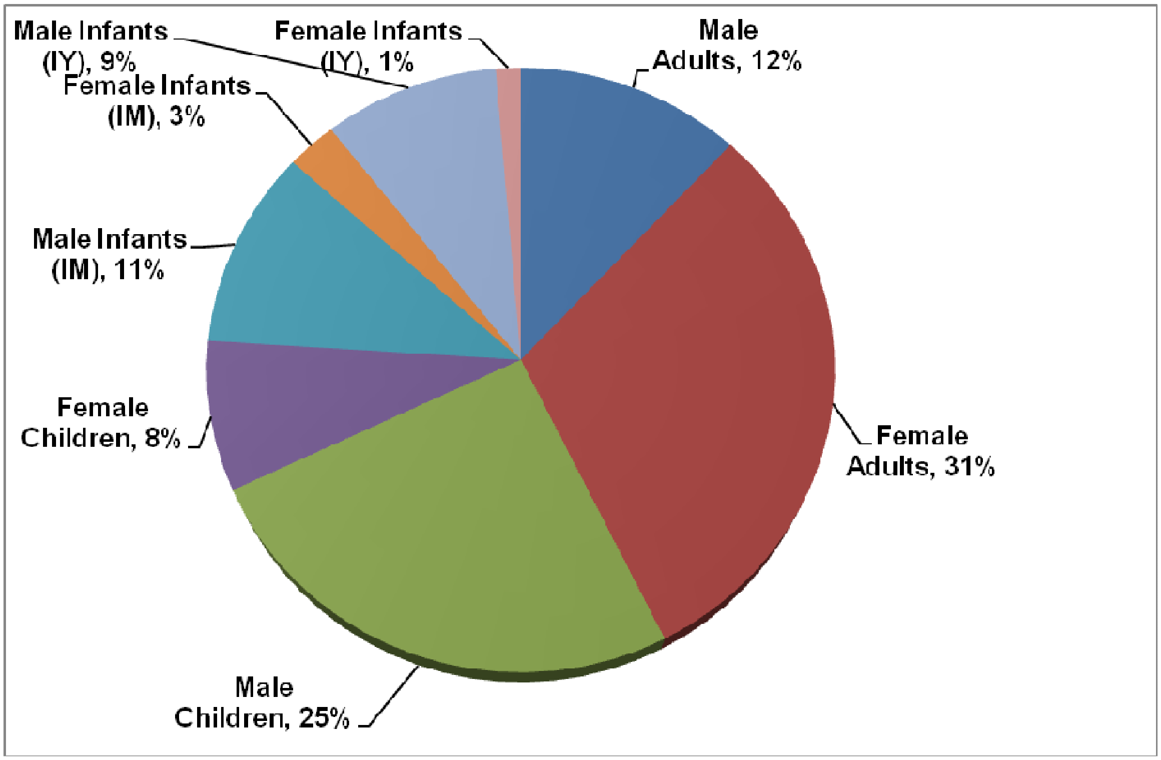


Figure 3.1: Pie chart of the percentage cases in the age investigation according to gender.

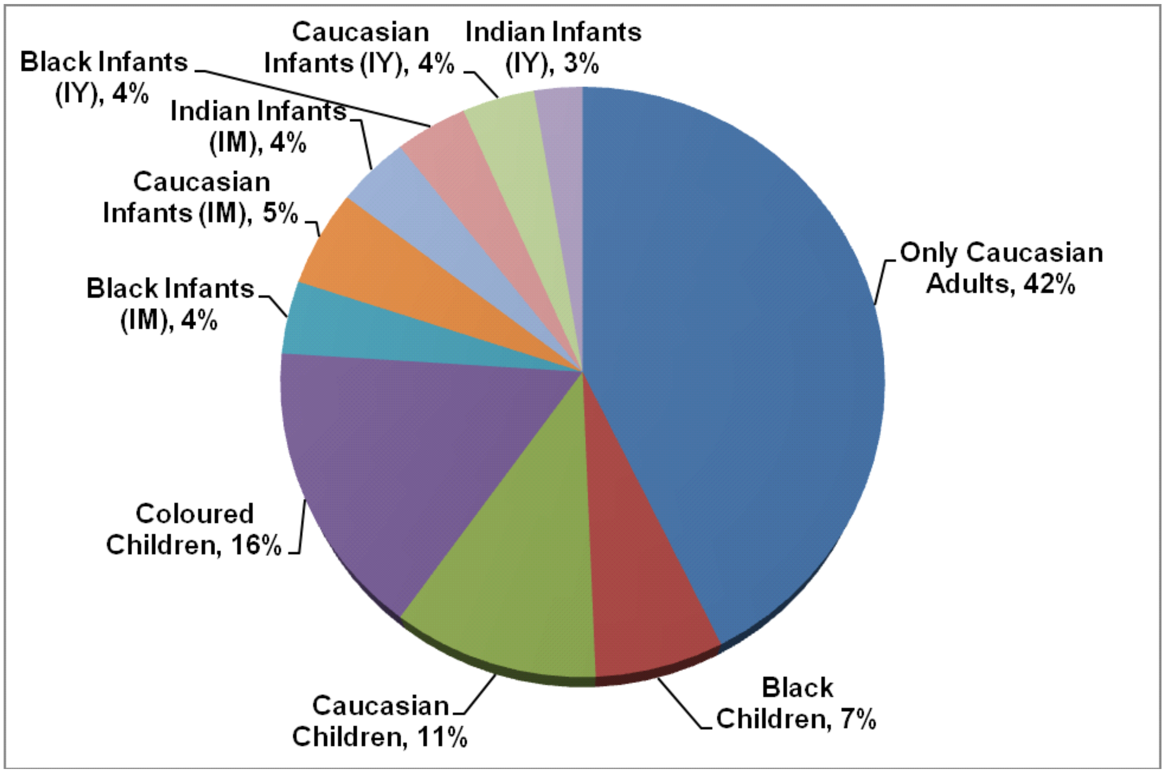


Figure 3.2: Pie chart of the percentage cases in the age investigation according to ethnicity.

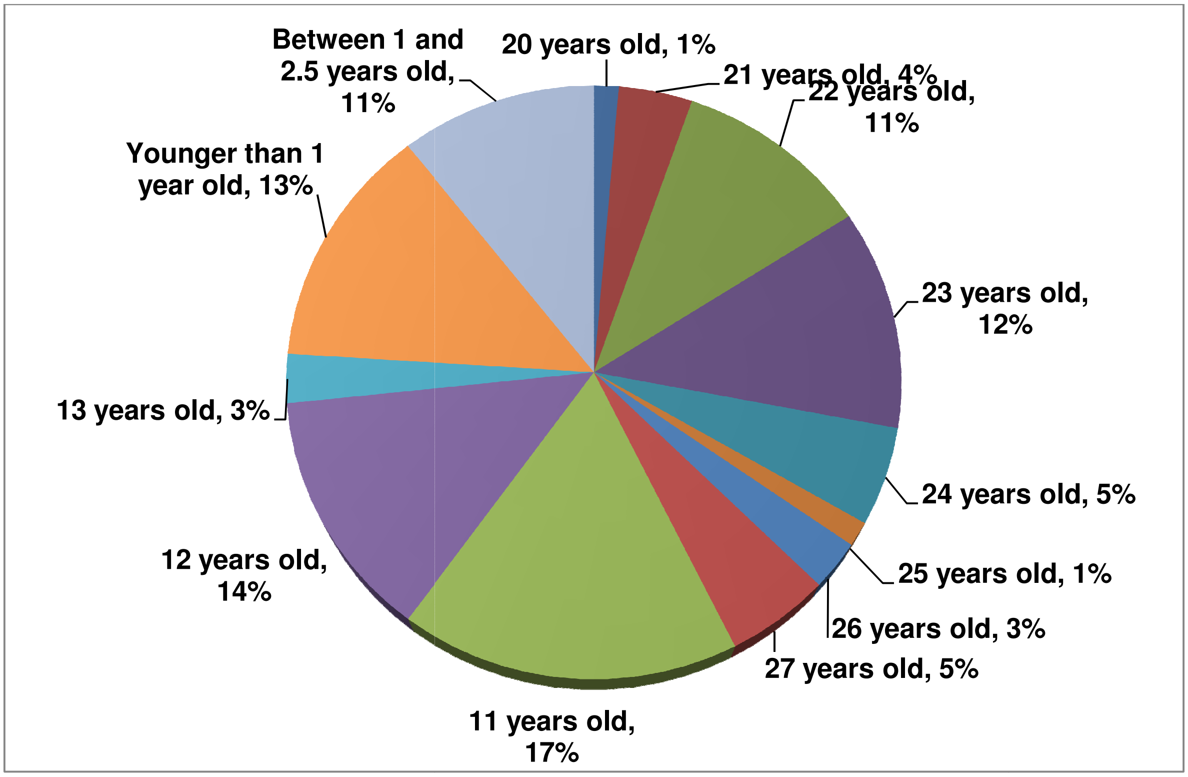


Figure 3.3: Pie chart of the percentage cases in the age investigation according to age.

Figure 3.1 depicts the percentage cases for the four age groups according to gender. Most of the young and older infants are male whereas the child and adult groups are mostly female. The percentage cases for the different age groups according to ethnicity are illustrated in Figure 3.2, with most of the young (IM) and older (IY) infants and adults being Caucasian and the child group Coloured. The percentage cases according to age are represented in Figure 3.3 with most infants being younger than a year and the rest of the infants being older than a year. The child group included mostly eleven-year-olds and the adult group mostly twenty-three-year-olds. However, the available samples are heterogeneous as far as gender and ethnicity are concerned.

3.3 Experimental design

Section 3.3 discusses the experimental design utilised for each perturbation, specifically the sample collection and sample storage for each perturbation before and after analysis of the urine samples. The analytical and statistical methods are discussed in Sections 3.5.2 and 3.7. The urine samples obtained were analysed in the Laboratory for Inherited Metabolic Defects at the Potchefstroom campus for organic acids by means of the standardised method.

3.3.1 Menstrual Cycle

Firstly the concentrations of the hormones were determined in blood and saliva over a single menstrual cycle for one of the participants, in order to establish the duration of each of the three phases. In total ten blood samples were drawn at the Drs. Du Buisson, Bruinette, Kramer Inc. laboratory situated next to the Mooi Med Hospital, Potchefstroom and sent to their main laboratory in Pretoria, where the concentrations of the FSH, LH, progesterone and estradiol hormones were determined. Table 3.6 lists the days when blood samples were taken during the course of the study, where day 1 was the first day of menstrual period. Ten saliva samples were collected on the day the blood samples were drawn and kept at -20℃. After the results for the various blood samples were received, three saliva samples were selected and sent away for estradiol and progesterone determination. These saliva samples were selected as they were the best representative of each cycle, specifically day 2 (follicular phase), day 16 (ovulatory phase) and day 27 (luteal phase). Hence one of the participants provided ten blood and three saliva samples obtained on days 2, 5, 10, 12, 14, 16, 18, 20, 23 and 27. This covers the whole period of one menstrual cycle.

Each participant started providing early morning urine samples of approximately 20 ml, the day after the menstrual period ended and stopped collection when the menstrual period began again. The participants indicated the day of the cycle when samples were collected on the urine container provided. The total number of urine samples collected for each participant was twenty-two. The urine samples were kept cold at about 4℃, until its delivery and were then stored at -20℃ before and after analysis. The samples were analysed within two to three months of the collection date.

3.3.2 Pregnancy

For the untargeted metabolomics approach for this perturbation we obtained an early morning urine sample of approximately 20 ml, from pregnant women and non-pregnant women. In addition to the urine samples blood samples were also collected as they were required for the targeted metabolomics approach, which will be discussed in Chapter 5. The blood samples were drawn at Drs. Du Buisson, Bruinette, Kramer Inc. or Pathcare laboratories in Potchefstroom and were delivered, with the early morning urine samples to the Laboratory for Inherited Metabolic Defects. The samples were delivered within twenty four hours of collection to the Laboratory of Inherited Metabolic Defects and were kept cold. The urine samples were then stored at -20℃ before and after analysis. The samples were analysed within one to five months of collection, depending on how many samples were collected in a given time, for example if ten samples were collected in a single month the samples were analysed within one month. DNA was immediately extracted from the various blood samples by means of a QIAamp blood mini kit from QIagen and was then kept at 4℃ before and after polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP).

3.3.3 Age

The participants in the adult group collected early morning urine samples. The samples were kept cold at about 4℃ until delivery, after which it was kept at -20℃ before and after analysis. The samples were analysed within a month after collection, as these samples were the easiest to obtain.

3.4 Material

The following section contains a list (Table 3.6) of all the material used during the study, the company where it was bought as well the catalogue number.

Table 3.6: Reagents, laboratory apparatus and the supplier or manufacturer used during organic acid analysis.

Reagents	**Supplier/Manufacturer**
Hydrochloric acid (HCl)	Merck: Cat no. 100319
4-Phenyl butyric acid (mw 164.20)	Fluka: Cat no. 78243
Ethyl acetate HPLC grade/distilled	Sigma: Cat no. 494518
Diethyl ether HPLC grade/distilled	Sigma: Cat no. 309966
Sodium sulphate (Na_2SO_4) anhydrous	Merck: Cat no. 1.06649
Bis(trimethylsilyl)-trifluoracetamid (BSTFA)	Supelco: Cat no. 33027
Trimethylchlorosilane (TMCS)	Sigma: Cat no. H T 4252
Pyridine	Merck: Cat no. 51124060LC
Hexane	Sigma: Cat no. 52767
Laboratory Apparatus	
Kimax culture tubes (Large): 16 x 125mm	Lasec: Cat no. GIMK 45066A16125
Kimax culture tubes (Large): 13 x 100mm	Lasec: Cat no. GIMK 45066A13100
Distiller Pasteur pipettes	Merck: Cat no. 612-1702
Roto-torque (Rotator)	Labotech: Cat no. 67003
Graduated pipettes – 10-100ul	Merck: 3111000149
Graduated pipettes – 100-1000ul	Merck: 3111000165
Centrifuge	Optolabor (BHG 1100)
Heating block	Pierce: Cat no. 18840
Evaporating adaptor	Pierce: Cat no. 18817
Nitrogen	Pierce: Cat no. 18785
Hamilton syringes (100 µl)	Separations: Cat no. 80391 and 80366
Sample vials	Separations: Cat no. 11090500
Sample inserts	Separations: Cat no. 09151819
Sample caps	Separations: Cat no. 06090357
Capillary GC-MS column: VF1-MS (30m x 0.32 x 0.25ID)	SMM-Instruments: Cat no. CP8924
Gas chromatograph	Hewlett Packard 5880
Mass spectrometer	Hewlett Packard 5988A
Data analysis software	Automated mass spectral deconvolution and identification system (AMDIS 2.65)

3.5 Methods

3.5.1 Organic Acid Analysis

This method is used on a daily basis in the Laboratory for Inherited Metabolic Defects at the Potchefstroom campus and Standard Operational Procedures for this method is available. Below is the description of the organic acid analysis as used in this investigation.

Urinary creatinine values are used as a reference metabolite to compensate for the differences in concentration between urine samples and to obtain an organic acid profile to be comparable regardless of these differences. The creatinine values of the participants' urine samples were determined by Drs. Du Buisson, Bruinette, Kramer Inc. Laboratory, Potchefstroom and according to the values a variable amount of urine and internal standard was used for organic acid extraction. Urine samples were thawed at room temperature and urine was added to a silanised glass tube (Kimax) according to the creatinine values found in Table 3.7.

Table 3.7: Volume of urine used according to the creatinine values.

Creatinine Value (mmol/l)	**Volume of Urine**
Creatinine > 8.8	0.5 ml
Creatinine < 8.8 and > 0.44	1 ml
Creatinine < 0.44 and > 0.18	2 ml
Creatinine < 0.18	3 ml

The internal standard (26.25 mg of 3-phenylbutyric acid dissolved in a few drops of NaOH and then diluted to 50ml with distilled H_2O) was added to the urine, to a final concentration of 180 mmol/mol creatinine. This ensures a fairly constant ratio between the internal standard and urinary organic acids, which leads to a more constant extraction efficiency of the internal standard and the organic acids. This specific internal standard was used as it is absent from normal urine as well as the urine of patients with known pathological conditions, it co-elutes with a very small number of other organic acids and elutes more or less in the middle of the organic acid profile. The urine was acidified by adding approximately 5 drops of 5N HCl to it, since it stabilises the protonated forms of the molecular structure of the organic acids and increases its solubility. Six ml of ethyl acetate was added to the sample and shaken on the rotary wheel for twenty minutes, after which it was centrifuged for two minutes at 1300 x g. The organic (upper) phase was aspirated into a clean glass tube and three ml diethylether was added to the aqueous (lower) phase. The tube was shaken for ten minutes and was then centrifuged for ten[8] minutes at 1300 x g. The organic phase was aspirated into the tube containing the ethylacetate phase. Thus organic acids are isolated from biological fluids as a group with organic acid solvents namely ethylacetate and diethylether, this represents a liquid-liquid extraction method (Potchefstroom Laboratory for Inherited Metabolic Defects). A small amount of Na_2SO_4 was added to the ethylacetate/diethylether phase in order to remove any residual water. It was centrifuged and transferred into a clean glass tube.

[8] This step differs from the standard operating protocol in the Laboratory for Inherited Metabolic Defects as the tube is centrifuged for about three minutes and not ten.

The organic phase is evaporated to dryness under nitrogen for about an hour in a 37℃ heating block. The dried extract is derivatised with O-bis(trimethylsislyl)-trifluoracetamid (BSTFA), Trimethylchlorosilane (TMCS) and piridine. Derivatisation is an important step since it converts organic acids to thermally stable, volatile and chemically inert derivatives that permit gas-phase resolution (Kuhara, 2005; Lehotay *et al.,* 1995). BSTFA is a silylation reagent that is an effective trimethylsilyl (TMS) donor which stabilizes a wide variety of polar parent compounds for volatilization during gas chromatography (Kuhara, 2005; Thermo Fisher Scientific, Inc.). BSTFA is the preferred silylation reagent for organic acids because of its reactivity, solvent properties, volatility as well as availability. It gives better chromatographic separation than other reagents since its by-products (mono(trimethylsilyl) trifluoro-acetamide and trifluoroacetamide) usually co-elute with the solvent front and not with the derivatised products (Kuhara, 2005; Thermo Fisher Scientific Inc.). TMCS acts as a silylation catalyst for the reaction with HCl as a byproduct of the reaction (Kuhara, 2005, www.instrument. com) and pyridine is added to the BSTFA and TMCS mixture since it acts as a polar solvent for the TMS reaction (Sigma Aldrich, Inc.). The volume urine initially used gave a creatinine concentration equal to 21 µmol/ml derivatisation agent, hence BSTFA, TMCS and pyridine was added to the tubes according to a 5:1:1 ratio. The tubes were then incubated in an 85℃ sand bath for 45 minutes. The derivatised mixture was finally transferred to a 1.5 ml vial for analysis on the GC/MS.

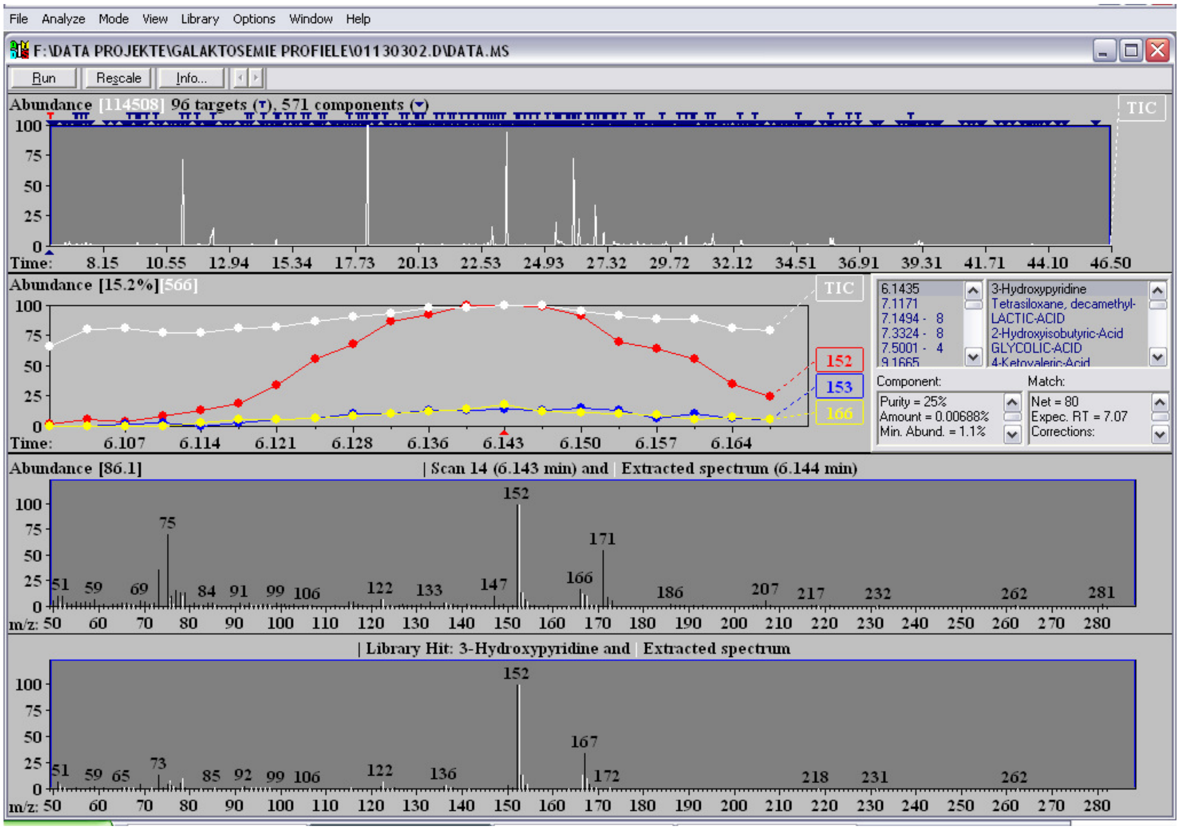
3.5.2. Gas chromatography and mass spectrometry specifications

The volatile derivatives are analysed on GC-MS which verifies the presence of any metabolites. The prepared samples were injected into the Hewlett-Packard GC-MS for analysis. The organic acids were analysed on a Hewlett Packard 5880 gas chromatograph linked to a 5988A mass selective detector. The sample (1 ul) was injected onto the GC using split injection at a split ratio of 15:1. The injection port temperature was kept at 240℃, the initial oven temperature was 60℃, and it was subsequently increased with 6℃ per minute after one minute. The final temperature was 270℃ which was maintained for 5 minutes. This system was equipped with a VF1-MS (30m x 0.32 x 0.25ID) capillary column which has a constant flow rate of helium (5.08 psi), the carrier gas at 2.1 ml/min. The column is responsible for fractionation of the various metabolites. Ionisation was attained via electron impact (EI) with a potential of 70eV and the mass spectrometer was in scan mode for acquisition.

3.6 Identification and quantification of the metabolites

The data generated on the GC-MS for the respective perturbations were analysed and deconvoluted via AMDIS 2.65 (Figure 3.4), which was linked to the NIST (National Institute of Standards and Technology) mass spectral search programme for the commercial

NIST/EPA/NIH mass spectral library as well as an organic acid specific library that was further developed by Prof LJ Mienie. Figure 3.4 gives an example of a chromatogram in AMDIS. The software extracts pure component mass spectra from GC/MS data files and then uses these spectra to identify the compounds in the chromatogram via a reference library (Stein, 1999). For all the compounds identified via AMDIS, a response factor of 1 was selected, with a minimum match factor of 70% and was analysed via the use of an internal standard for RI. The analytic settings for deconvolution had a medium adjacent peak subtraction, resolution, sensitivity and shape requirements as well as a component width of 12.


File
Analyze
Mode
View
Library
Options
Window
Help
F:\DATA PROJEKTE\GALAKTOSEMIE PROFIELE\01130302.D\DATA.MS
Run
Rescale
Info...
Abundance [114508] 96 targets (T), 571 components (▾)
TIC
100
75
50
25
0
Time: 8.15 10.55 12.94 15.34 17.73 20.13 22.53 24.93 27.32 29.72 32.12 34.51 36.91 39.31 41.71 44.10 46.50
Abundance [15.2%][566]
TIC
152
153
166
Time: 6.107 6.114 6.121 6.128 6.136 6.143 6.150 6.157 6.164
Abundance [86.1]
| Scan 14 (6.143 min) and | Extracted spectrum (6.144 min)
152
75
171
51 59 69 84 91 99 106 122 133 147 166 186 207 217 232 262 281
m/z: 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280
| Library Hit: 3-Hydroxypyridine and | Extracted spectrum
152
167
51 59 65 73 85 92 99 106 122 136 172 218 231 262
m/z: 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280

6.1435	3-Hydroxypyridine
7.1171	Tetrasiloxane, decamethyl-
7.1494 - 8	LACTIC-ACID
7.3324 - 8	2-Hydroxyisobutyric-Acid
7.5001 - 4	GLYCOLIC-ACID
9.1665	4-Ketovaleric-Acid

Component:
Purity = 25%
Amount = 0.00688%
Min. Abund. = 1.1%

Match:
Net = 80
Expec. RT = 7.07
Corrections:

Figure 3.4: Example of a chromatogram in AMDIS.

The data obtained was exported to Excel and the relative metabolite concentrations were determined according to the concentration of 3-phenylbutyric acid[9]. If compounds produced more than one peak, the sum of the area of the peaks was used if the components were well established, e.g. urea, hippuric acid and 3-methylglutaconic acid. Some unidentified compounds with low concentrations were excluded from further analyses, as the objective of this study was not to identify such unknown compounds. The concentrations of the different organic acids

[9] The concentration was calculated with the following formula = Area of the compound/Area of the IS x concentration of the IS. The concentration may be calculated in either mg/g or mmol/mol creatinine.

identified for the three perturbations were expressed as mmol/mol creatinine. It should be noted that quantification of organic acids is more complex than the identification of the organic acids. As this is a rather complex aspect, it is not elaborated on here, but is presented as supplementary material in Appendix D. This material is however, of key importance in relation to Aim 1 of my investigation. Statistical analyses were performed on the respective matrixes obtained for the different perturbations studied.

3.7 Statistical Methods

To address the research questions the following statistical methods were used:

- Descriptive statistics,
- Data smoothing,
- Effect sizes,
- Data transformation and data imputation,
- Multivariate methods such as principal component analysis (PCA) and partial least square-discriminant analysis (PLS-DA).

Subsequently a short description of the methods above and their applicability are discussed.

3.7.1 Descriptive statistics and data smoothing

Descriptive statistics summarises the data set quantitatively with summary statistics and graphical representation in order to describe the main features of a collection of data (Trochim, W.M.K., 2006). Amongst others, summary statistics summarise data via measures of central tendency and measures of dispersion. Measures of central tendency provide information about the expected value of a variable and can be determined with the mean or the median. Measures of dispersion provide information about the variation of a variable and are usually expressed as the standard deviation. The graphical representation in this dissertation includes scatterplots, boxplots, errorbars, pie charts and scores plots. The following smoothing methods were also used, namely moving average and moving medians and are discussed in more detail in Section 4.2.1.

3.7.4 Effect sizes (Group comparison and variable selection)

The multivariate data that are generated in this dissertation are marred by dimensionality i.e. a lot of variables with few cases. In order to remove unnecessary noise in the data, prior to multivariate analysis, effect sizes are used. This reduces the dimension of the data so that the statistical models e.g. PCA and PLS-DA can only consider variables that have a potential to discriminate between groups. Effect sizes is a measure of practical significance (Ellis *et al.,* 2003), hence if the difference between two groups is important with respect to a specific variable the effect size for that variable will be large. The effect size (d) for this type of comparison is usually interpreted as follows: d=0.2 (small effect), d=0.5 (medium effect) and d=0.8 (large effect). The effect size used in this dissertation is as follows:

$$d_{i,j,k} = \frac{[\bar{X}_{ik} - \bar{X}_{jk}]}{max\,(S_{ik}, S_{jk})},$$

where i and j is the indexes representing the groups and k represents the variable under consideration. Also $\bar{X}_{ik}$ and $\bar{X}_{jk}$ is the mean value for variable k of groups i and j respectively. Similarly S_{ik} and S_{jk} is the standard deviation for variable k of the respective groups.

The effect size $d_{i,j,k}$ is calculated for all possible combinations of selecting two groups from the number of groups under consideration e.g. for three groups there will be three effect sizes per variables namely $d_{1,2,k}$, $d_{1,3,k}$ and $d_{2,3,k}$ comparing groups 1 with 2, 1 with 3, and 2 with 3. In the case where effect sizes are used for variable selection, a specific variable is considered for further multivariate statistical analysis if at least one of these effect sizes is of medium importance.

3.7.5 Data transformation and data imputation

The data that are generated in this dissertation are further quarred by (i) huge scale differences amongst variables and (ii) a significant number of zero values which is caused by the limit of detection of the GC-MS system used. To address issue (i) the variables are shifted log transformed (i.e. Y = ln(X+1)) in order to make the variables scales more comparable. Data transformation provides a means of modifying variables in order to correct violations of the statistical assumptions underlying multivariate analysis or to improve the relationship between variables (Hair *et al.,* 2006).

In addition, the zero values are replaced by a random sample from a beta(0.1,1) distribution bounded between zero and the detection limit. These values are extremely close to zero and do not alter the statistical results when replaced by a different set of random values.

3.7.6 Multivariate methods

Multivariate analysis refers to a type of statistical technique where two or more variables or observations which are dependent on each other are studied simultaneously (Hair *et al.,* 2006:4). According to them true multivariate analysis is where all the variables must be random and interrelated in such ways that their different effects cannot meaningfully be interpreted separately. Thus its character lies not only in the number of variables being described but also in the multiple combinations of variables (Hair *et al.,* 2006).

3.7.6.1 Principal component analysis (PCA)

According to Johnson and Wichern (1998) a principal component analysis is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables. Its general objectives are data reduction and interpretation. The first principal component (PC1) describes the largest variation in the data set, whilst the second principal component (PC2) describes the next-largest variation and so on. The principal component scores (lower-dimensional mapping) are presented visually using a scores plot. This plot indicates inherent clustering of groups of data which are based on the similarity of their input coordinates (Coen *et al.,* 2005). Hence, in metabolomic research PCA is often used as an unsupervised pattern recognition tool, where the main aim is to discriminate amongst various groups and to find the variables causing this separation. These variables are often calculated from the principal component loadings or by ranking the variables important in the PCA projection (VIP).

3.7.6.2 Partial least squares-Discriminant analysis (PLS-DA

In metabolomic research PLS-DA is used as a supervised pattern recognition procedure. It many ways, PLS can be regarded as a substitute for the method of multiple regression, especially when the number of predictor variables is large. PLS-DA aims to predict group membership (dependent variable - coded as dummy variables) from the various metabolite concentrations (predictor variables). Potential biomarkers can also be identified as those variables important in the prediction model. The reader is referred to Kettaneh *et al.,* (2005) for a more elaborate discussion of PLS-DA as well as PCA.

Chapter 4 – Results of an untargeted metabolomics analysis of three natural perturbations

4.1 Introduction

Since metabolomics is a relatively new research field not many studies have been done on normal human metabolism as they usually focus on metabolic defects or physiological conditions which present with a major perturbation. Natural variations might, however, influence the metabolomics results as was discussed previously. Three natural perturbations were chosen to investigate this possibility, of which the results are presented here. This study could lead to a better understanding of the human metabolome with respect to these perturbations (the menstrual cycle, pregnancy and age), as well as of the use of case-control studies of larger perturbations. The following sections will cover the results obtained for each of the selected perturbations. The statistical analysis and discussion of the menstrual cycle perturbation is presented in Section 4.2. Sections 4.3 and 4.4 contain the results of the pregnancy and age perturbations respectively. The statistical methods used to analyse the data are discussed in Section 3.7.

In addition a biological filter was applied to the data (where applicable) in order to reduce the dimension of the data matrix. This filter is used to exclude non-physiological metabolites thus reducing the number of variables being analyzed statistically. In addition, exclusion is obtained by comparing the metabolites or variables in the data set with the physiological importance of these metabolites as described in literature, specifically the human metabolome database (www.hmdb.ca) and the information from Blau *et al.,* (2005). The latter was used in the application of the biological filter as it is an internationally used reference with regard to the diagnosis of metabolic defects, whereas the human metabolome database is a freely available electronic database that contains comprehensive information about small molecular weight metabolites found in human biological fluids. A limitation of the use of the biological filter is that it may exclude possible biomarkers as only metabolites with a known physiological importance or influence in the human body are selected. It was nevertheless used since the research focused on normal human metabolism as well as the possible differences that may occur in the groups of the selected perturbations being studied.

The reference ranges used to compare the different metabolites with one another were either obtained from Blau *et al.,* (2005) or the human metabolome database (www.hmdb.ca). Blau *et al.,* (2005) derived their reference values from 30 premature infants, 34 term newborns, 33 children (13 younger than 5 years and 20 older than 5 years) and finally 9 adults. The reference values were reported according to the minimal, maximal and mean mmol/mol creatinine value obtained for the different metabolites in each age group. The human metabolome database lists different reference values obtained from various studies performed on different population groups, hence it is not reported if the values are listed according to the mean or median values for the specific metabolite (see Appendix E for an example of how reference values are listed in the database).

4.2 Menstrual cycle

This perturbation was studied by comparing the organic acid profile of urine samples collected in the pre- to postmenstrual period. An important aspect of this experiment is the effect of time on the metabolomics profile, which is clearly shown in Figure 2.4. Before any statistical analyses of the urinary organic acids were done, I had to determine the duration of each phase in days. I monitored the change in estradiol, FSH, LH and progesterone concentrations over a period of one month for one of the cases (participant 1) via blood and saliva samples (see Section 3.1.1). The results obtained from Drs. Du Buisson, Bruinette, Kramer Inc. (Table 4.1) were compared with their reference ranges (Table 4.2) for each of the phases. The hormone concentrations for each day were compared to the reference values and resulted in the placement of the day in a particular phase. The days of each phase was then grouped together and an average value for each phase for each hormone was obtained, defining the three phases of the postmenstrual period. For statistical analysis of the urine samples the three phases were thus grouped as follows: Phase I stretched from day eight to fourteen; Phase II from day fifteen to nineteen and Phase III from day twenty to twenty-seven.

Table 4.1: Changes in hormonal concentration during the menstrual cycle determined in blood samples for participant 1.

Phase	Blood samples	Estradiol (pmol/l)	FSH (IE/l)	LH (IE/l)	Progesterone (nmol/l)
Follicular	**Day 2**	94	6.6	5.1	2.5
	Day 5	116	6.3	4.7	2.1
	Day 10	140	8	9.7	2.1
	Day 12	183	6.8	9.6	1.9
	Average	**161.5**	**7.4**	**9.65**	**2**
Ovulatory	**Day 14**	208	7.3	13	2.5
	Day 16	330	6.3	14.8	1.9
	Day 18	798	7.3	27.2	2
	Average	**564**	**6.8**	**21**	**1.95**
Luteal	**Day 20**	356	3.2	4.5	19.5
	Day 23	373	2.5	7.1	31.2
	Day 27	304	2.3	2	17.1
	Average	**344.33**	**2.67**	**4.53**	**22.6**
Phase	**Saliva samples**				
Follicular	**Day 2**	12.9	-	-	0.2
Ovulatory	**Day 16**	15.1	-	-	0.2
Luteal	**Day 27**	14.0	-	-	0.6

Table 4.2: Reference ranges for each of the four menstrual cycle hormones, for each phase according to Drs. Du Buisson, Bruinette, Kramer Inc.

Phase	Estradiol (pmol/l)	FSH (IE/l)	LH (IE/l)	Progesterone (nmol/l)
Follicular	<657	2.9-14.6	1.9-14.6	<9.3
Ovulatory	173-1902	4.7-23.2	12-118	2.4-9.4
Luteal	146-1045	1.4-8.9	0.7-12.9	4.5-111

A multivariate analysis (PCA and PLS) was not performed on the data since there were insufficient cases for such an analysis. The data is also time-dependent which complicates the statistical analysis. In addition, the duration of each cycle can vary between participants, for example the ovulatory phase is between 16 and 32 hours, but to ensure that the ovulatory phase was investigated the data gathering stretched over five days. The data was also smoothed (see 3.7.2 and 4.2.1) according to the moving average and moving median values for each metabolite for the three phases as this removes irregular variation and takes into account the overlap in phases. We developed a bioinformatics protocol to compare the three successive phases during the menstrual cycle. This statistical design is shown in Figure 4.1 and is discussed in Section 4.2.1.

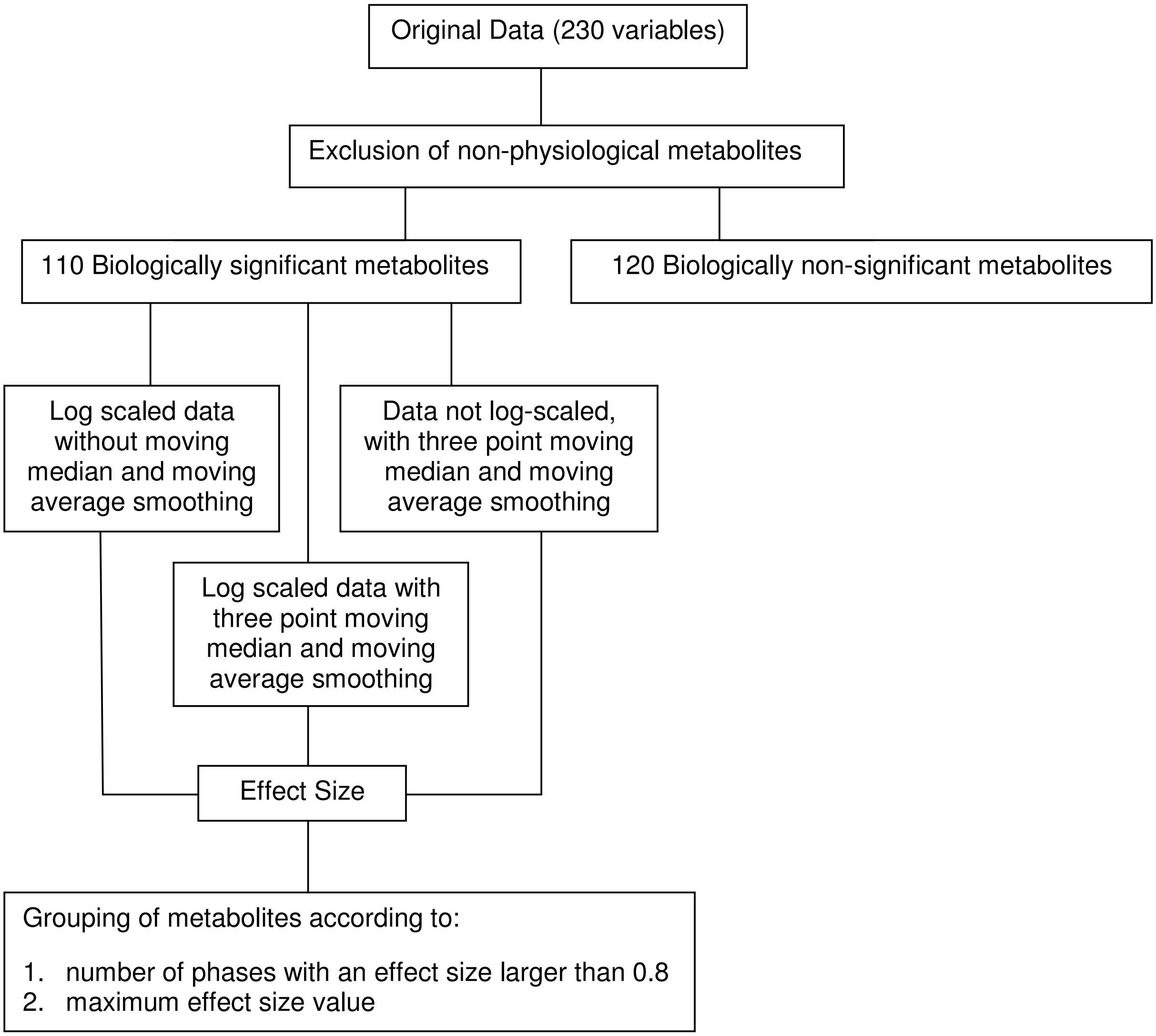

Original Data (230 variables)
Exclusion of non-physiological metabolites
110 Biologically significant metabolites
120 Biologically non-significant metabolites
Log scaled data without moving median and moving average smoothing
Data not log-scaled, with three point moving median and moving average smoothing
Log scaled data with three point moving median and moving average smoothing
Effect Size
Grouping of metabolites according to:
1. number of phases with an effect size larger than 0.8
2. maximum effect size value

Figure 4.1: Flowchart of statistical analysis for the menstrual cycle.

4.2.1 Statistical analysis for the smoothing of the metabolite data on the menstrual cycle

Let $X_{t,i,j}$ be the data where j = 1,…,p indicates a specific metabolite; i = 1,…,n represents the cases (where *n*=3) and ***t = 8,…,27*** the time of the measurement. Also, let w be the window width over which the smoothing is performed. Also, let

$$\boldsymbol{Y}_{t,j,w} = \{\boldsymbol{X}_{t,1,j}, \boldsymbol{X}_{t+1,1,,j}, \ldots \boldsymbol{X}_{t+(w-1),1,j}; \boldsymbol{X}_{t,2,j}, \boldsymbol{X}_{t+1,2,j}, \ldots \boldsymbol{X}_{t+(w-1),2,j}; \boldsymbol{X}_{t,n,j}, \boldsymbol{X}_{t+1,n,j} \ldots, \boldsymbol{X}_{t+(w-1),n,j}\}$$

The following subsets can thus be constructed:

$$\boldsymbol{Y}_{t=8,j,w}; \boldsymbol{Y}_{t=9,j,w}; \ldots \boldsymbol{Y}_{t=27-(w-1),j,w}$$

The moving average and moving medians can therefore be defined as the arithmetic averages and medians of the subsets described directly above. These moving averages and moving medians are calculated for w=3 and w=5 and are then associated with times $t + \frac{w-1}{2}$ where t = 8, 9,..., 27-(w-1). The effect size as described in Section 3.7.4 was then calculated with the smoothed and unsmoothed data as input. The effect sizes were subsequently ranked according to the number of phases with effect sizes larger than 0.8 and then according to the maximum effect size. To illustrate the effect of data smoothing vs. no smoothing, the data for Lactic acid are visualised graphically in Figures 4.2 to 4.6 using scatterplots.

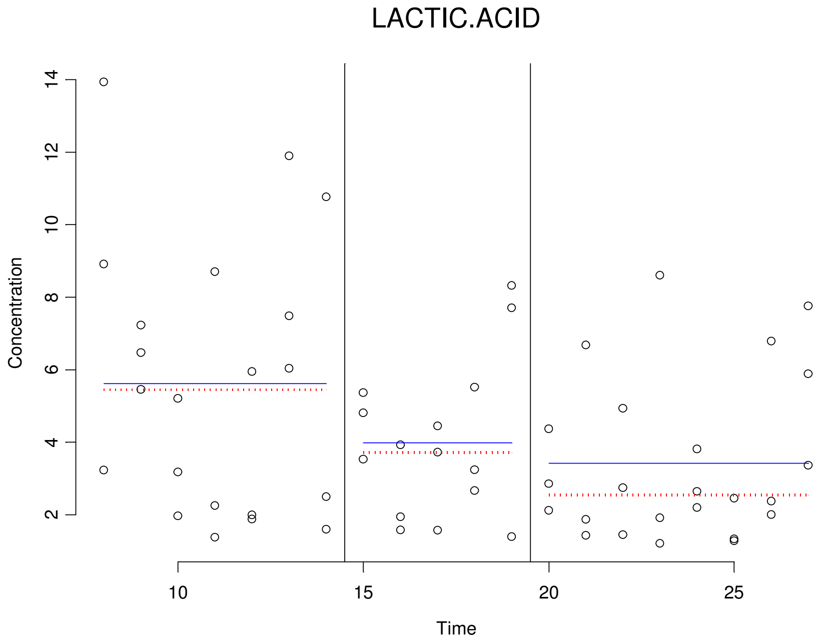


Figure 4.2: Mean (blue) and median (red) concentrations for the three phases for the participants with no data smoothing.

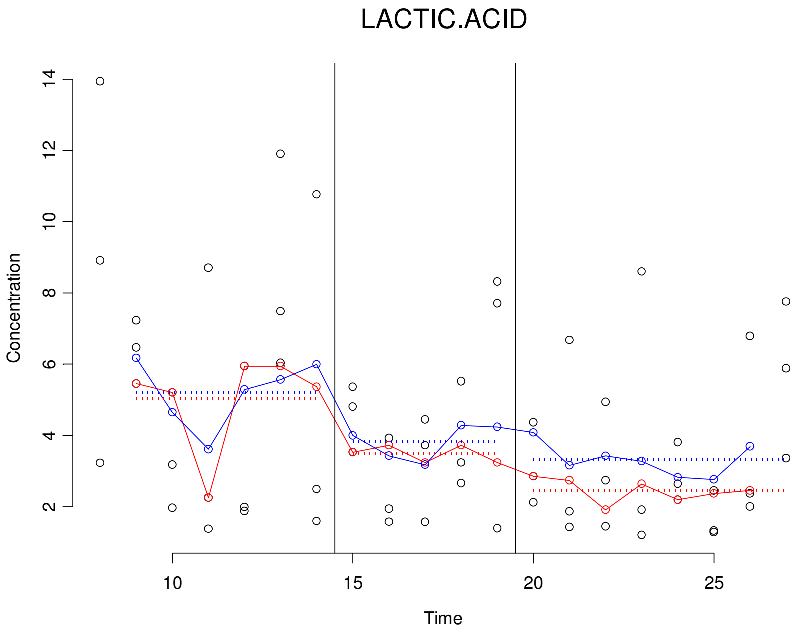


Figure 4.3: Mean (blue) and median (red) concentrations for the three phases for the participants with a three point moving average and moving median smoothing. The dotted lines indicate the mean value of the moving average and moving median data. The circles indicate the data points of the participants.

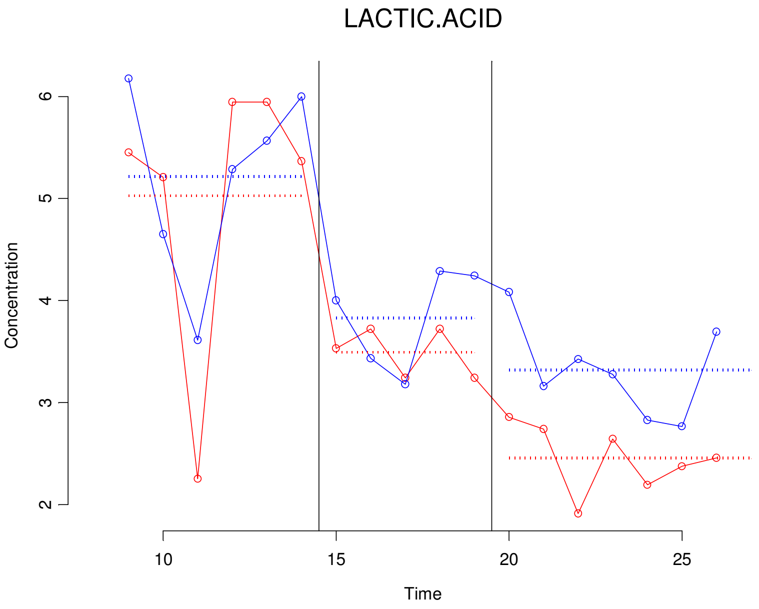


Figure 4.4: Mean (blue) and median (red) concentrations for the three phases for the participants with a three point moving average and moving median smoothing. The dotted lines indicate the mean value of the moving average and moving median data. No data points are indicated on this graph to emphasise the trend over time.

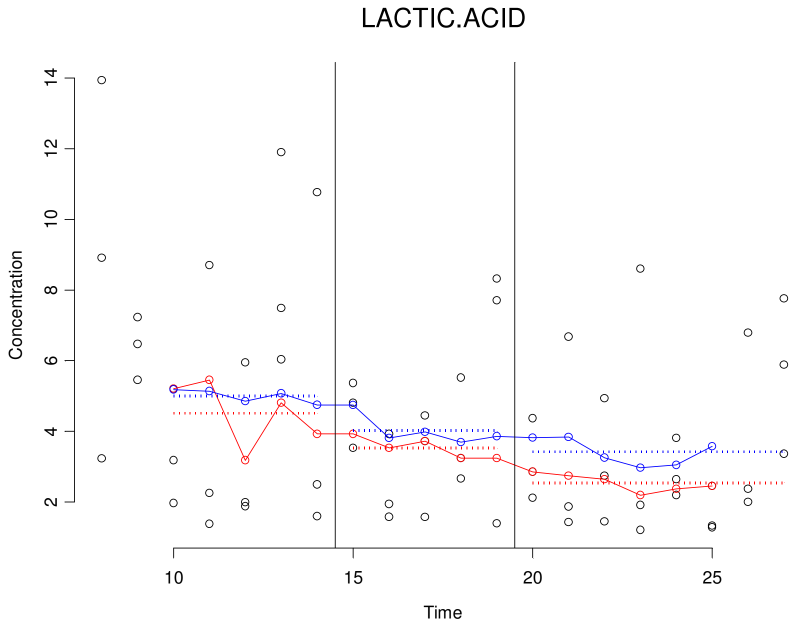


Figure 4.5: Mean (blue) and median (red) concentrations for the three phases for the participants with a five point moving average and moving median smoothing. The dotted lines indicate the mean value of the moving average and moving median data. The circles indicate the data points of the participants.

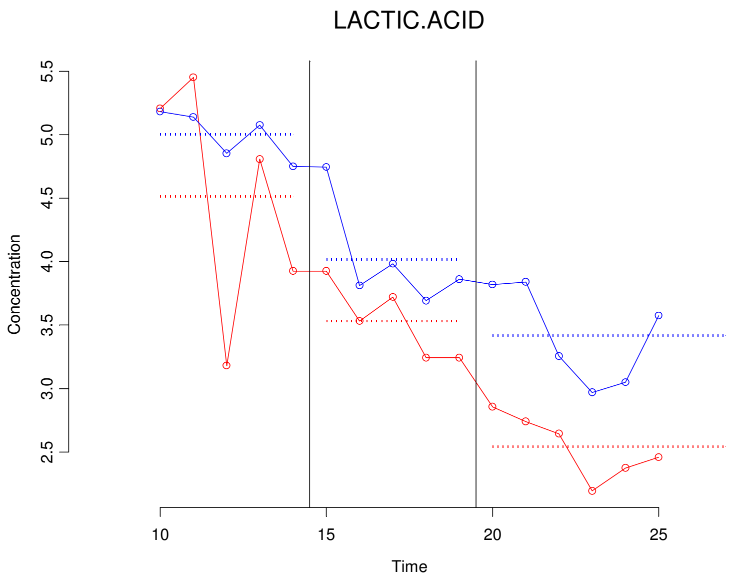


Figure 4.6: Mean (blue) and median (red) concentrations for the three phases for the participants with a five point moving average and moving median smoothing. The dotted lines indicate the mean value of the moving average and moving median data. No data points are indicated on this graph to emphasise the trend over time.

It is evident when Figure 4.2 is compared to Figure 4.3 or Figure 4.5 that the higher the data smoothing value (window width) e.g. three or five, the smaller the variation of the resulting smoothed data becomes. I only report the results based on the three point moving average and moving median smoothing and not the five point moving average and moving median smoothing, since these results were similar. A PCA or PLS-DA could not be performed due to the limited number of cases in comparison to the number of variables. We deemed it important to be able to rank the variables according to their variation between the three phases. This is accomplished via the determination of the effect size (see 3.7.4.1). The effect size values were obtained by comparing the first phase with the second phase, the first phase with the third phase and the second phase with the third phase and are reported in Table 4.3. The data also shows the inherent scale differences amongst variables as is typical for metabolomics data. In order to address this issue the data was log-scaled (see 3.7.5) and compared to data that was not log-scaled.

Subsequently, the following comparisons are investigated

1. Effect size of log-scaled data with zero data smoothing;
2. Effect size of the three point moving average and moving median smoothed data, where the smoothing was performed on the original data; and
3. Effect size of the three point moving average and moving median smoothed, where the smoothing was performed on the log-transformed data.

The metabolites in Table 4.3 are ranked according to (1) their occurrence in the three phases and (2) the maximum effect size value. Therefore if one metabolite has an effect size greater than 0.8 for each phase it is listed before a metabolite that has an effect size greater than 0.8 for two of the phases even though this metabolite may have a greater maximum effect size value (see Appendices E to G). The ranking, as described above, was performed using the median smoothed data, as the mean smoothing is affected by very high or low concentration values.

Table 4.3: Ranking of variables according to their variation (3 point smoothing) log and no log.

Log-scaled data with no smoothing[10]		Log-scaled data with three point smoothing[11]			Data not log-scaled with three point smoothing[12]		
Metabolite	**Effect Size**	**Metabolite**	**Effect Size based on:**		**Metabolite**	**Effect Size based on:**	
			Median smoothing	**Mean smoothing**		**Median smoothing**	**Mean smoothing**
2,5-Furandicarboxylic acid	0.99	Lactic acid	2.72	2.18	Lactic acid	3.16	1.99
2,3-Di-OH-butanoic acid	0.69	Phosphoric acid13	2.19	1.57	Palmitic acid	3.25	1.08
Palmitic acid14	0.69	Succinic acid	1.67	1.39	2,5-Furandicarboxylic acid	3.22	2.90
Octadecanoic acid	0.66	Palmitic acid	3.10	1.29	Ethylmalonic acid	3.21	1.07
3-OH-phenylacetic acid	0.66	2,3-DiOH-butanoic acid	3.08	3.53	Octadecanoic acid	3.07	1.08
Citramalic acid	0.66	Octadecanoic acid	2.95	1.32	2,3-Di-OH-butanoic acid	2.79	3.71
3-OH-isobutyric acid	0.64	Ethylmalonic acid	2.90	1.79	Isocitric lactone	2.49	1.97
2,5-Di-OH-benzoic acid	0.63	Isocitric lactone15	2.30	1.91	3-meth-4-OH-cinnamic acid	2.34	3.64
4-OH-mandelic acid	0.62	3-meth-4-OH-cinnamic acid	2.29	3.49	Adipic acid	2.22	2.06
Lactic acid	0.62	Adipic acid	2.19	1.95	Phosphoric acid	2.00	1.97

Hydroxy- = OH-; Methoxy- = meth; Methyl = met-

[10] See Appendix F for the log-scaled dataset with no smoothing.

[11] See Appendix G for the log-scaled dataset with three point smoothing.

[12] See Appendix H for the three point smoothing dataset that was not log-scaled

[13] Palmitic acid is one of the most common saturated fatty acids found in animals and plants.

[14] Phosphoric acid is not an organic acid but an inorganic acid which is extracted during the analysis.

[15] Isocitric lactone is caused by the derivatisation process and it is difficult to differentiate between citric acid and isocitric acid on a chromatogram generated via AMDIS.

It is evident from Table 4.3 that four metabolites are listed in all three comparisons, six in only two of the comparisons and six in one of the comparisons and are grouped as follow:

Metabolites present in all three comparisons:

- 2,3-Dihydroxybutanoic acid, palmitic acid, octadecanoic acid and lactic acid.

Metabolites present in any two comparisons:

- Phosphoric acid, ethylmalonic acid, isocitric lactone, adipic acid, 3-methoxy-4-hydroxycinnamic acid and 2,5-furandicarboxylic acid.

Metabolites present in any one comparison:

- 3-Hydroxyphenylacetic acid, citramalic acid, 3-hydroxyisobutyric acid, 2,5-dihydroxybenzoic acid, 4-hydroxymandelic acid and succinic acid.

Observations:

For the purposes of this study I only focused on the metabolites identified in the second comparison i.e. log-scaled data with three point smoothing. This particular list was chosen as the metabolites are mostly endogenous, the data are comparable according to scale and the irregular variation in the dataset is reduced. Table 4.4 consists of the mean, median and standard deviation values of the ten metabolites for each phase according to the original data i.e. zero data smoothing and no log-scaling. In addition the reference values and biological significance of these metabolites are also reported in Table 4.4.

The ten metabolites identified were grouped as follows:

1. Long chain fatty acids (octadecanoic acid and palmitic acid)
2. Short chain hydroxy acid (lactic acid)
3. Short chain dihydroxy acid (2,3-dihydroxybutanoic acid)
4. Aromatic dicarboxylic acid (succinic acid, ethylmalonic acid and adipic acid)
5. Phenolic acid (3-methoxy-4-hydroxycinnamic acid)
6. Inorganic compound (phosphoric acid)

Most of the metabolites (lactic acid, octadecanoic acid, palmitic acid, 2,3-dihydroxybutanoic acid, phosphoric acid, isocitric lactone and adipic acid) listed in Table 4.4 showed an overall decrease in mean concentration value over a menstrual cycle, whereas ethylmalonic acid increased in mean concentration value for the cycle, even though the increase was not that pronounced. Succinic acid showed a decrease between phase 1 and 2 and an increase between phase 2 and 3, whereas 3-methoxy-4-hydroxycinnamic acid showed an increase between phase 1 and 2 and a decrease between phase 2 and 3. Four of the metabolites listed (lactic acid, octadecanoic acid, palmitic acid and 3-methoxy-4-hydroxycinnamic acid) had values that were lower than the reported reference values, whereas one metabolite (octadecanoic acid) had mean concentration values that were larger than the reference values. The mean concentration values for succinic acid, ethylmalonic acid and adipic acid fell within the reference range.

Conclusion:

A discussion of the results presented in this section is given in Chapter 5, but it does not appear that the menstrual cycle has an effect on the organic acid metabolism.

Table 4.4: Summary Statistics of the ten variables listed in the log-scaled dataset with three point smoothing.

Metabolite	Phase I (mmol/mol creatinine)			Phase II (mmol/mol creatinine)			Phase III (mmol/mol creatinine)			Reference Range [16] (mmol/mol creat)	Biological significance
	Mean	Median	Standard deviation	Mean	Median	Standard deviation	Mean	Median	Standard deviation		
Lactic acid	5.6	5.5	3.7	4.0	3.7	2.1	3.4	2.6	2.2	13-46[b]	Component of cysteine, propanoate and pyruvate metabolism.
Octadecanoic acid	2.52	0.94	2.97	0.99	0.80	0.84	0.81	0.39	0.95	0.10 +/- 0.03[a]	A type of saturated fatty acid (found in vegetable fats and oils).
Palmitic acid	3.66	1.566	5.55	1.41	1.18	0.94	1.12	0.81	0.84	0.30 +/- 0.09[a]	Palmitic acid is the first fatty acid produced during fatty acid synthesis and from which longer fatty acids can be produced.
2,3-Dihydroxy-butanoic acid	4.9	3.7	3.0	4.1	3.9	1.7	3.1	2.7	1.3	25.0 +/- 10.0[a]	A normally occurring carboxylic acid in humans.

[16] Reference ranges (mmol/mol) are obtained from either Blau *et al.*, (2005) or the Human Metabolome Database (www.hmdb.com)

Metabolite	Phase I (mmol/mol creatinine)			Phase II (mmol/mol creatinine)			Phase III (mmol/mol creatinine)			Reference Range (mmol/mol creat)	Biological significance
	Mean	Median	Standard deviation	Mean	Median	Standard deviation	Mean	Median	Standard deviation		
2,3-Dihydroxy-butanoic acid	4.9	3.7	3.0	4.1	3.9	1.7	3.1	2.7	1.3	25.0 +/- 10.0[a]	A naturally occurring carboxylic acid in humans.
Phosphoric acid	20.43	7.37	31.14	16.87	3.77	23.52	10.41	4.26	14.33	n.d.	It is an osmolyte, enzyme cofactor and is involved with signalling.
Succinic acid	17.60	11.86	14.40	12.19	8.047	7.68	15.34	9.18	15.91	0.5-16[b]	It is a component of arginine, proline, butanoate, and glutamate and propanoate metabolism.
Ethylmalonic acid	1.01	0.74	0.81	1.20	1.12	0.53	1.31	0.85	0.92	0.4-4.2[b]	Ethylmalonic acid is identified in the urine of patients with SCAD (a disorder of the fatty acid metabolism).

Metabolite	Phase I (mmol/mol creatinine)			Phase II (mmol/mol creatinine)			Phase III (mmol/mol creatinine)			Reference Range (mmol/mol creat)	Biological significance
	Mean	Median	Standard deviation	Mean	Median	Standard deviation	Mean	Median	Standard deviation		
Isocitric lactone	1.12	0.89	1.08	0.89	0.79	0.46	0.68	0.56	0.41	n.d.	It is an isocitric acid without an H_2O group. It is a component of the glutathione metabolism.
3-Methoxy-4-Hydroxy-cinnamic acid	0.1	0	0.16	0.12	0.11	0.12	0.05	0	0.07	36.3 +/- 11.2[a]	Trans-ferulic acid is a highly abundant phenolic phytochemical which is present in plant cell walls.
Adipic acid	2.13	1.63	1.57	1.83	1.83	0.85	1.53	1.27	0.74	0.8-35[b]	Adipic acid is derived from the oxidation of various fats.

[a] Reference range obtained from the Human Metabolome Database (www.hmdb.com).

[b] Reference range obtained from Blau *et al.,* (2005: 32-38) according to the minimum to maximum values in mmol/mol creatinine.

n.d.: not determined according to public database[a] and tables published in the relevant literature[b].

# Note that the median value for one of the groups is much lower than their mean concentration, meaning that one or more cases had a high mean concentration compared to the other cases.

4.3 Pregnancy

4.3.1 Determination of the organic acid profile for the control and pregnant samples

Pregnancy as a perturbation on the metabolic profile was studied by comparing the organic acid profile in urine samples from twenty-nine cases which were divided into four groups namely a control group i.e. non-pregnant women and pregnant women in their first, second and third trimester. Time as an influence on pregnancy could not be analysed since we did not obtain samples from the same individuals over the pregnancy period; hence the study was not a longitudinal study. The study groups consisted of the following:

1. 11 controls,
2. 4 subjects in trimester one,
3. 7 subjects in trimester two,
4. 5 subjects in trimester three.

The different subjects in each trimester are all independent except for one subject for which one sample was obtained in each trimester. However, due to the limited number of subjects, the samples from this subject were retained in the statistical analysis. Subsequently, we compared the control group with each of the three pregnant groups separately. It should be noted that the number of study subjects are extremely small. For this reason no generalisation of the findings are made, hence the current study should be viewed as only exploratory. The variable reduction as well as the statistical methods applied to the dataset are schematically presented in the flowchart (see Figure 4.7).

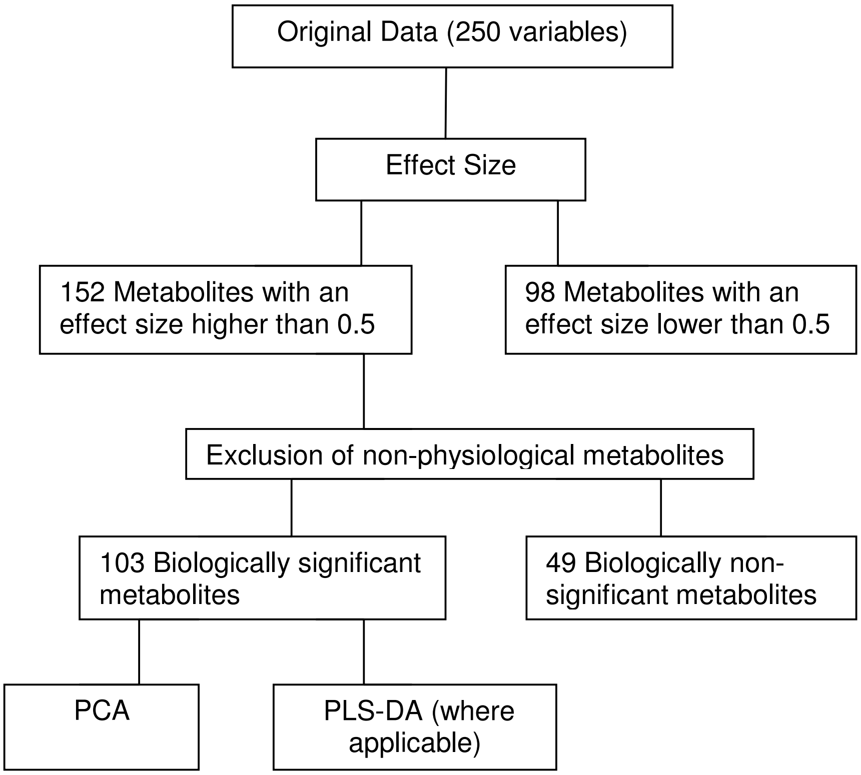

Original Data (250 variables)
Effect Size
152 Metabolites with an effect size higher than 0.5
98 Metabolites with an effect size lower than 0.5
Exclusion of non-physiological metabolites
103 Biologically significant metabolites
49 Biologically non-significant metabolites
PCA
PLS-DA (where applicable)

Figure 4.7: Flowchart of statistical analysis for pregnancy.

The GC-MS data are presented in Figure 4.8 (log-scaled data) and show the typical metabolomics profile of variables with varying concentrations. This figure also enables a visual comparison of the four groups. It does appear that the groups compare visually with one another according to small and big concentration peaks. The metabolites visualised in Figure 4.8 and Figure 4.9 were ranked according to the names (x2-hydroxybutyric acid) given by the statistical computer programme (S-Plus) as well as some alphabetical ranking. This ranking does not compare with the retention reference value found in the original GC-data and does not influence the multivariate analysis.

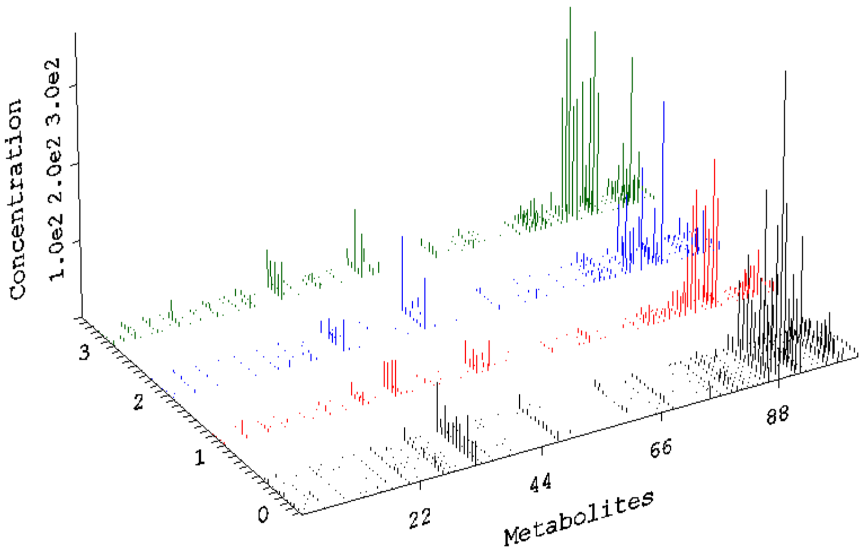


Figure 4.8: The GC-MS profiles of the control group (black); pregnant women in their first trimester (red); pregnant women in their second trimester (blue) and pregnant women in their third trimester (green).

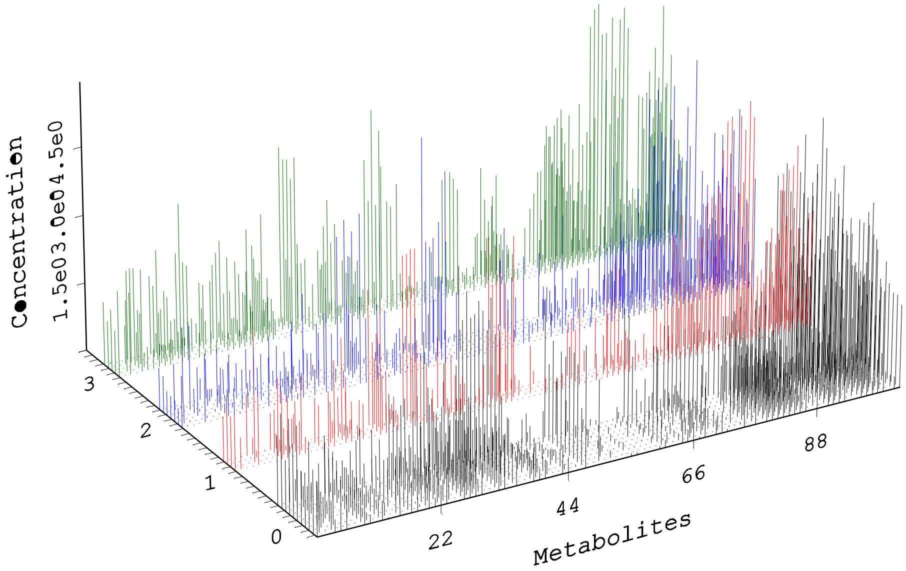


Figure 4.9 Log-scaled data of the GC-MS profiles of the control group (black); pregnant women in their first trimester (red); pregnant women in their second trimester (blue) and pregnant women in their third trimester (green).

The data were log-scaled and centered prior to multivariate statistical analysis. A PCA (see 3.7.6.1) was performed on the log data as an unsupervised pattern recognition method. The

PCA scores plot of PC1 and PC2 (see Figure 4.10) gives an indication of the separation between the control group and the three pregnant groups that were investigated i.e. trimester one, two and three.

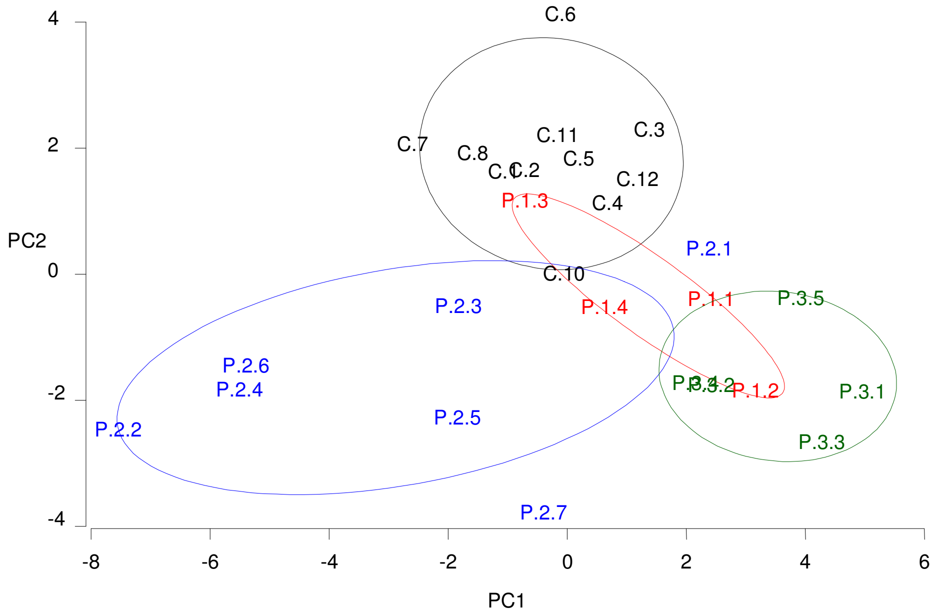


Figure 4.10: A PCA scores plot to compare the control group (C - black) with pregnant women in their first trimester (P1 - red), second trimester (P2 - blue) and third trimester (P3 - green).

The first three principal components explained 56.9% of the variation in the data. From the PCA scores plot we conclude that the control group separates completely from the pregnant women in their second and third trimester and that there is an overlap of the first trimester group with the control group. For the PCA and PLS-DA analyses to follow, three and two principal components (see 3.7.6) were extracted for calculation of the VIP values respectively. Subsequently, the following group comparisons are investigated:

- Control group vs. pregnant women in their first trimester.
- Control group vs. pregnant women in their second trimester.
- Control group vs. pregnant women in their third trimester.

The variance explained by the components produced in the multivariate methods are presented in Table 4.5. Note that PLS-DA analysis is aimed at finding a discrimination model to predict group membership. Therefore this method should be applied with caution when the PCA scores plots indicate an overlap between groups, since a PLS-DA model regularly overfit the data in this case. The PLS-DA model was not validated and was only used as an explorative statistical method.

Table 4.5: Percentage of variation explained.

Comparison	PCA	PLS-DA
Control group vs. pregnant women in their thirst trimester	56.1%	Not done[17]
Control group vs. pregnant women in their second trimester	61.9%	89.2%
Control group vs. pregnant women in their third trimester	60.8%	96.3%

Subsequently, the results of PCA and PLS-DA (where applicable) are presented and discussed. First, the scores plots of the PCA comparisons are presented in Figure 4.11, Figure 4.12 and Figure 4.14.

When the control group was compared to pregnant women in their first trimester there was no definite separation between these two groups (Figure 4.11). Therefore the identification of important variables via PCA and PLS-DA is not performed for this comparison. There is a clear separation between the control group and the pregnant women in their second trimester as well as the control group and pregnant women in their third trimester (Figure 4.12 and Figure 4.14). The PLS-DA scores plots for the pregnant women in their second and third trimester are shown in Figure 4.13 and 4.15 respectively. Tables 4.6 and 4.7 contain the ten most important variables identified via PCA and PLS-DA for these two comparisons. The second and third columns of tables 4.6 and 4.7 show the rankings (one being top-ranked) of the variables according to the VIP values obtained from a PCA and PLS-DA respectively. In addition, descriptive statistics, reference ranges, and comments are presented in Tables 4.6 and 4.7. Certain metabolites were excluded and are not listed in Tables 4.6 and 4.7, even if they had a high VIP value. The reasons for exclusion were the following: individual variation caused by diet (3-methylphenol), it is a by-product of the extraction and/or derivitization method (phosphoric acid) or if the origin and physiological role of the metabolite is not known (3,4-

[17] The PLS-DA was not done since PCA did not show separation between the groups.

dihydroxyfuranone[18]). Subsequently these metabolites were excluded since they may not be responsible for the variation investigated in the dataset i.e. metabolome from a biological point of view. As a result metabolites were only included if the physiological function was known. The metabolites that were excluded are listed in Appendix I along with the reason for exclusion. The VIP ranking presented in Tables 4.6 and 4.7 were changed accordingly.

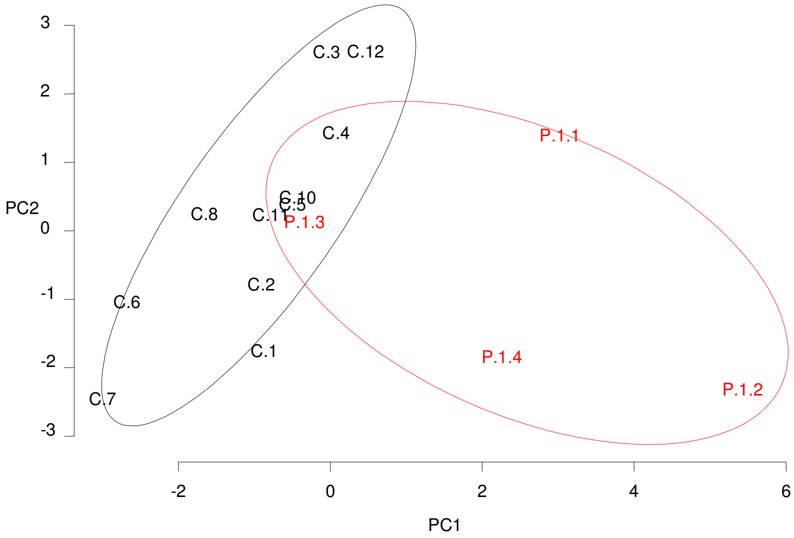


Figure 4.11: PCA scores plot to compare control group (black) with pregnant women in their first trimester (red).

[18] Note added in proof: 3,4-dihydroxyfuranone was not excluded by the application of the biological filter due to an individual oversight. This substance does not occur in either Blau *et al.* (2005) or the Human Metabolome Database (www.hmdb.com).

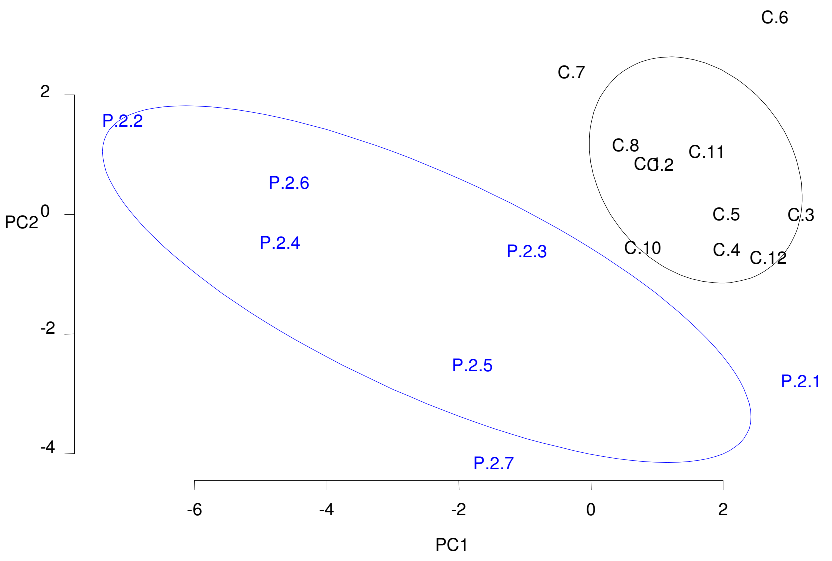


Figure 4.12: PCA scores plot to compare control group (black) with pregnant women in their second trimester (blue).

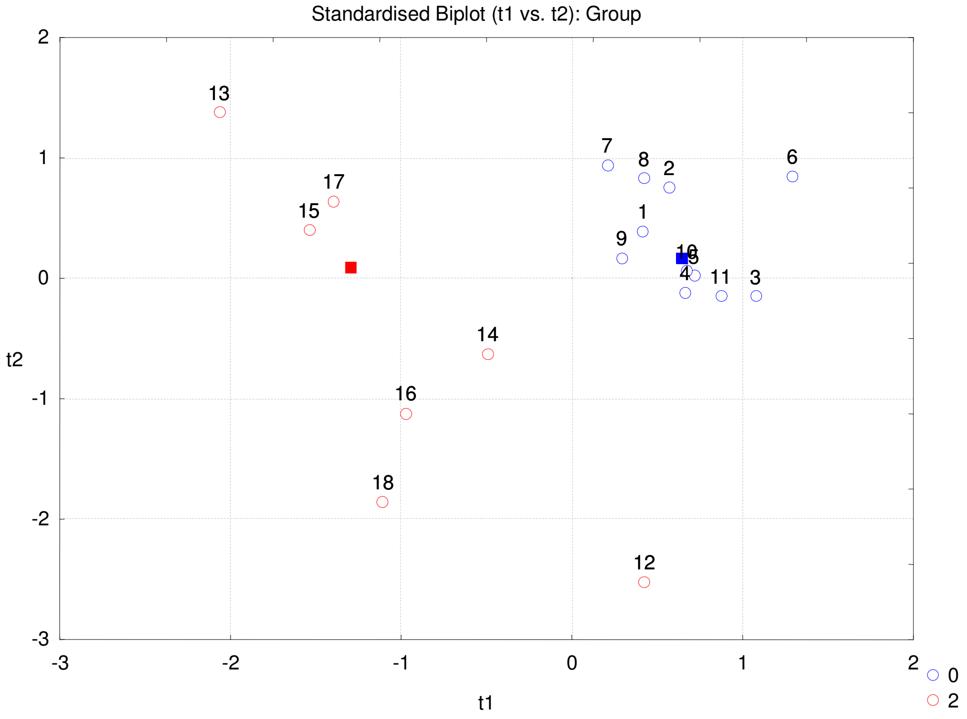


Figure 4.13: PLS scores plot to compare control group (blue) with pregnant women in their second trimester (red).

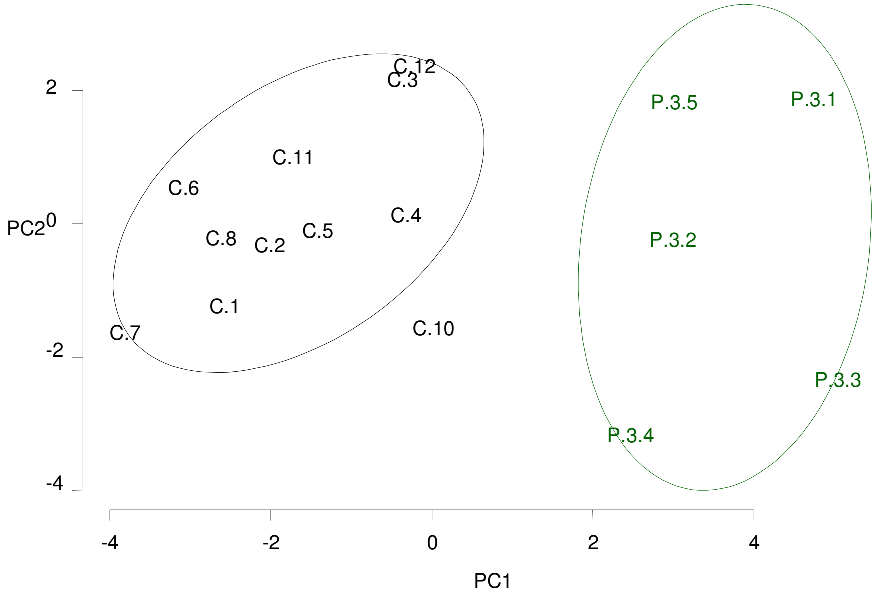


Figure 4.14: PCA scores plot to compare control group (black) with pregnant women in their third trimester (green).

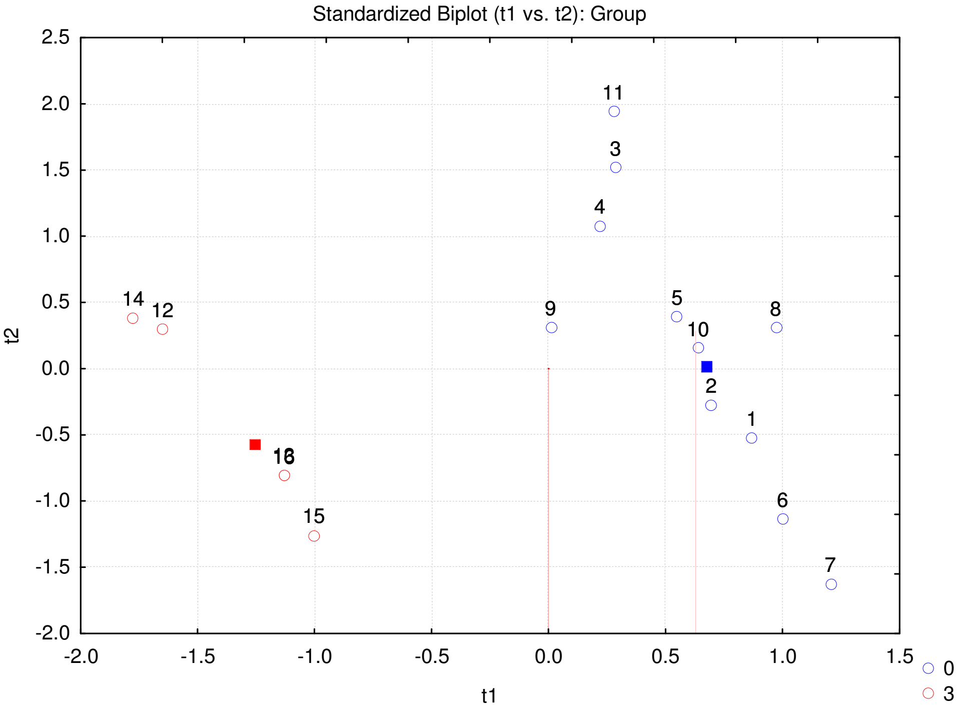


Figure 4.15: PLS-DA scores plot to compare control group (blue) with pregnant women in their third trimester (red).

Table 4.6: Comparison of the control group with pregnant woman in their second trimester.

Variables **(VIP – PCA)**	**VIP Ranking**		**Control Group**			**Second Trimester**			**Reference Range**[19]	**Comment, according to the mean concentration values of the groups in comparison with the reference values**
	PCA	**PLS**	**Mean (mmol/ mol creat)**	**Median (mmol/ mol creat)**	**Standard deviation (mmol/ mol creat)**	**Mean (mmol/ mol creat)**	**Median (mmol/ mol creat)**	**Standard deviation (mmol/ mol creat)**	**Adults**	
Pyroglutamic acid	1	8	5.1	4.6	1.7	3.1	1.2	4.6	3.4-54.2[a]	-
Hippuric acid	2	1	135.2	102.6	95.3	50.9	20.2	75.8	170-390[b]	#
Vanillylmandelic acid	3		5.4	5.5	1.4	4.0	3.6	3.5	2.5-16.1[b]	-
Succinic acid	4		15.9	16.5	7.2	20.8	18.2	17.9	0.5-16[b]	-
3-Hydroxyisobutyric acid	5		7.2	6.1	3.4	9.1	7.9	7.9	4.1-19[a]	-
4-Hydroxyhippuric acid	6	3	7.1	5.1	5.4	1.9	0.7	3.2	n.d.	-
3-Hydroxyisovaleric acid	7	10	3.9	3.5	1.9	1.5	1.6	1.2	6.9-25[b]	-
3-Hydroxy-3-methylglutaric acid	8	4	1.5	1.5	0.4	0	0	0	0.7-3.0[a]	-
Aconitic acid	9	2	35.7	30.6	13.4	17.5	16.3	15.0	2.7-44[b]	-
Malic acid	10	5	0.2	0	0.7	2.3	0	3.5	0.7-5.3[b]	-

[19] Reference ranges (mmol/mol) are obtained from either Blau *et al.*, (2005) or the Human Metabolome Database (www.hmdb.com)

Variables (VIP – PCA)	VIP Ranking		Control Group			Second Trimester			Reference Range	Comment, according to the mean concentration values of the groups in comparison with the reference values
	PCA	PLS	Mean (mmol/ mol creat)	Median (mmol/ mol creat)	Standard deviation (mmol/ mol creat)	Mean (mmol/ mol creat)	Median (mmol/ mol creat)	Standard deviation (mmol/ mol creat)	Adults	
Lactic acid		6	8.3	7.4	4.8	28.1	9.1	38.3	13-46[b]	#
Citric acid		7	85.0	82.3	48.3	57.1	40.3	47.2	70-226[b]	#
2-Hydroxyisobutyric acid		9	3.2	2.9	1.5	0.9	0.5	0.7	n.d.	The reference value is not determined for adults, but the reference value for adolescents is between 2.9-10.3 mmol/mol[b].

[a] Reference range obtained from the Human Metabolome Database (www.hmdb.com).

[b] Reference range obtained from Blau *et al.,* (2005: 32-38) according to the minimum to maximum values in mmol/mol creatinine.

n.d.: not determined according to public databases[a] and tables published in the relevant literature[b].

# Note that the median value for one of the groups is much lower than their mean concentration, meaning that one or more cases had a high mean concentration compared to the other cases.

Table 4.7: Comparison of the control group with pregnant women in their third trimester.

Variables **(VIP – PCA)**	**VIP Ranking**		**Control Group**			**Third Trimester**			**Reference Range**[20]	**Comment, according to the mean concentration values of the groups in comparison with the reference values**
	PCA	**PLS-DA**	**Mean (mmol/mol creat)**	**Median (mmol/mol creat)**	**Standard deviation (mmol/mol creat)**	**Mean (mmol/mol creat)**	**Median (mmol/mol creat)**	**Standard deviation (mmol/mol creat)**	**Adults**	
Fumaric acid	1	3	0.5	0.4	0.3	6.9	6.4	4.3	0.2-0.8^{b}	-
3-Hydroxypropionic acid	2		1.6	1.5	0.9	5.1	4.9	1.8	3-10^{a}	-
Uracil	3		0.4	0.3	0.2	1.0	0.9	0.4	2.37-3.5^{a}	-
2-Ketoglutaric acid	4	2	0.3	0.1	0.6	9.8	7.2	10.2	0.42-18.3^{a}	-
Succinic acid	5	4	15.9	16.5	7.2	78.7	65.6	63.3	0.5-16^{b}	#
2-Hydroxyglutaric acid	6		1.3	1.4	0.4	5.1	4.2	3.1	0.8-52^{b}	-
3-Hydroxyisobutyric acid	7		7.2	6.1	3.4	17.8	19.2	6.0	4.1-19^{a}	-
Ethylmalonic acid	8		0.6	0.4	0.8	1.7	1.6	1.6	0.4-4.2^{b}	-

[20] Reference ranges (mmol/mol) are obtained from either Blau *et al.*, (2005) or the Human Metabolome Database (www.hmdb.com)

Variables (VIP – PCA)	VIP Ranking		Control Group			Third Trimester			Reference Range	Comment, according to the mean concentration values of the groups in comparison with the reference values
	PCA	PLS-DA	Mean (mmol/mol creat)	Median (mmol/mol creat)	Standard deviation (mmol/mol creat)	Mean (mmol/mol creat)	Median (mmol/mol creat)	Standard deviation (mmol/mol creat)	Adults	
Quinolinic acid	9		0.3	0.3	0.2	1.1	1.0	0.2	2.5 +/- 1.1[a]	-
Citric acid	10		85.0	82.3	48.3	202.9	195.3	59.0	70-226[b]	#
3-Methylglutaconic acid		1	3.5	2.1	4.8	21.0	21.8	7.4	0.-9.0[a]	-
2-Ethyl-3-Hydroxypropionic acid		5	1.3	0.8	1.8	6.4	7.1	2.3	n.d.	-
Malic acid		6	0.2	0	0.7	5.0	5.8	4.8	0.7-5.3[b]	-
3-Methylglutaric acid		7	0.3	0	0.7	2.9	3.0	1.1	1.0-6.5[a]	-
Lactic acid		8	8.3	7.4	4.8	36.5	23.3	32.2	13-46[b]	#
Malonic acid		9	0.6	0	1.5	6.9	0.8	9.2	0-2[a]	-
Octadecanoic acid		10	3.71	2.21	3.60	9.03	10.73	5.04	0.10 +/- 0.03[a]	-

[a] Reference range obtained from the Human Metabolome Database (www.hmdb.com).

[b] Reference range obtained from Blau *et al.,* (2005: 32-38) according to the minimum to maximum values in mmol/mol creatinine.

n.d.: not determined according to public databases[a] and tables published in the relevant literature[b].

# Note that the median value for one of the groups is much lower than their mean concentration, meaning that one or more cases had a high mean concentration compared to the other cases.

Discussion of results:

- Discussion of Table 4.6:

  Nine of the variables showed a decrease in mean concentration value for the second trimester whereas five of the variables showed an increase in mean concentration after comparison with the control group. One variable, 3-hydroxy-3-methylglutaric acid showed a zero value in the second trimester group. Seven variables were common in the PCA and PLS-DA VIP lists, namely 4-hydroxyhippuric acid, hippuric acid, pyroglutamic acid, malic acid, 3-hydroxyisovaleric acid, aconitic acid, and 3-hydroxy-3-methylglutaric acid. The metabolites identified were grouped as follows:

  a. Short chain dicarboxylic acid (succinic acid, malic acid and 3-hydroxy-3-methylglutaric acid)
  b. Short chain hydroxy acid (3-hydroxyisovaleric acid, 3-hydroxyisobutyric acid, 2-hydroxyisobutyric acid, lactic acid)
  c. Short chain tricarboxylic acid (aconitic acid, citric acid)
  d. Amino acid conjugate (pyroglutamic acid)
  e. Catecholamine (vanillylmandelic acid)
  f. Glycine conjugate (hippuric acid)

Most of the variables are either part of an amino acid metabolic pathway or are TCA intermediates. It may be speculated that the metabolites excreted also indicate an increased strain put on the kidneys.

- Discussion of Table 4.7:

  All the variables listed in Table 4.7 showed an increase in the mean concentration value when the third trimester group was compared with the control group. Three common variables were identified by PCA and PLS-DA, i.e. fumaric acid, succinic acid and 2-ketoglutaric acid. The metabolites identified were grouped as follows:

  a. Long chain fatty acids (octadecanoic acid)
  b. Short chain dicarboxylic acid (fumaric acid, succinic acid, 2-hydroxyglutaric acid, malic acid, 3-methylglutaric acid, 3-methylglutaconic acid, malonic acid, 2-ketoglutaric acid, malonic acid, ethylmalonic acid)
  c. Short chain hydroxy acids (3-hydroxypropionic acid, 2-ethyl-3-hydroxypropionic acid, 3-hydroxyisobutyric acid, lactic acid)
  d. Short chain tricarboxylic acid (citric acid)
  e. Aromatic dicarboxylic acid (quinolinic acid)
  f. Pyrimidine derivative (uracil)

Most of the variables were either involved in the TCA cycle or the amino acid metabolism. A discussion of the results is given in Chapter 5.

4.4 Age

The extraction method used (see 3.5.1) to study the age perturbation was again aimed at isolating the organic acids. However additional metabolites with the same chemical properties were also identified. The dimension of the data generated was again reduced by using effect sizes (see 3.7.4.1) as well as the biological filter (see 4.1). Age as a perturbation on the metabolic profile was studied by comparing the organic acid profile in seventy-seven cases of urine samples which were divided into four age groups namely infants younger than a year (3 days – 7 months old), infants older than a year (1 – 2.5 years), children (11 – 13 years old) and young adults (20 – 27 years old). The variable reduction as well as the statistical methods applied to the reduced dataset is schematically presented in Figure 4.16.

It should be noted that all infants and child groups were of mixed ethnicity, while the adult group consisted only of Caucasian subjects. The effect of ethnicity on the normal metabolite profile was not an aim of the present study. For control purposes, we included a comparison on the metabolomics profiles of different ethnic groups which will be shown below (see Figures4.17 and 4.18). The child group was selected for this comparison given that it was the biggest group containing three ethnic groups i.e. 25 participants which were grouped as follows: 12 Coloured (Co/Co C), 8 Caucasian (Ca C) and 5 Black (B C/B L) children. A PCA was performed for this comparison on the log-scaled data using all 457 variables (Figure 4.17) and 114 selected variables (Figure 4.18).

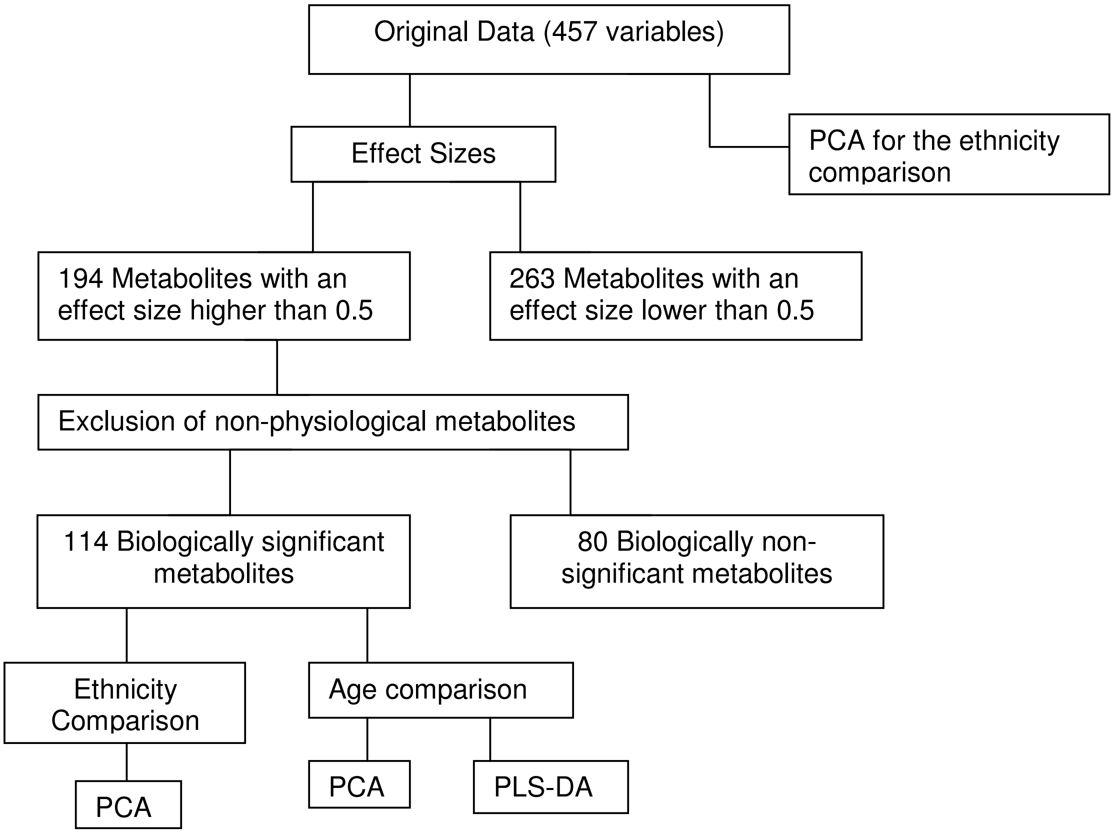

Original Data (457 variables)
Effect Sizes
PCA for the ethnicity comparison
194 Metabolites with an effect size higher than 0.5
263 Metabolites with an effect size lower than 0.5
Exclusion of non-physiological metabolites
114 Biologically significant metabolites
80 Biologically non-significant metabolites
Ethnicity Comparison
Age comparison
PCA
PCA
PLS-DA

Figure 4.16: Flowchart of statistical analysis for age as a perturbation.

The age group comparison was performed on a data matrix containing 114 variables and seventy-seven cases which consisted of eighteen infants, twenty-five children and thirty-two adults. The reader is referred to Section 3.2.3 for a complete profile of the study group and to Section 3.7 for a discussion of the various statistical methods used.

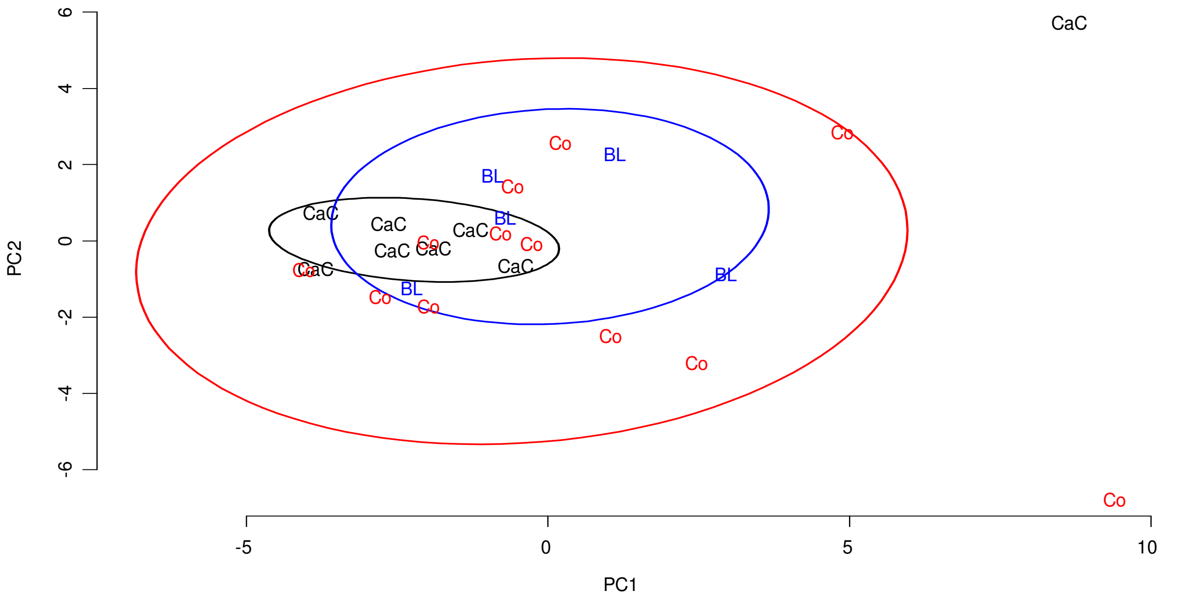


Figure 4.17: PCA scores plot of three different ethnic groups (Black – blue, Caucasian – black and Coloured – red) for the log-scaled data using all 457 variables.

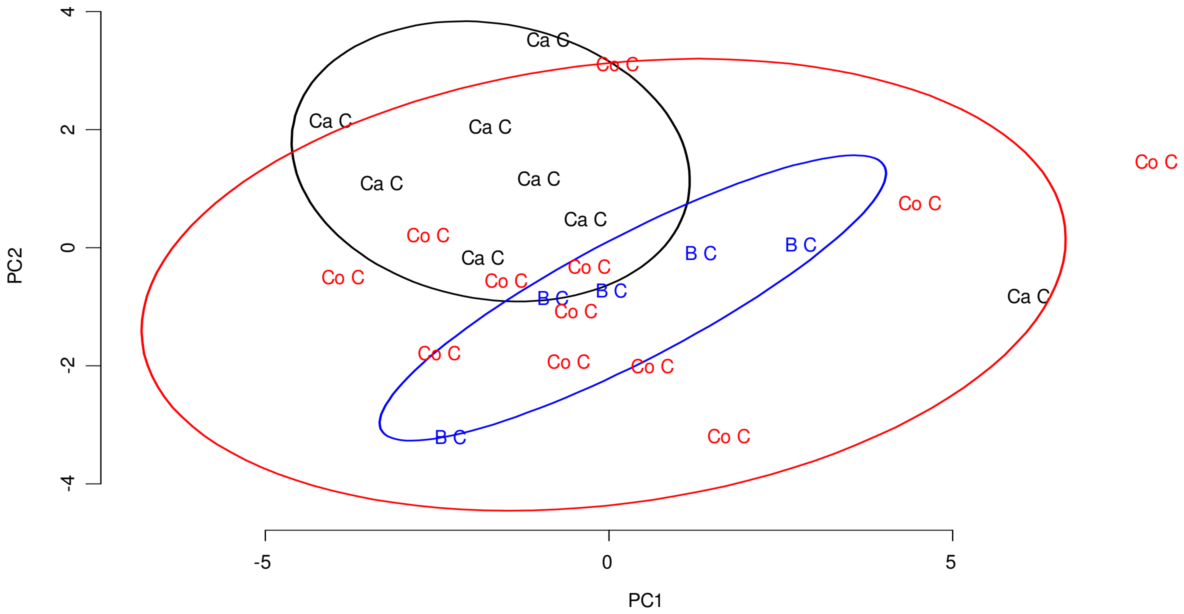


Figure 4.18: PCA scores plot of three different ethnic groups (Black - blue, Caucasian - black and Colored – red) based on the 114 selected log-scaled variables (see Figure 4.16).

The PCA scores plots for the ethnicity comparison show that there is an overlap in the three groups for both the original dataset and the reduced dataset. Hence ethnicity will not play a significant role in the results obtained for the age comparison. The results obtained should be seen as a pilot study, since there were only 25 participants. Consequently this study should be performed on a larger group, each ethnic group with the same male to female ratio, containing participants of the same age etc.

Subsequently, the discrimination of the four age groups based on the organic profile is investigated using the log-scaled data. PCA as well as PLS-DA are used in this regard. Although it was found (see Figure 4.21) that there is a natural discrimination between the four study groups, *separate* group comparisons were performed to identify possible important variables causing the *specific* group separation under investigation. The GC-MS data and the log-scaled GC-MS data are presented in Figures 4.19 and 4.20. The order of the metabolites in Figures 4.19 and 4.20 do not compare with the retention reference value found in the original GC-data and does not influence the multivariate analysis.

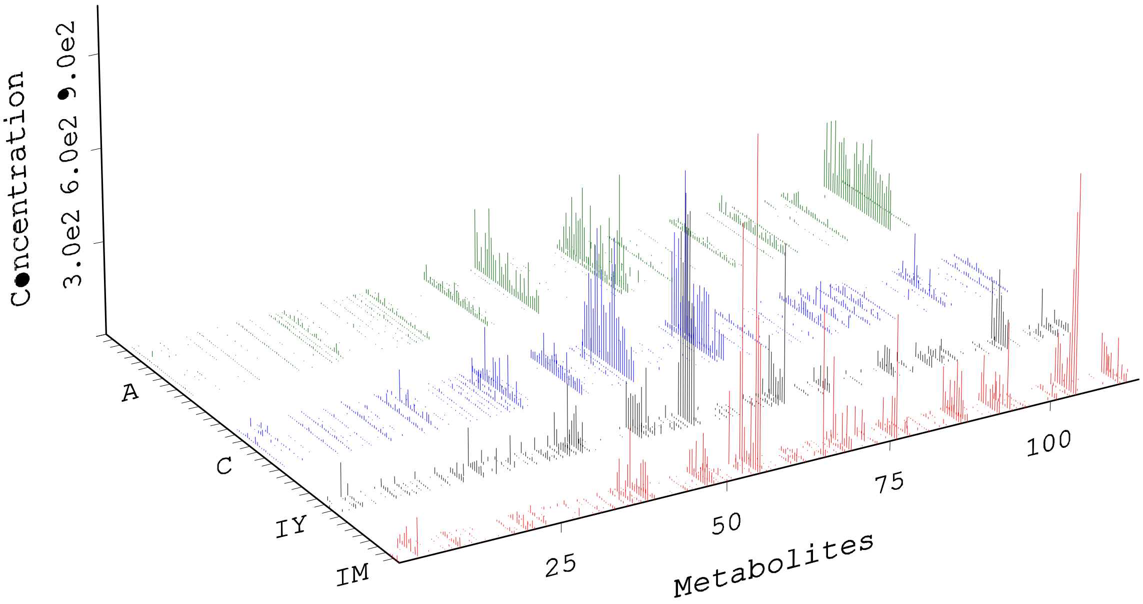


Figure 4.19: The GC-MS profiles of IM - infants younger than a year (red); IY - infants older than a year (black); children (blue); adults (green).

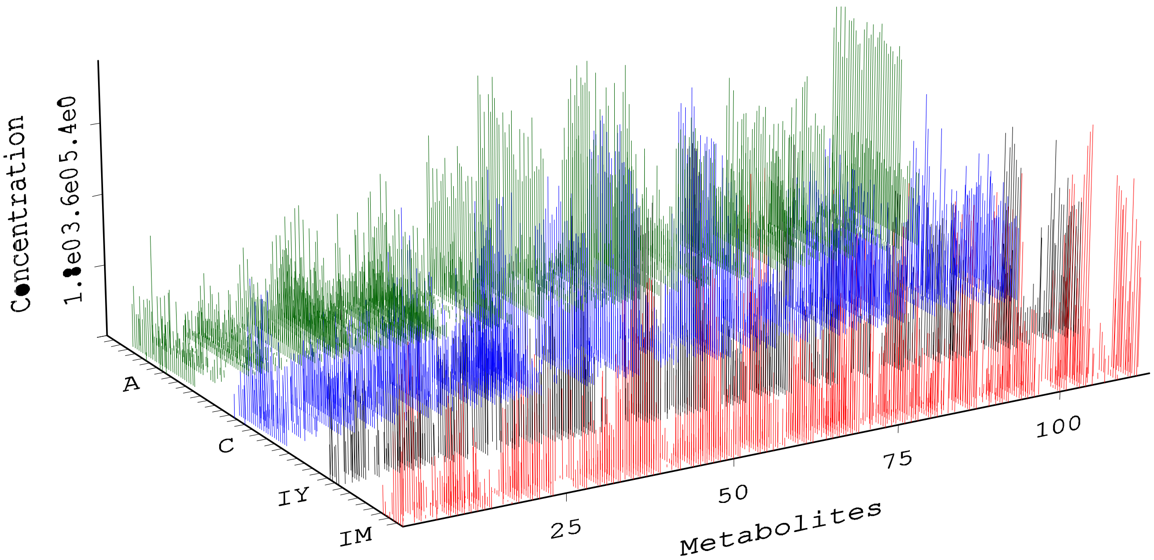


Figure 4.20: The log-scaled GC-MS profiles of IM - infants younger than a year (red); IY - infants older than a year (black); children (blue); adults (green).

Figure 4.21 illustrates the natural discrimination between the four age groups when compared simultaneously using PCA, as an unsupervised method of analysis.

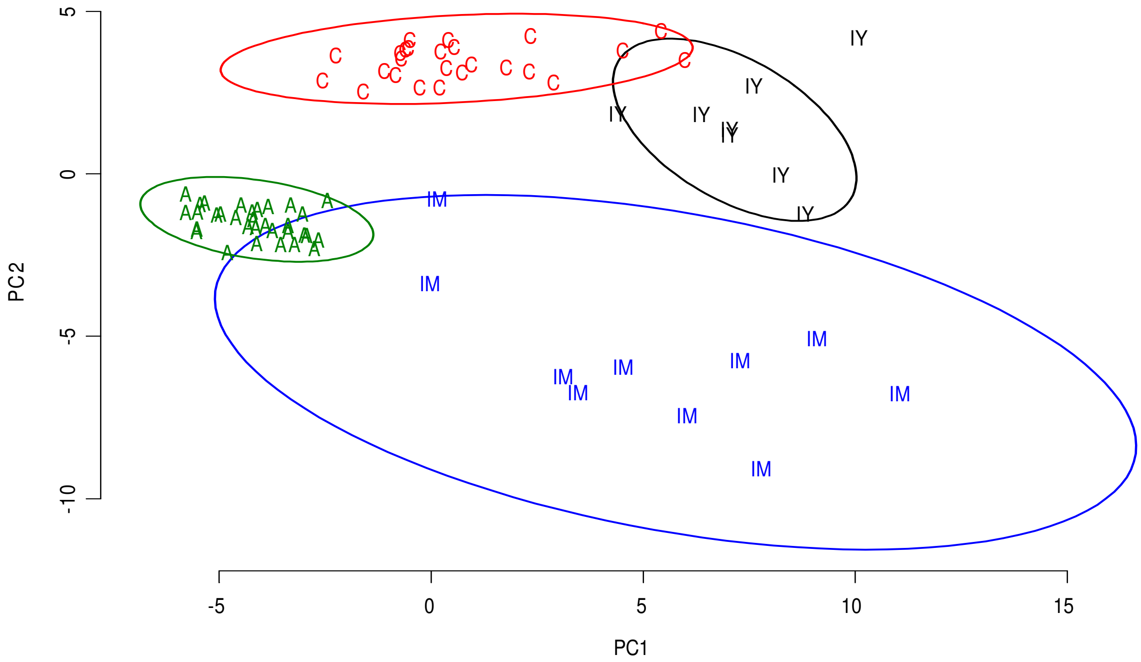


Figure 4.21: PCA scores plot of the infants (IM – blue and IY - black) vs. children (red) vs. adults (green) groups.

The following group comparisons were investigated:

- Infants younger than a year (IM) vs. infants older than a year (IY),
- Infant group (IM) vs. children,
- Infant group (IY) vs. children,
- Children vs. adults.

Throughout, three and two components were respectively extracted for calculation of the VIP values (see Section 3.7.6) with PCA and PLS-DA. The percentage of variation explained by the components extracted for the PCA and PLS-DA are presented in Table 4.8.

Table 4.8: Percentage of variation explained.

Comparison	**PCA**	**PLS-DA**
Infant Comparison (IY vs. IM)	55.5%	97.8%
Infants vs. Children (IM vs. C)	60.2%	95.4%
Infants vs. Children (IY vs. C)	61.1%	90.2%
Children vs. Adults (C vs. A)	60.8%	98.4%

The PCA and PLS-DA scores plots of the infant comparison (IY vs. IM); infant (IM) compared to the child group, infants (IY) compared to the child group, and the child group compared to the adult group are displayed in Figures 4.22 to 4.29.

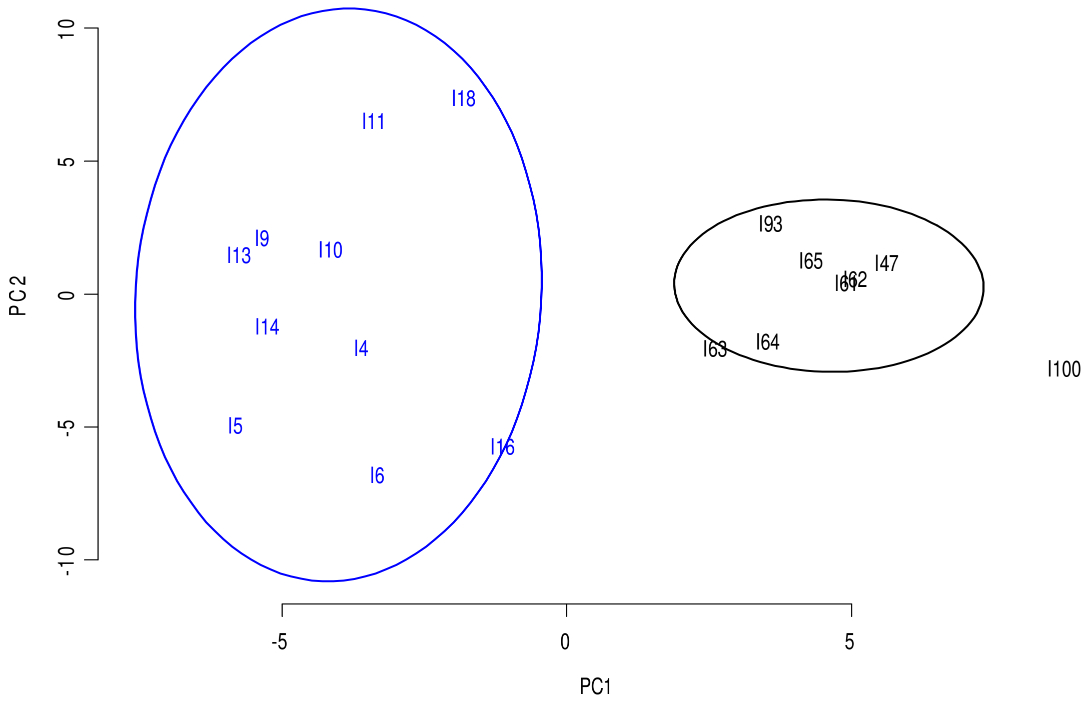


Figure 4.22: PCA scores plot of the two infant groups with IY (black) and IM (blue).

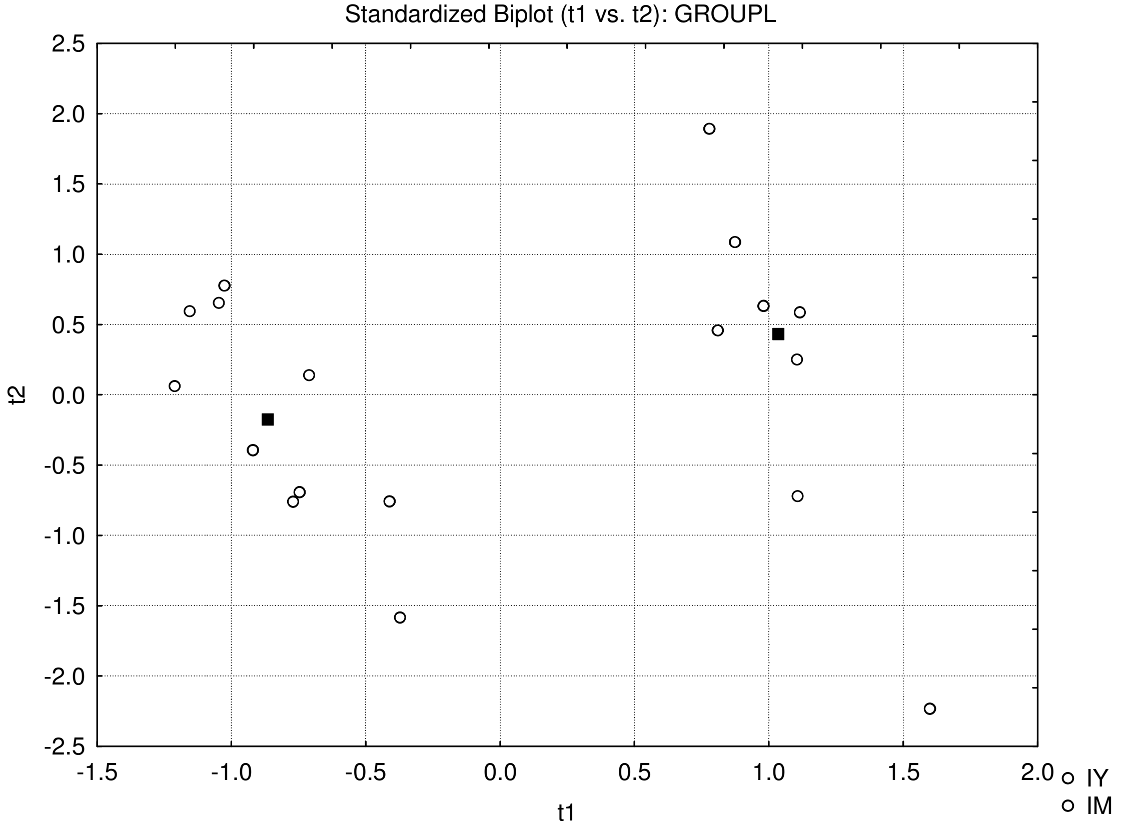


Figure 4.23: PLS-DA scores plot of the two infant groups with IY (blue) and IM (red).

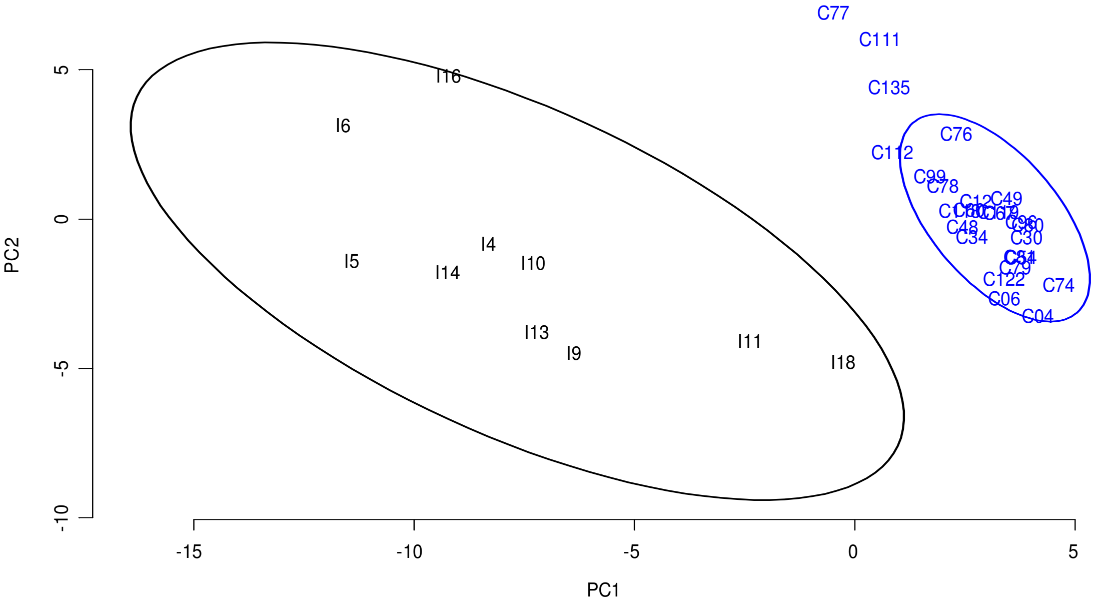


Figure 4.24: PCA scores plot of IM (black) compared to the child group (blue).

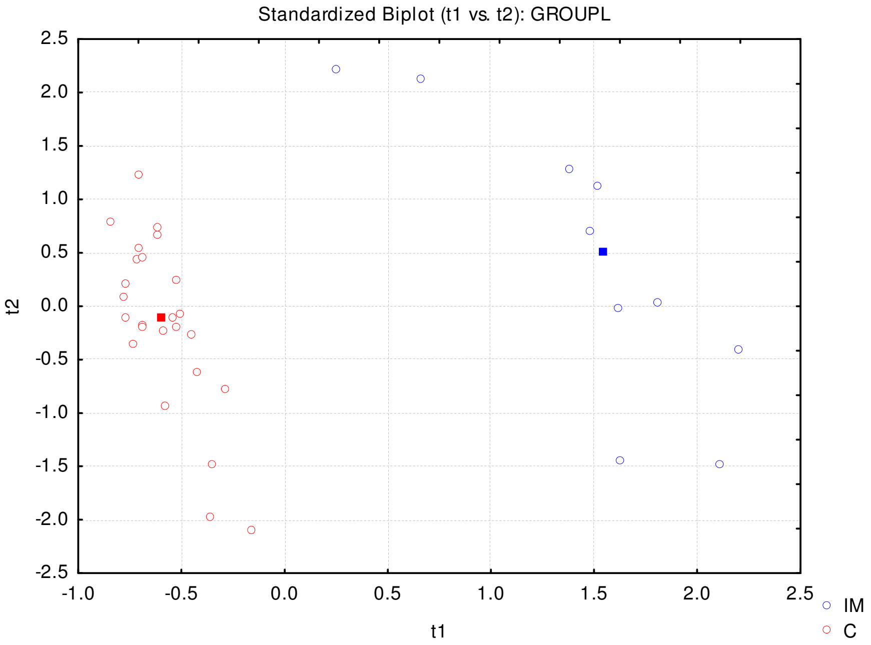


Figure 4.25: PLS-DA scores plot of IM (black) compared to the child group (blue).

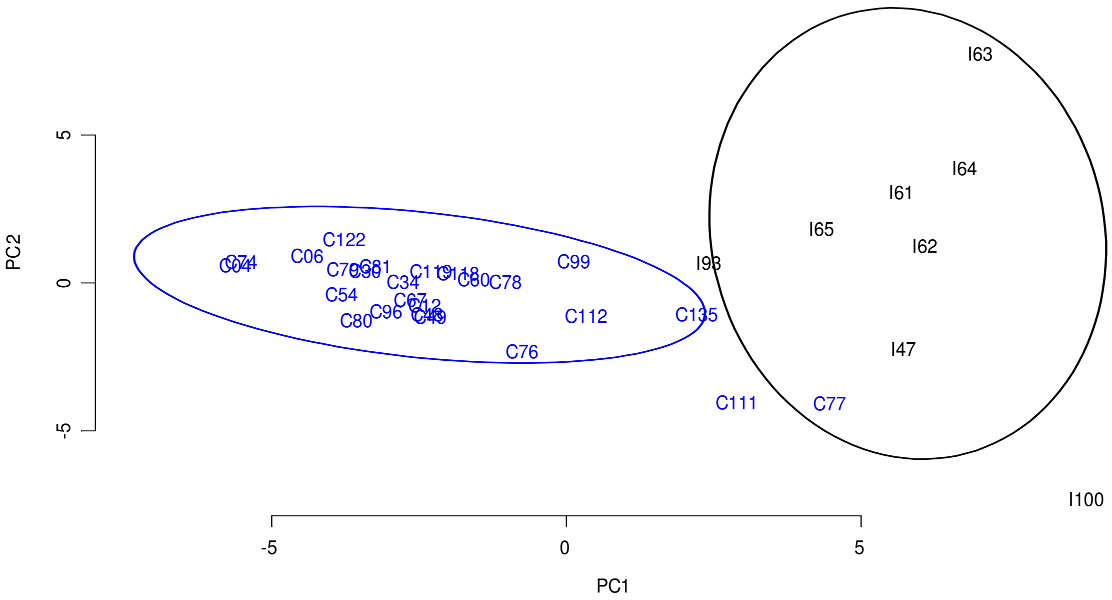


Figure 4.26: PCA scores plot of IY (black) compared to the child group (blue).

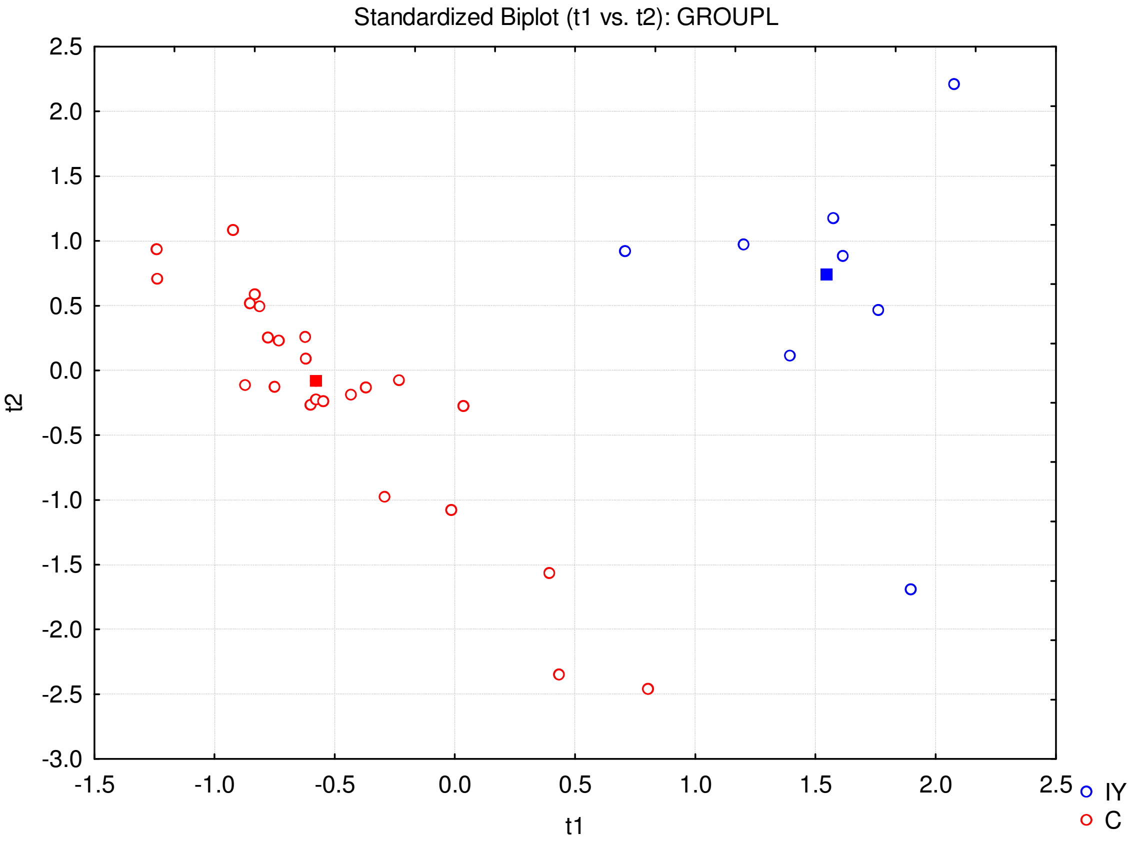


Figure 4.27: PLS-DA scores plot of IY (blue) compared to the child group (red).

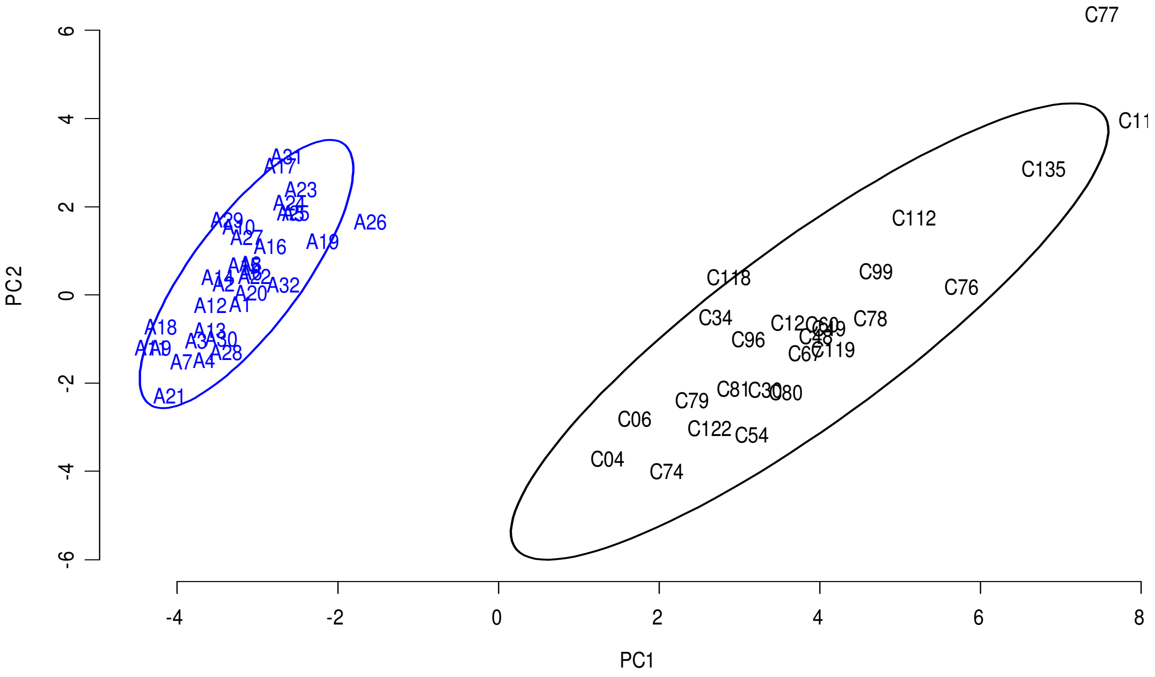


Figure 4.28: PCA scores plot of the child group (black) compared to the adult group (blue).

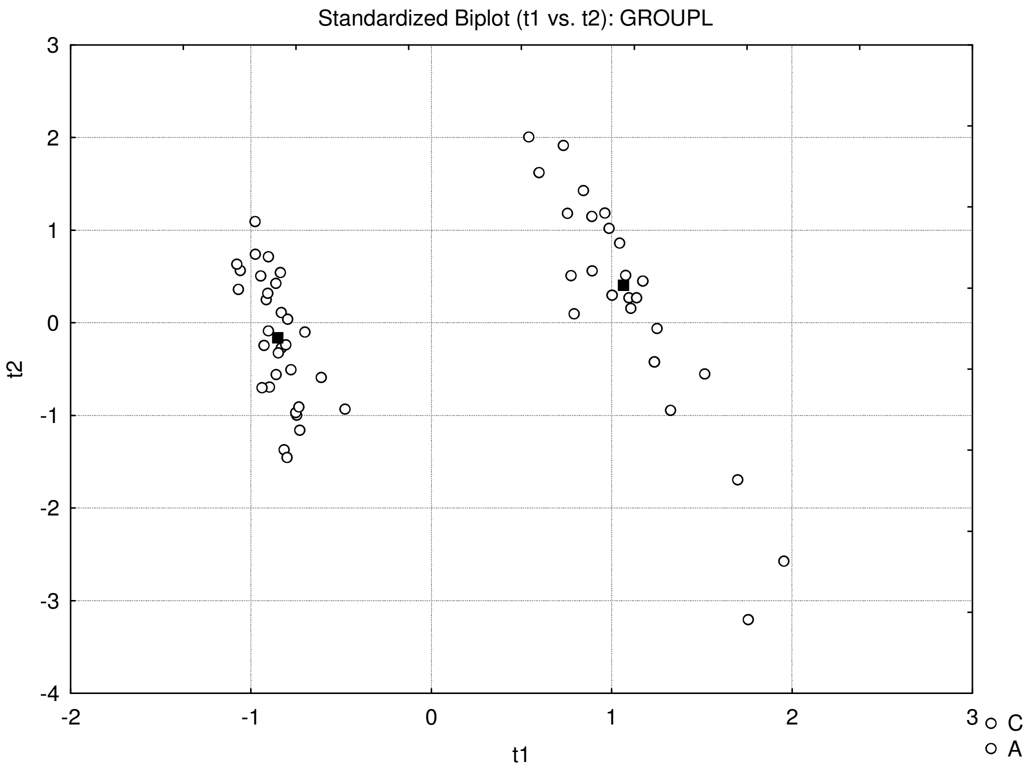


Figure 4.29: PLS-DA scores plot of the child group (blue) compared to the adult group (red).

It is evident from Figure 4.22 and Figure 4.23 that there is a distinct separation between the two infant groups (IY vs. IM). This separation is most probably age-related seeing that the group of Prof. I. Smuts (i.e., IY, the reader is referred to Section 3.2.3 for more detail) ranged from 1 to 2.5 years old, whereas the infants from the Laboratory for Inherited Metabolic Defects (IM) were younger than a year old. The reader is referred to Section 3.2.3 for more detail concerning the origin of the samples.

The infant group (IM) and the child group separated from one another (Figures 4.24 and 4.25). The child group had more samples (i.e. 25) than the infant group (i.e. 10) but both groups had more male than female samples, specifically nineteen for the child group and eight for the infant group. Figures 4.26 and 4.27 compare the infant group (IY), older than a year with the child group. The separation between these two groups was not as pronounced as was seen for the infant group (IM), younger than a year compared to the child group. Figures 4.28 and 4.29 demonstrate that there is a definite separation between the adult group and the child group.

From the group comparisons discussed above, we conclude that there exists a clear separation amongst groups of different ages. Hence, the identification of important variables in both the PCA and PLS-DA could lead to an explanation of the group separation based on the GC-MS profile of the various groups. To accomplish this, the VIP values (see Section 3.7.6) are calculated and ranked. Tables 4.8, 4.9 and 4.10 contain a consolidated list of the ten important variables (metabolites) in the PCA and PLS-DA projection (ranked according to the VIP values) as well as some descriptive statistics (calculated from the untransformed data), normal reference values and some additional comments. As previously mentioned for the pregnancy data certain metabolites were excluded and were not listed in Tables 4.9, 4.10, 4.11 and 4.12, even if they had a high VIP value. The metabolites that were excluded are listed in Appendix J along with the reason for exclusion. Note: Columns 2 and 3 in these tables give the ranking of the identified variables using PCA and PLS-DA respectively. These rankings indicate the importance of each variable in the multivariate analysis performed, with one being a top-ranked variable.

Discussion of results:

- Discussion of Table 4.9: Infant comparison (IM vs. IY)

  Five of the variables listed in Table 4.9 had zero values for one of the groups namely acetylaspartic acid, 2-methyl-3-hydroxybutyric acid, 3-methoxy-4-hydroxybenzoic acid, malic acid and uric acid. Five common variables were identified using PCA and PLS-DA, namely malic acid, acetylaspartic acid, 2-methyl-3-hydroxybutyric acid, 3-hydroxyisobutyric acid and 2-hydroxyisobutyric acid. Ten metabolites listed in 4.9 showed an increase in the mean concentration value, whereas four showed a decrease in mean concentration value. The metabolites identified were grouped as follow:

a. Long chain fatty acids (palmitic acid)
b. Short chain dicarboxylic acid (malic acid and adipic acid)
c. Short chain hydroxy acid (3-hydroxyisobutyric acid, 3-hydroxyvaleric acid, 2-methyl-3-hydroxybutyric acid and 2-hydroxyisobutyric acid)
d. Short chain dihydroxy acid (3,4-dihydroxybutyric acid)
e. Hydroxy purine (uric acid)
f. Amino acid conjugate (acetylaspartic acid and pyroglutamic acid)
g. Short chain tricarboxylic acid (aconitic acid)
h. Vanillylmandelic acid (catecholamine)
i. Phenolic acid (3-hydroxyphenylhydracrylic acid)

Most of the variables were either involved in amino acid metabolism or the TCA cycle.

- Discussion of Table 4.10: Infant (IM) vs. Child comparison

Two of the variables listed in Table 4.10 had zero values for one of the groups, namely malic acid and 3-hydroxyadipyllactone. Six common variables in Table 4.10 were identified, namely malic acid, vanillylmandelic acid, glutaric acid, fumaric acid, succinic acid and homovanillic acid. Most of the variables showed a decrease in mean concentration value when the child group was compared to the infant (IM) group. Twelve of the fourteen metabolites listed in Table 4.10 showed a decrease in mean concentration value, whereas two showed an increase. The metabolites identified were grouped as follow:

a. Long chain fatty acids (octadecenoic acid and palmitic acid)
b. Short chain dicarboxylic acid (fumaric acid, malic acid, 2-Hydroxyglutaric acid, glutaric acid, 3-hydroxy-3-methylglutaric acid, succinic acid and adipic acid)
c. Short chain hydroxy acid (lactic acid)
d. Catecholamine (vanillylmandelic acid and homovanillic acid)
e. Amino acid conjugate (acetylaspartic acid)
f. Phenolic acid (3-hydroxyphenylhydracrylic acid)
g. (3-hydroxyadipyllactone)

Most of the variables were either involved in amino acid metabolism or the TCA cycle.

Table 4.9: Comparison of infant groups (IY vs. IM).

Variables	VIP ranking		Group 1 (IM)			Group 2 (IY)			Reference Range[21]		Comment/Note
	PCA	PLS-DA	Mean	Median	Standard deviation	Mean	Median	Standard deviation	Infants (younger than one year)	Infants (older than one year)	
Acetylaspartic acid	1	6	4.68	3.65	3.32	0.00	0.00	0.00	5-34[b]	7-40.8[b]	-
Malic acid	2	2	11.50	7.31	11.80	0.00	0.00	0.00	5-38[b]	2.2-16.2[b]	#
2-Hydroxyisobutyric acid	3	7	1.24	0.24	1.71	8.72	7.41	6.54	0.1-8.2[a]	3.7-19.5[a]	-
2-Methyl-3-Hydroxybutyric acid	4	4	0.00	0.00	0.00	6.52	4.25	7.13	<2-7.5[b]	3.2-26.6[b]	-
3-Hydroxyisobutyric acid	5	8	9.38	7.19	8.15	36.91	26.65	35.08	<5-38[b]	20.2-118[b]	-
3-Hydroxyvaleric acid	6	10	17.24	7.76	20.17	34.11	24.72	36.91	n.d.	n.d.	# Reference values are unknown, but the reference value given for adults lies between 0 and 2 mmol/mol creatinine.

[21] Reference ranges (mmol/mol) are obtained from either Blau *et al.*, (2005) or the Human Metabolome Database (www.hmdb.com)

Variables	VIP ranking		Group 1 (IM)			Group 2 (IY)			Reference Range		Comment/Note
	PCA	PLS-DA	Mean	Median	Standard deviation	Mean	Median	Standard deviation	Infants (younger than one year)	Infants (older than one year)	
Pyroglutamic acid	7		4.22	3.71	2.72	14.16	10.63	10.14	3.4-54.2[a]	n.d.	-
Aconitic acid	8		65.48	60.82	48.41	122.70	117.05	22.64	10-54[b]	26.87-189[b]	-
3,4-Dihydroxybutyric acid	9		5.66	5.50	3.55	16.74	14.26	11.59	14-142[b]	109-454[b]	-
Adipic acid	10		20.71	8.45	29.61	12.59	11.57	7.70	<2.8-32[b]	<5.9-34.3[b]	-
Vanillylmandelic Acid		1	1.02	0.00	2.75	21.88	22.83	12.64	<2.92-14[b]	0.7-17[b]	-
Uric Acid		3	0.00	0.00	0.00	12.34	11.92	12.66	118.09 +/- 114.57[a]	n.d.	
Palmitic Acid		5	18.38	13.42	16.88	3.73	1.20	6.53	6-26.1[a]	0.8-8.2[a]	-
3-Hydroxyphenylhydracrylic acid		9	1.05	0.00	2.96	17.18	3.40	29.85	n.d.	n.d.	-

[a] Reference range obtained from the Human Metabolome Database (www.hmdb.com).

[b] Reference range obtained from Blau *et al.,* (2005: 32-38) according to the minimum to maximum values in mmol/mol creatinine.

n.d.: not determined according to public databases[a] and tables published in the relevant literature[b].

# Note that the median value for one of the groups is much lower than their mean concentration, meaning that one or more cases had a high mean concentration compared to the other cases.

Table 4.10: Comparison of infant group (IM) with child group.

Variables	VIP Ranking		Infants (IM)			Children (C)			Reference Range[22]		Comment/Note
	PCA	PLS-DA	Mean	Median	Standard deviation	Mean	Median	Standard deviation	Infants	Children	
Acetylaspartic Acid	1		4.68	3.65	3.32	0.06	0.00	0.15	5-34^{b}	6-21.6^{b}	-
Malic Acid	2	2	11.50	7.31	11.80	0.00	0.00	0.00	5-38^{b}	<2.3.-5.5^{b}	-
3-Hydroxyadipyllactone	3		7.15	6.87	7.24	0.00	0.00	0.00	n.d.	n.d.	-
Vanillylmandelic Acid	4	3	1.02	0.00	2.75	10.20	8.28	5.43	<2.92-14^{b}	1-15^{b}	-
Glutaric Acid	5	6	53.55	7.16	144.05	0.70	0.00	1.34	<0.3-3^{b}	<0.6-3.8^{b}	#
Octadecenoic Acid	6		0.81	0.74	0.71	0.02	0.00	0.07	n.d.	n.d.	-
Fumaric Acid	7	4	13.50	11.67	12.99	0.37	0.30	0.49	1-14^{b}	<1.5-3.7^{b}	-
Succinic acid	8	1	248.22	145.87	230.83	25.69	13.50	37.00	13-125^{b}	4.9-81.3^{b}	#
Homovanillic Acid	9	8	78.00	72.53	46.91	10.87	8.33	9.61	2.5-18.5^{b}	0.7-10.3^{b}	-
Adipic acid	10		20.71	8.45	29.61	1.28	1.29	0.83	<2.8-32^{b}	<1.1-5.3^{b}	#
3-Hydroxyphenylhydracrylic acid		5	1.05	0.00	2.96	15.92	6.38	22.67	n.d.	n.d.	#
Palmitic Acid		7	18.38	13.42	16.88	2.10	0.94	4.54	6.0-26.1^{a}	0.8-8.2^{a}	-
Lactic acid		9	90.44	39.84	117.74	8.41	5.22	6.04	0.5-156^{b}	35-131^{b}	#

[22] Reference ranges (mmol/mol) are obtained from either Blau *et al.*, (2005) or the Human Metabolome Database (www.hmdb.com)

Variables	VIP Ranking		Infants (IM)			Children (C)			Reference Range		Comment/Note
	PCA	PLS-DA	Mean	Median	Standard deviation	Mean	Median	Standard deviation	Infants	Children	
3-Hydroxy-3-Methylglutaric acid		10	15.41	13.70	13.03	1.57	0.00	2.81	15-43b	<10.3-28[b]	-

[a] Reference range obtained from the Human Metabolome Database (www.hmdb.com).

[b] Reference range obtained from Blau *et al.,* (2005: 32-38) according to the minimum to maximum values in mmol/mol creatinine.

n.d.: not determined according to public databases[a] and tables published in the relevant literature[b].

# Note that the median value for one of the groups is much lower than their mean concentration, meaning that one or more cases had a high mean concentration compared to the other cases.

Table 4.11: Comparison of infant group (IY) with child group.

Variables	VIP Ranking		Infants (IY)			Children (C)			Reference Range[23]		Comment/Note
	PCA	PLS-DA	Mean	Median	Standard deviation	Mean	Median	Standard deviation	Infants	Children	
Pyroglutamic acid	1		14.16	10.63	10.14	4.88	4.36	3.75	n.d.	2.9-10.4^{a}	Reference values are unknown, but the reference value for infants (0 and 1 year old), according to the Human Metabolome Database is between 3.4-54.2 mmol/mol creatinine.
Aconitic acid	2		122.70	117.05	22.64	46.31	38.97	24.59	26.87-189^{b}	20.5-135^{b}	-
2-Hydroxyglutaric acid	3		11.72	9.50	6.75	3.59	2.82	2.49	5-26.8^{b}	1.3-13.9^{b}	-
3-Hydroxyisobutyric acid	4	8	36.91	26.65	35.08	8.92	7.94	7.25	20.2-118^{b}	12.8-137^{b}	-
Glyceric acid	5		2.07	2.02	1.46	0.12	0.00	0.22	4.2-3.2.2^{b}	2.6-28.2^{b}	-
3-Hydroxyvaleric acid	6	10	34.11	24.72	36.91	7.10	5.98	5.02	n.d.	n.d.	-
4-Hydroxybenzoic acid	7		11.73	6.50	11.28	2.63	2.10	2.20	0.7-13.7^{a}	0.6-4.2^{a}	

[23] Reference ranges (mmol/mol) are obtained from either Blau *et al.*, (2005) or the Human Metabolome Database (www.hmdb.com)

Variables	VIP Ranking		Infants (IY)			Children (C)			Reference Range		Comment/Note
	PCA	PLS-DA	Mean	Median	Standard deviation	Mean	Median	Standard deviation	Infants	Children	
3-Hydroxy-3-Methylglutaric acid	8	2	17.87	16.60	13.28	1.57	0.00	2.81	6.2-49.7[a]	<10.3-28[b]	-
3-Hydroxysebacic acid	9	7	16.54	1.46	31.94	0.04	0.00	0.20	<2.3-9.1[b]	<0.2-2.0[b]	#
Suberic acid	10	3	9.20	6.40	9.47	0.47	0.00	1.00	<2.2-10.1[b]	<1.4-8.8[b]	-
Succinic acid		1	134.52	124.41	58.48	25.69	13.50	37.00	17.6-79.2[b]	4.9-81.3[b]	
Adipic acid		4	12.59	11.57	7.70	1.28	1.29	0.83	<5.9-34.3[b]	<1.1-5.3[b]	#
Uric acid		5	12.34	11.92	12.66	0.99	0.00	1.53	n.d.	524.75 +/-249.57[a]	The reference values for the child are for ages ranging from 1 to 13 years old.
3,4-Dihydroxybutyric acid		6	16.74	14.26	11.59	3.27	2.18	3.47	109-454[b]	58-320[b]	-
3-Hydroxyphenylhydracrylic acid		9	17.18	3.40	29.85	15.92	6.38	22.67	n.d.	n.d.	#

[a] Reference range obtained from the Human Metabolome Database (www.hmdb.com).

[b] Reference range obtained from Blau *et al.,* (2005: 32-38) according to the minimum to maximum values in mmol/mol creatinine.

n.d.: not determined according to public databases[a] and tables published in the relevant literature[b].

Note that the median value for one of the groups is much lower than their mean concentration, meaning that one or more cases had a high mean concentration compared to the other cases.

Table 4.12: Comparison of child group with adult group.

Variables	VIP Ranking		Children (C)			Adults (A)			Reference Range[24]		Comment, according to the mean concentration values of the groups in comparison with the reference values
	PCA	PLS-DA	Mean	Median	Standard deviation	Mean	Median	Standard deviation	Children	Adults	
3-Hydroxyvaleric acid	1	1	7.10	5.98	5.02	0.00	0.00	0.00	n.d.	0-2^{a}	-
2-Hydroxyglutaric acid	2	6	3.59	2.82	2.49	0.80	0.77	0.37	1.3-13.9^{b}	0.8-52^{b}	-
Vanillylmandelic acid	3	9	10.20	8.28	5.43	4.67	4.47	1.38	1-15^{b}	2.5-16.1^{a}	-
3-Hydroxyisovaleric Acid	4	4	0.00	0.00	0.00	2.11	1.72	1.23	9.8-50.2^{b}	6.9-25^{b}	-
4-Hydroxyphenylacetic Acid	5	2	45.01	26.93	43.31	9.81	8.03	6.53	7.4-30.1^{b}	3.5-22^{b}	#
3-Hydroxyphenylhydracrylic acid	6	10	15.92	6.38	22.67	7.38	3.78	7.66	n.d.	n.d.	
3-Hydroxyisobutyric Acid	7		8.92	7.94	7.25	4.30	3.35	2.39	12.8-137^{b}	4.1-19^{b}	-
Homovanillic Acid	8	3	10.87	8.33	9.61	2.14	1.98	1.28	0.7-10.3^{b}	0.9-5.5^{b}	-

[24] Reference ranges (mmol/mol) are obtained from either Blau *et al.*, (2005) or the Human Metabolome Database (www.hmdb.com)

Variables	**VIP Ranking**		**Children (C)**			**Adults (A)**			**Reference Range**		**Comment, according to the mean concentration values of the groups in comparison with the reference values**
	PCA	**PLS-DA**	**Mean**	**Median**	**Standard deviation**	**Mean**	**Median**	**Standard deviation**	**Children**	**Adults**	
2-Hydroxyisobutyric acid	9	8	4.26	3.75	2.36	1.66	1.49	0.83	3.8-14.1[a]	n.d.	-
4-Hydroxybenzoic acid	10		2.63	2.10	2.20	1.16	1.00	0.73	0.1-11.1[a]	0.6-4.2[a]	-
Citric Acid		5	181.39	154.60	107.31	77.30	73.45	51.70	120-582[b]	70-226[b]	-
Aconitic acid		7	46.31	38.97	24.59	24.41	22.61	13.81	20.5-135[b]	2.7-44[b]	-

[a] Reference range obtained from the Human Metabolome Database (www.hmdb.com).

[b] Reference range obtained from Blau *et al.,* (2005: 32-38) according to the minimum to maximum values in mmol/mol creatinine.

n.d.: not determined according to public databases[a] and tables published in the relevant literature[b].

# Note that the median value for one of the groups is much lower than their mean concentration, meaning that one or more cases had a high mean concentration compared to the other cases.

- Discussion of Table 4.11: Infant (IY) vs. Child comparison

Four common variables were identified with PCA and PLS-DA, namely 3-hydroxyisobutyric acid, 3-hydroxy-3-methylglutaric acid, suberic acid and 3-hydroxysebacic. All the variables listed in the table showed a decrease in mean concentration value for the comparison between IY and C. The metabolites identified were grouped as follows:

a. Phenolic acids (3-hydroxyphenylhydracrylic acid and 4-hydroxybenzoic acid)
b. Short chain dicarboxylic acid (2-hydroxyglutaric acid, adipic acid, 3-hydroxy-3-methylglutaric acid and succinic acid)
c. Short chain hydroxy acid (3-hydroxyvaleric acid and 3-hydroxyisobutyric acid)
d. Short chain dihydroxy acid (3,4-dihydroxybutyric acid and glyceric acid)
e. Short chain tricarboxylic acid (aconitic acid)
f. Amino acid conjugate (pyroglutamic acid)
g. Long chain dicarboxylic acid (3-hydroxysebacic acid and suberic acid)
h. Hydroxy purine (uric acid)

Most of the variables were either involved in amino acid metabolism or the TCA cycle.

- Discussion of Table 4.12: Child vs. Adult comparison

Two of the variables present in Table 4.12 had zero values in one of the groups namely 3-hydroxyvaleric acid and 3-hydroxyisovaleric acid. Eight common variables were identified via PCA and PLS-DA, namely 3-hydroxyvaleric acid, 2-hydroxyglutaric acid, vanillylmandelic acid, 3-hydroxyisovaleric acid, 4-hydroxyphenylacetic acid, 3-hydroxyphenylhydracrylic acid, homovanillic acid and 2-hydroxybutyric acid. Eleven of the metabolites listed showed a decrease in mean concentration value, whereas only one showed an increase. The metabolites identified were grouped as follows:

a. Phenolic acids (4-hydroxyphenylacetic acid and 4-hydroxybenzoic acid)
b. Short chain dicarboxylic acid (2-hydroxyglutaric acid)
c. Short chain hydroxy acid (2-hydroxyisobutyric acid, 3-hydroxyvaleric acid, 3-hydroxyisovaleric acid and 3-hydroxyisobutyric acid)
d. Catecholamine (homovanillic acid and vanillylmandelic acid)
e. Short chain tricarboxylic acid (aconitic acid and citric acid)
f. Phenolic acids (3-hydroxyphenylhydracrylic acid)

Most of the variables were either involved in amino acid metabolism or the TCA cycle. A conclusion will be given in Chapter 5 with regard to the results discussed in Section 4.4.

Chapter 5 – Discussion

5.1 Introduction

This Chapter was not designed to discuss the details of each result obtained in the dissertation, as presented in the previous Chapter, but to focus on aspects which are related to the three aims cited in Chapter 2. This will be presented in Sections 5.2, 5.3 and 5.4. I will attempt to make a clear recommendation linked to each aim based on the results obtained, and each recommendation will then be motivated in Section 5.5.

5.2 The first aim

As previously stated in Chapter 2 the first aim of this study was to:

> **"Become acquainted with the technology of metabolomics to generate analytical data with sufficient chromatographic richness and resolution for multivariate statistical analysis and for metabolite identification and quantification"** (Harrigan *et al.*, 2005).

With regard to this aim it is important to note that an important requirement for multivariate analyses of analytical data is an adequate number of cases in the control and experimental groups. For statistical analysis a data matrix must first be compiled containing a list of the cases as well as their respective metabolites (variables). Table 5.1 gives an example of such a data matrix. I will also discuss possible reasons for the reduction of a data matrix with regard to the number of variables present.

Table 5.1: Example of a general data matrix with n cases and p variables, illustrated in the table below for selective data from the age perturbation.

$$\underset{(n \times p)}{\mathbf{X}} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

Case *i* in *X*	**2,3-Dihydroxy-butane (j=10)**	**1,2-Dihydroxy-benzene (j=25)**	**1,2-Dihydroxy-ethane (j=31)**	**1-Indole-3-Acetic Acid (j=125)**
C04 (i = *4*)	0 (=$X_{4,10}$)	2.012 (=$X_{4,25}$)	0 (=$X_{4,31}$)	0 (=$X_{4,125}$)
C06 (i = 6)	0 (=$X_{6,10}$)	2.254 (=$X_{6,25}$)	0 (=$X_{6,31}$)	0 (=$X_{6,125}$)
C81 (i=*81*)	0 (=$X_{81,10}$)	1.320 (=$X_{81,125}$)	0 (=$X_{81,31}$)	1.154 (=$X_{81,125}$)

In the table presented above, i=1,2,...n represents a specific case, and j=1,2,…,p represents a specific variable (metabolite). Firstly I will focus on the selection and availability of cases (*n* in the data matrix), with regard to the investigation of the three perturbations:

1. Menstrual cycle

- Due to the sensitive nature of this particular study it was difficult to approach women for their participation; subsequently we only had two participants who gave samples over the post-menstrual period. Multivariate analysis was not possible for this data, owing to the lack of a sufficient number of cases. For this reason only univariate analysis was performed.

2. Pregnancy

- The original intention of this study was to obtain ten urine and blood samples at various stages during pregnancy i.e. for a longitudinal study. The blood samples would have been used to determine the *SULT1A1* genotype each person possesses as well as the concentrations of the thyroid hormones and their sulfonated conjugates, whereas the urine samples would have been for organic acid analysis. This study was not possible since only one participant gave more than one sample over the pregnancy period (Table 5.2).

- Subsequently the current study focuses on the difference between the three trimesters with regard to the control group i.e. instead of a longitudinal study it became a cross-sectional study. It was possible to determine the thyroid hormone concentrations in blood, but the method used is not validated and the concentrations obtained were lower than the reference values and in some cases even so low that it was not detected by the LC-MS. For this reason I did not use these results in this dissertation, but focused only on the results obtained from the urinary organic acid.

- Table 5.2 shows the composition of the control and experimental groups, specifically from whom samples were obtained. Only one participant gave more than one sample, specifically for each of the trimesters. Her samples were included in the analysis since each trimester was compared to the control group, separately and not collectively.

Table 5.2: Composition of the control and experimental groups for the pregnancy study.

Participants	Control	Trimester 1	Trimester 2	Trimester 3
C1	√			
C2	√			
C3	√			
C4	√			
C5	√			
C6	√			
C7	√			
C8	√			
C9	√			
C10	√			
C11	√			
C12	√			
P.1.1		√		
P.1.2.		√		
P.1.3		√	√	√
P.1.4		√		
P.2.1			√	
P.2.2			√	
P.2.4			√	
P.2.5			√	
P.2.6			√	
P.2.7			√	
P.3.2				√
P.3.3				√
P.3.4				√
P.3.5				√

3. Age

- For this perturbation there were no complications in getting sufficient samples for the adult group. However, I was unable to get any, albeit sufficient, samples for a metabolomics study of the infant or child groups. Consequently I had to obtain information from existing datasets from routine diagnostic studies on possible defects in organic acid metabolism, done on infants and children. In all these cases the respective infants or children were diagnosed as “normal”, and this material was thus used for the comparable analyses of the effect of aging on the metabolite profiles.
- This study had the most cases and number of variables of all three the perturbations investigated, hence if was possible to perform multivariate analyses.

Secondly I will discuss the variables (= *p* in the in the data matrix above) with regard to the identification and quantification of urinary organic acids before statistical analysis is performed. Variables are included in a data matrix after they have undergone several methodical processes such as sample preparation, extraction, derivatisation, separation on the GC/MS system, identification by means of AMDIS and manual quantification using the raw data given by AMDIS for each metabolite. Hence it is important for metabolites to be unequivocally identified, especially when automated identification methods are employed. Seeing that metabolites can be falsely identified or not identified at all, this can lead to a false negative or false positive result and ultimately a misconceived deduction. The automated identification system (AMDIS) I employed has been used for years in the Laboratory for Inherited Metabolic Defects with great confidence and as such made it possible to identify numerous metabolites. However, *AMDIS is good for identification but much less so for quantification of metabolites.* Hence quantification was done manually which made data pre-processing time-consuming. The reader is referred to Section 3.6 for a discussion on the method of identification and quantification of the metabolites.

The number of variables present in a data matrix can be reduced for several reasons e.g. the removal of variables with no variation in the data matrix, or variables that do not have a known physiological importance and finally variables that do not contain group separation information. Figure 5.1 shows how the age data matrix (see Figure 4.16) was reduced so that only about 25% of the original variables were used for possible biomarker discovery. The reasons for data reduction will be discussed with regard to Figure 5.1, since this particular data matrix contained the most cases and variables of all three perturbations.

Initially the data matrix comprised 457 (100%) variables, but since it is vital to remove unnecessary noise in the data matrix, effect sizes were used (see Section 3.7.4). Consequently 58% of the data had an effect size lower than 0.5 i.e. a medium effect. The number of variables with effect sizes higher than 0.5 were further reduced via a biological filter (see Section 4.1); this filter is utilised in order to exclude metabolites of non-physiological importance. After the biological filter was applied to the data matrix only 25% of the original data matrix was used for multivariate analysis. For the multivariate analyses an unsupervised (PCA) and supervised (PLS-DA) pattern recognition method was used on the remaining data matrix (25%). These two methods rank variables according to their importance in the projection (see Section 3.7.6). From this ranking, approximately 3% of the variables contained the new information (potential biomarkers) about this particular perturbation. To conclude even though the data matrix was reduced, it was still an adequate data matrix for statistical analysis since a total of 75 cases were compared to one another, for 114 variables (8550 data points).

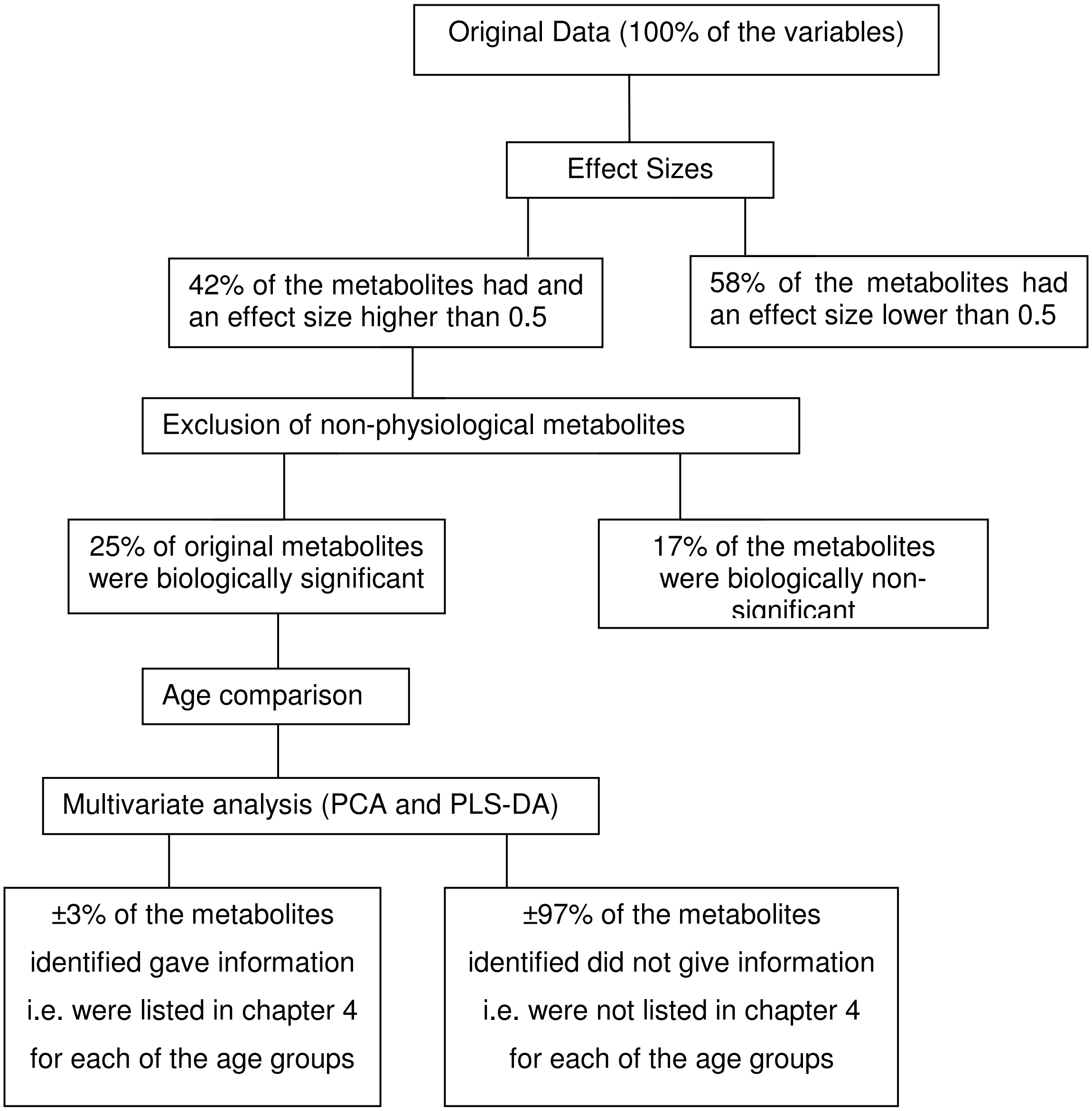

Original Data (100% of the variables)
Effect Sizes
42% of the metabolites had and an effect size higher than 0.5
58% of the metabolites had an effect size lower than 0.5
Exclusion of non-physiological metabolites
25% of original metabolites were biologically significant
17% of the metabolites were biologically non-significant
Age comparison
Multivariate analysis (PCA and PLS-DA)
±3% of the metabolites identified gave information i.e. were listed in chapter 4 for each of the age groups
±97% of the metabolites identified did not give information i.e. were not listed in chapter 4 for each of the age groups

Figure 5.1: Flowchart of an example of data reduction in the age study by means of statistical analysis for age as a perturbation

In conclusion multivariate statistical analysis is only sensible if there are a sufficient number of cases as was seen for the pregnancy and age studies, otherwise univariate analysis has to be performed as was the case in the menstrual cycle study. The number of variables in a metabolomics investigation is not a problem, seeing that the metabolome i.e. multiple metabolites are investigated and not a small number of specific metabolites.

5.3 The second aim

From Chapter 2 it is clear that the second aim was:

> **"To investigate three natural perturbations not formerly subjected to a detailed metabolomics study and of sufficient complexity to become acquainted with the biostatistical research methods to generate new biological information on these perturbations."**

This investigation, where possible must try and lead to the generation of new biological information, for example biomarkers on these perturbations via biostatistical research methods.

Where possible, I will compare my results obtained by means of the metabolomics approach with the results of other studies mentioned in Chapter 2, which were not metabolomics-orientated. The following information was obtained for each study:

1. Menstrual cycle

- In Section 2.4.1.2 it is mentioned that no specific study has of yet been reported on the urinary organic acid profile and its potential role in the menstrual cycle. For this reason I cannot compare my results with the results of previous studies or related studies such as Kochhar *et al*., (2006), as Kochhar *et a*l., (2006) only collected urine samples from women during day ten and fifteen of the menstrual cycle and not for each phase. As stated in Section 4.2 most of the metabolites identified (seven of the ten) showed an overall decrease in mean concentration value over the menstrual cycle, and one metabolite showed an increase. The other two metabolites had mixed trends. All of the metabolites identified were naturally occurring metabolites (see Table 4.4, Section 4.2).

- This study also had another purpose: Urine samples were obtained randomly from the participants in the non-pregnant group and as such it was not known if a specific phase influenced the organic acid profile. From the results obtained for the menstrual cycle study, we accordingly did not find it necessary to stratify the pregnancy control group according to the menstrual cycle.

- In conclusion, from this study it does not seem that inference or generalisation with regard to the perturbation of the menstrual cycle could be made. This could, however, be due to the limited number of cases used, and this investigation should consequently only be seen as a pilot study.

2. Pregnancy

- For this perturbation I will firstly compare my results with the results of Christie (1982), since this study also focuses on the urinary organic acid profile of pregnant women. Christie (1982) obtained more samples from pregnant and non-pregnant women than my study; hence it was possible for her to make certain conclusions with regard to the role of some organic acids during the progression of pregnancy, given that she obtained samples from the same pregnant women for two pregnancy intervals, whereas I had only one pregnant woman who gave more than one sample over the pregnancy period.

- The study of Christie (1982) was not a metabolomics study and as such only focused on a couple of metabolites that were unequivocally identified, whereas my study identified numerous urinary organic acid metabolites (250). It was possible to identify more metabolites that might play a possible role during pregnancy than Christie (1982) because of the advances made in science with regard to technological development (e.g. GC/MS and AMDIS). In Table 5.3 I will compare my findings with the findings of Christie with regard to the metabolites that both studies had in common. I will also refer to some metabolites that were not discussed by Christie (1982).

- The results of Christie (1982) and my investigation should be interpreted separately, since there was a difference in the extraction, identification and quantification methods employed.

Table 5.3: Comparison between the findings of Christie (1982) and this investigation.

Metabolite	**Christie (1982)** **umol/24 h excretion**			**My findings** **mmol/mol creatinine**		
(Normal reference range)	Non-pregnant (n=15)	25.5 – 27.5 weeks pregnant (n=25)	Comment	Non-pregnant (n=11)	Third trimester (n=5)	Comment
	Mean values					
Lactic acid (13-46 mmol/mol creatinine[25])	240	1310	P<0.001 Significantly increased	8.3	36.45	Increased from non-pregnant state to the third trimester
Glycolic acid (15-122 mmol/mol creatinine[26])	570	1420	P<0.001 Significantly increased	18.4	28.13	Increased from non-pregnant state to the third trimester
Erythronic acid (unknown)	780	1040	P<0.001 Significantly increased	0.14	0	Decreased from non-pregnant state to the third trimester
4-Deoxytetronic acid (unknown)	530	1160	P<0.001 Significantly increased	-	-	Not detected in analysis

- Christie (1982) identified 10 metabolites, unambiguously. Her study focused on only four (glycolic acid, erythronic acid, 4-deoxytetronic acid and lactic acid) of the metabolites,

[25] Reference range obtained from Blau *et al.,* (2005: 32-38) according to the minimum to maximum values in mmol/mol creatinine.

[26] Reference range obtained from the Human Metabolome Database (www.hmdb.com).

whereas I focused on a total of twenty-five metabolites identified via PCA and PLS-DA (see Tables 4.6 and 4.7).

- The four metabolites which Christie (1982) finally discussed showed an increase as pregnancy progressed. My investigation showed the same trend as most of the twenty-five metabolites showed an increase in concentration when the participants in the third trimester were compared to the participants in the second trimester (Table 4.7).

- Note that at least four (succinic acid, fumaric acid, malic acid, aconitic acid, and citric acid) of the TCA cycle intermediates were present in both Tables 4.6 and 4.7. The TCA cycle is involved in aerobic cellular respiration, the synthesis of ATP via oxidative phosphorylation and produces intermediates for many other metabolic pathways. Consequently the presence of the TCA intermediates in urine, gives an indication of energy production i.e. the greater the concentration of the intermediates the greater the energy utilisation.

- In conclusion the increase in metabolite excretion is most likely attributed to the metabolic adaptations present during pregnancy, which is needed for homeostasis.

- Potential biomarkers were not identified in view of the fact that the study did not have many cases, even though there were many metabolites that were identified.

3. Age

- For the purposes of this discussion I will compare my results with the results of Guneral and Bachmann (1994) (see Section 2.4.3.2). For the age perturbation a total of twenty-nine metabolites were identified by means of the four group comparisons i.e. IM vs. IY, IM vs. C, IY vs. C, and C vs. A (see tables 4.9 to 4.12). Twenty-two of these metabolites were also identified by Guneral and Bachmann (1994). Tables 5.4.1 and 5.4.2 contain the findings of both studies. I will also include those metabolites that were only identified in my study and not in the study of Guneral and Bachmann (1994) (Table 5.4.3). It should, however, be noted that the age groupings reported by Guneral and Bachmann (1994) overlaps with my groupings, but are not identical for comparable purposes. I will thus present only the trends observed in both studies in Tables 5.4.1 and 5.4.2. The results coming from these studies should thus be interpreted separately.

Table 5.4.1 First group of metabolites decreasing with age, according to Guneral and Bachmann (1994).

Metabolite/Variable	**Guneral and Bachmann (1994)**[27]	**My findings**[28]
2-Hydroxyglutaric acid	Decreasing with age	Decreasing with age
Acetylaspartic acid	Decreasing with age	Decreasing with age
Lactic acid	Decreasing with age with a pattern of non-significant fluctuations	Decreasing with age
Succinic acid	Decreasing with age with a pattern of non-significant fluctuations	Decreasing with age
Glutaric acid	Decreasing with age with a pattern of non-significant fluctuations	Decreasing with age
Fumaric acid	Decreasing with age with a pattern of non-significant fluctuations	Decreasing with age
Malic acid	Decreasing with age with a pattern of non-significant fluctuations	Decreasing with age
Adipic acid	Decreasing with age with a pattern of non-significant fluctuations	Decreasing with age
3-Hydroxyisovaleric acid	Decreasing with age with a significant increase at ages 1-6 months	Decreasing pattern with an increase between the child and the adult groups
Homovanillic acid	Decreasing with age with a significant increase at ages 1-6 months	Decreasing with age

- The metabolites in Table 5.4.1 showed primarily a decrease in mean concentration values for my four age groups (IM, IY, C and A), except for 3-hydroxyisovaleric acid which showed an increase between my child and adult groups.

[27] These samples were grouped according to 5 defined age groups namely: newborns between 2 and 28 days, infants between 1 and 6 months old, children between 2 and 6 years old, children between 6 and 10 years old and children older than 10 years. This study did not contain samples from adults.

[28] For my investigation I defined the following 4 experimental groups: a young infant group (3 days – 7 months old), an older infant group (1 year - 2.5 years), a children group (11 – 13 years old) and a young adults group (20 – 27 years old).

Table 5.4.2: Second group of metabolites with an increasing pattern and a decrease from infants older than a year to adults, according to my findings.

Metabolite/Variable	**Guneral and Bachmann (1994)**	**My findings**
2-Hydroxyisobutyric acid	Increasing with age	Increasing pattern with a decrease from infants older than a year to adults
3-Hydroxy-2-methylbutyric acid	Increasing with age	Increasing pattern with a decrease from infants older than a year to adults
Citric acid	Decreasing with age	Increasing pattern with a decrease from infants older than a year to adults
Glyceric acid	Decreasing with age with a pattern of non-significant fluctuations	Increasing pattern with a decrease from infants older than a year to adults
Suberic acid	Decreasing with age with a pattern of non-significant fluctuations	Increasing pattern with a decrease from infants older than a year to adults
4-Hydroxyphenylacetic acid	Decreasing with age with a pattern of non-significant fluctuations	Increasing pattern with a decrease from infants older than a year to adults
Vanillylmandelic acid	Decreasing with age with a pattern of non-significant fluctuations	Increasing pattern with a decrease from infants older than a year to adults
3-Hydroxyisobutyric acid	Decreasing with age with a significant increase at ages 1-6 months	Increasing pattern with a decrease from infants older than a year to adults
Pyroglutamic acid	Decreasing with age with a significant increase at ages 1-6 months	Increasing pattern with a decrease from infants older than a year to adults
3-Hydroxy-3-methylglutaric acid	Decreasing with age with a significant increase at ages 1-6 months	Increasing pattern with a decrease from infants older than a year to adults
3-Hydroxysebacic acid	Decreasing in age with a pattern of non-significant fluctuations	Increasing pattern with a decrease from infants older than a year to adults

Table 5.4.3: Third group of metabolites which were not identified by Guneral and Bachmann (1994) or showed other trends.

Metabolite/Variable	Guneral and Bachmann (1994)	My findings
4-Hydroxybenzoic acid	Unchanged pattern	Increasing pattern with a decrease from infants older than a year to adults
3-Hydroxyvaleric acid	-	Increasing pattern with a decrease from infants older than a year to adults
Aconitic acid	-	Increasing pattern with a decrease from infants older than a year to adults
3,4-Dihydroxybutyric acid	-	Increasing pattern with a decrease from infants older than a year to adults
Uric acid	-	Increasing pattern with a decrease from infants older than a year to adults
Palmitic acid	-	Decreasing pattern with an increase from child to adults
3-Hydroxyphenyl-hydracrylic acid	-	Increasing pattern with a decrease from infants older than a year to adults
Octadecenoic acid	-	Decreasing with age
3-Hydroxyadipyllactone		Decreasing with age (This substance was probably formed due to the derivatisation conditions)

- The metabolites in my study mostly showed an increase in mean concentration value from infants younger than a year to infants older than a year. From the infants older than a year to the adult group it showed a decrease (Table 5.4.2).

- Table 5.4.3 lists metabolites that showed other patterns e.g. an unchanged pattern (Guneral and Bachmann, 1994) or metabolites (seven) that were not identified by Guneral and Bachmann (1994). They identified a total of 69 organic acids; only the twenty-nine metabolites identified with the multivariate analyses used in my study were listed in Tables 5.4.1 to 5.4.3. Most of the metabolites identified by Guneral and Bachmann (1994) were also identified in my study except for the following: erythro-4-deoxytetronic acid, threo-4-deoxytetronic acid, 2-oxoglutaric acid, hydroxypyruvic acid, and hexadecanoic acid.

- Note that some of the metabolites had the same trend i.e. decreasing with age for both studies, in particular 2-hydroxyglutaric acid, acetylaspartic acid, lactic acid, succinic acid, glutaric acid, fumaric acid, malic acid, adipic acid, 3-hydroxyisovaleric acid, and homovanillic acid.

- At present it is not possible to identify potential biomarkers of the aging process, since it is a complex process incorporating many metabolic processes, but what is evident is that there is a distinct difference between varying age groups according to the organic acid profile.

5.4 The third aim

The final aim of this study is to try and formulate a possible hypothesis, based on the metabolomics of natural perturbations as well as an approach to test the hypothesis. According to Lawrence (2005) a hypothesis is defined as a proposed explanation of a scientific problem or phenomenon that has to be tested experimentally before it is accepted as scientific law. Kell (2004) states that metabolomics is especially hypothesis-generating, and not hypothesis testing.

It was only possible to formulate one hypothesis from the study specifically for the age perturbation since the menstrual cycle study had too few cases, the pregnancy study had a sufficient number of cases for at least one of the trimesters, but it did not show a distinct difference between the control and pregnant groups. Subsequently the following hypothesis is formulated:

> **“Based on the difference in metabolites as a result of age one can hypothesise that the best experimental approach for biomarker discovery and identification would be to select the group with the perturbation at random and the control group according to the age profiles of the perturbation group.”**

This hypothesis differs from the kind of hypothesis expected from a metabolomics study which is formulated as a consequence of a strong perturbation, like in an inherited metabolic disorder. This hypothesis would be important for metabolomics studies, however, as shown in Figures 5.2 and 5.3. I chose my illustration of this statement from the results of a current PhD-study by Mrs M Dercksen in our laboratory. Her study focuses on a possible difference in the organic acid profiles of children diagnosed with Isovaleric acideamia (IVA), an inherited metabolic defect, when compared to controls and to their parents and other siblings who are obligate heterozygotes. An obligate heterozygote as defined by Intergenetics.co.uk is an individual who is clinically unaffected, but must carry a particular mutant allele based on pedigree analysis. In Figure 5.2 the participants who are heterozygous for the defect (mostly adults) is compared to the children and infants who were the control group in the IVA-study. This figure showed an overlap in the data of the two groups. Figure 5.3 compares the obligate heterozygotes with the adult group used in my study. This comparison shows a distinct separation between the two groups, most probably because they were closer in age than the groups compared with one another in Figure 5.2. This study illustrates that further studies should be performed, based on the hypothesis, before any generalisations are made; even when complex perturbations are investigated such as genetic or metabolic disorders.

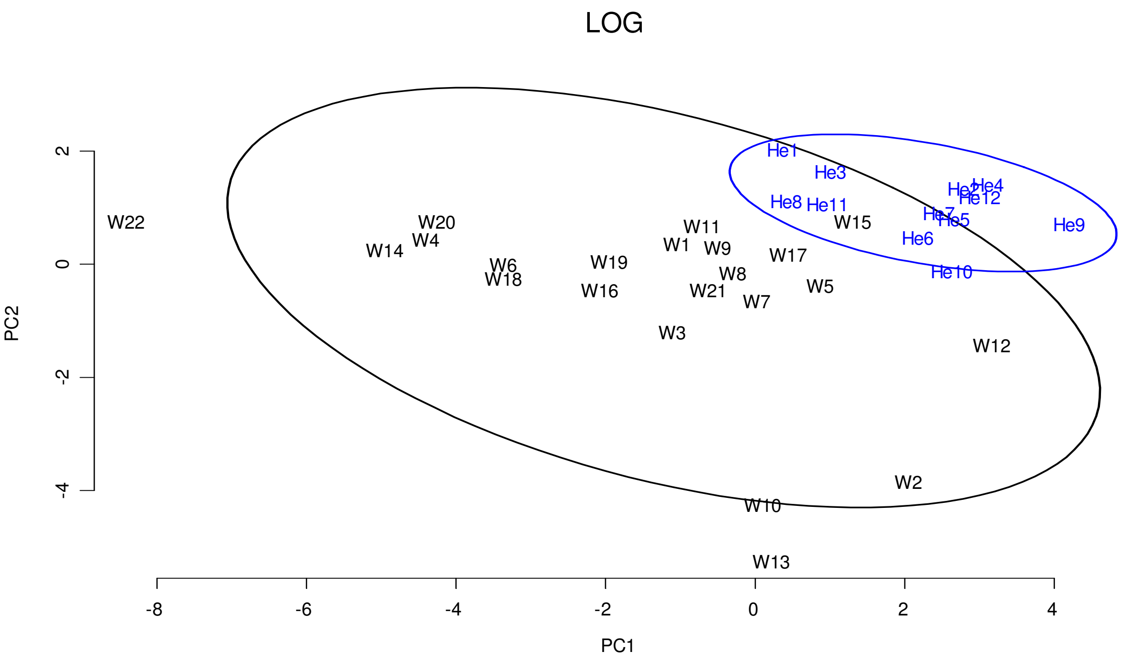


Figure 5.1: PCA scores plot of a control group of infants and children (W – black) compared to obligate heterozygotes for IVA (He – blue).

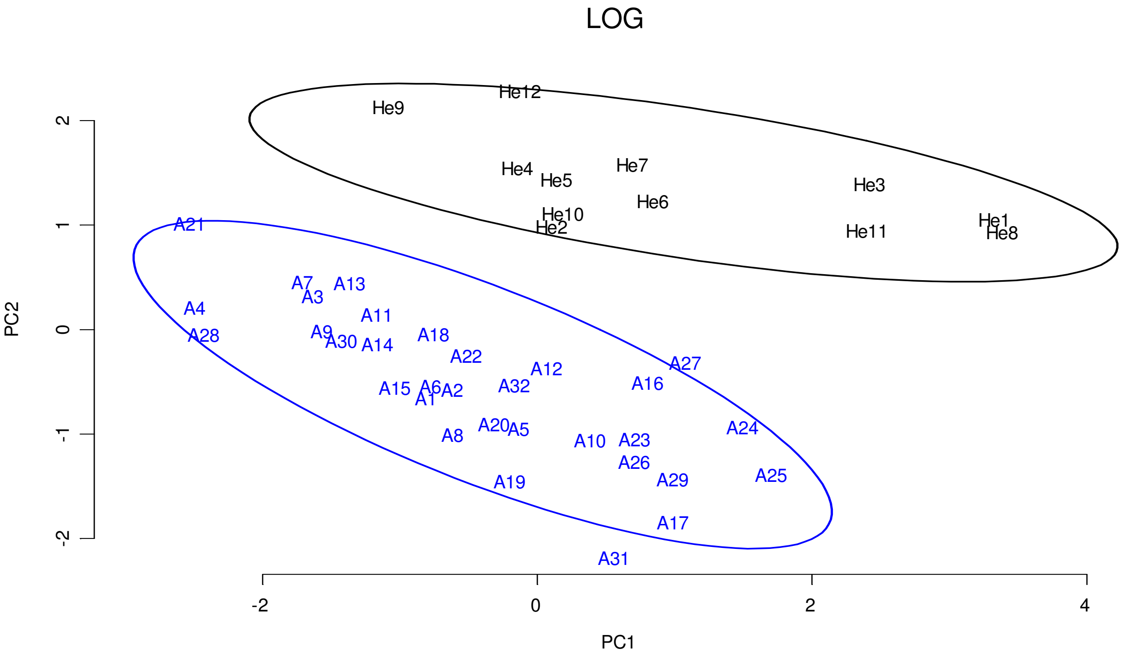


Figure 5.2: PCA scores plot of obligate heterozygotes for IVA (He – black) and a control adult group (A – blue).

5.5 Recommendations

Based on the discussion of the results presented in this dissertation, I would like to make the following recommendations in the interest of future studies in the field of metabolomics:

1. My investigation could not identify any possible biomarkers involved with the three perturbations I investigated by means of a metabolomics approach since the number of participants in each study was not sufficient. Subsequently before any investigation can generalise their findings a sufficient amount of data i.e. cases versus variables is needed. The number of cases needed will be determined by the specific research question, as well as by the willingness of individuals to participate in a certain study. This is, however, not always possible as was seen in my Masters study where the experimental subjects were humans and as such participated voluntarily.

   Recommendation one: If it appears that an insufficient number of participants can be generated for a metabolomics study, such a study should be discarded in the interest of another more feasible investigation.

2. Secondly it is important with any investigation that the analytical methods utilised for identification and quantification of metabolites present in complex biological matrixes such as urine should be used with a high level of confidence i.e. the results should be, for example, repeatable and accurate within the limits of quantification. This is especially important as all the samples used in a study are not analysed on the same day or even in the same analytical run.

   Recommendation two: It is advisable that a number of appropriate analytical validation parameters should be incorporated in the early stages of a metabolomics study, specifically linked to the context of the perturbation chosen for the investigation.

3. Finally when studies are undertaken that compare a control group with an experimental group (case-control studies) that is influenced by a specific perturbation such as a genetic or metabolic defect certain strategies should be employed with regard to the selection process of participants.

   Recommendation three: The control and experimental groups should be as homogenous that is to say as comparable as possible with regard to age, ethnicity, diet, and gender, lifestyle habits and other possible confounding influences, except for the specific perturbation being studied. In a perfect world this would be possible, specifically when hypothesis formulation, testing and finally the expansion of scientific knowledge will be a desired outcome of the investigation.

Bibliography

ALLEN, R.G. & BALIN, A.K. 2003. Metabolic rate, free radicals and aging. [*In* Cutler, R.G. & Rodriques, H. Critical reviews of oxidative stress and aging: advances in basic science, diagnostics and intervention. World Scientific. New Jersey. p. 4-29.]

ASHOK, B.T. & R. ALI. 1999. The aging paradox: free radical theory of aging. *Experimental gerontology*, 34:293-303.

BAGGOT, P.J., ELISEO, A.Y., DENICOLA, N.G., KALAMARIDES, J.A. & SHOEMAKER, J.D. 2008. Organic acid concentrations in amniotic fluid found in normal and down syndrome pregnancies. *Fetal diagnosis and therapy*, 23:245-248.

BAKER, F.C. & DRIVER, H.S. 2007. Circadian rhythms, sleep and the menstrual cycle. *Sleep medicine*, 8:613-622.

BLACKBURN, S.T. & LOPER, D.L. 1992. Carbohydrate, fat, and protein metabolism. (*In* Blackburn, S.T. & Loper, D.L. Editors. Maternal, fetal and neonatal physiology: a clinical perspective. Philadelphia: W.B. Saunders. p. 583-613).

BOLLARD, M.E., STANLEY, E.G., LINDON, J.C., NICHOLSON, J.K. & HOLMES, E. 2005. NMR-based metabonomic approaches for evaluating physiological influences on biofluid composition. *NMR in biomedicine*, 18:143-162.

CHRISTIE, E.J. 1982. Profiling of urinary metabolites in human pregnancy. Saskatoon: University of Saskatchewa. (Thesis – M.Sc.) 151 p.

COEN, M., O'SULLIVAN, M., BUBB, W.A., KUCHEL, P.W. & SORRELL, T. 2005. Proton nuclear magnetic resonance-based metabonomics for rapid diagnosis of meningitis and ventriculitis. *Clinical infectious diseases*, 41: 1582-1590.

DETTMER, K., ARONOV, P.A. & HAMMOCK, B.D. 2007. Mass spectrometry-based metabolomics*. Mass Spectrometry Reviews*, 26(1):51-78.

DUNN, W.B. & ELLIS, D.I. 2005. Metabolomics: current analytical platforms and methodologies. *Trends in analytical chemistry*, 24(4):285-294.

ELLIS, S.M. & STEYN, H.S. 2003. Practical significance (effect sizes) versus or in combination with statistical significance (p-values). *Management Dynamics*, 12(4): 51-53.

ESCUDERO, F., GONZALES, G.F. & GONEZ, C. 1996. Hormone profile during the menstrual cycle at high altitude. *International journal of gynecology & obstetrics*, 55:49-58.

FIEHN, O. 2002. Metabolomics: the link between genotypes and phenotypes. *Plant Molecular Biology*, 48: 155–171.

FINKEL, T. & HOLBROOK, N. 2000. Oxidants, oxidative stress and the biology of ageing. *Nature*, 408: 239-247.

GANDARA, B.K., LERESCHE, L. & MANCL, L. 2007. Patterns of salivary estradiol and progesterone across the menstrual cycle. *Annals of the New York Academy of Sciences*, 1098:446-450.

GARRETT, R.H. & GRISHAM, C.M. 2005. Biochemistry. 3rd ed. Thomson. 1086 p.

GATES, S.C. & SWEELEY, C.C. 1978. Quantitative metabolic profiling based on gas chromatography. *Clinical Chemistry,* 24:1663-1673.

GOODACRE, R. Data analysis standards in metabolomics. http://msi-workgroups.sourceforge.net/data-processing/reports/DAstandardsVer2.pdf Date of Access: 1 Dec. 2009.

GOODACRE, R., VAIDYANATHAN, S., DUNN, W.B., HARRIGAN, G.G. & KELL, D.B. 2004. Metabolomics by numbers: acquiring and understanding global metabolite data. *TRENDS in Biotechnology*, 22(5): 245-252.

GOODACRE,R., BROADHURST, D., SMILDE, A.K., KRISTAL, B.S., BAKER, J.D., BEGER, R., BESSANT, C., CONNOR, S., CAPUANI, G., CRAIG, A., EBBELS, T., KELL, D.B., MANETTI, C., NEWTON, J., PATERNOSTRO, G., SOMORJAI, R., SJöSTRöM, M., TRYGG, J. & WULFERT, F. 2007. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, 3:231-241.

GRIFFIN, J.L. 2005. The Cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball? *Philosophical Transactions of the Royal Society B*, 361:147-161.

GRIFFIN,J.L., NICHOLLS, A.W., DAYKIN, C.A., HEALD, S., KEUN, H.C., SCHUPPE-KOISTINEN, I., GRIFFITHS, J.R., CHENG, L.L., ROCCA-SERRA, P., RUBTSOV, D.V. & ROBERTSON, D. 2007. Standard reporting requirements for biological samples in metabolomics experiments: mammalian/in vivo experiments. *Metabolomics*, 3:179-188.

GUNERAL, F. & BACHMANN, C. 1994. Age-related reference values for urinary organic acids in a healthy Turkish pediatric population. *Clinical Chemistry,* 40(6):862-868.

HADDEN, D.R. & MCLAUGHLIN, C. 2008. Normal and abnormal maternal metabolism during pregnancy. *Seminars in Fetal & Neonatal Medicine*, 14:66-71.

HAIR, J.F., BLACK, W.C., BABIN, B.J., ANDERSON, R.E. & TATHAM, R.L. 2006. Multivariate data analysis. 6th ed. New Jersey: Pearson. 899 p.

HARRIGAN, G.G. & GOODACRE, R. 2003. Metabolic profiling: its role in biomarker discovery and gene function analysis. London: Kluwer academic publishers.

HARRIGAN, G.G., BRACKET, D.J. & BOROS, L.G. 2005. Medicinal chemistry, metabolic profiling and drug discovery: a role for metabolic profiling in reverse pharmacology and chemical genetics. *Mini Reviews in Medicinal Chemistry*, 5: 13-20.

HOFFMAN, G.F. & FEYH, P. 2003. Organic acid analysis. [*In* Blau, N., Duran, M., Blaskovics, M.E. & Gibson, K.M. Physician's guide to the laboratory diagnosis of metabolic diseases. Springer, Berlin. p. 27-44.]

INTERGENETICS.CO.UK. 2002. Obligate heterozygotes. http://www.intergenetics.co.uk/glossary/o/obligate-heterozygote.html Date of Access: 25 May 2010.

JOHNSON, R.A. & WICHERN, D.W. 1998. Applied multivariate statistical analysis. 4th ed. New Jersey: Prentice-Hall. 816 p.

KALHAN, S.C. 2000. Protein metabolism in pregnancy. *American Journal of Clinical Nutrition*, 71:1249S–55S.

KATAJAMAA, M. & ORESIC, M. 2007. Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography A*, 1158: 318–328.

KELL, D.B. Metabolomics and systems biology: making sense of the soup. 2004. *Current opinion in microbiology*, 7: 296-307.

KETTANEH, N., BERGLUND, A. & WOLD, S. 2005. PCA and PLS with very large data sets. *Computational Statistics & Data Analysis*, 48:69-85.

KIND, T. 2003. Welcome to www.amdis.net. http://www.amdis.net/. Date of Access: 20 Nov. 2009.

KING, J.C. 2000. Physiology of pregnancy and nutrient metabolism. *The American Journal of Clinical Nutrition*, 71:1918S-1925S.

KOCHHAR, S., JACOBS, D.M., RAMADAN, Z., BERRUEX, F., FUERHOLZ, A. & FAY, L.B. 2006. Probing gender-specific metabolism differences in humans by nuclear magnetic resonance-based metabonomics. *Analytical Biochemistry*, 352:274-281.

KREGEL, K.C. & ZHANG, H.J. 2007. An integrated view of oxidative stress in aging: basic mechanisms, functional effects and pathological considerations. *The American Journal of Physiology – Regulatory, Integrative and Comparative Physiology*, 292: 18-36.

KUHARA, T. 2005. Metabolome profiling of human urine with capillary gas chromatography/mass spectrometry [*In* Metabolomics: The frontier of systems biology. Springer. p. 53-74.]

LAWRENCE, E. 2005. Henderson’s dictionary of biology. 13th ed. Pearson Education Limited. 748 p.

LEHOTAY, D. & CLARKE, T.R. 1995. Organic acidurias and related abnormalities. *Critical reviews in clinical laboratory sciences*, 32(4):377-429.

MARTINS, A.M., CAMACHO, D., SHUMAN, J., SHA, W., MENDES, P. & SHULAEV, V. 2004. A systems biology study of two distinct growth phases of *Saccharomyces cerevisiae* cultures. *Current genomics*, 5: 649-663.

MENDES, P. 2002. Emerging bioinformatics for the metabolome. *Briefings in Bionformatics*, 3(2):134-145.

MOCK, D.M., QUIRK, J.G. & MOCK, N.I. 2002. Marginal biotin deficiency during normal pregnancy. *American Journal of Clinical Nutrition*, 75:295-299.

MOCO, S., BINO, R.J., DE VOS, R.C.H. & VERVOORT, J. 2007. Metabolomics technologies and metabolite identification. *Trends in Analytical Chemistry*, 26(9):855-866.

NICHOLSON, J.K. & WILSON, I.D. 2003. Understanding ‘global’ systems biology: metabonomics and the continuum of metabolism. *Nature Reviews Drug Discovery,* 2: 668–676.

OATS, J. & ABRAHAM, S. 2005. Fundamentals of obstetrics and gynaecology. 8th ed. Edinburgh: Elsevier Mosby. 365 p.

OLIVER, S.G., WINSON, M.K., KELL, D.B. & BAGANZ, F. 1998. Systematic functional analysis of the yeast genome. *Tibtech*, 16:373-378.

PASIKANTI, K.K., HO, P.C. & CHAN, E.C.Y. 2008. Gas chromatography/mass spectrometry in metabolic profiling of biological fluids. *Journal of chromatography B*, 871:202-211.

ROESSNER, U., WAGNER, C., KOPKA, J., TRETHEWEY, R.N. & WILLMITZER, L. 2000. Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *The Plant Journal*, 23(1):131-142.

ROESSNER, U., LUEDEMANN, A., BRUST, D., FIEN, O., LINKE, T., WILLMITZER, L. & FERNIE, A.R. 2001. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *The plant cell*, 13:11-29.

ROSENBLATT, P.L. 2007. Menstrual cycle: Biology of the female reproductive system. http://www.merck.com/ Date of Access: 1 Nov. 2009.

SCALBERT, A., BRENNAN, L., FIEHN, O., HANKEMEIER, T., KRISTAL, B.S., VAN OMMEN, B., PUJOS-GUILLOT, E., VERHEIJ, E., WISHART, D. & WOPEREIS, S. 2009. Mass-spectrometry-based metabolomics: limitations and recommendation for future progress with particular focus on nutrition research. *Metabolomics*, 5:435-458.

SEYMOUR, C.A., THOMASON, M.J., CHALMERS, R.A., ADDISON, G.M., BAIN, M.D., COCKBURN, F., LITTLEJOHNS, P., LORD, J. & WILCOX, A.H. 1997. Newborn screening for inborn errors of metabolism: a systematic review. *Health Technology Assessment*, 1(11):16-18.

SHULAEV, V. 2006. Metabolomics technology and bioinformatics. *Briefings in bioinformatics*, 7(2):128-139.

SIGMA ALDRICH, INC. BSTFA + TMCS: Product Specification. http://www.sigmaaldrich.com/etc/medialib/docs/Aldrich/General_Information/bstfa_tmcs.Par.0001.File.tmp/bstfa_tmcs.pdf Date of Access: 21 Oct. 2009.

SILVERTHORN, D.U. 2010. Human physiology: an integrated approach. 5th ed. Pearson. 844-851 p.

STEIN, S.E. 1999. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of American Social Mass Spectrometry*, 10:770-781.

STEUER, R. 2006. Review: On the analysis and interpretation of correlations in metabolomic data. *Briefings in Bioinfromatics*, 7(2):151-158.

STROTT, C.A. 2002. Sulfonation and molecular action. *Endocrine Reviews*, 23(5):703-732.

THERMO FISHER SCIENTIFIC, INC. 2010. Instructions for BSTFA, N,O-Bis(trimethylsilyl)trifluoroacetamide. http://www.piercenet.com/files/0255dh5.pdf Date of Access: 21 Oct. 2009].

TOMITA, M. 2005. Overview. [*In* Metabolomics: The frontier of systems biology. Springer, p. 1-6.]

TROCHIM, W.M.K. 2006. Descriptive statistics. http://www.socialresearchmethods.net/kb/statdesc.php. Date of Access: 21 Nov. 2009.

USBIOTEK INTERNATIONAL, INC. The necessity of gender-specific reference ranges. http://www.usbiotek.com/Downloads/information/gender_ranges_advertisement.pdf Date of Access: 3 Feb. 2010].

VAN DEN BERG., R.A., HOEFSLOOT, H.C.J., WESTERHUIS, J.A., SMILDE, A.K. & VAN DER WERF, M.J. 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 4:142.

VAN DER WERF, M., TAKORS, R., SMEDSGAARD, J., NIELSEN, J., FERENCI, T., PORTAIS, J.C., WITTMANN. C., HOOKS, M., TOMASSINI, A., OLDIGES, M., FOSTEL, J. & SAUER, U. 2007. Standard reporting requirements for biological samples in metabolomics experiments: microbial and in vitro biology experiments. *Metabolomics*, 3:189-194.

VILLAS-BOAS, S.G., RASMUSSEN, S. & LANE, G.A. 2005. Metabolomics or metabolite profiles. *Trends in Biotechnology*, 23(8):385.

WEINERT, B.T. & TIMIRAS, P.S. 2003. Invited review: theories of aging. *Journal of Applied Physiology*, 96: 1706-1716.

WERNER, E., HEILIER, J., DUCRUIX, C., EZAN, E., JUNOT, C. & TABET, J. 2008. Mass spectrometry for the identification of the discriminating signals from metabolomics: Current status and future trends. *Journal of Chromatography B*, 871: 143–163.

WITTEN, T.A., LEVINE, S.P., KING, J.O. & MARKEY, S.P. 1973. Gas-chromatographic-mass-spectrometric determination of urinary acid profiles of normal young adults on a controlled diet. *Clinical Chemistry*, 19(6):586-589.

Appendix A

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

INFORMED CONSENT TO PARTICIPATE IN A RESEARCH STUDY

STUDY TITLE: **A Metabolomics Study of Selected Perturbations of Normal Metabolism**

INVESTIGATORS NAME: Me Elmarie Davoren

SUPERVISOR: Prof. C.J. Reinecke, Centrum for Human Metabonomics (Acting Head)

INVESTIGATOR SITE NAME & ADDRESS: School of Physical and Chemical Sciences
NWU, Potchefstroom Campus
Private Bag X6001
POTCHEFSTROOM 2520
South Africa
Tel: 018 299 2309
Fax: 018 293 5248

INTRODUCTION

The North-West University recently established a Centre for Human Metabonomics. **Metabonomics** is the most contemporary new field that emanated from Biochemistry and Molecular Biology. In the case of humans, it is the study of normal physiology as well as responses of the body towards invasions like diseases, drugs, viral infections and environmental changes. Metabonomics is an extension of the field of genomics (a comprehensive study of genomes - DNA) and metabolomics (a comprehensive study of low molecular weight biomolecules, commonly known as metabolites). The study of metabonomics significantly contributes to our understanding of the normal processes and of diseases in the human body, but can also lead to more efficient drug discovery and individualized patient treatment for inherited or acquired disorders. Knowledge of the normal profile of metabolites is important with regard to clinical studies. The new state-of-the-art equipment of the Centre of Human Metabonomics opens the possibility to study metabolism with an exceptional high degree of sensitivity and specificity.

PURPOSE OF THE STUDY

The aim of this study is to determine if and to what degree selected perturbations of normal metabolism may influence the metabolome. The perturbations under investigation include:

- The menstrual cycle.
- Pregnancy.
- Aging.

BIOLOGICAL SPECIMENS

The biological specimens required for this study is an early morning urine sample taken for a month period, beginning the day after menstrual bleeding has ended and ending when menstrual bleeding has begun. The participant will need to document the day when the samples were taken in order for the researchers to determine in which phase the samples fall.

INFORMED CONCENT PROCEDURES

Participation in the project is fully voluntary. You are free to enquire on the project through the researcher and/or supervisor if agreed to participate, the participant will be asked to sign this informed consent form when the urine sample is due to be collected. All information gained through the research will be available to the participant upon request.

BENEFITS ASSOCIATED WITH THE STUDY

This project is basic research and the primary benefit is a better understanding of a specific aspect of normal physiology. As in all cases, improved knowledge of the normal physiology eventually benefit a better understanding and treatment of any deviation from the normal physiology. The outcome of this research will be used by the researcher for a M.Sc.-thesis and no reference will be included in the thesis regarding any individual who participated in the study.

PAYMENT OR REIMBURSEMENT

Participants will not be paid for their participation and don’t contribute to the cost of the study.

CONFIDENTIALITY

All research records are confidential unless law requires disclosure. No name or other personal identifying information of the participant will be used in any reports or publications resulting from this study. Data from this study will be used in an anonymous statistical analysis and reported as such by the NWU. No patient identification detail will be reported or made known to other parties.

CONTACTS

If you have any questions about this study you may contact Prof. C.J. Reinecke, the supervisor of the study.

VOLUNTARY PARTICIPATION AND CONDITIONS OF WITHDRAWAL

Your participation in this study is completely voluntary. You may choose not to participate in this study to which you are otherwise entitled.

CONSENT

I, ____ have read and understood the preceding information describing this research study and my questions have been answered to my satisfaction. "I voluntary consent to participate in this research study, as set forth under the "Confidentiality" section.

I do not waive my legal rights by signing this consent form. I will receive a signed and dated copy of this consent form.

I would like feedback on the outcome of our contribution to the research: YES / NO "

PARTICIPANT:

__________________________ ______________________ ________________

Printed name **Signature** **Date**

Appendix B

INFORMED CONSENT TO PARTICIPATE IN A RESEARCH STUDY

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

STUDY TITLE: **A Metabolomics Study of Selected Perturbations of Normal Human Metabolism**

INVESTIGATORS NAME: Me Elmarie Davoren

SUPERVISOR: Prof. C.J. Reinecke, Centrum for Human Metabonomics (Acting Head)

INVESTIGATOR SITE NAME & ADDRESS: School of Physical and Chemical Sciences
NWU, Potchefstroom Campus
Private Bag X6001
POTCHEFSTROOM 2520
South Africa
Tel: 018 299 2309
Fax: 018 293 5248

INTRODUCTION

The North-West University recently established a Centre for Human Metabonomics. **Metabonomics** is the most contemporary new field that emanated from Biochemistry and Molecular Biology. In the case of humans, it is the study of normal physiology as well as responses of the body towards invasions like diseases, drugs, viral infections and environmental changes. Metabonomics is an extension of the field of genomics (a comprehensive study of genomes - DNA) and metabolomics (a comprehensive study of low molecular weight biomolecules, commonly known as metabolites). The study of metabonomics significantly contributes to our understanding of the normal processes and of diseases in the human body, but may also lead to more efficient drug discovery and individualized patient treatment for inherited or acquired disorders. One of the projects in the Centre deals with detoxification. Sulfonation forms part of the Phase II detoxification pathway in the liver, which is catalyzed by the cytosolic sulfotransferases (SULTs) enzymes. SULTs are of the most important detoxification enzymes in the endogenous substances, like the thyroid hormones, in the foetal metabolism It is well established that all humans contain SULT1A1 genes of one of three kinds (known as polymorphisms):

- Homozygous (wild type allele)
- Heterozygous (both wild and polymorphic alleles)
- Homozygous (both alleles are polymorphic)

Knowledge of the normal profile of metabolites is important with regard to clinical studies. The new state-of-the-art equipment of the Centre of Human Metabonomics opens the possibility to study metabolism with an exceptional high degree of sensitivity and specificity.

PURPOSE OF THE STUDY

The aim of this study is to determine if and to what degree selected perturbations of normal metabolism may influence the metabolome. The perturbations under investigation include:

- The menstrual cycle.
- Pregnancy.
- Aging.

BIOLOGICAL SPECIMENS

The biological specimens required for this study is a blood and urine sample from pregnant and non-pregnant women. All samples will be supplied anonymously to the research laboratory. All information on the mothers will be handled confidentially by the physician, and no information on the name of the mother will be given to the researchers. Information on the stage (in weeks) of the pregnancy and the ethnicity will be provided to the researchers, as this is important for the analysis of the data.

INFORMED CONCENT PROCEDURES

Participation in the project is fully voluntary. You are free to enquire on the project through the physician and if agreed to participate, the participating mother will be asked to sign this informed consent form when the blood and urine samples are due to be collected. All samples (early morning) will be collected at the laboratory of DuBuisson, Bruinette and Kramer, Inc., at Mooi Med Hospital or Patcare at MediClinic. The form for informed consent, as well as the outcome of the analysis will be reported by the researchers to the physician who will file it for record purposes. All information gained through the research will be available to the participants upon request. This can be provided by the physician or from the researchers, according to their preference.

BENEFITS ASSOCIATED WITH THE STUDY

This project is basic research and the primary benefit is a better understanding of a specific aspect of the normal physiology during pregnancy. As in all cases, improved knowledge of the normal physiology eventually benefit a better understanding and treatment of any deviation from the normal physiology. The outcome of this research will be used by the researcher for a M.Sc.-thesis and no reference will be included in the thesis regarding any individual who participated in the study.

PAYMENT OR REIMBURSEMENT

Participants will not be paid for their participation and don’t contribute to the cots of the study.

CONFIDENTIALITY

All research records are confidential unless law requires disclosure. No name or other personal identifying information of the participating mothers will be used in any reports or publications resulting from this study. Data from this study will be used in an anonymous statistical analysis and reported as such by the NWU. No patient identification detail will be reported or made known to other parties.

CONTACTS

If you have any questions about this study, you may contact the physician who will take this up with Prof. C.J. Reinecke, the supervisor of the study.

VOLUNTARY PARTICIPATION AND CONDITIONS OF WITHDRAWAL

Your participation in this study is completely voluntary. You may choose not to participate in this study to which you are otherwise entitled.

CONSENT

I, ____ have read and understood the preceding information describing this research study and my questions have been answered to my satisfaction. "I voluntary consent to participate in this research study, as set forth under the "Confidentiality" section.

I do not waive my legal rights by signing this consent form. I will receive a signed and dated copy of this consent form.

I would like feedback on the outcome of our contribution to the research: YES / NO "

PARTICIPANT:

__________________________ ______________________ ________________

Printed name **Signature** **Date**

PHYSICIAN:

________________________ _____________________ _______________

Printed name **Signature** **Date**

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

Questionnaire of Clinical Information of Participant in a Masters Study - A Metabolomics Study of Selected Perturbations of Normal Human Metabolism

1. How many weeks are you pregnant: ________________
2. Age (Date of birth): ________________
3. Ethnicity: _____________
4. Medical History (any diagnosed medical problems for example: diabetes, allergies, heart disease etc.):

   ____

   ____

   ____

   ____

5. Do you currently take any medication? _____________________
   a. If so please name:

      ____
      ____
      ____
      ___

PARTICIPANT:

______________________ __________________ ________________

Printed name **Signature**

Appendix C

INFORMED CONSENT TO PARTICIPATE IN A RESEARCH STUDY

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

STUDY TITLE: **A Metabolomics Study of Selected Perturbations of Normal Metabolism**

INVESTIGATORS NAME: Me Elmarie Davoren

SUPERVISOR: Prof. C.J. Reinecke, Centrum for Human Metabonomics (Acting Head)

INVESTIGATOR SITE NAME & ADDRESS: School of Physical and Chemical Sciences
NWU, Potchefstroom Campus
Private Bag X6001
POTCHEFSTROOM 2520
South Africa
Tel: 018 299 2309
Fax: 018 293 5248

INTRODUCTION

The North-West University recently established a Centre for Human Metabonomics. **Metabonomics** is the most contemporary new field that emanated from Biochemistry and Molecular Biology. In the case of humans, it is the study of normal physiology as well as responses of the body towards invasions like diseases, drugs, viral infections and environmental changes. Metabonomics is an extension of the field of genomics (a comprehensive study of genomes - DNA) and metabolomics (a comprehensive study of low molecular weight biomolecules, commonly known as metabolites). The study of metabonomics significantly contributes to our understanding of the normal processes and of diseases in the human body, but can also lead to more efficient drug discovery and individualized patient treatment for inherited or acquired disorders. Knowledge of the normal profile of metabolites is important with regard to clinical studies. The new state-of-the-art equipment of the Centre of Human Metabonomics opens the possibility to study metabolism with an exceptional high degree of sensitivity and specificity.

PURPOSE OF THE STUDY

The aim of this study is to determine if and to what degree selected perturbations of normal metabolism may influence the metabolome. The perturbations under investigation include:

- The menstrual cycle.
- Pregnancy.
- Aging.

BIOLOGICAL SPECIMENS

The biological specimens required for this study is an early morning urine sample.

INFORMED CONCENT PROCEDURES

Participation in the project is fully voluntary. You are free to enquire on the project through the researcher and/or supervisor if agreed to participate, the participant will be asked to sign this informed consent form when the urine sample is due to be collected. All information gained through the research will be available to the participant upon request.

BENEFITS ASSOCIATED WITH THE STUDY

This project is basic research and the primary benefit is a better understanding of a specific aspect of normal physiology. As in all cases, improved knowledge of the normal physiology eventually benefit a better understanding and treatment of any deviation from the normal physiology. The outcome of this research will be used by the researcher for a M.Sc.-thesis and no reference will be included in the thesis regarding any individual who participated in the study.

PAYMENT OR REIMBURSEMENT

Participants will not be paid for their participation and don't contribute to the cost of the study.

CONFIDENTIALITY

All research records are confidential unless law requires disclosure. No name or other personal identifying information of the participant will be used in any reports or publications resulting from this study. Data from this study will be used in an anonymous statistical analysis and reported as such by the NWU. No patient identification detail will be reported or made known to other parties.

CONTACTS

If you have any questions about this study you may contact Prof. C.J. Reinecke, the supervisor of the study.

VOLUNTARY PARTICIPATION AND CONDITIONS OF WITHDRAWAL

Your participation in this study is completely voluntary. You may choose not to participate in this study to which you are otherwise entitled.

CONSENT

I, ______________________________________ have read and understood the preceding information describing this research study and my questions have been answered to my satisfaction. “I voluntary consent to participate in this research study, as set forth under the “Confidentiality” section.

I do not waive my legal rights by signing this consent form. I will receive a signed and dated copy of this consent form.

I would like feedback on the outcome of our contribution to the research: YES / NO “

PARTICIPANT:

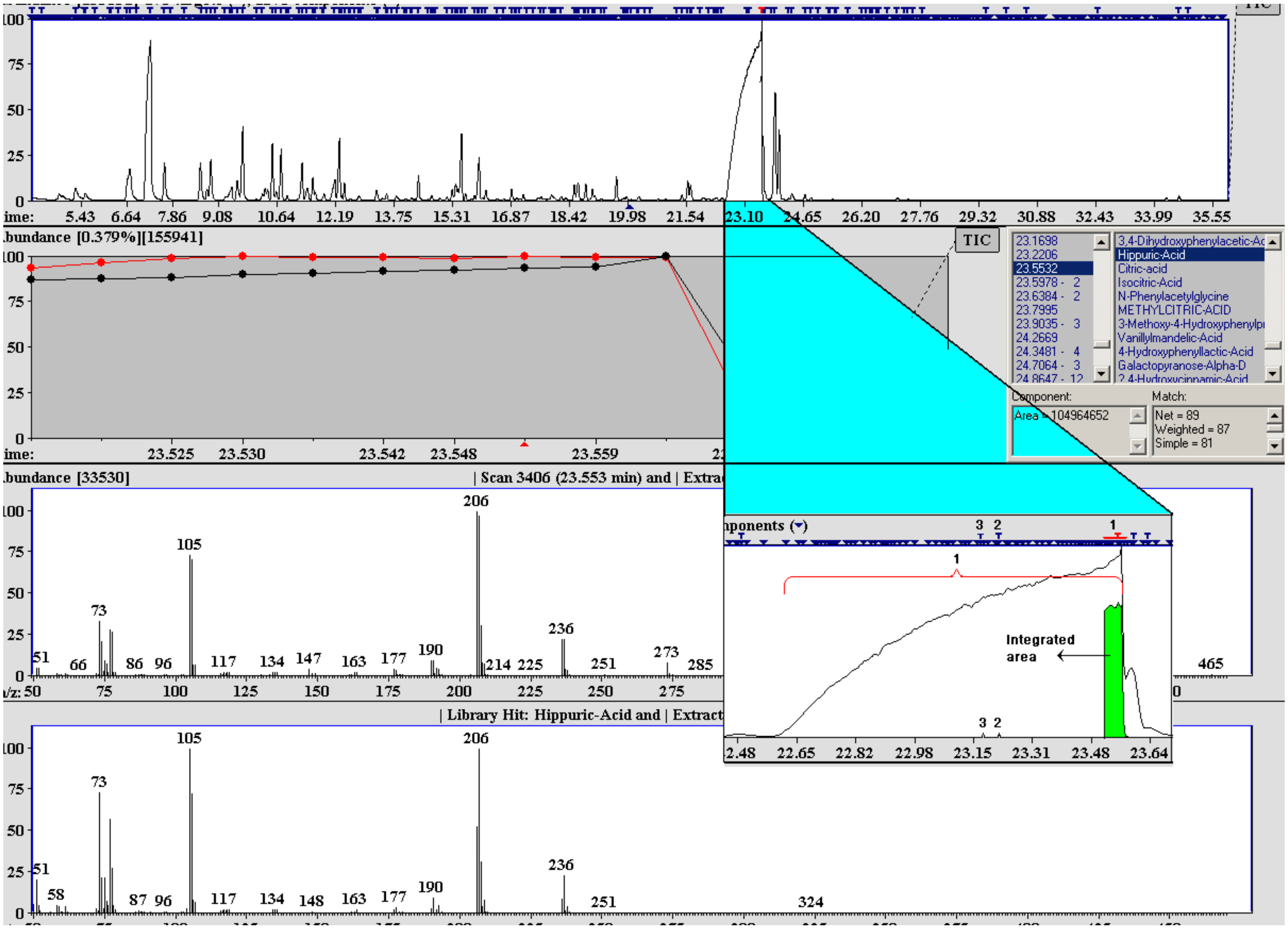
Printed name   **Signature**   **Date**

Appendix D

As stated in Section 2.5.2 the data analysis process in AMDIS involves noise analysis, component perception, spectrum deconvolution and compound identification. I will discuss these steps briefly: the signal characteristics are extracted from the data file, individual chromatographic components are perceived and a model peak shape for each component is determined, followed by “purified” spectra extracted from the individual ion chromatograms by means of the model shape. Finally the extracted spectra are compared to the spectra in the reference library, after which a “hit list” is produced (Stein, S.E., 1999). Deconvolution will be discussed in broader terms, since it is an important step with regards to metabolite identification.

Deconvolution includes the subtraction of signals from nearby components by means of their model peak characteristic i.e. the subtracted spectra of two adjacent spectra in AMDIS are determined. This method makes it possible to identify two compounds that contaminate i.e overlap each other on the chromatogram. Hence AMDIS can be used to “separate” more than one compound with the same retention time, which leads to better identification of the total metabolite profile. Deconvolution influences the integration (abundance) of peaks negatively, seeing that compounds with a high concentration value has broad spectrums. Subsequently, the extracted spectra of the same compound is determined which leads to a spectra with no ions. These spectra are not recognised as a compound, it is not integrated and only the last spectra of the compound/ion chromatogram will give a recognizable ion profile. This ion profile will then be integrated.

In section 3.6 I showed the AMDIS information on 3-hydroxypyridine that occurs in very low concentrations in urine. For this discussion I have chosen the information on hippuric acid, which occurs in high concentrations in urine. Figure D.1 shows that, hippuric acid (number 1) gives a broad spectrum due to its high excretion value in urine. AMDIS subtracts the ions of one component from the adjacent component of the same peak, this leads to spectra with zero ions. This subtraction process takes place until an ion is subtracted that does not give a zero ion profile. Hence the most prominent of these ion chromatograms is used as the actual chromatographic component and is integrated (indicated in green). Two compounds (numbers 2 and 3) were still identified during this subtraction process, given that they had different model peak shapes.



100
75
50
25
0
ime: 5.43 6.64 7.86 9.08 10.64 12.19 13.75 15.31 16.87 18.42 19.98 21.54 23.10 24.65 26.20 27.76 29.32 30.88 32.43 33.99 35.55
bundance [0.379%][155941]
TIC
ime: 23.525 23.530 23.542 23.548 23.559
bundance [33530]
| Scan 3406 (23.553 min) and | Extrac
51 66 73 86 96 105 117 134 147 163 177 190 206 214 225 236 251 273 285 465
/z: 50 75 100 125 150 175 200 225 250 275
nponents
3 2 1
Integrated area
2.48 22.65 22.82 22.98 23.15 23.31 23.48 23.64
| Library Hit: Hippuric-Acid and | Extract
51 58 73 87 96 105 117 134 148 163 177 190 206 236 251 324

23.1698
23.2206
23.5532
23.5978 - 2
23.6384 - 2
23.7995
23.9035 - 3
24.2669
24.3481 - 4
24.7064 - 3
24.8647 - 12

3,4-Dihydroxyphenylacetic-Ac
Hippuric-Acid
Citric-acid
Isocitric-Acid
N-Phenylacetylglycine
METHYLCITRIC-ACID
3-Methoxy-4-Hydroxyphenylp
Vanillylmandelic-Acid
4-Hydroxyphenyllactic-Acid
Galactopyranose-Alpha-D
2,4-Hydroxycinnamic-Acid

Component:
Area = 104964652

Match:
Net = 89
Weighted = 87
Simple = 81

Figure D1: Example of the deconvolution of hippuric acid.

Appendix E

Succinic acid

Biofluid	**CSF**
Value	150.0 +/- 110.0 uM
Age	Elderly:>65 yrs old
Sex	Both
Patient information	Normal
Comments	Not Available
References	• Redjems-Bennani N, Jeandel C, Lefebvre E, Blain H, Vidailhet M, Gueant JL: Abnormal substrate levels that depend upon mitochondrial function in cerebrospinal fluid from Alzheimer patients. Gerontology. 1998;44(5):300-4. [PubMed]

Biofluid	**CSF**
Value	3.0 +/- 2.0 uM
Age	Adult:>18 yrs old
Sex	Both
Patient information	Normal
Comments	Not Available
References	• Hoffmann GF, Meier-Augenstein W, Stockler S, Surtees R, Rating D, Nyhan WL: Physiology and pathophysiology of organic acids in cerebrospinal fluid. J Inherit Metab Dis. 1993;16(4):648-69. [PubMed]

Biofluid	**Saliva**
Value	2260 (60.0-4460) uM
Age	Adult:>18 yrs old
Sex	Both
Patient information	Normal
Comments	Not Available
References	• Silwood CJ, Lynch E, Claxson AW, Grootveld MC: 1H and (13)C NMR spectroscopic analysis of human saliva. J Dent Res. 2002 Jun;81(6):422-7. [PubMed]

Biofluid	**Urine**
Value	7.7 (1.9-20.0) umol/mmol creatinine
Age	Adolescent:13-18 yrs old
Sex	Both
Patient information	Normal
Comments	Not Available
References	• Guneral F, Bachmann C: Age-related reference values for urinary

organic acids in a healthy Turkish pediatric population. Clin Chem. 1994 Jun;40(6):862-6. [PubMed]

Biofluid	**Urine**
Value	197.2 (29.4-486.2) umol/mmol creatinine
Age	Newborn:0-30 days old
Sex	Both
Patient information	Normal
Comments	Not Available
References	• Guneral F, Bachmann C: Age-related reference values for urinary organic acids in a healthy Turkish pediatric population. Clin Chem. 1994 Jun;40(6):862-6. [PubMed]

Biofluid	**Urine**
Value	185.4 (6.0-342.6) umol/mmol creatinine
Age	Infant:0-1 yr old
Sex	Both
Patient information	Normal
Comments	Not Available
References	• Guneral F, Bachmann C: Age-related reference values for urinary organic acids in a healthy Turkish pediatric population. Clin Chem. 1994 Jun;40(6):862-6. [PubMed]

Appendix F

Table F.1: Log-scaled dataset with no smoothing.

Metabolite	1 vs. 2	1 vs. 3	2 vs. 3	Maximum effect size	No. of phases with an effect size higher than 0.8
2,5-Furandicarboxylic acid	0.015	0.856	0.988	0.988	2
2,3-Dihydroxybutanoic acid	0.242	0.694	0.603	0.694	0
Palmitic acid	0.520	0.692	0.340	0.692	0
Octadecanoic acid	0.490	0.659	0.293	0.659	0
3-Hydroxyphenylacetic acid	0.659	0.377	0.291	0.659	0
Citramalic acid	0.448	0.656	0.209	0.656	0
3-Hydroxyisobutyric acid	0.087	0.398	0.643	0.643	0
2,5-Dihydroxybenzoic acid	0.632	0.021	0.496	0.632	0
4-Hydroxymandelic acid	0.061	0.478	0.624	0.624	0
Lactic acid	0.377	0.621	0.309	0.621	0
Fumaric acid	0.507	0.616	0.148	0.616	0
Glyceric acid	0.208	0.608	0.493	0.608	0
3-Methoxy-4-hydroxycinnamic acid	0.117	0.376	0.604	0.604	0
Glycolic acid	0.241	0.239	0.563	0.563	0
3-Hydroxysebacic acid	0.256	0.544	0.401	0.544	0
3-Hydroxyadipyllactone	0.314	0.540	0.356	0.540	0
3-Methyladipic acid	0.248	0.529	0.529	0.529	0
Octenedioic acid	0.528	0.493	0.083	0.528	0
4-Hydroxyphenylacetic acid	0.055	0.473	0.522	0.522	0
Methylsuccinic acid	0.130	0.423	0.520	0.520	0
3-Hydroxyphenylpropionic acid	0.043	0.499	0.349	0.499	0

Appendix G

Table G.1: Log-scaled dataset with three point smoothing based on median effect size.

Metabolite	1 vs. 2	1 vs. 3	2 vs. 3	Maximum effect size	No. of phases with an effect size higher than 0.8
Lactic acid	0.910	1.822	2.722	2.722	3
Phosphoric acid	1.504	2.185	0.853	2.185	3
Succinic acid	1.674	1.012	0.879	1.674	3
Palmitic acid	0.738	1.244	3.095	3.095	2
2,3-Dihydroxybutanoic acid	0.142	1.877	3.078	3.078	2
Octadecanoic acid	0.634	1.256	2.949	2.949	2
Ethylmalonic acid	2.901	0.865	0.286	2.901	2
Isocitric lactone	0.463	1.534	2.295	2.295	2
3-Methoxy-4-hydroxycinnamic acid	1.452	0.195	2.292	2.292	2
Adipic acid	1.162	0.561	2.192	2.192	2
2,5-Furandicarboxylic acid	0.725	1.415	2.050	2.050	2
4-Hydroxymandelic acid	0.554	0.805	1.815	1.815	2
Oxalic acid	1.752	0.293	0.945	1.752	2
4-Hydroxyhippuric acid	0.381	1.556	1.193	1.556	2
Fumaric acid	0.668	1.450	1.488	1.488	2
2,5-Dihydroxybenzoic acid	1.471	0.586	1.464	1.471	2
3-Hydroxyphenylacetic acid	1.098	1.372	0.533	1.372	2
3-Methyladipic acid	0.163	1.311	1.359	1.359	2
Citric acid	1.284	0.761	1.240	1.284	2
4-Hydroxyphenyllactic acid	0.552	0.869	1.272	1.272	2
Glyceric acid	0.299	1.215	1.172	1.215	2
Citramalic acid	1.194	1.211	0.331	1.211	2
Uracil	1.136	0.316	0.893	1.136	2
Pimelic acid	1.088	0.378	0.843	1.088	2
3-Hydroxyadipyllactone	0.014	0.885	1.038	1.038	2
2-Hydroxyglutaric acid	0.839	0.189	1.024	1.024	2
3-Hydroxyphenylpropionic acid	0.910	0.910		0.910	2
2-Ketoglutaric acid		0.800	0.800	0.800	2
3-Hydroxyisobutyric acid	0.694	0.734	3.327	3.327	1
Glycolic acid	0.608	0.517	2.574	2.574	1
3,4-Dihydroxybutyric acid	0.649	0.175	2.142	2.142	1
3-Hydroxyisvaleric acid	0.290	0.279	2.056	2.056	1
Erythronic acid	0.645	0.756	1.450	1.450	1
2-Methyl-3-hydroxybutyric acid	0.519	0.232	1.252	1.252	1

Pyroglutamic acid	0.647	0.568	1.109	1.109	1
Methylsuccinic acid	0.544	0.512	1.104	1.104	1
2-Hydroxyglutaryllactone	0.433	0.712	1.034	1.034	1
3-Hydroxy-3-methylglutaric acid	0.484	0.245	1.015	1.015	1
4-Hydroxyphenylacetic acid	0.428	0.653	0.996	0.996	1
3,5-Dihydroxybenzoic acid	0.786	0.184	0.989	0.989	1
Hippuric acid	0.688	0.967	0.245	0.967	1
Vanillylmandelic acid	0.122	0.872	0.683	0.872	1
3-Hydroxypropionic acid	0.092	0.864	0.788	0.864	1
Glycerol	0.643	0.855	0.508	0.855	1
2,3,4-Trihydroxybutyric acid	0.386	0.697	0.851	0.851	1
N-isovalerylglycine	0.484	0.839	0.583	0.839	1
3-Methylglutaconic acid	0.432	0.838	0.633	0.838	1
4-Hydroxbenzoic acid	0.019	0.778	0.830	0.830	1
1,2-Dihydroxyethane		0.800	0.800	0.800	0
N-acetylaspartic acid	0.525	0.729	0.295	0.729	0
Glutaric acid	0.335	0.315	0.702	0.702	0
Octenedioic acid	0.645	0.645		0.645	0
Methylcitric acid	0.165	0.606	0.446	0.606	0
Phenylacetylglutamine	0.533	0.270	0.328	0.533	0
Hydantoinpropionic acid	0.512	0.008	0.519	0.519	0
Benzoic acid	0.403	0.452	0.046	0.452	0

Table G.2: Log-scaled dataset with three point smoothing based on mean effect size

Metabolite	1 vs. 2	1 vs. 3	2 vs. 3	Maximum effect size	No. of phases with an effect size higher than 0.8
2,3-Dihydroxybutanoic acid	0.837	3.532	3.309	3.532	3
2,5-Dihydroxybenzoic acid	2.043	1.022	3.347	3.347	3
3-Hydroxyadipyllactone	0.907	2.263	1.807	2.263	3
Lactic acid	1.495	2.177	1.047	2.177	3
Monostearylglycerol	0.859	1.020	1.792	1.792	3
Levulinic acid	1.540	0.922	1.746	1.746	3
3-Hydroxy-3-methylglutaric acid	0.846	1.459	0.892	1.459	3
2-Hydroxyhippuric acid	1.026	0.913	1.382	1.382	3
Palmitic acid	1.013	1.288	1.153	1.288	3
3-Hydroxyglutaric acid	1.076	1.250	1.095	1.250	3
Uracil	0.819	1.130	0.988	1.130	3

Methylmalonic acid	1.083	0.895	1.076	1.083	3
Benzoic acid	1.032	0.870	1.026	1.032	3
3-Methoxy-4-hydroxycinnamic acid	0.766	0.862	3.494	3.494	2
N-acetylaspartic acid	0.009	1.328	3.475	3.475	2
Azelaic acid	2.506	0.246	3.069	3.069	2
3-Hydroxyphenylpropionic acid	0.255	3.056	1.186	3.056	2
Glycolic acid	0.424	0.865	2.917	2.917	2
3-Methyladipic acid	0.382	1.579	2.882	2.882	2
Methylsuccinic acid	0.054	0.816	2.550	2.550	2
3-Hydroxyisobutyric acid	0.106	1.343	2.307	2.307	2
Glyceric acid	0.495	2.181	1.524	2.181	2
4-Hydroxymandelic acid	0.023	1.429	2.170	2.170	2
3-Hydroxysebacic acid	0.469	1.726	2.144	2.144	2
Adipic acid	0.074	1.346	1.952	1.952	2
2,5-Furandicarboxylic acid	0.174	1.772	1.942	1.942	2
Pyroglutamic acid	0.151	1.579	1.942	1.942	2
Isocitric lactone	0.512	1.906	1.867	1.906	2
2,4-Dihydroxybutyric acid	1.190	1.882	0.419	1.882	2
4-Hydroxyphenyllactic acid	1.197	1.831	0.716	1.831	2
1,2-Dihydroxyethane	1.821	0.686	1.049	1.821	2
N-hexanoylglycine	1.774	1.123	0.392	1.774	2
Fumaric acid	1.355	1.707	0.777	1.707	2
N-acetyltrheonine	0.355	1.139	1.672	1.672	2
Oxalic acid	1.086	1.617	0.539	1.617	2
N-tiglylglycine	1.066	0.246	1.576	1.576	2
Phosphoric acid	0.267	1.568	1.355	1.568	2
N-isovalerylglycine	0.452	1.513	1.212	1.513	2
Nonanoic acid	0.448	1.429	1.124	1.429	2
Methylcitric acid	0.335	1.414	1.189	1.414	2
Octenedioic acid	1.204	1.412	0.768	1.412	2
3,4-Dihydroxybenzoic acid	1.380	0.198	1.403	1.403	2
3-Hydroxyphenylacetic acid	1.354	0.528	1.024	1.354	2
Octadecanoic acid	0.980	1.317	0.742	1.317	2
Hydantoinpropionic acid	0.455	0.863	1.307	1.307	2
4-Hydroxyphenylacetic acid	0.166	1.307	1.231	1.307	2
1,2-Dihydroxypropane	1.168	1.222	0.447	1.222	2
Glycerol	0.162	1.205	1.055	1.205	2
2,3,4-Trihydroxybutyric acid	0.154	0.979	1.175	1.175	2
2-Keto-octanoic acid	0.750	1.139	0.892	1.139	2
Glutaric acid	0.822	1.121	0.659	1.121	2
Oleic acid	0.873	1.099	0.777	1.099	2
2-Hydroxysebacic acid	1.095		1.095	1.095	2
2-Ketoisovaleric acid	1.095		1.095	1.095	2

4-Hydroxyhippuric acid	0.306	0.927	1.081	1.081	2
3-Methoxy-4-Hydroxyphenylpropionic acid	1.066	0.492	0.966	1.066	2
3-Hydroxypropionic acid	0.171	0.918	0.962	0.962	2
Citramalic acid	0.926	0.910	0.323	0.926	2
4-Hydroxybutyric acid	0.851	0.079	0.817	0.851	2
2-Hydroxybenzoic acid		0.802	0.802	0.802	2
3-Hydroxypyridine	0.505	0.291	4.076	4.076	1
Ethylmalonic acid	1.786	0.616	0.192	1.786	1
3,5-Dihydroxybenzoic acid	0.353	0.638	1.629	1.629	1
Isocitric acid	0.540	1.398	0.796	1.398	1
Succinic acid	1.394	0.183	0.377	1.394	1
Pimelic acid	0.360	0.328	1.287	1.287	1
2-Methyl-2-hydroxybutyric acid	0.793	0.202	1.100	1.100	1
Citraconic acid	0.241	0.735	1.096	1.096	1
Octanoic acid	0.407	0.678	1.062	1.062	1
2-Hydroxybutyric acid	0.972	0.715	0.355	0.972	1
2-Ketoglutaric acid	0.650	0.969	0.432	0.969	1
2-Methyl-3-ketobutyric acid	0.730	0.960	0.662	0.960	1
Malic acid	0.308	0.917	0.469	0.917	1
N-acetyltyrosine	0.845	0.376	0.741	0.845	1
2-Hydroxyglutaric acid	0.817	0.170	0.796	0.817	1
3-Methylpimelic acid	0.810	0.437	0.365	0.810	1
3-Hydroxyadipic acid	0.447	0.798	0.407	0.798	0
2-Hydroxy-3-methylvaleric acid	0.790	0.790		0.790	0
4-Ketovaleric acid	0.363	0.519	0.772	0.772	0
2-Hydroxyisovaleric acid	0.585	0.771	0.121	0.771	0
3-Hydroxyisovaleric acid	0.205	0.205	0.768	0.768	0
2-Methyl-3-hydroxybutyric acid	0.242	0.582	0.763	0.763	0
Dodecamethylpentasiloxane		0.738	0.738	0.738	0
2-Hydroxyphenylacetic acid	0.287	0.130	0.716	0.716	0
Maleic acid	0.064	0.713	0.568	0.713	0
Aconitic acid	0.078	0.598	0.671	0.671	0
Vanillylmandelic acid	0.335	0.656	0.603	0.656	0
2-Hydroxy-3-methylbutryic acid	0.049	0.610	0.589	0.610	0
2-Ketobutyric acid	0.447	0.586	0.176	0.586	0
2,3-Dihydroxybutane	0.006	0.561	0.370	0.561	0
Erythronic acid	0.001	0.558	0.484	0.558	0
Malonic acid	0.482	0.015	0.556	0.556	0
Treonic acid	0.258	0.382	0.550	0.550	0
3-Methylglutaconic acid	0.337	0.528	0.544	0.544	0
Suberic acid	0.326	0.197	0.524	0.524	0
2-Hydroxyisobutyric acid	0.287	0.451	0.521	0.521	0
Pyruvic acid	0.497	0.091	0.439	0.497	0

Appendix H

Table H.1: Dataset, not log-scaled with three point smoothing based on median effect size.

Metabolite	1 vs. 2	1 vs. 3	2 vs. 3	Maximum effect size	No. of phases with an effect size higher than 0.8
Palmitic acid	0.710	1.017	3.250	3.250	2
2,5-Furandicarboxylic acid	0.586	1.092	3.218	3.218	2
Ethylmalonic acid	3.211	0.827	0.201	3.211	2
Octadecanoic acid	0.640	1.037	3.073	3.073	2
2,3-Dihydroxybutanoic acid	0.176	1.613	2.786	2.786	2
Isocitric lactone	0.437	1.421	2.485	2.485	2
3-Methoxy-4-hydroxycinnamic acid	1.460	0.199	2.343	2.343	2
Adipic acid	1.187	0.556	2.219	2.219	2
Phosphoric acid	1.465	2.003	0.738	2.003	2
Oxalic acid	1.616	0.136	0.858	1.616	2
2,5-Dihydroxybenzoic acid	1.304	0.565	1.546	1.546	2
Fumaric acid	0.664	1.387	1.464	1.464	2
4-Hydroxyhippuric acid	0.364	1.115	1.419	1.419	2
3-Methyladipic acid	0.187	1.190	1.375	1.375	2
Citramalic acid	1.179	1.357	0.266	1.357	2
Citric acid	1.285	0.745	1.289	1.289	2
3-Hydroxyphenylacetic acid	1.025	1.239	0.507	1.239	2
4-Hydroxyphenyllactic acid	0.548	0.840	1.221	1.221	2
Succinic acid	1.212	0.789	0.866	1.212	2
Glyceric acid	0.310	1.199	1.147	1.199	2
Uracil	1.109	0.325	0.872	1.109	2
Pimelic acid	1.085	0.378	0.843	1.085	2
3-Hydroxyadipyllactone	0.013	0.875	1.015	1.015	2
2-Hydroxyglutaric acid	0.816	0.156	1.007	1.007	2
3-Hydroxyphenylpropionic acid	0.910	0.910		0.910	2
3-Hydroxyisobutyric acid	0.660	0.734	2.936	2.936	1
Glycolic acid	0.552	0.571	2.482	2.482	1
3,4-Dihydroxybutyric acid	0.634	0.220	2.092	2.092	1
3-Hydroxyisovaleric acid	0.181	0.384	1.999	1.999	1
4-Hydroxymandelic acid	0.525	0.767	1.664	1.664	1
Erythronic acid	0.645	0.759	1.447	1.447	1
2-Methyl-3-hydroxybutyric acid	0.464	0.272	1.330	1.330	1
Methylsuccinic acid	0.528	0.511	1.066	1.066	1
Pyroglutamic acid	0.659	0.503	1.063	1.063	1
4-Hydroxyphenylacetic acid	0.471	0.653	1.039	1.039	1

2-Hydroxyglutaryllactone	0.418	0.702	1.027	1.027	1
Hippuric acid	0.595	1.003	0.105	1.003	1
3-Hydroxy-3-methylglutaric acid	0.426	0.265	0.985	0.985	1
3,5-Dihydroxybenzoic acid	0.761	0.212	0.965	0.965	1
3-Hydroxypropionic acid	0.103	0.888	0.794	0.888	1
N-isovalerylglycine	0.427	0.847	0.570	0.847	1
2,3,4-Trihydroxybutyric acid	0.372	0.671	0.844	0.844	1
Glycerol	0.647	0.829	0.496	0.829	1
2-Ketoglutaric acid		0.800	0.800	0.800	0
1,2-Dihydroxyethane		0.799	0.799	0.799	0
3-Methylglutaconic acid	0.464	0.778	0.564	0.778	0
N-acetylaspartic acid	0.552	0.770	0.328	0.770	0
Vanillylmandelic acid	0.119	0.759	0.697	0.759	0
4-Hydroxbenzoic acid	0.027	0.727	0.742	0.742	0
Glutaric acid	0.317	0.311	0.682	0.682	0
Octenedioic acid	0.645	0.645		0.645	0
Methylcitric acid	0.174	0.618	0.449	0.618	0
Hydantoinpropionic acid	0.560	0.025	0.580	0.580	0
Phenylacetylglutamine	0.527	0.237	0.362	0.527	0
4-Ketovaleric acid	0.459	0.346	0.185	0.459	0

Table H.2: Dataset, not log-scaled with three point smoothing based on mean effect size.

Metabolite	1 vs. 2	1 vs. 3	2 vs. 3	Maximum effect size	No. of phases with an effect size higher than 0.8
2,3-Dihydroxybutanoic acid	0.818	2.927	3.710	3.710	3
2,5-Dihydroxybenzoic acid	1.941	1.108	3.138	3.138	3
Isocitric acid	0.825	2.450	1.100	2.450	3
Lactic acid	1.452	1.986	1.024	1.986	3
3-Hydroxyadipyllactone	0.810	1.805	1.682	1.805	3
Levulinic acid	1.527	0.929	1.800	1.800	3
Monostearylglycerol	0.861	1.021	1.766	1.766	3
3-Hydroxyglutaric acid	1.053	1.192	1.095	1.192	3
Palmitic acid	0.858	0.996	1.077	1.077	3
Methylmalonic acid	1.074	0.885	1.029	1.074	3
2-Hydroxyhippuric acid	1.066	0.837	0.980	1.066	3
Benzoic acid	0.954	0.893	0.920	0.954	3
3-Methoxy-4-hydroxycinnamic acid	0.751	0.846	3.638	3.638	2
N-acetylaspartic acid	0.065	1.274	3.149	3.149	2
3-Methyladipic acid	0.390	1.416	3.026	3.026	2

3-Hydroxyphenylpropionic acid	0.281	2.983	1.184	2.983	2
2,5-Furandicarboxylic acid	0.006	1.178	2.904	2.904	2
Azelaic acid	2.373	0.252	2.590	2.590	2
Glycolic acid	0.331	0.823	2.418	2.418	2
3-Hydroxyisobutyric acid	0.061	1.180	2.251	2.251	2
Glyceric acid	0.495	2.227	1.484	2.227	2
3-Hydroxysebacic acid	0.473	1.643	2.107	2.107	2
Adipic acid	0.109	1.193	2.059	2.059	2
4-Hydroxymandelic acid	0.051	1.247	2.004	2.004	2
Isocitric.Lactone	0.511	1.708	1.967	1.967	2
Phosphoric acid	0.213	1.965	1.293	1.965	2
Pyroglutamic acid	0.120	1.333	1.874	1.874	2
N-hexanoylglycine	1.773	1.120	0.391	1.773	2
2,4-Dihydroxybutyric acid	1.167	1.766	0.455	1.766	2
N-acetyltrheonine	0.352	1.129	1.691	1.691	2
1,2-Dihydroxyethane	1.671	0.546	1.071	1.671	2
4-Hydroxyphenyllactic acid	1.126	1.658	0.733	1.658	2
N-isovalerylglycine	0.469	1.627	1.265	1.627	2
Fumaric acid	1.292	1.602	0.778	1.602	2
N-tiglylglycine	1.051	0.265	1.559	1.559	2
Nonanoic acid	0.447	1.450	1.129	1.450	2
Oxalic acid	1.048	1.423	0.535	1.423	2
3,4-Dihydroxybenzoic acid	1.406	0.190	1.402	1.406	2
Methylcitric acid	0.339	1.355	1.215	1.355	2
3-Hydroxy-3-methylglutaric acid	0.793	1.277	0.926	1.277	2
4-Hydroxyphenylacetic acid	0.194	1.254	1.129	1.254	2
Octenedioic acid	1.090	1.254	0.763	1.254	2
Hydantoinpropionic acid	0.453	0.827	1.223	1.223	2
2,3,4-Trihydroxybutyric acid	0.203	0.900	1.216	1.216	2
Glycerol	0.141	1.212	1.070	1.212	2
1,2-Dihydroxypropane	1.148	1.198	0.447	1.198	2
3-Hydroxyphenylacetic acid	1.159	0.490	0.996	1.159	2
Glutaric acid	0.823	1.132	0.641	1.132	2
2-Hydroxysebacic acid	1.095		1.095	1.095	2
2-Ketoisovaleric acid	1.095		1.095	1.095	2
Octadecanoic acid	0.878	1.082	0.686	1.082	2
2-Keto-octanoic acid	0.788	1.081	0.908	1.081	2
Uracil	0.798	1.069	0.980	1.069	2
3-Methoxy-4-hydroxyphenylpropionic acid	1.066	0.494	0.971	1.066	2
Oleic acid	0.823	1.003	0.772	1.003	2
Citramalic acid	0.917	0.935	0.257	0.935	2
3-Hydroxypropionic acid	0.178	0.819	0.911	0.911	2
4-Hydroxybutyric acid	0.850	0.081	0.816	0.850	2
2-Hydroxybenzoic acid		0.802	0.802	0.802	2
3-Hydroxypyridine	0.500	0.352	3.901	3.901	1
Methylsuccinic acid	0.012	0.773	2.672	2.672	1

Ethylmalonic acid	1.913	0.643	0.114	1.913	1
3,5-Dihydroxybenzoic acid	0.322	0.635	1.565	1.565	1
Pimelic acid	0.311	0.408	1.384	1.384	1
Succinic acid	1.203	0.052	0.496	1.203	1
Citraconic acid	0.252	0.731	1.107	1.107	1
2-Methyl-2-hydroxybutyric acid	0.761	0.214	1.090	1.090	1
Octanoic acid	0.407	0.678	1.060	1.060	1
4-Hydroxyhippuric acid	0.332	0.788	1.007	1.007	1
2-Methyl-3-ketobutyric acid	0.730	0.958	0.664	0.958	1
2-Ketoglutaric acid	0.622	0.949	0.434	0.949	1
2-Hydroxybutyric acid	0.932	0.690	0.361	0.932	1
Malic acid	0.284	0.868	0.469	0.868	1
N-acetyltyrosine	0.845	0.379	0.741	0.845	1
3-Hydroxyadipic acid	0.447	0.798	0.406	0.798	0
3-Hydroxyisovaleric acid	0.116	0.240	0.786	0.786	0
2-Hydroxyglutaric acid	0.785	0.134	0.778	0.785	0
2-Hydroxy-3-methylvaleric acid	0.784	0.784		0.784	0
3-Methylpimelic acid	0.784	0.444	0.355	0.784	0
4-Ketovaleric acid	0.376	0.514	0.777	0.777	0
2-Methyl-3-hydroxybutyric acid	0.250	0.565	0.770	0.770	0
2-Hydroxyisovaleric acid	0.586	0.758	0.125	0.758	0
2-Hydroxyphenylacetic acid	0.293	0.100	0.713	0.713	0
Maleic acid	0.053	0.706	0.566	0.706	0
Dodecamethylpentasiloxane		0.706	0.706	0.706	0
Aconitic acid	0.091	0.580	0.646	0.646	0
Vanillylmandelic acid	0.344	0.611	0.625	0.625	0
2-Hydroxy-3-methylbutryic acid	0.053	0.588	0.565	0.588	0
2-Ketobutyric acid	0.447	0.586	0.176	0.586	0
Erythronic acid	0.013	0.568	0.487	0.568	0
Treonic acid	0.258	0.384	0.551	0.551	0
3-Methylglutaconic acid	0.361	0.522	0.536	0.536	0
Suberic acid	0.323	0.205	0.531	0.531	0
2,3-Dihydroxybutane	0.057	0.524	0.438	0.524	0
2-Hydroxyisobutyric acid	0.330	0.466	0.523	0.523	0
Pyruvic acid	0.495	0.091	0.447	0.495	0

Appendix I

Table I.1: The metabolites excluded for the comparison of the control group with pregnant women in their second trimester.

Metabolite	VIP Ranking		Reason for exclusion
	PCA	PLS-DA	
Phosphoric acid	1	1	By-product of extraction/derivatisation method
2,3-Dihydroxybutane		6	Origin unknown
Citramalic acid	8		Origin unknown
1,2-Dihydroxybenzene	10		Diet
Isocitric acid		12	Difficult to differentiate between citric acid and isocitric acid on chromatogram
4-Hydroxymandelic acid		13	Origin unknown/Diet

Table I.2: The metabolites excluded for the comparison of the control group with pregnant women in their third trimester.

Metabolite	VIP Ranking		Reason for exclusion
	PCA	PLS	
Isocitric acid	7	12	Difficult to differentiate between citric acid and isocitric acid on chromatogram
2,3-Dihydroxybutanoic acid	8		Origin unknown/Diet
Palmitic acid	9	9	One of the most common saturated fatty acids

Appendix J

Table J.1: The metabolites excluded for the infant comparison (IM vs. IY).

Metabolite	VIP Ranking		Reason for exclusion
	PCA	PLS	
Indole-3-acetic Acid	2	5	Microbial metabolite
Citramalic acid	8		Origin unknown
Urea	10	1	Constitutes about one half of the total urinary solids
Isocitric acid		8	Difficult to differentiate between citric acid and isocitric acid on chromatogram
3-Methylphenol		9	Diet
3-Methoxy-4-hydroxybenzoic acid		10	Food additive

Table J.2: The metabolites excluded for the infant vs. child comparison (IM vs. C).

Metabolite	VIP Ranking		Reason for exclusion
	PCA	PLS	
Urea	4	1	Constitutes about one half of the total urinary solids
Acetylthreonine	7		Origin unknown/Diet
3,4-Dihydroxyfuranone	8		Origin unknown/Diet
3-Methylphenol		4	Diet
1,2-Dihydroxypropane		7	Exogenous compound

Table J.3: The metabolites excluded for the infant vs. child comparison (IY vs. C).

Metabolite	VIP Ranking		Reason for exclusion
	PCA	PLS	
Azelaic acid		5	Diet
4-Hydroxyhippuric acid		6	Diet
3-Methoxy-4-hydroxybenzoic acid		7	Food additive
2-Furancarboxylic acid		11	Appears in the urine of workers exposed to furfural, is marker of exposure to this compound

Table J.4: The metabolites excluded for the children vs. adult comparison (C vs. A).

Metabolite	VIP Ranking		Reason for exclusion
	PCA	PLS	
Urea	1	1	Constitutes about one half of the total urinary solids
3-Methylphenol	8	2	Diet
4-Hydroxyhippuric acid	9	8	Diet
2-Furancarboxylic acid		4	Origin unknown/diet
Phosphoric acid		9	By-product of extraction/derivatisation method
2,3-Dihydroxysuccinic acid		12	Food additive
Indole-3-acetic Acid		13	Microbial metabolite
2,3-Dihydroxybutanoic acid		14	Origin unknown/Diet
1,2-Dihydroxypropane		16	Exogenous compound
Hippuric acid		19	Normal urinary component which is typically increased with increased consumption of phenolic tea, wine, fruit juices